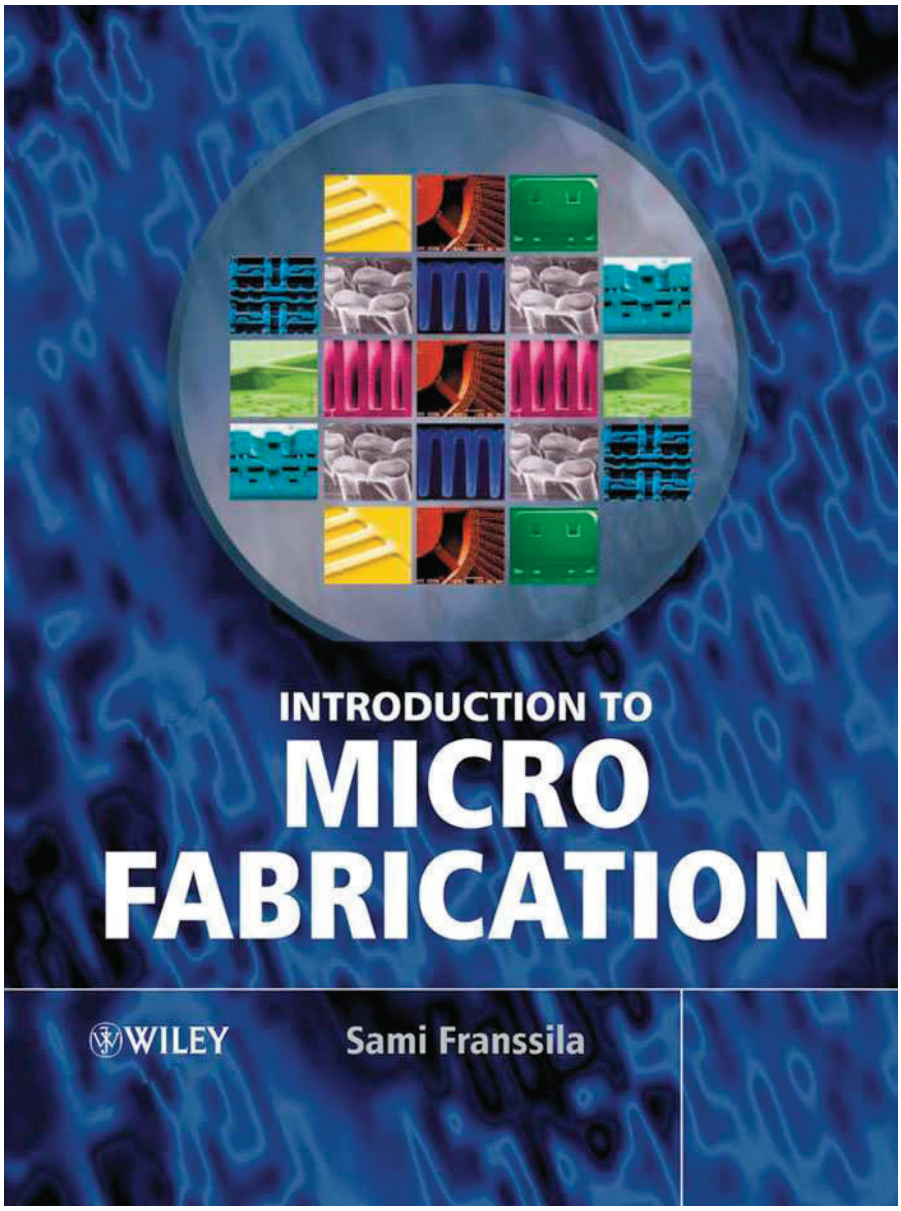


Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Introduction to Microfabrication

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Introduction to Microfabrication

Sami Franssila
Director of Microelectronics Centre,
Helsinki University of Technology, Finland



John Wiley & Sons, Ltd

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Copyright © 2004 John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester,
West Sussex PO19 8SQ, England

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wileyeurope.com or www.wiley.com

All Rights Reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except under the terms of the Copyright, Designs and Patents Act 1988 or under the terms of a licence issued by the Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London W1T 4LP, UK, without the permission in writing of the Publisher. Requests to the Publisher should be addressed to the Permissions Department, John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, or emailed to permreq@wiley.co.uk, or faxed to (+44) 1243 770620.

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the Publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 33 Park Road, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada M9W 1L1

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Library of Congress Cataloging-in-Publication Data

Franssila, Sami.
Introduction to microfabrication / Sami Franssila.
p. cm.
Includes bibliographical references and index.
ISBN 0-470-85105-8 (cloth : alk. paper) – ISBN 0-470-85106-6 (pbk. : alk. paper)
1. Microelectromechanical systems. 2. Electronic apparatus and appliances. 3. Microfabrication. I. Title.
TK7875.F73 2004
621.3 – dc22 2004004940

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0-470-85105-8 (HB)
ISBN 0-470-85106-6 (PB)

Typeset in 9/11pt Times by Laserwords Private Limited, Chennai, India
Printed and bound in Great Britain by Antony Rowe Ltd, Chippenham, Wiltshire
This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

**Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030**

Contents

Preface xv

Acknowledgements xix

PART I: INTRODUCTION **1**

1 Introduction 3
1.1 Microfabrication disciplines 3
1.2 Substrates 4
1.3 Materials 4
1.4 Surfaces and interfaces 5
1.5 Processes 5
1.6 Lateral dimensions 7
1.7 Vertical dimensions 7
1.8 Devices 8
1.9 MOS transistor 11
1.10 Cleanliness and yield 12
1.11 Industries 12
1.12 Exercises 14
References and related readings 15

2 Micrometrology and Materials Characterization 17
2.1 Microscopy and visualization 17
2.2 Lateral and vertical dimensions 17
2.3 Electrical measurements 19
2.4 Physical and chemical analyses 20
2.5 XRD (X-ray diffraction) 20
2.6 TXRF (total reflection X-ray fluorescence) 21
2.7 SIMS (secondary ion mass spectrometry) 21
2.8 Auger electron spectroscopy (AES) 22
2.9 XPS (X-ray photoelectron spectroscopy)/ESCA 22
2.10 RBS (Rutherford backscattering spectrometry) 22
2.11 EMPA (electron microprobe analysis)/EDX (energy dispersive X-ray analysis) 23
2.12 Other methods 24
2.13 Analysis area and depth 24
2.14 Practical issues with micrometrology 25
2.15 Exercises 26
References and related readings 26

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

3 Simulation of Microfabrication Processes 27
3.1 Types of simulation 27
3.2 1D simulation 28
3.3 2D simulation 29
3.4 3D simulation 30
3.5 Exercises 31
References and related readings 32

PART II: MATERIALS 33

4 Silicon 35
4.1 Silicon material properties 35
4.2 Silicon crystal growth 36
4.3 Silicon crystal structure 39
4.4 Silicon wafering process 40
4.5 Defects and non-idealities in silicon crystals 43
4.6 Exercises 44
References and related readings 45

5 Thin-Film Materials and Processes 47
5.1 Thin films versus bulk materials 47
5.2 Physical vapour deposition (PVD) 49
5.3 Evaporation and molecular beam epitaxy 49
5.4 Sputtering 50
5.5 Chemical vapour deposition (CVD) 51
5.6 Other deposition technologies 53
5.7 Metallic thin films 56
5.8 Dielectric thin films 58
5.9 Properties of dielectric films 59
5.10 Polysilicon 62
5.11 Silicides 63
5.12 Exercises 64
References and related readings 64

6 Epitaxy 65
6.1 Heteroepitaxy 66
6.2 CVD homoepitaxy of silicon 67
6.3 Simulation of epitaxy 69
6.4 Advanced applications of epitaxy 70
6.5 Exercises 70
References and related readings 71

7 Thin-film Growth and Structure 73
7.1 General features of thin-film processes 73
7.2 PVD-film growth and structure 74
7.3 CVD-film growth and structure 77
7.4 Surfaces and interfaces 79
7.5 Adhesion layers and barriers 81
7.6 Multilayer films 82
7.7 Stresses 83

**Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030**

7.8 Thin films over topography: step coverage 86
7.9 Simulation of deposition 88
7.10 Exercises 90
References and related readings 90

PART III: BASIC PROCESSES 91

8 Pattern Generation 93
8.1 Beam writing strategies 93
8.2 Electron beam physics 94
8.3 Photomask fabrication 94
8.4 Photomasks as tools 95
8.5 Photomask inspection, defects and repair 96
8.6 Exercises 97
References and related readings 97

9 Optical Lithography 99
9.1 Lithography tools (alignment and exposure) 99
9.2 Resolution 101
9.3 Basic pattern shapes 102
9.4 Alignment and overlay 103
9.5 Exercises 104
References and related readings 104

10 Lithographic Patterns 107
10.1 Resist application 107
10.2 Resist chemistry 108
10.3 Thin film optics in resists 110
10.4 Extending optical lithography 112
10.5 Lithography simulation 113
10.6 Lithography practice 114
10.7 Photoresist stripping/ashing 116
10.8 Exercises 117
References and related readings 117

11 Etching 119
11.1 Wet etching 120
11.2 Electrochemical etching 123
11.3 Anisotropic wet etching 125
11.4 Plasma etching 125
11.5 Characterization of etch processes 128
11.6 Etch processes for common materials 128
11.7 Etch time and spacers 129
11.8 Comparison of wet etching, anisotropic wet etching and plasma etching 130
11.9 Exercises 130
References and related readings 131

12 Wafer Cleaning and Surface Preparation 133
12.1 Contamination forms 133
12.2 Wet cleaning 135

**Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030**

12.3 Particle contamination 136
12.4 Organic contamination 138
12.5 Metal contamination 138
12.6 Rinsing and drying 140
12.7 Physical cleaning 140
12.8 Exercises 141
Suggested further reading 141

13 Thermal Oxidation 143
13.1 Oxidation process 143
13.2 Deal–grove oxidation model 143
13.3 Oxide structure 145
13.4 Simulation of oxidation 146
13.5 Local oxidation of silicon (LOCOS) 147
13.6 Stress and pattern effects in oxidation 148
13.7 Exercises 150
References and related readings 150

14 Diffusion 153
14.1 Diffusion mechanisms 154
14.2 Doping profiles in diffusion 155
14.3 Simulation of diffusion 156
14.4 Diffusion applications 157
14.5 Exercises 158
References and related readings 158

15 Ion Implantation 159
15.1 The implant process 159
15.2 Implant damage and damage annealing 161
15.3 Ion implantation simulation 162
15.4 Tools for ion implantation 162
15.5 SIMOX: SOI by ion implantation 164
15.6 Exercises 164
References and related readings 164

16 CMP: Chemical–Mechanical Polishing 165
16.1 CMP process and tool 165
16.2 Mechanics of CMP 167
16.3 Chemistry of CMP 168
16.4 Applications of CMP 169
16.5 CMP control measurements 170
16.6 Non-idealities in CMP 170
16.7 Exercises 171
References and related readings 172

17 Bonding and Layer Transfer 173
17.1 Silicon fusion bonding 174
17.2 Anodic bonding 176
17.3 Other bonding techniques 177

**Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030**

17.4 Bonding mechanics 178
17.5 Bonding of structured wafers 179
17.6 Bonding for SOI wafer fabrication 180
17.7 Layer transfer 180
17.8 Exercises 181
References and related readings 181

18 Moulding and Stamping 183
18.1 Moulding 183
18.2 2D surface stamping 186
18.3 3D-volume stamping 187
18.4 Comparison with lithography 189
18.5 Exercises 189
References 189

PART IV: STRUCTURES **191**

19 Self-aligned Structures 193
19.1 Self-aligned MOS gate 193
19.2 Self-aligned twin well 194
19.3 Spacers and self-aligned silicide (salicide) 194
19.4 Self-aligned junctions 196
19.5 Exercises 197
References and related readings 197

20 Plasma-etched Structures 199
20.1 Multi-step etching 199
20.2 Multi-layer etching 200
20.3 Resist effects on etching 201
20.4 Non-masked etching 201
20.5 Pattern size and pattern density effects 202
20.6 Etch residues and damage 203
20.7 Exercises 203
References and related readings 204

21 Wet-etched Silicon Structures 205
21.1 Basic structures on <100> silicon 205
21.2 Etchants 205
21.3 Etch masks and protective coatings 206
21.4 Etch rate and etch stop 207
21.5 Diaphragm fabrication 208
21.6 Complex shapes by <100> etching 209
21.7 Front side bulk micromachining 211
21.8 Corner compensation 212
21.9 <110> Etching 212
21.10 <111> silicon etching 213
21.11 Comparison of <100>, <110> and <111> etching 215
21.12 Exercises 215
References and related readings 216

**Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030**

22 Sacrificial and Released Structures 217
22.1 Structural and sacrificial layers 217
22.2 Single structural layer 218
22.3 Stiction 219
22.4 Two structural–layer processes 220
22.5 Rotating structures 222
22.6 Hinged structures 222
22.7 Sacrificial structures using porous silicon 223
22.8 Exercises 223
References and related readings 224

23 Structures by Deposition 227
23.1 Plated structures 227
23.2 Lift-off metallization 228
23.3 Special deposition applications 229
23.4 Localized deposition 230
23.5 Sealing of cavities 232
23.6 Exercises 233
References and related readings 233

PART V: INTEGRATION 235

24 Process Integration 237
24.1 Process integration aspects of a solar-cell process 237
24.2 Wafer selection 238
24.3 Patterns 241
24.4 Design rules 242
24.5 Contamination budget 247
24.6 Thermal processes 248
24.7 Thermal budget 249
24.8 Metallization 249
24.9 Reliability 250
24.10 Exercises 252
References and related readings 253

25 CMOS Transistor Fabrication 255
25.1 5 μm polysilicon gate CMOS process 255
25.2 MOS transistor scaling 258
25.3 Advanced CMOS issues 260
25.4 Gate module 262
25.5 Contact to silicon 265
25.6 Exercises 266
References and related readings 267

26 Bipolar Technology 269
26.1 Fabrication process of SBC bipolar transistor 269
26.2 Advanced bipolar structures 272
26.3 BiCMOS technology 275
26.4 Exercises 275
References and related readings 276

**Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030**

27 Multilevel Metallization 277
27.1 Two-level metallization 277
27.2 Multilevel metallization 278
27.3 Damascene metallization 280
27.4 Metallization scaling 280
27.5 Copper metallization 281
27.6 Low-k dielectrics 282
27.7 Exercises 284
References and related readings 285

28 MEMS Process Integration 287
28.1 Double-side processing 287
28.2 Membrane structures 291
28.3 Through-wafer structures 293
28.4 Patterning over severe topography 294
28.5 DRIE versus anisotropic wet etching 295
28.6 IC–MEMS integration 296
28.7 Exercises 298
References and related readings 298

29 Processing on Non-silicon Substrates 301
29.1 Substrates 301
29.2 Thin-film transistors, TFTs 302
29.3 Exercises 304
References and related readings 304

PART VI: TOOLS **307**

30 Tools for Microfabrication 309
30.1 Batch processing versus single-wafer processing 309
30.2 Equipment figures of merit 310
30.3 Tool life cycles 311
30.4 Process regimes: temperature–pressure 311
30.5 Simulation of process equipment 312
30.6 Measuring fabrication processes 312
30.7 Exercises 314
References and related readings 314

31 Tools for Hot Processes 315
31.1 High temperature equipment: hot wall versus cold wall 315
31.2 Furnace processes 315
31.3 Rapid-thermal processing/rapid-thermal annealing 316
31.4 Exercises 319
References and related readings 319

32 Vacuum and Plasmas 321
32.1 Vacuum-film interactions 321
32.2 Vacuum production 322
32.3 Plasma etching 324
32.4 Sputtering 325

**Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030**

32.5 PECVD 327
32.6 Residence time 327
32.7 Exercises 327
References and related readings 327

33 Tools for CVD and Epitaxy 329
33.1 CVD rate modelling 329
33.2 CVD reactors 330
33.3 ALD (Atomic Layer Deposition) 331
33.4 MOCVD 332
33.5 Silicon CVD epitaxy 333
33.6 Epitaxial reactors 334
33.7 Exercises 335
References and related readings 336

34 Integrated Processing 337
34.1 Ambient control 337
34.2 Dry cleaning 338
34.3 Integrated tools 339
34.4 Exercises 339
References and related readings 339

PART VII: MANUFACTURING 341

35 Cleanrooms 343
35.1 Cleanroom standards 343
35.2 Cleanroom subsystems 345
35.3 Environment, safety and health (ESH) aspects 346
35.4 Exercises 348
References and related readings 348

36 Yield 349
36.1 Yield models 349
36.2 Process step effect 352
36.3 Yield ramping 352
36.4 Exercises 352
References and related readings 352

37 Wafer Fab 355
37.1 Historical development of IC manufacturing 356
37.2 Manufacturing challenges 357
37.3 Cycle time 357
37.4 Cost-of-ownership (CoO) 358
37.5 Cost of processed silicon 359
37.6 Exercises 360
References and related readings 360

**Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030**

PART VIII: FUTURE **361**

38 Moore's Law 363
38.1 From transistor to integrated circuit 363
38.2 Moore's law 364
38.3 Extending optical lithography: phase-shift masks (PSM) 366
38.4 Alternatives to optical lithography 368
38.5 Fundamental and practical limits 369
38.6 IC industry 371
38.7 Exercises 372
References and related readings 372

39 Microfabrication at Large 373
39.1 New materials 373
39.2 High aspect ratio structures 374
39.3 Tools of microfabrication 375
39.4 Bonding and layer transfer 376
39.5 Devices 376
39.6 Microfabrication industries 378
39.7 Exercises 379
References and related readings 380

Appendix A: Comments and Hints to Selected Problems **381**

Appendix B: Constants and Conversion Factors **387**

Index **391**

**Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030**

Preface

Microfabrication is generic: its applications include integrated circuits, MEMS, microfluidics, micro-optics, nanotechnology and countless others. Microfabrication is encountered in slightly different guises in all of these applications: electroplating is essential for deep submicron IC metallization and for LIGA-microstructures; deep-RIE is a key technology in trench DRAMs and in MEMS; imprint lithography is utilized in microfluidics where typical dimensions are 100 μm, as well as in nanotechnology, where feature sizes are down to 10 nm. This book is unique because it treats microfabrication in its own right, independent of applications, and therefore it can be used in electrical engineering, materials science, physics and chemistry classes alike.

Instead of looking at devices, I have chosen to concentrate on microstructures on the wafer: lines and trenches, membranes and cantilevers, cavities and nozzles, diffusions and epilayers. Lines are sometimes isolated and sometimes in dense arrays, irrespective of linewidths; membranes can be made by timed etching or by etch stop; source/drain diffusions can be aligned to the gate in a mask aligner or made in a self-aligned fashion; oxidation on a planar surface is easy, but the oxidation of topographic features is tricky. The microstructure-view of microfabrication is a solution against outdating: alignment must be considered for both 100 μm fluidic channels and 100 nm CMOS gates, etch undercutting target may be 10 nm or 10 μm, but it is there; dopants will diffuse during high temperature anneals, but the junction depth target may be tens of nanometres or tens of micrometres.

A common feature of older textbooks is concentration on physics and chemistry: plasma potentials, boundary layers, diffusion mechanisms, Rayleigh resolution, thermodynamic stability and the like. This is certainly a guarantee against outdating in rapidly evolving technologies, but microfabrication is an engineering discipline, not physics and chemistry. CMOS scaling trends have in fact been more reliable than basic physics and chemistry in the past 40 years: optical lithography was predicted to be unable to print submicron lines and gate oxides today are thinner than the ultimate limits conceived in the 1970s. And it is pedagogically better to show applications of CVD films before plunging into pressure dependence of deposition rate, and to discuss metal film functionalities before embracing sputtering yield models.

In this book, another major emphasis is on materials. Materials are universal, and not outdated rapidly. New materials are, of course, being introduced all the time, but the basic materials properties like resistivity, dielectric constant, coefficient of thermal expansion and Young's modulus must always be considered for low-k and high-k dielectrics, SnO_2 sensor films, diamond coatings and 100 μm-thick photoresists alike. Silicon, silicon dioxide, silicon nitride, aluminium, tungsten, copper and photoresist will be met again in various applications: nitride is used not only in LOCOS isolation, but also in MEMS thermal isolation; aluminium not only serves as a conductor in ICs but also as a mirror in MOEMS; copper is used for IC metallization and also as a sacrificial layer under nickel in metal MEMS; photoresist acts not only as a photoactive material but also as an adhesive in wafer bonding.

Devices are, of course, discussed but from the fabrication viewpoint, without thorough device physics. The unifying idea is to discuss the commonalities and generic features of the fabrication processes. Resistors and capacitors serve to exemplify concepts like alignment sequence and design rules, or interface stability. After basic processes and concepts have been introduced, process integration examples show a wide spectrum of full process flows: for example, solar cell, piezoresistive pressure sensor, CMOS, AFM cantilever tip, microfluidic out-of-plane needle and super-self-aligned bipolar transistor. Small process-sequence examples include, similarly, a variety of structures: replacement gate, cavity sealing, self-aligned rotors and dual damascene-low-k options are among the others.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Older textbooks present microfabrication as a toolbox of MEMS or as the technology for CMOS manufacturing. Both approaches lead to unsatisfactory views on microfabrication. Ten years ago, chemical–mechanical polishing was not detailed in textbooks, and five years ago discussion on CMP was included in multilevel metallization chapter. Today, CMP is a generic technology that has applications in CMOS front-end device isolation and surface micromechanics, and is used to fabricate photonic crystals and superconducting devices. It therefore deserves a chapter of its own, independent of actual or potential applications. Similarly, wafer cleaning used to be presented as a preparatory step for oxidation, but it is also essential for epitaxy, wafer bonding and CMP. Device-view, be it CMOS or some other, limits processes and materials to a few known practices, and excludes many important aspects that are fruitful in other applications.

The aim of the book is for the student to feel comfortable both in a megafab and in a student lab. This means that both research-oriented and manufacturing-driven aspects of microfabrication must be covered. In order to keep the amount of material manageable, many things have had to be left out: high density plasmas are mentioned, but the emphasis is on plasma processing in general; KOH and TMAH etching are both described, but commonalities rather than differences are shown; imprint lithography and hot embossing are discussed but polymer rheology is neglected; alternatives to optical lithography are mentioned, but discussed only briefly. Emphasis is on common and conceptual principles, and not on the latest technologies, which hopefully extends the usable life of the book.

STRUCTURE OF THE BOOK

The structure of this book differs from the traditional structure in many ways. Instead of discussing individual process steps at length first and putting full processes together in the last chapter, applications are presented throughout the book. The chapters on equipment are separated from the chapters on processes in order to keep the basic concepts and current practical implementations apart.

The introduction covers materials, processes, devices and industries. Measurements are presented next, and more examples of measurement needs in microfabrication are presented in almost every chapter. A general discussion of simulation follows, and more specific simulation cases are presented in the chapters that follow.

Materials of microfabrication are presented next: silicon and thin films. Silicon crystal growth is shortly covered but from the very beginning, the discussion centres on wafers and structures on wafers: therefore, silicon wafering process, and resulting wafer properties are emphasized. Epitaxy, CVD, PVD, spin coating and electroplating are discussed, with resulting materials properties and microstructures on the centre stage, rather than equipment themselves. Lithography and etching then follow. This order of presentation enables more realistic examples to be discussed early on.

The basic steps in silicon technology, such as oxidation, diffusion and ion implantation are discussed next, followed by CMP and bonding. Moulding and stamping techniques have also been included. In contrast to older books, and to books with CMOS device emphasis, this book is strong in back-end steps, thin films, etching, planarization and novel materials. This reflects the growing importance of multilevel metallization in ICs as well as the generic nature of etch and deposition processes, and their wide applicability in almost all microfabrication fields. Packaging is not dealt with, again in line with wafer-level view of microfabrication. This also excludes stereomicrolithography and many miniaturized traditional techniques like microelectrodischarge machining.

Microfabrication is an engineering discipline, and volume manufacturing of microdevices must be discussed. Discussions on process equipment have often been bogged by the sheer number of different designs: should the students be shown both 13.56 MHz diode etcher, triode, microwave, ECR, ICP and helicon plasmas, and should APCVD, LPCVD, SA-CVD, UHV-CVD and PECVD reactors all be presented? In this book, the process equipment discussion is again tied to structures that result on wafers, rather than in the equipment *per se*: base vacuum interaction with thin-film purity is discussed; the role of RTP temperature uniformity on wafer stresses is considered; and surface reaction versus transport controlled growth in different CVD reactors is analysed. Cleanroom technology, wafer fab operations, yield and cost are also covered. Moore's law and other trends expose students to some current and future issues in microfabrication processes, materials and applications.

In many cases, treatment has been divided into two chapters: for example, Chapter 5 treats thin film basics, and Chapter 7 deals with more advanced topics. Lithography and etching have been divided similarly. This enables short or long course versions to be designed around the book. The figures from the book are available to teachers via the Internet. Please register at Wiley for access www.wileyeurope.com/go/microfabrication.

**Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030**

ADVICE TO STUDENTS

This book is an introductory text. Basic university physics and chemistry suffices for background. Materials science and electronics courses will of course make many aspects easier to understand, but the structure of the book does not necessitate them. The book contains 250 homework problems, and in line with the idea of microfabrication as an independent discipline, they are about fabrication processes and microstructures; not about devices. Problems fall mainly in three categories: process design/analysis, simulations and back-of-the-envelope calculations. The problems that are designed to be solved with a simulator are marked by "S". A simple one-dimensional simulator will do. The "ordinary" problems are designed to develop a feeling for orders of magnitude in the microworld: linewidths, resistances, film thicknesses, deposition rates, stresses etc. It is often enough to understand if a process can be done in seconds, minutes or hours; or whether resistance range is milliohms, ohms or kiloohms. You must learn to make simplifying assumptions, and to live with uncertain data. Searching the Internet for answers is no substitute to simple calculations that can be done in minutes because the simple estimates are often as accurate (or inaccurate) as answers culled from Internet. It should be borne in mind that even constants are often not well known: for instance, recent measurements of silicon melting point have resulted in values 1408°C by one group, 1410°C by one, 1412°C by seven groups, 1413°C by eight groups and 1416°C by three groups, and if older works are encountered, values range from 1396°C to 1444°C. With thin film materials properties are very much deposition process dependent, and different workers have measured widely different values for such basic properties as resistivity or thermal conductivity. Even larger differences will pop up, if, for instance, the phase of metal film changes from body-centered cubic to β-phase: temperature coefficient of resistivity can then be off by a factor of ten. Polymeric materials, too, exhibit large variation in properties and processing. There are also calculations of economic aspects of microfabrication: wafer cost, chip size and yield. A bit of memory costs next to nothing, but the fabs (fab is short for fabrication facility) that churn out these chip are enormously expensive.

Comments and hints to selected homework problems are given in Appendix A. In Appendix B you can find useful physical constants, silicon material properties and unit conversion factors.

**Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030**

Acknowledgements

Writing a book takes a lot of time, and numerous people have contributed their time and effort at various stages of this project. Jyrki Kaitila, Andreas Englmüller, Olli Anttila, Risto Mutikainen, Joni Mellin, Ari Lehto and Tarja Rahikainen read through the manuscript in its nascent state, and provided essential input into organization of the book. Their interest in both details and overall structure is much appreciated.

A far larger group of people have contributed to selected parts of the book by providing me with data, micrographs and photos; they have led me to useful sources, pointed out gaps and corrected my text. Thanks are due to Bo Bängtsson, Martin Kulawski, Klas Hjort, Arturo Ayon, Pekka Seppälä, Robert Eichinger-Heue, Marin Alexe, Markku Tilli, Juha Rantala, Jyrki Kiihamäki, Weileun Fang, Mikko Ritala, Martti Blomberg, Jaakko Saarilahti, Hannu Kattelus, Mikko Kiviranta, Veli-Matti Airaksinen, Paula Heikkilä, Harri Pohjonen, Jouni Ahopelto, Antti Lipsanen, Jari Likonen, Eero Haimi, Ulrika Gyllenberg, Kestas Grigoras and Victor Ovtchinnikov. Charlotta Tuovinen has provided assistance with computers on countless occasions.

My students and teaching assistants Tuuli Juvonen, Antti Niskanen, Santeri Tuomikoski, Esa Tuovinen and Seppo Marttila have been guinea pigs for the reading of the text and exercises. They have lived to tell the tale!

Pekka Kuivalainen and Ari Sihvola are acknowledged for their encouragement in teaching, in general, and in textbook writing, in particular.

Peter Mitchell, Kathryn Sharples, Céline Durand and Susan Barclay at Wiley have brought the project to completion through face-to-face meetings and numerous e-mails.

Omissions and factual errors remain my sole responsibility.

Sami Franssila
Helsinki, February 29, 2004

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Part I

Introduction

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

1

Introduction

1.1 MICROFABRICATION DISCIPLINES

Integrated circuits industry and related industries such as microsystems/MEMS, solar cells, flat-panel displays and optoelectronics rely on microfabrication technologies. Typical dimensions are around 1 μm in the plane of the wafer (the range is rather wide; from 0.1 μm to 100 μm). Vertical dimensions range from atomic-layer thickness (0.1 nm) to hundreds of micrometres but thicknesses from 10 nm to 1 μm are typical.

The historical development of microfabrication-related disciplines is shown below (Figure 1.1). Invention of the transistor in 1947 sparked a revolution. The transistor was born out of fusion of radar technology (fast crystal detectors for electromagnetic radiation) and solid-state physics. Adoption of microfabrication methods enabled fabrication of many transistors on a single piece of semiconductor, and a few years later, the fabrication of integrated circuits; that is, transistors were connected with each other on the wafer rather than being separated from each other and reconnected on the circuit board.

Microelectronic and optoelectronic devices make use of the semiconducting properties of silicon. Doping of silicon can change its resistivity by eight orders of magnitude, enabling a great number of microstructures and devices to be made. Silicon microelectronic devices today are characterized by their immense complexity and miniaturization; a hundred million transistors fit on a chip the size of a fingernail.

Gallium arsenide and other III–V compound semiconductors are used to make light emission devices like lasers. Silicon optoelectronic devices can be used as light detectors, but, recently, light transmission from silicon has been demonstrated in laboratory experiments. Micro-optics makes use of silicon in another way: silicon surfaces act as mirrors, or as extremely flat and smooth supports for metallic or dielectric mirrors. Silicon can be machined to make movable mirrors and adaptive optical elements. Silicon dioxide and silicon nitride can be deposited and etched to form waveguides with graded or stepped refractive indices like optical fibres.

Micromechanics makes use of mechanical properties of silicon. Silicon is extremely strong, and flexible beams and diaphragms can be made from it. Pressure sensors, resonators, gyroscopes, switches and other mechanical and electromechanical devices utilize the excellent mechanical properties of silicon.

Micromachines, as well as many microsensors and actuators, make use of active materials, for example, piezoelectric materials or shape memory alloys. Silicon has the role of precise platform on which these devices can be built. Superconducting devices are made on silicon because silicon is compatible with a plethora of processing technologies.

Nanotechnology is an outgrowth and extension of microfabrication. Some of the tools are same, like the electron-beam lithography machines, which have been used to draw nanometre-sized structures long before the term *nanotechnology* was coined. Some of the methods are based on scanning probe devices such as the atomic force microscope (AFM), which is an important instrument for microstructure characterization. Thin films down to atomic-layer thicknesses have been grown and deposited in the microfabrication communities for decades. Novel ways of depositing films, like self-assembled monolayers (SAMs), have been introduced by nanotechnologists, and some of those techniques are being investigated by the established microfabrication community as tools for continued downscaling of microstructures.

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

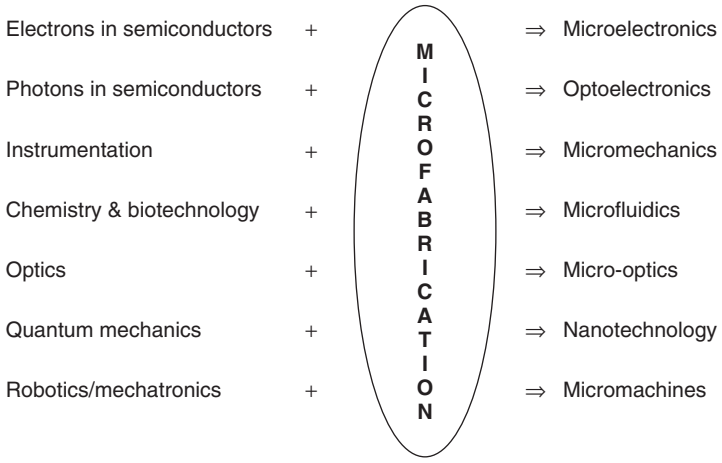


Figure 1.1 Microtechnology subfields

1.2 SUBSTRATES

Silicon is the workhorse of microfabrication. Integrated circuits (IC) utilize the electrical properties of silicon, but many microfabrication disciplines use silicon for convenience: silicon is available in a wide variety of sizes, shapes and resistivities; it is smooth, flat, mechanically strong and fairly cheap. What is more, silicon wafers are by default compatible with microfabrication equipment because most of the machinery for microfabrication was originally developed for silicon ICs.

Bulk silicon wafers are single-crystal pieces cut and polished from larger single-crystal ingots. Silicon is extremely strong, on par with steel, and it also retains its elasticity at much higher temperatures than metals. However, single-crystalline silicon (SCS) wafers are fragile: once fracture starts, it immediately develops across the wafer because covalent bonds do not allow dislocation movements.

Resistivities of silicon-wafer range from 0.001 to 20 000 ohm-cm. High-resistivity silicon can sometimes be used instead of dielectric wafers, but this depends on application. Silicon-on-insulator wafers offer the best of both worlds: an insulator layer (usually SiO_2) between two silicon pieces provides dielectric isolation. The oxide in between can act as a stop layer so that the two silicon parts can be processed independently. Thin layers can be cut from silicon-wafer surface, and transferred to another substrate, which may be altogether a different material.

Silicon wafers are available in 3″, 100, 125, 150, 200 and 300 mm diameters. In addition to size, resistivity and dopant type, wafer specifications include thickness and its variation, crystal orientation, particle counts and many others.

Wafers can be single crystalline, polycrystalline or amorphous. Silicon, quartz (SiO_2) gallium arsenide (GaAs), silicon carbide (SiC), gallium arsenide (GaAS), lithium niobate ($LiNbO_3$) and sapphire (Al_2O_3) are examples of single-crystalline substrates. Polycrystalline silicon is widely used in solar cell production, and thin-film transistors have been made on steel. Amorphous substrates are also common: glass (which is SiO_2 mixed with metal oxides like Na_2O); fused silica (SiO_2, chemically it is identical to quartz) and alumina (Al_2O_3), which is a common substrate for microwave circuits. Even plastic sheets have been used as substrates. Exotic substrates must be evaluated for available sizes, purities, smoothness, thermal stability, mechanical strength, and so on. Round substrates are easy to accommodate but square and rectangular ones need special processing because tools for microfabrication are geared for round silicon wafers.

1.3 MATERIALS

Just like substrate wafers, the grown and deposited thin films can be

- single crystalline,
- polycrystalline,
- amorphous.

During wafer processing, single-crystalline films usually stay single crystalline, but they can be amorphized by, for example, ion bombardment; polycrystalline

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

films experience grain growth, for instance, during heat treatments; amorphous films can stay amorphous or they can crystallize, usually into polycrystalline state and under very special circumstances into single-crystalline state.

Elemental substrates and elemental thin films are simple and they have various uses; silicon, aluminium, copper and tungsten are widely used. Compounds introduce new possibilities and challenges: silicon dioxide (SiO_2), silicon nitride (Si_3N_4), hafnium dioxide (HfO_2), titanium silicide ($TiSi_2$), titanium nitride (TiN) and aluminium nitride (AlN) are not necessarily stoichiometric when deposited. For instance, titanium nitride is more accurately described as TiN_x, with the exact value of x determined by the details of the deposition process.

In addition to elemental and compound materials, alloys are widely used. Instead of using elemental aluminium for metallization, it is beneficial to use Al–1% Si or Al–0.5% Si–2% Cu alloy, for metallization stability, as will be seen in Chapter 24. Alloys of dissimilar-sized atoms often result in amorphous films, and in some applications, it is beneficial to maintain amorphousness upon annealing and to prevent crystallization.

Deposition conditions strongly affect thin-film properties, for example via impurity incorporation or process temperature: silicon will be amorphous if deposited at low temperature, polycrystalline at medium temperatures and single-crystalline material can be obtained at high temperatures under tightly controlled conditions. Materials in microfabrication must be amenable to micropatterning technologies, which translates to either etching or polishing. Sometimes it is enough to deposit films on flat, planar wafers, but most often the films have to extend over steps and into trenches, which may be 40 times deeper than wide. These severe topographies introduce further deposition process–dependent subtleties.

1.4 SURFACES AND INTERFACES

The general material structure of a microfabricated device is shown below. Interfaces between thin-film and bulk, and between two films, are important for stability of structures. Wafers experience a number of thermal treatments during their fabrication, and various chemical and physical processes are operative at interfaces: for example, reactions or diffusion.

Film 1 of Figure 1.2 might present for example an aluminium conductor, and film 2 is the passivation layer of silicon nitride, or film 1 is flash-memory tunnel oxide and film 2 is the polysilicon floating gate, or film 1 is oxide insulation and film 2 is a gas-sensitive SnO_2 film.

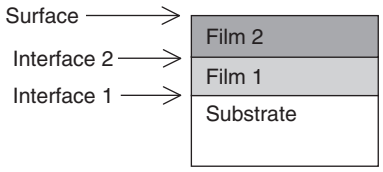


Figure 1.2 Materials and interfaces in a schematic microstructure

Surface physical properties like roughness and reflectivity are material and fabrication process dependent. The chemical nature of the surface is equally important: many surfaces are covered by native oxide films (e.g., silicon, aluminium and titanium form surface oxides readily) and by residual films. Adsorbed gases and moisture affect processing via adhesion or nucleation changes.

Thick substrates are not immune to thin films: a thin film of a few tens of nanometres may have such a high stress that a 500 μm thick silicon wafer is curved; or minute iron contamination on the surface will diffuse through a 500 μm thick wafer during a fairly moderate thermal treatment.

1.5 PROCESSES

Microfabrication processes consist of four basic operations:

1. High-temperature processes
2. Thin-film deposition processes
3. Patterning
4. Layer transfer and bonding.

Surface preparation and wafer cleaning could be termed the fifth basic operation but unlike the four others, wafer cleaning is never done in isolation: it is always closely connected with both the preceding and the following process steps. Under each basic operation, there are many specific technologies, which are suitable for certain devices, certain substrates, certain linewidths or certain cost levels.

High-temperature steps modify dopant atom distributions inside silicon, and they are crucial for transistor characteristics. Devices like piezo-resistive pressure sensors also rely on high-temperature steps, with epitaxy and resistor diffusion as the key processes. High-temperature steps can be simulated extensively, by solving diffusion equations on a computer. High-temperature regime in microfabrication is ca. 900 °C and upwards, temperatures where dopants readily diffuse.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Low-temperature processes leave metal-to-silicon interface stable, and generally, 450 °C is regarded as the upper limit for low temperatures. In between 450 and 900 °C, there is a middle range that must be discussed with specific materials and interfaces in mind.

High-temperature regime is also known as *front-end of the line* (FEOL) in silicon IC business, and low-temperature regime as *back-end of the line* (BEOL). But these terms have other meanings as well: for many people in the electronics industry outside silicon-wafer fabrication plants, front-end includes all processing on wafers, and back-end is dicing, testing, encapsulation and assembly. We will use the first definition.

Thin-film steps are used to make structures of metallic, dielectric and semiconducting films. Many thin-film steps can be carried out identically on silicon wafers and other substrates; by definition they are layers deposited on top of a substrate. Thin-film steps do not affect dopant distribution inside silicon, that is, diodes and transistors are unaffected by them.

Processes act on whole wafers; this is the basic premise. If materials are not needed everywhere, it has to be etched or polished away locally. Patterning processes define structures usually in two steps: photolithographic patterning of resist film, which then acts as a mask for etching or modification of the underlying material (Figure 1.3). Photomask defines areas where the photosensitive film (the photoresist) will be exposed. This photoresist will then serve as a mask for subsequent steps.

Wafer bonding and layer transfer enable more complex structures to be made. Stacks of wafers are used in

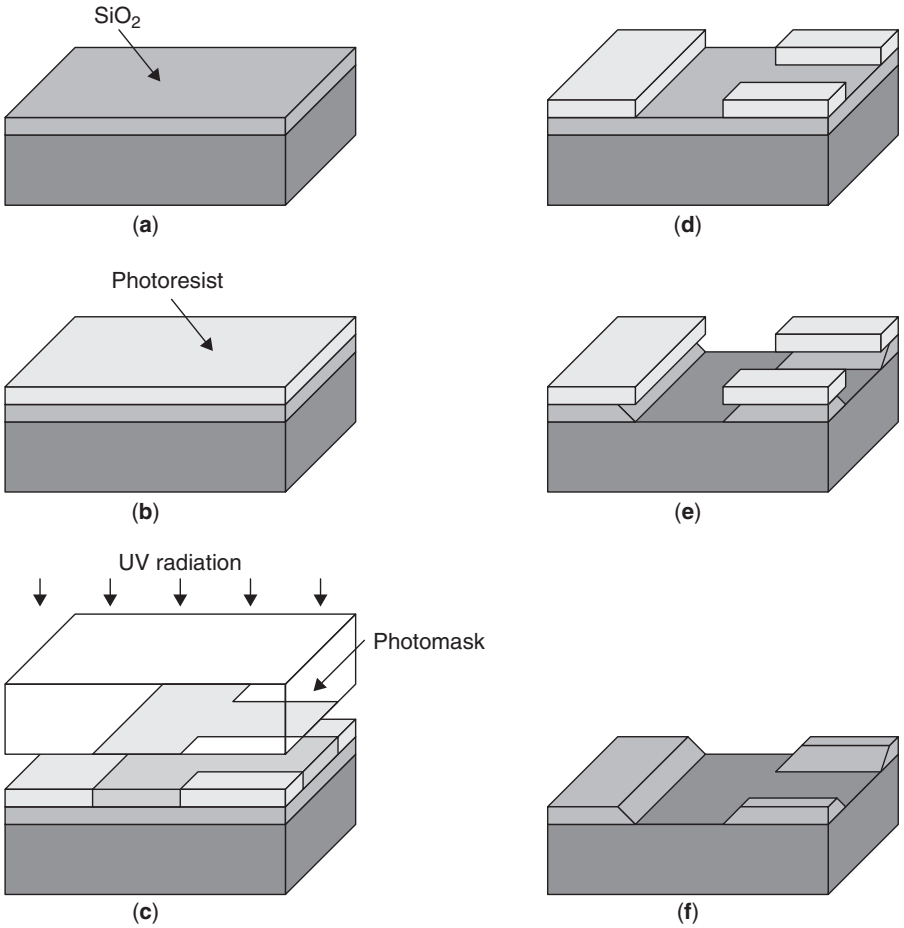


Figure 1.3 Lithographic patterning process: (**a**) oxide-film deposition; (**b**) photoresist application; (**c**) UV exposure through a photomask; (**d**) development of resist image; (**e**) etching of oxide and (**f**) photoresist removal. Drawing courtesy Esa Tuovinen, Helsinki University of Technology

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

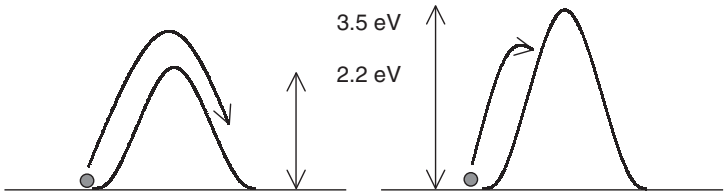


Figure 1.4 Diffusion process: 2.2 eV barrier can be crossed at ease at 900 °C but the frequency of crossing the 3.5 eV barrier is low. Higher temperature, for example, 1050 °C, would be needed for the 3.5 eV barrier to be crossed at ease

fluidic devices for channel enclosure, in microelectromechanical systems (MEMS) bonding forms sealed cavities for resonating devices, and bonding enables single-crystal silicon to be attached on amorphous oxide for electrical insulation.

These elementary operations are combined many times over to create devices. Process complexity is often discussed in terms of the number of lithography steps: six lithography steps are enough for a simple P-Type Metal-Oxide Semiconductor (PMOS) transistor (late 1960s technology, and still used as a student lab process in many universities), and many MEMS, solar cell and flat-panel display devices can be made with two to six photolithography steps even today but the 0.18 μm CMOS (Complementary Metal Oxide Semiconductor) circuits of year 2000 need 25 lithography steps. Systems which combine CMOS with other functionalities, like bipolar transistors, integrated displays or sensors, use for example, 0.5 to 0.8 μm CMOS with 15 mask levels, and add half a dozen lithography steps in addition to the CMOS process.

1.5.1 Arrhenius behaviour

Many chemical and physical processes are exponentially temperature dependent. Arrhenius equation is a very general and useful description of the rates of thermally activated processes. Activation energy can be illustrated as a jumping process over a barrier (Figure 1.4). According to Boltzman distribution, an atom at the temperature T has an excess of energy E_a with a probability $\exp(-E_a/kT)$. Higher temperature leads higher barrier crossing probability

$$\text{rate} = z(T)\exp(-E_a/kT) \tag{1.1}$$

$k = 1.38 \times 10^{-23}$ J/K or 8.62×10^{-5} eV/K.

A great many microfabrication processes show Arrhenius-type dependence: etching, resist development, oxidation, epitaxy, chemical vapor deposition (which are chemical processes) are all governed by exponential temperature dependencies, as are diffusion, electromigration and grain growth (which are physical processes).

The magnitude of the pre-exponential factor $z(T)$ and the activation energy E_a vary a lot. In etching reactions, activation energy is below 1 eV, in polysilicon deposition E_a is 1.7 eV, in substitutional dopant diffusion it is 3.5 to 4 eV and in silicon self-diffusion it is 5 eV.

1.6 LATERAL DIMENSIONS

Microfabricated systems have dimensions around 1 μm: some devices perform well with 5 or 10 μm structures, and others need 100 nm for good performance (Figure 1.5). But almost every device includes structures with ca. 100 μm dimension. These are needed to interface the microdevices to the outside world: most devices need electrical connections (by wire bonding or bumping process); microfluidic devices must be connected to capillaries or liquid reservoirs; solar cells and power semiconductors must have thick and large metal areas to bring out the high currents involved, and connections to and from optical fibres require structures about the size of fibres, which is also of the order of 100 μm.

Narrow individual lines can be made by a variety of methods; what really counts is resolution; the power to resolve two neighboring structures. It determines device-packing density. The resolution usually gets most of attention when microscopic dimensions are discussed, but alignment between structures in different lithography steps is equally important. Alignment is, as a rule of thumb, one-third of the minimum linewidth. High resolution but poor alignment can result in inferior device-packing density compared with poorer resolution but tighter alignment.

1.7 VERTICAL DIMENSIONS

As a rule of thumb, vertical and lateral dimensions of microdevices are similar. If the height-to-width,

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

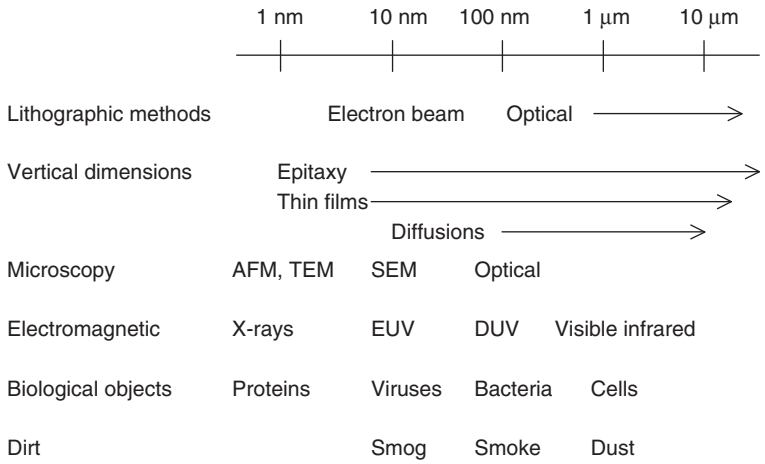


Figure 1.5 Dimension in the microworld. Note: $1\,\mu m = 10^{-6}\,m$; $1\,nm = 10^{-9}\,m$; $1\,Å = 10^{-10}\,m$; $1\,nm = 10\,Å$

or aspect ratio, is more than 2:1, special processing is needed, and new phenomena need to be addressed in such three-dimensional devices. Highly three-dimensional structures are used extensively in both deep submicron ICs and in MEMS.

Oxide thicknesses below 5 nm are used in CMOS manufacturing as gate oxides and as flash-memory tunnel oxides. Epitaxial layer thicknesses go down to an atomic layer, and up to 100 μm in the thick end. There are also self-limiting deposition processes, which enable extremely thin films to be made, often at the expense of deposition rate. Chemical vapor deposition (CVD) can be used for anything from a few nanometres to a few micrometres. Sputtering also produces films from 0.5 nm to 5 μm. Spin coating is able to produce films as thin as 100 nm, or as thick as 100 μm. Typical applications include polymer spinning, both photoresist as well as polymers that form permanent parts of devices. Electroplating (galvanic deposition) can produce metal layers of almost any thickness, up to 100 μm.

Photoresist thickness is an important parameter in determining resolution: it is easier to make small structures in thin photoresist layers (this is the same reason why slide films have better resolution than negatives). Typical resist thickness for ICs is 1 μm, but for MEMS devices, 10 μm, 100 μm or even 500 μm resist thicknesses are required, and nanodevices fabricated by e-beam often use 100 nm thick resist, and SAMs that are one molecule thick are not uncommon.

Etching of thin films can produce structures equal to thin film thickness. Etching of silicon wafers can produce structures with heights equal to wafer thickness, in the 500 μm range. Depth is one thing, profile is another: vertical walled structures are much more difficult to make than sloped walls. When two or more wafers are bonded together, structural heights of several millimetres are encountered.

1.8 DEVICES

Microfabricated device can be classified by many ways:

- material: silicon, III–V, wide band gap (SiC, diamond), polymer, glass;
- integration: monolithic integration, hybrid integration, discrete devices;
- active vs passive: transistor vs resistor; valve vs sieve;
- interfacing: externally (e.g., sensor) vs internally (e.g., processor).

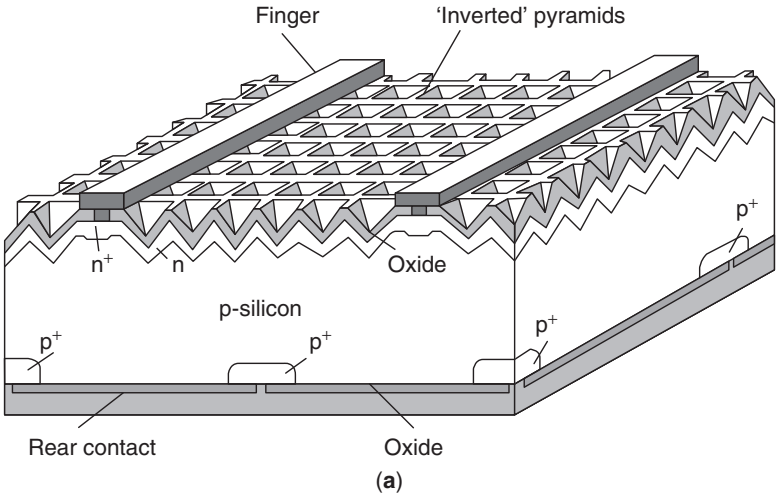
The above classifications are based on device functionality. In this book, we are concentrating on fabrication technologies, and then the following classification is more useful:

- volume (or bulk) devices;
- surface devices;
- thin film devices;
- stacked devices.

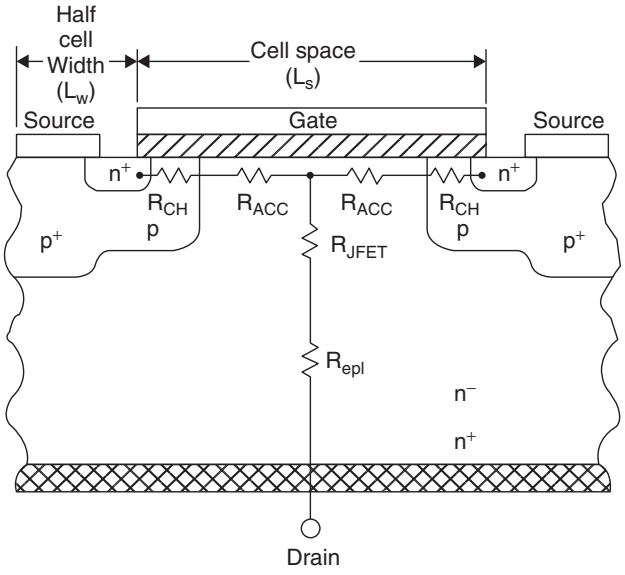
1.8.1 Volume devices

Power transistors, thyristors, radiation detectors and solar cells are volume devices: currents are generated

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030



(**a**)



(**b**)

Figure 1.6 Volume devices: (**a**) passivated emitter, rear-locally diffused solar cell. Reproduced from Green, A.M.: (1995), by permission of University of New South Wales. (**b**) n-channel power MOSFET cross section. Reproduced from Yilmaz, H. *et al*. (1991), by permission of IEEE

and transported (vertically) through the wafer (Figure 1.6), or alternatively, device structures extend through the wafer, like in many bulk micromechanical devices. The starting wafers for volume devices need to be uniform throughout. Patterns are often made on both sides of the wafer, and it is important to note that some processes affect both sides of the wafer and some are one sided.

1.8.2 Surface devices

Surface devices make use of the materials properties of the substrate but generally only a fraction of wafer thickness is utilized in making the devices. However, device structure or operation is connected with the properties of the substrate. Most ICs fall under this category: metal oxide semiconductor (MOS) and bipolar transistors, photodiodes and CCD image sensors.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

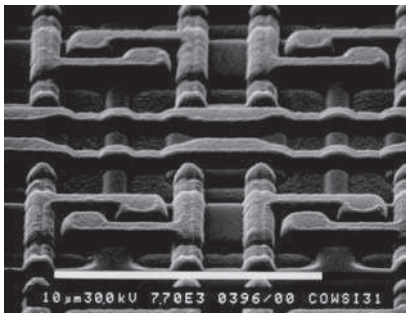


Figure 1.7 Surface devices: a 0.5 µm CMOS in a scanning electron microscope view

In silicon CMOS (Figure 1.7), only the top 5 µm layer of the wafer is used in making the active device, and the remaining 500 µm of wafer thickness is for support: mechanical strength and impurity control. Surface devices can have very elaborate three-dimensional structures, like multilevel metallization in logic circuits, which can be 10 µm thick but this is still only a fraction of wafer thickness; therefore the term surface device applies.

1.8.3 Thin-film devices

Devices can be built by depositing and patterning thin films on the wafers, and the wafer has no role in device operation. Wafer properties like thermal conductivity or transparency may be important (Figure 1.8), but the substrate is not machined or modified. Thin-film transistors (TFTs) are most often fabricated on non-semiconductor substrates: glass, plastic or steel. Surface micromechanical devices like switches, relays, DNA arrays, fluidic channels and gas sensors are often fabricated on silicon wafers for convenience but they could be fabricated on glass substrates as well.

1.8.4 Membrane devices

Membrane devices are a sub-class of thin-film devices: again, all functionality is in the thin top layer, but instead of full wafer mechanical support, only a thin membrane supports the structures. Many thermal devices are membrane devices for thermal isolation: thermopiles, bolometers, chemical microreactors and mass flow meters (Figure 1.9). Many acoustic devices also utilize bulk removal. Optical paths can be opened by removing the bulk semiconductor. X-ray lithography masks are gold or tungsten microstructures on a micrometre-thick membrane.

1.8.5 Stacked devices

Stacked devices are made by layer transfer and bonding techniques. Two or more wafers are joined together permanently. Devices with vacuum cavities, for example, absolute pressure sensors, accelerometers and gyroscopes are stacked devices made of bonded silicon/glass wafer pairs. Micropumps and valves, and

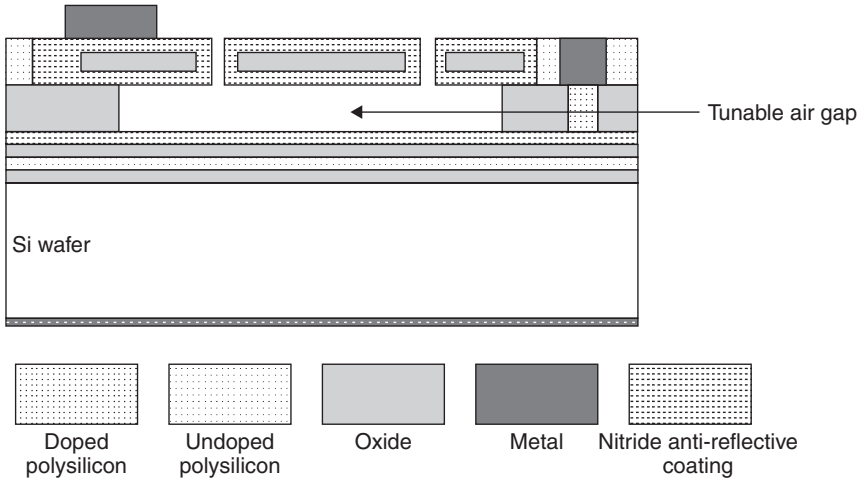


Figure 1.8 Surface micromachined Fabry–Perot interferometer: thick oxide has been etched away to create a tunable air gap. Silicon is transparent at infrared wavelengths, and radiation can enter the device through the wafer. Redrawn from Blomberg, M. *et al*. (1997), by permission of Royal Swedish Academy of Sciences

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

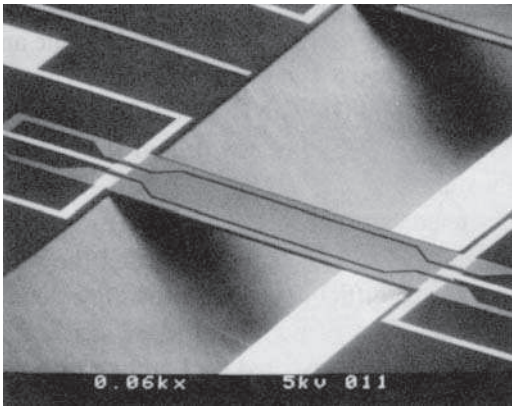


Figure 1.9 Mass flow sensor: a resonating bridge over an etched channel. Reproduced from Bouwstra, S. *et al*. (1990), by permission of Elsevier

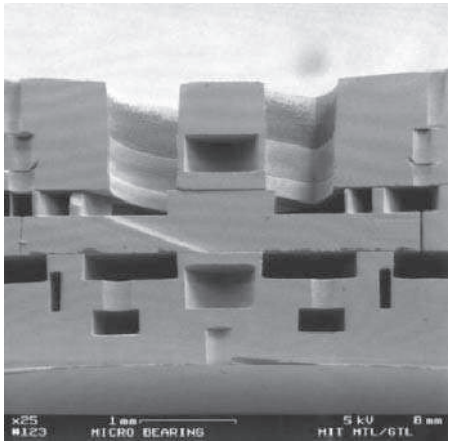


Figure 1.10 A microturbine by silicon-to-silicon bonding. Reproduced from Lin, C.-C. *et al*. (1999), by permission of IEEE

many micropower devices like turbines and thrusters are stacked devices with up to six wafers bonded together (Figure 1.10). More and more layer transfer and wafer bonding techniques are being developed, and stacked devices of various sorts are expected to appear; for example, GaAs optical devices bonded to Si-based electronics, or MEMS devices bonded to ICs.

1.9 MOS TRANSISTOR

The metal-oxide-semiconductor transistor, MOS, has been the driving force of microfabrication industries. It is the number one device by all measures: number of devices sold, silicon area consumed, the narrowest linewidths and the thinnest oxides in mass production, as well as dollar value of production. Most equipment for microfabrication have originally been designed for MOS IC fabrication, and later adapted to other applications.

The MOS transistor is a capacitor with silicon substrate as the bottom electrode, the gate oxide as the capacitor dielectric and the gate metal as the top electrode. Despite the name MOS, the gate electrode is usually made of phosphorus-doped polycrystalline silicon, not metal (Figure 1.11). The basic function of a MOS transistor is to control the flow of electrons from the source to the drain by the gate voltage and the field it generates in the channel. A positive voltage on the gate pulls electrons from the p-type channel to Si/SiO_2 interface where inversion occurs, enabling electron flow from n^+ source to n^+ drain.

The transistors are isolated electrically from the neighbouring transistors by silicon dioxide field oxide areas. This isolation eats up a lot of area, and therefore transistor-packing density on a chip does not depend on transistor dimensions alone.

Scaling down MOS transistor channel length makes the transistors faster. The other main aspect is area scaling: factor N linear dimension scaling reduces

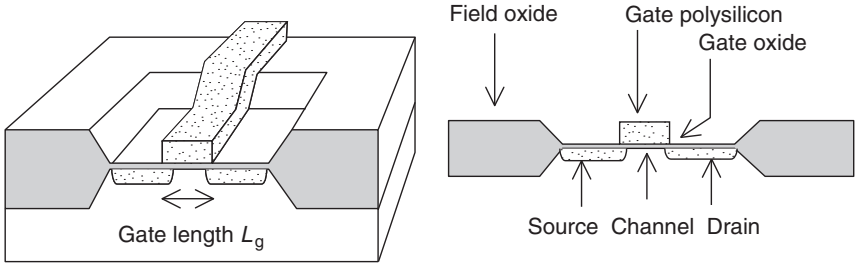


Figure 1.11 Schematic of a 5 μm gate length (L_g) MOS transistor: exploded view and cross section. Source/drain-diffusion depth is ca. 1 μm and gate oxide thickness ca. 0.1 μm. Field oxide thickness is ca. 1 μm and polysilicon gate thickness is 0.5 μm. Note that the z-scale has been exaggerated for clarity

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

area to A/N^2. Gate width, gate oxide thickness and source/drain-diffusion depths are closely related, and the ratios are more or less unchanged when transistors are scaled down. As a rough guide, for gate length of L, oxide thickness is $L/45$, and source/drain junction depth is $L/5$.

1.10 CLEANLINESS AND YIELD

Microfabrication takes place under carefully controlled conditions of particle purity, temperature, humidity and vibration because otherwise micrometre scale structures would be destroyed by particles or else lithography process would be ruined by vibrations or temperature and humidity fluctuations. Two cleanroom designs are shown in Figure 1.12: high-efficiency filters can be placed locally or they can have 100% coverage, offering improved cleanliness and laminar (unidirectional) airflow. Wafers are cleaned actively during processing: hundreds of litres of ultrapure water (de-ionized water, DIW) are used for each wafer during its fabrication. This is the dynamic part of particle cleanliness: the passive part comes from careful selection of materials for cleanroom walls, floors and ceilings, including sealants and paints, plus process equipment, wafer storage boxes and all associated tools, fixtures and jigs.

Even though extreme care is taken to ensure cleanliness during microprocessing, some devices will always be defective. As the number of process steps increases, the yield goes down as $Y = Y_o^n$, where Y_o is the yield of a single process step and n is the number of steps. With 100 process steps and 99% yield in each individual step, this results in 37% yield (representative of 64 kbit Dynamic random access memory (DRAM) chip) but 99% yield for a 500 step process (representative of 16 Mbit DRAM) results in <1% yield. Clearly, 99% yield is not enough for modern memory fabrication. Chip design also affects yield through area: $Y = \exp(-DA)$ where A is chip area and D is the defect density: making small chips is much easier than making big chips.

Yield has two major components: stochastic and systematic. Stochastic (random) defects are unpredictable occurrences of pinholes in protective films, particle adhesion on the wafer, corrosion of metal lines, and so on. Systematic defects come from equipment and operator failures, impurities in starting materials and design errors: two features are placed so close to each other that they will inadvertently touch, or impurities in chemicals do not allow low enough leakage currents.

Integrated circuit wafers contain typically a hundred or hundreds of chips (also called die), Figure 1.13. This number has remained more or less unchanged over decades because chip size and wafer size have grown in parallel: 0.2 cm^2 chips were made on 100 mm wafers while 2 cm^2 chips are usual on 300 mm wafers. In extreme cases, only one chip fits the wafer, for example, a solar cell, a thyristor or a position-sensitive radiation detector. Microfluidic separation devices with 5 cm long channels and optical waveguide devices with large radii of curvature can have a handful of devices per wafer. With standard logic chips or with micromechanical pressure sensors, thousands can be crammed to fit into a wafer.

1.11 INDUSTRIES

The electronics industry is based on semiconductor devices, which are based on silicon.

In 2002, ca. 10^{18} transistors were shipped, some 150 million for each and every human on earth. As recently as 1968, it was one transistor per year per person. The price, of course, explains a lot: in 1968, transistors cost ca. $1 a piece; in 2002, the cost was $0.000 0001.

Worldwide, about $6 billion is spent on silicon wafers annually. These are used to make $150 billion worth of semiconductor devices, which fuel the 1000 billion electronics industry. Other related businesses include the $25 billion semiconductor manufacturing equipment industry and the $15 billion materials industry (which includes for example chemicals, gases, photomasks and sputtering targets).

Microsystems industry as such does not exist: microsystems are rather a technology more than an industry; therefore, statistics are erratic. Some estimates put microsystems sales at $13 billion (2000), but this presents module prices (e.g., ink-jet cartridge; not just the silicon nozzle chip). Chip sales might be 10% of module prices, because microsystems packaging and testing are very complex. The flat-panel displays industry has sales of some $23 billion in 2000. It has more and more of its own suppliers for process equipment, and of course, for the glass plates used as substrates.

Device density on chips is quadrupling in three-year intervals, a trend known as *Moore's law*. Scaling has continued relentlessly for the past 40 years. Linewidths were in the 30 μm range in early 1960s, and they are 0.18 μm in the year 2000. Lithographic scaling has thus improved packing density by a factor $(30/0.18)^2 \approx 30\,000$. The number of transistors on a chip has

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

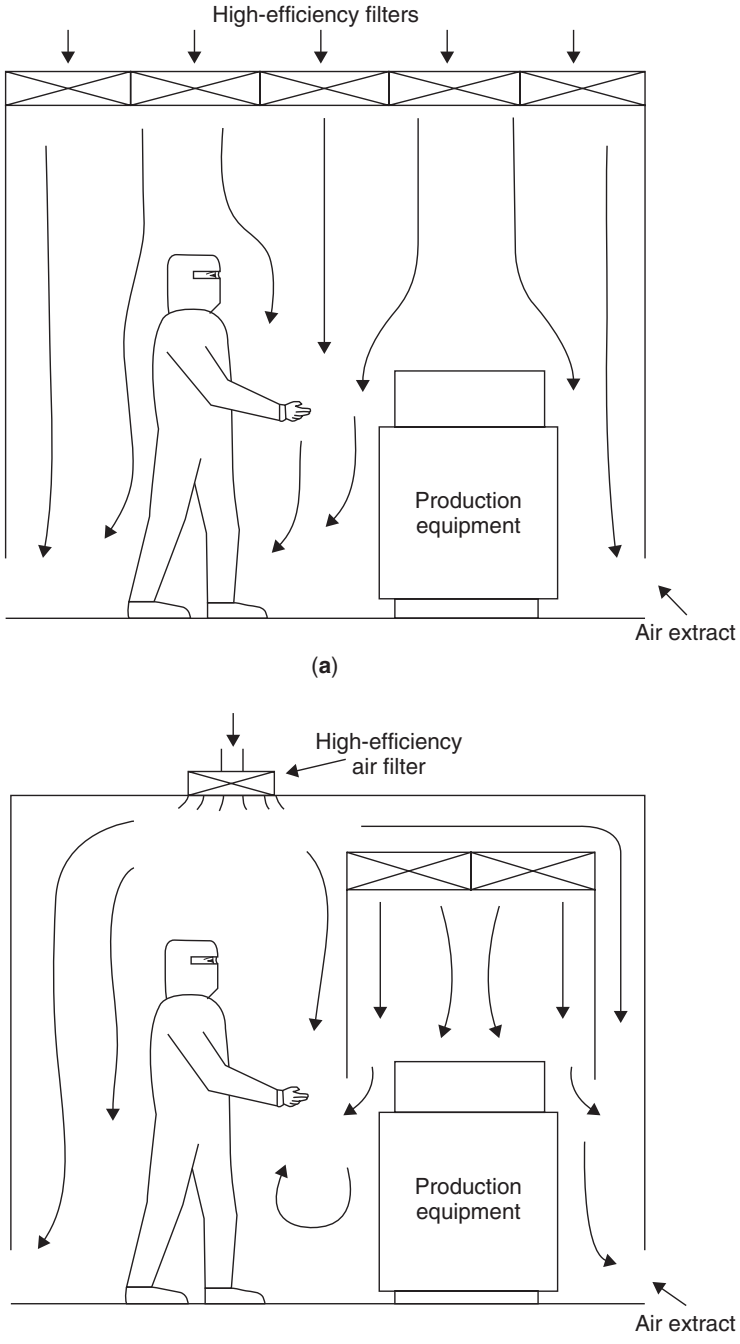


Figure 1.12 Two cleanroom designs: (**a**) laminar airflow in the whole room with 100% filter coverage and (**b**) laminar flow above process equipment only. Source: Cleanroom Design, 2nd edition, W. Whyte, 1999, © John Wiley & Sons, Limited

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

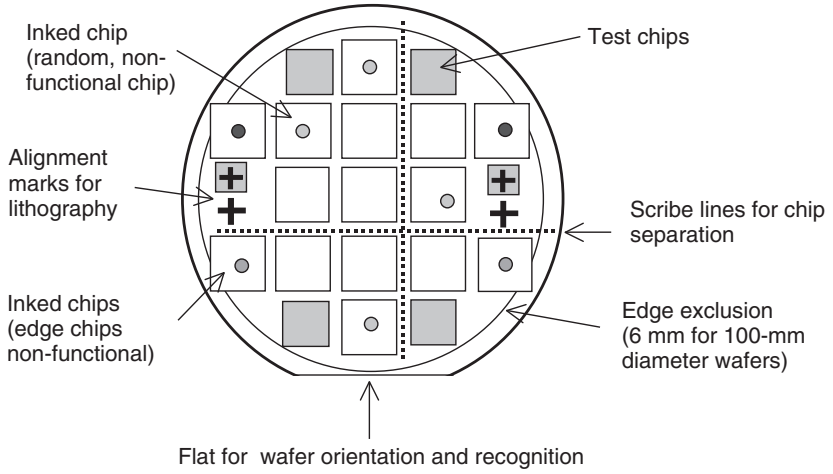


Figure 1.13 Silicon wafer with chips, test chips and alignment marks. Edge exclusion adds to non-saleable area. Non-functional chips have been 'inked'

increased form one to 100 000 000, however. The terms VLSI and ULSI, for Very Large Scale Integration and Ultra Large Scale Integration, respectively, are used today as synonyms for advanced chips, but historically they were measures of integration density: VLSI density was ca. 10^5 to 10^7 devices per chip, and ULSI referred to 10^7 to 10^9 devices per chip. The other two main factors have been chip-size increase, which has been possible by improvements in manufacturing techniques, and yield. This has contributed a factor of ca. 200 as chip size has increased from 1 mm^2 in 1960 to 2 cm^2 in 2000. The remaining factor of 10 has come from device and circuit cleverness: new designs, new fabrication processes and novel materials that use less area for same functionality.

IC technology generations are classified by their linewidths and each new generation has dimensions roughly 30% smaller than the previous. In the year 2003, the minimum linewidth in production is 0.13 μm but this presents just a fraction of all IC's manufactured. In fact, when counted as wafer starts, the distribution of linewidths was as follows:

≤0.13 μm	0.18–0.25 μm	0.35–0.5 μm	0.65–1 μm	>1.0 μm
15%	20%	20%	15%	30%

When counted as silicon area, the smaller linewidths gain importance because linewidth scaling has been accompanied by wafer-size increase which means that 0.13 μm devices are fabricated on 300 mm wafers but 1 μm devices on 100 mm wafers.

1.11.1 Note on drawings

The *z*-dimension is enlarged relative to *xy*-directions to make drawings easier to read. MOS transistor gate oxide is usually 2% of gate thickness, and if it were drawn to scale, it would not be seen. In bulk micromechanics, the diaphragm of a piezoresistive sensor is, for example, 20 μm, or 5% of wafer thickness, and the piezoresistor diffusion depth is 5% of diaphragm thickness, that is 1 μm. If the drawing is to scale, it will be specifically notified; all other figures in this book have *z*-scale enlarged for readability.

1.12 EXERCISES

1. The silicon atom density is 5×10^{22} cm^{-3}. If dopant concentration is 10^{15} cm^{-3} of boron, how far are the boron atoms from each other?
2. IC chips are getting larger even though the linewidths are scaled down because more functions are integrated on a chip. Calculate the signal path resistance for
   (a) 3 μm wide, 1 μm thick aluminium conductors, 500 μm long (resistivity 3 μohm-cm)
   (b) 0.3 μm wide, 0.5 μm thick, 1 mm long copper conductors (2 μohm-cm)
3. Silicon dioxide can sustain 10 MV/cm electric field. Calculate oxide thickness regimes for
   (a) CMOS ICs where operating voltages are 1 to 5 V
   (b) capillary electrophoresis (CE) microfluidic chips where 500 to 5000 V are used

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

4. Silicon is etched in plasma according to reaction Si (s) + $2Cl_2$ (g) → $SiCl_4$ (g). What is the theoretical maximum etch rate of a 200 mm diameter silicon wafers when chlorine flow is 100 sccm (standard cubic centimetres per minute)?
5. Accelerated tests for chips are run at elevated temperatures in order to find out failures faster. Acceleration factor temperature (AFT) is given by Arrhenius formula $\mathrm{AFT} = \exp(E_a/(1/kT_{\mathrm{operation}} - 1/kT_{\mathrm{test}})$. Use activation energy, 0.7 eV. What acceleration factor does 175 °C present? Temperatures are junction temperatures, and typical values are 55 °C for consumer and 85 °C for industrial electronics.
6. Aluminium wires do not tolerate current densities higher than 1 MA/cm^2. What are maximum currents that can run in micrometre aluminium wiring?
7. CMOS linewidths have been scaled down steadily by 30% every three years. In the year 2000, linewidths were in the range of 0.18 μm. When will linewidth equal atomic dimensions?

Comments, hints and answers to selected problems are presented in appendix A.

REFERENCES AND RELATED READINGS

Blomberg, M. *et al*: Electrically tunable micromachined Fabry-Perot interferometer in gas analysis, *Physica Scripta*, **T69** (1997), 119.

Bouwstra, S. *et al*: Resonating microbridge mass flow sensor, *Sensors Actuators*, **A21–A23** (1990), 332.

Green, A.M.: *Silicon Solar Cells*, University of New South Wales, Sydney, 1995.

Lin, C.-C. *et al*: Fabrication and characterization of a micro turbine/bearing rig, *Proc. MEMS '99* (1999), p. 529.

Whyte, W.: (ed.): *Cleanroom Design*, 2nd ed., Wiley, 1999.

Yilmaz, H. *et al*: 2.5 million cell/in^2, low voltage DMOS FET technology, *Proc. IEEE APEC* (1991), p. 513.

Solid State Technology Magazine: http://sst.pennwellnet.com/home.cfm

Semiconductor International Magazine: http://www.reed-electronics.com/semiconductor/

Materials database at http://www.memsnet.org/material/

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

2

Micrometrology and Materials Characterization

When micrometre lines are patterned and nanometre films are grown, measurement tools have to be available to characterize those processes. In addition to seeing and measuring those structures, we sometimes have to see details of the structures, and sometimes atomic level analysis is required, for example, to understand thin-film nucleation and interface quality. This is possible but time consuming, and it should not be mixed up with quick and simple methods that are used in everyday process monitoring.

2.1 MICROSCOPY AND VISUALIZATION

Optical microscopy resolution is similar to wavelength, that is, in the micrometre range. This is useful in many applications because we can always include test structures of any dimensions, irrespective of actual device dimensions. Dark field microscopes have illumination from the side, which gives an enhanced detection of steps and edges that reflect light up, and in confocal microscopy, light from focus depth alone is collected by the optical system. Fluorescence microscopy can be used to see organic residues on the wafer and Nomarski interference contrast images provide enhanced information about surface-height differences.

Scanning electron microscopy (SEM) has minimum resolution down to 5 nm, which makes it applicable to almost all microfabricated structures. In top view imaging, SEM is like optical microscope, except for the higher resolution. Its real power comes into play in tilted and cross-sectional views (Figure 2.1). Cross-sectional images can be used to obtain topographic information (photoresist sidewall angle, deposition step coverage) but at the expense of sample destruction and associated increase in analysis time. SEM resolution is, however, not enough for thickness determination of, for example, CMOS gate oxides.

Transmission electron microscope (TEM) provides ultimate image resolution, down to atomic imaging (Figure 2.2). High-resolution TEM (HRTEM) has a special advantage in calibration: lattice spacing of atoms can be used as accurate internal calibration standards.

2.2 LATERAL AND VERTICAL DIMENSIONS

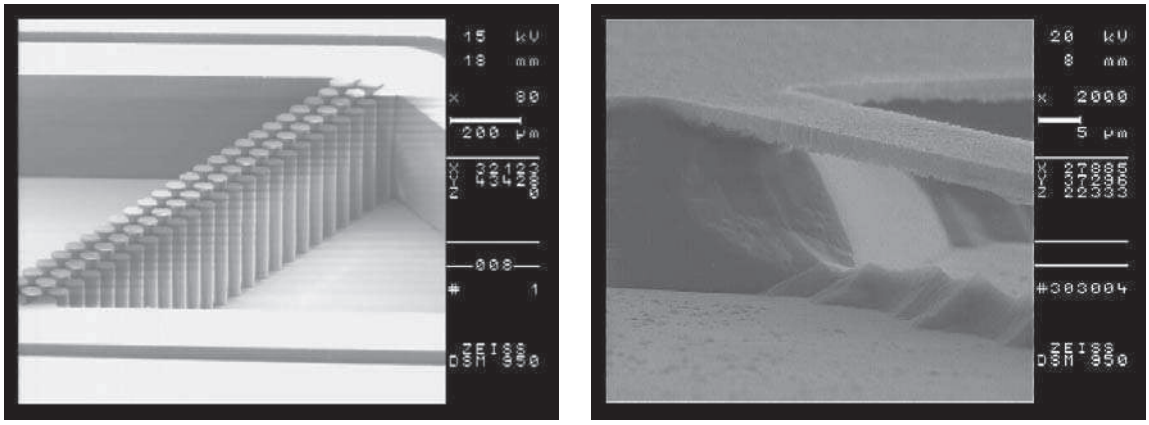
For device lateral dimensions, 10% deviation is usually accepted as fabrication tolerance. Measurement precision should be 10% of that variation, that is, 10 nm for 1 μm structures. For 100 nm structures, this translates to 1 nm, which is very difficult indeed.

Linewidth is often known as *critical dimension* (CD). All major CD measurements rely on scanning: an optical slit or aperture, a laser or electron beam spot or a mechanical stylus is scanned over the line. Linewidth measurement depends on edge detection in all these methods. This has both inherent and microstructure-related limitations. A signal from the edge is not a delta function even in the case of perfectly vertical sidewall. Beam spot and mechanical stylus alike have dimensions that are similar to microstructure dimensions and these lead to systematic errors in linewidth measurement. Needle radius of curvature determines the minimum line/space (pitch) that can be resolved. Both electromechanical stylus systems (known as *surface profilers*) and atomic force microscopes (AFM) can be used, but as can be seen from Figure 2.3, they seldom provide information about profile. The former have needle radius of curvature 1 to 10 μm, and the latter 1 to 10 nm.

Film thicknesses range from one atomic layer to hundreds of micrometres, and no single method can

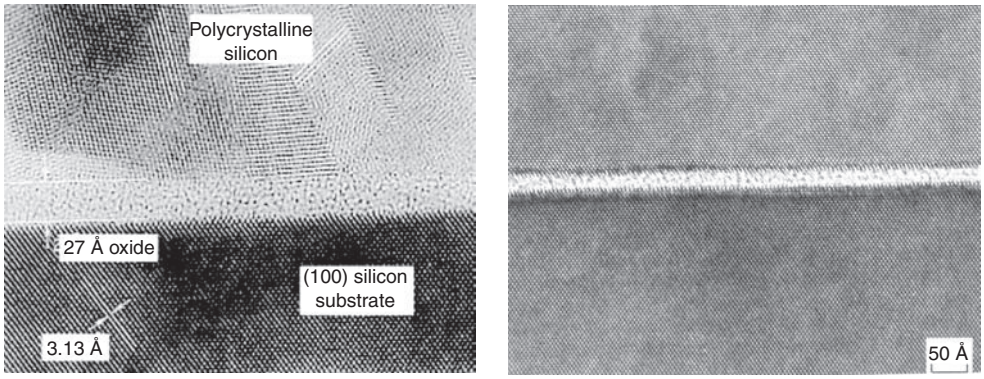
Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030



(**a**) (**b**)

Figure 2.1 Scanning electron microscopy: (**a**) a 400 µm thick SU-8 pillars in a microfluidic bead trap. Photo courtesy Santeri Tuomikoski, Helsinki University of Technology; (**b**) a heavily boron-doped silicon bridge. Photo courtesy Kestas Grigoras, Helsinki University of Technology



(**a**) (**b**)

Figure 2.2 High-resolution transmission electron micrographs (HRTEM): (**a**) single-crystal silicon/silicon oxide/polycrystalline silicon structure. From Buchanan, M. (1999), by permission of IBM; (**b**) bonded wafer interface: amorphous native oxide is seen between two single-crystal wafers. Source: Tong, Q.Y. & U. Gösele, Semiconductor Bonding, © Wiley, 1999. This material is used by permission of John Wiley & Sons, Inc

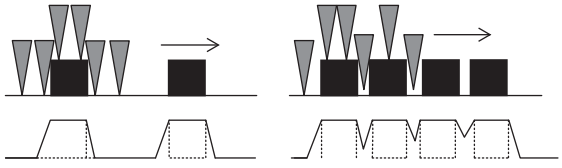


Figure 2.3 Scanning probe over vertical walled, isolated and dense lines. The scan profile is shown below. Linewidths of isolated lines are measured but the shape of the probe tip affects the line profile. In dense array, linewidth cannot be measured but pitch (line + space) can be

cover such a thickness range. Conductive and dielectric films must often be measured by different techniques but scanning probe methods are quite universal: a step is formed by etching and a probe-tip scans over the step. Z-scale precision can be 1 nm or even down to 1 Å, but in most practical cases, surface roughness sets the lower limit for step height/film thickness measurement.

Scanning tunnelling microscope (STM) can have atomic resolution. It is a research tool for surface science, but its relative, the atomic force microscope (AFM), which has nanometre resolution, is becoming a favourite metrology tool in microfabrication

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

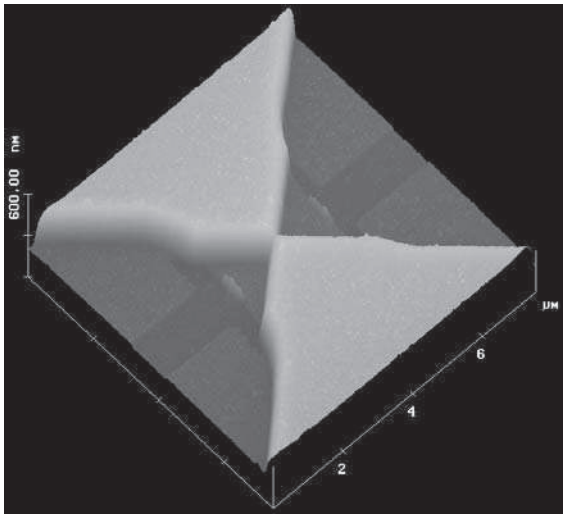


Figure 2.4 Atomic force microscope (AFM) tapping mode image of a quantum point contact structure on a SOI wafer. Thickness is ca. 100 nm and the neck lateral dimension is 20 nm. Picture courtesy Jouni Ahopelto, VTT

(Figure 2.4). AFM images provide not only surface images but also step height and linewidth data. AFM is also the standard method for measuring wafer-surface roughness.

Commonly used optical thickness measuring methods are ellipsometry and reflectometry. In ellipsometry, the complex reflection ratio and phase change are measured in a single measurement, and film thickness can be calculated when substrate optical constants are known from independent measurement. In reflectometry, a wavelength scan is made (e.g., 300–800 nm) and this is fitted to a reflection model. For very thin films, uncertainty is introduced because optical constants are not really constants, but depend on film thickness. X-ray reflection (XRR) can be used to measure film thickness. Unlike optical methods, XRR is insensitive to refractive index change. Measurement time, however, is in minutes or even hours, compared with seconds for optical tools.

2.3 ELECTRICAL MEASUREMENTS

A number of electrical measurements can be used to characterize substrates and deposited thin films: resistivity, conductivity type, carrier density and lifetime, mobility, contact resistance or barrier height. Resistivity is an important property of conducting layers but resistance is the property that can be measured easily. For

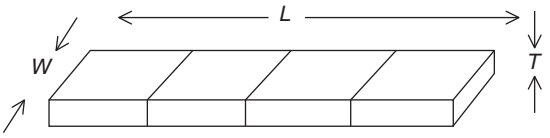


Figure 2.5 Conceptualizing metal line as a number of four square elements: $R = 4R_s$

a rectangular piece of conducting material, resistance is given by

$$R = \rho L/WT \tag{2.1}$$

where ρ is resistivity, L, length, T, thickness and W, width (Figure 2.5).

If we consider a square piece of metal, $L = W$, we can then define sheet resistance, R_s,

$$R_s \equiv \rho/T \tag{2.2}$$

where R_s is in units of ohm/square.

Sheet resistance is independent of square size. Resistance of a conductor line can now be easily calculated by breaking down the conductor into n squares: $R = nR_s$. Sheet resistances of doped semiconductor layers will be discussed in Chapter 14.

Measurement of R_s can be done in several ways: direct measurement necessitates the fabrication of metal line (lithography and etching steps), but the result follows easily:

$$R_s = R/n = V/nI \tag{2.3}$$

The four-point probe method uses two outer probe needles to feed current through the sample, and two inner needles to measure voltage, see Figure 2.6.

In semi-infinite case, resistivity is given by

$$\rho = (V/I)2\pi s \tag{2.4}$$

In the case of a thin-film of thickness T on an insulating substrate (e.g., Al film on SiO_2), resistivity is

$$\rho = (V/I)T(\pi/\ln 2) = 4.53(V/I)T \text{ or}$$
$$R_s = 4.53(V/I) \tag{2.5}$$

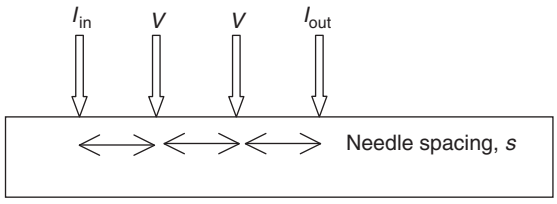


Figure 2.6 A four-point probe measurement set-up with identically spaced needles

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

When the sample size is 15 times larger than the probe spacing, resistivity is correct within 1%. For smaller samples, geometric correction factors need to be applied.

Thickness has to be measured independently. Alternatively, sheet resistance can be used to calculate thickness after thin-film resistivity is known (bulk values cannot usually be used).

Many electrical test structures have been devised for conductive films and doping structures. These are fast measurements, ideally suited for wafer mapping: sheet resistance measurement requires four pads for probe needles, and electrical linewidth measurements also require the same. Contact chains make do with two pads but generally 4-pad measurements, with separate feeds for current and voltage measurements, eliminate contact resistance parasitics. A combined 6-pad structure (Figure 2.7) can be used to measure both sheet resistance R_s and electrical linewidth.

In the six-terminal structure, sheet resistance is measured by driving current I_c through terminals 2 and 3 and measuring the voltage drop V_c across terminals 5 and 6.

$$R_s = (\pi/\ln 2)(V_c/I_c) \tag{2.6}$$

Bridge resistance R_b is the voltage drop between terminals 4 and 5, V_{45}, divided by current I_{13} driven through terminals 1 and 3. Linewidth is then simply,

$$W = R_s \cdot L/R_b \tag{2.7}$$

Assumption of a square cross-sectional profile usually holds fairly well for plasma-etched lines. Line length L is fixed on the photomask, and if $L >> W$, minor inaccuracies in lithography (for example, corner rounding) can be ignored. Diffusions can be measured similarly, but the assumption of profile needs to be accounted for.

Electrical test structures are implemented on test chips on the wafer, or alternatively, they can be embedded in the scribelines between chips. Test structures for wafer fab measurements can thus be discarded after the fabrication is completed. This saves area because the dicing saw requires a margin of ca. one hundred micrometres between the chips anyway, as shown in Figure 1.13.

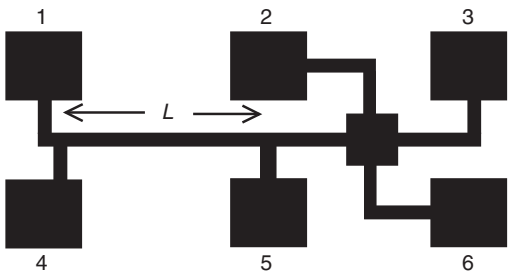


Figure 2.7 An electrical six-terminal test structure for sheet resistance and linewidth

2.4 PHYSICAL AND CHEMICAL ANALYSES

The measurement and characterization of microstructures differs from macroscopic structures and bulk materials in many respects. Small analysis areas and volumes limit available methods and sensitivities. Signal-to-noise ratio, S/N, is proportional to square root of the number of atoms probed:

$$S/N \propto \sqrt{\text{number of atoms probed}} \propto R\sqrt{z} \tag{2.8}$$

where R is the probing radius and z is the depth of analysis (cylinder volume $\propto R^2 z$)

The above formula explains why no single method can fulfil all microcharacterization needs.

One special aspect of semiconductor materials is their extreme purity: impurities are specified even at parts per trillion (ppt; 10^{-12} relative abundance) level. This is a relief in some cases because background signals are very low, but if the impurities themselves need to be measured, then we are in for some tough challenges.

Elemental concentrations are often needed: nitrogen in TiN thin films (50% for stoichiometric film), copper in aluminium (Al-0.5%Cu), phosphorous in oxide (5% by weight), boron in silicon wafers ($1 \times 10^{16}\,\text{cm}^{-3}$), oxygen in silicon (10–20 ppma, parts per million atoms), sodium impurity in tungsten sputtering target (ppb, parts per billion), or iron in silicon (ppt). These different concentration levels result in a fairly wide range of analytical methods that must be employed.

Elemental detection can be accomplished with many methods quite readily, but quantification is often difficult. Comparative results are often presented: treatments A, B, C versus reference sample. Treatments might represent new plasma CVD oxide processes and thermal oxide is used as reference; or the treatments are different annealing conditions with the unannealed sample as a reference.

2.5 XRD (X-RAY DIFFRACTION)

Structural information, that is, crystal orientation, texture and grain size, is important in a number of cases. Resistivity of metal film can increase by an order of magnitude upon phase change, and polycrystalline silicon final grain size distribution after annealing is dependent on

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

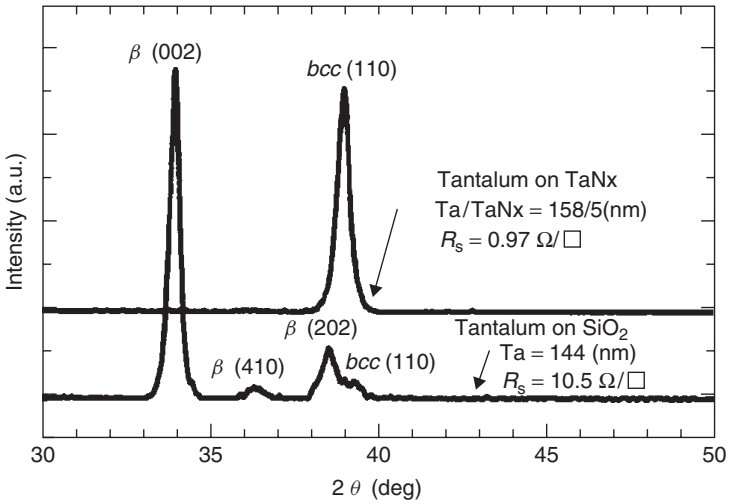


Figure 2.8 X-ray diffraction of tantalum thin films: the underlying material has a major effect on film crystal structure and resistivity. Reproduced from Ohmi, T. (2001), by permission of IEEE

the initial state: amorphous and polycrystalline silicon behave differently upon subsequent annealing. X-ray diffraction provides structural information (Figure 2.8). TEM also provides similar information, but TEM analysis area is in tens of nanometres, whereas XRD gives an average over hundreds of micrometres.

2.6 TXRF (TOTAL REFLECTION X-RAY FLUORESCENCE)

If minute amounts of matter on wafer surface must be analysed, total reflection can be used. A method known as *total reflection X-ray fluorescence* (TXRF) provides atomic identification by X-ray fluorescence, that is, characteristic X-ray radiation. TXRF can measure surface impurities at a level of $10^{10}\,\mathrm{cm}^{-2}$.

2.7 SIMS (SECONDARY ION MASS SPECTROMETRY)

In SIMS, the surface to be analysed is bombarded by ions that detach secondary ions. These secondary ions are mass-analysed, giving their identity. SIMS is thus a surface-sensitive technique, but another important SIMS application is depth profiling: the ion beam erodes the surface, and layers beneath the surface become available

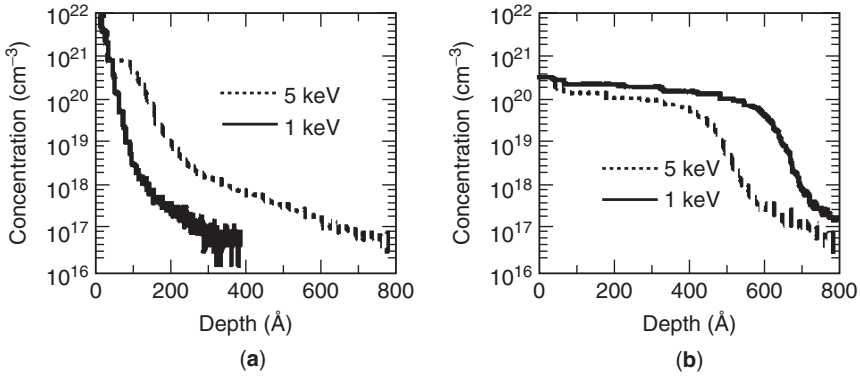


Figure 2.9 SIMS data of low-energy arsenic implantation into silicon with two different energies: (**a**) immediately after implantation; (**b**) after 1050 °C, 10 s heat treatment. Reproduced from Plummer, J.D. & P.B. Griffin (2001), by permission of IEEE

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

for analysis. When the erosion rate is known, SIMS data provides information about atomic concentrations as a function of depth.

SIMS measurement is slow and expensive, but it is the accepted standard for dopant depth distribution measurement (even though we are most often interested in electrically active dopants, whereas SIMS only counts atoms). SIMS offers nanometre depth resolution and 10^6 dynamic range (Figure 2.9).

2.8 AUGER ELECTRON SPECTROSCOPY (AES)

In Auger measurement an electron beam (3–5 keV) hits the surface, and an inner core electron is ejected. An electron from an outer shell fills the hole, and gives off excess energy during transition. Another outer shell electron receives this energy and escapes. The energy of this Auger electron is uniquely determined by the atomic structure, and therefore the identity of the element giving rise to the signal can be determined. The escape depth of low energy Auger electrons is of the order of nanometer, which makes Auger a truly surface sensitive technique. Auger can identify surface atoms, be they residues from previous steps or contaminants from processes. Auger is therefore a tool for surface chemical analysis (Figure 2.10).

With the aid of sample erosion technique (similar to SIMS), Auger can be transformed into a depth-profiling technique: after surface analysis, sputtering removes some material, and the Auger measurement of the newly formed surface is made. This is continued until the desired sample depth is probed.

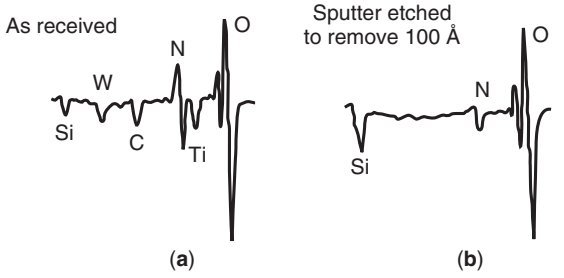


Figure 2.10 Auger analysis of silicon dioxide surface: (**a**) evidence of titanium and tungsten residues; (**b**) after sputter etching has removed 100 Å (10 nm) surface layer, the sample has been reanalysed and found free of Ti and W. Reproduced from Schaffner, T.J. (2000), by permission of IEEE

2.9 XPS (X-RAY PHOTOELECTRON SPECTROSCOPY)/ESCA

The X-ray photoelectron spectroscopy (XPS) is closely related to Auger in two senses: low-energy electrons are analysed, and because their escape depth is so small, the method is surface-sensitive, but XPS excitation is by X-rays. This has an important ramification for the analysis area: X-ray spots are fairly large, in the hundred micrometre range, and large areas are needed for analysis.

Primary X-rays (a few kilovolts) eject electrons from the sample. The energy of ejected electrons is related to their binding energy, and this enables not only elemental identification but also chemical bond identification. Electron energy is slightly different depending on bonding, and, for example, C–O, C–F and C–C bonds can be distinguished. The other name for XPS, ESCA, (electron spectroscopy for chemical analysis) emphasizes this important feature of XPS.

2.10 RBS (RUTHERFORD BACKSCATTERING SPECTROMETRY)

Rutherford backscattering spectrometry (RBS) is based on elastic recoil collisions. Helium ions (alpha particles) penetrate matter and slow down, but one ion in a million experiences 180° elastic recoil, and bounces

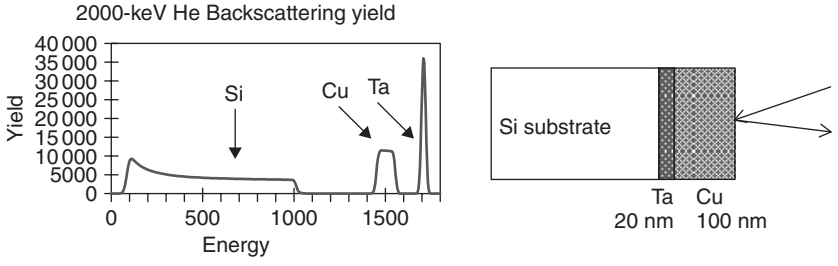


Figure 2.11 RBS spectrum of Si/Ta/Cu (20 nm/100 nm) sample: even though tantalum is beneath copper, its signal is at a higher energy because tantalum is so much heavier. Figure courtesy Jaakko Saarilahti, VTT

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

back towards the surface, slows down on the way back, and finally emerges from the solid and reaches the detector. All these steps can be handled calculationally, since RBS is a quantitative method. Elastic recoil from heavy atoms is more pronounced, and RBS is ideally suited for atoms like arsenic, tantalum, copper or tungsten.

Signal energy is sometimes confusing because it depends not only on the depth at which it originates but also on the mass of the atom that caused backscattering. In Figure 2.11, a tantalum barrier beneath copper has been measured by RBS. Silicon signal is weak because silicon is a light atom and beneath copper and tantalum. Copper is the topmost layer, but because it is lighter than tantalum, its peak is lower in energy.

RBS detectability depends on matrix: elements lighter than the matrix are not readily detectable. Oxygen and nitrogen analysis on top of silicon wafers are therefore difficult for RBS. Mass separation between neighbouring elements is poor in RBS, and therefore silicon, aluminium and phosphorous cannot readily be resolved. The RBS-detection limits are around $10^{20}\,\mathrm{cm}^{-3}$, but with heavy elements, it even goes down to $10^{17}\,\mathrm{cm}^{-3}$ (0.001%).

2.11 EMPA (ELECTRON MICROPROBE ANALYSIS)/EDX (ENERGY DISPERSIVE X-RAY ANALYSIS)

Electron beams can be focussed down to 5 nm spots, and the devices can be probed for localized analysis. The electron beam diverges as it interacts with the matter. The scattering of electrons spreads the beam to a volume much larger than the beam spot on the surface, as shown in the Figure 2.12. Auger electrons, which originate at the very surface, are unaffected by this spreading, but X-rays and backscattered electrons that are generated deep inside the sample can escape and reach the detector.

The radius of X-ray signals can be estimated by

$$R_x(\mu\mathrm{m}) = 0.04\,V^{1.75}/\rho \tag{2.9}$$

where the acceleration voltage is given in kilovolts and the density in grams/cm^3. The analysis radius R is given by

$$R = \sqrt{R_x^2 + d^2} \tag{2.10}$$

where d is the beam spot diameter.

This radius of electron microprobe analysis (EMPA) (a.k.a. *EDX* or *energy dispersive X-ray analysis*) can be orders of magnitude bigger than the electron beam spot size. EMPA/EDX can detect elemental concentrations at 1% level. Examples of suitable analytical tasks include phosphorous determination in doped oxide (5% wt typical) or copper concentration in aluminium film (0.5–4% Cu typical). EMPA/EDX is most often connected to a SEM, which is used to image the area of interest first, and then subjected to elemental analysis by EMPA/EDX. If the sample is made thin, of the order of 100 nm, electron scattering effects can be eliminated. This is utilized in transmission electron microscopy (TEM) and electron energy loss spectroscopy (EELS).

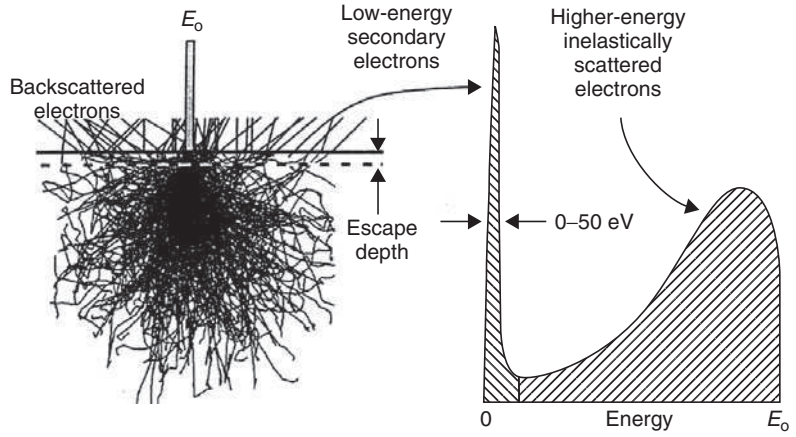


Figure 2.12 A finely focussed electron beam hits the sample surface, and low-energy secondary electrons escape from the surface only, but backscattered and inelastically scattered electrons contribute to signals deep inside the sample. Reproduced from Schaffner, T.J. (2000), by permission of IEEE

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

2.12 OTHER METHODS

Unfortunately, most methods are limited to certain elements only. The only exception is SIMS, which can detect every element from hydrogen to uranium. Auger spectroscopy cannot detect H, He or Li because of fundamental limitation of the three-electron Auger process, but all other elements that are detectable. X-ray methods are insensitive to light elements: depending on X-ray window design, boron ($m = 11$) can be detected, but sometimes fluorine ($m = 19$) or sodium ($m = 23$) is the lightest detectable element.

Infrared spectroscopy measures absorption due to molecular vibrations that are around 10 μm wavelength. It gives information about chemical bonds, because infrared vibrations are typically bond stretching and bending vibrations. Si–O bonds are desirable in silicon dioxide, but Si–H bonds indicate unwanted atomic arrangements and potential reliability problems. Si–F bonds on an etched surface hint at polymeric residue formation mechanism and help in designing the removal process. Infrared spectroscopy is most often practiced using an interferometric measurement set-up known as FTIR, for *Fourier-transform IR*. It is used to measure oxygen and carbon concentrations in silicon wafers, as revealed by optical absorption in 8 to 17 μm wavelength range.

Bulk wafers can be analysed by charge-carrier excitation methods such as microwave photoconductive decay (μPCD) and surface photovoltage (SPV). In μPCD, the sample is excited by a laser beam that creates excess-charge carriers. The amount of these carriers over time is measured in a non-contact arrangement by microwave reflection. Charge-carrier lifetime can be correlated with impurities and defects in the semiconductor material.

Neutron activation analysis (NAA) detects gamma quanta that have been excited by neutrons. NAA can detect selected elements at concentrations as low as $10^{11}\,\mathrm{cm}^{-3}$ (Cu, Ag, Au) and many others at concentrations $<10^{13}\,\mathrm{cm}^{-3}$ (Fe, Zn, Ni).

X-ray tomography (XRT) images full wafers with micron resolution. This is not enough for most crystallographic defects as such, but local stresses around defects often extend to many microns, so the method can indirectly see small defects.

If the material to be analysed can be extracted from the wafer, a much larger repertoire of analytical methods can be used. Thermal desorption spectroscopy (TDS) analyses desorption products upon heating. If the material can be dissolved in acid, atomic adsorption spectroscopy (AAS) and other methods of standard chemical analysis become available.

2.13 ANALYSIS AREA AND DEPTH

Analysis methods differ fundamentally in their analysis depth:

- surface-sensitive methods
- bulk methods
- micrometre methods.

Surface-sensitive methods probe only the topmost atomic layers, a nanometre or two.

Methods that analyse low-energy electrons are surface-sensitive because the escape depth of low-energy electrons is just a few nanometres. Auger electron spectroscopy and X-ray Photoelectron Spectroscopy are examples.

Diffusion depths and film thicknesses are often of the order of one micrometre. Analysis techniques that extend this deep would be very useful, but only a few exist. Rutherford backscattering spectrometry (RBS) has a typical analysis depth of around micron (for helium ion energy of 2 MeV). Electron beam–induced X-ray fluorescence also probes at ca. micron depth. The combination of sputter erosion and surface-sensitive analysis is commonly adopted for top micrometre analysis: ion-beam sputtering removes material and the newly formed surface is probed by, for example, Auger or SIMS.

Optical beam spots are micrometre-sized and they can be used to measure within a real device structure. However, some optical methods such as ellipsometry require ca. 100 μm analysis area. Because X-rays cannot be focussed, X-ray methods require typically rather large areas, in the millimetre range. Ion beams can be focussed to submicron spots in focussed ion beam (FIB) equipment, but most applications use broad beams, in the millimetre range.

Analysis must be done not only on microfabricated structures themselves but also on defects and non-idealities that are smaller than the device dimensions. If the chemical composition or structure of defects has to be identified, it is even more demanding than analysis of regular microstructures. Contaminants often come in quantities too small for even the best analytical methods. Vacancies and other point defects are smaller than the resolution of even the best microscopic methods. Indirect methods, such as carrier lifetime measurements (defects act as traps for charge carriers), positron annihilation spectroscopy (PAS) (positron lifetime is longer in material with voids) or photoluminescence (identification of defects by their recombination

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

radiation) or Raman spectroscopy (structural defects, implant damage, local stresses shift photon energy), must be used.

2.14 PRACTICAL ISSUES WITH MICROMETROLOGY

Many analytical methods can produce accurate results only at the expense of great time and effort: TEM can image individual atoms but the analysis time is days (it consists mostly of tedious sample preparation and also of complicated analysis). TEM analysis costs ca. $1000 to $2000 per sample if bought as a service.

Monitoring must be preferably so fast that whole wafer mapping can be performed for uniformity checking. Mapping measurement also requires that the analytical equipment can handle whole wafers. Many optical and electrical measurements are suited for mapping, but most physical and chemical methods require wafer breakage for sample preparation.

Uniformity can be defined across the wafer (a.k.a. within-wafer non-uniformity, WIWNU), wafer-to-wafer (WTWNU) and lot-to-lot. The standard definitions for uniformity are

$$U = (\max - \min)/2 \times \text{average}$$

$$U = (\max - \min)/(\max + \min) \qquad (2.11)$$

The former is applied when five measurements are taken, one at the wafer centre and four at 90° from each other at half-radius; the latter when the four points are at wafer edges.

Uniformity of 5% was long accepted as a typical process performance (thin-film thickness, etch rate), but some processes are inherently better, for example, thermal oxidation and photoresist spinning routinely produce better than 1% uniformity. On the other side, CMP (chemical–mechanical polishing) is notoriously non-uniform, with 10% as good uniformity.

2.14.1 Contact versus non-contact measurements

Measurements can be divided into two categories: contact and non-contact (non-invasive). Both modulated photoreflectance and four-point probe can be used to monitor ion implant dose, but 4PP makes physical contact to the wafer with metal (tungsten) needles, and the wafer is deemed contaminated. It is not allowed to continue into high-temperature steps.

Linewidth measurement by a SEM is non-contact as opposed to stylus profiler or AFM, which make contact with the wafer. Because full wafers are analysed in a linewidth SEM, only top view pictures are possible, and no cross-sectional information can be obtained.

2.14.2 Blanket versus patterned wafer analysis

Both in R&D and in production, analytical methods are bound by a number of practical constraints related to the number of data points, measurement spot size and speed of measurement. Blanket wafer measurements are simple to perform and many basic studies in film deposition, diffusion, ion implantation, polishing or bonding can be done on blanket wafers but in many cases structured wafers are indispensable. Linewidths and spacings need to be identical to product wafers, but more amenable to probing, by optical or electron beams, or by mechanical probes. Test-structure size needs to be matched to design complexity: if the product chip has 1 000 000 contact holes, how to extrapolate from 1000 hole test structure? The one-million contact test structure would probably be so large that no other test structures could be accommodated in the area allocated for testing.

2.14.3 Destructive versus non-destructive analysis

Cost of measurement can range from a few cents to a few dollars per wafer, but if the measurement is wafer destructive, its cost is at least the wafer cost, or $10 to $100 per sample. Many physical measurements are destructive, like SIMS, Auger depth-profiling and cross-sectional SEM. But care should be made between wafer destructive and sample destructive measurements. RBS analysis is performed on 1 cm^2 pieces; that is, the wafer has to be broken for RBS analysis. But after RBS analysis, other analyses can be done, for example, EMPA or SIMS. But after SIMS, depth profiling the sample is irrevocably lost.

2.14.4 Standards and reference materials

Calibration standards (with traceability to NIST, National Institute of Standards and Technology) and reference materials (which are supplier-certified) are available for all major wafer-level measurements: film thickness and step height, dimensions, electrical resistivity and particles. Reference materials are enough for daily work but they must be calibrated against traceable standards regularly.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

The standards and references are silicon wafers with dedicated test patterns for quantities in question. One wafer can provide a series of standards, such as different resistivity windows or steps heights. General step height standard is usually a quartz piece with etched steps; and not a separate piece for each specific material.

2.14.5 Devices as measuring instruments

It is not unusual that no analytical method is able to do a good job: either the quantities involved or the analysis areas are too small. Quite often it is possible to use devices themselves as measuring instruments: device performance degradation is attributed to minutiae effects that are not amenable to direct physical measurements. Metal Oxide Semiconductor (MOS) transistors are sensitive to metal contamination at levels below analytical detection limits (in the $10^9\,\mathrm{cm}^{-3}$ range). Microscopic vacuum cavities are created by wafer bonding or deposition, and no pressure gauge is small enough to probe these cavities. But mechanical quality factor, Q, of the microfabricated mechanical resonators in the cavities is indicative of cavity pressure.

2.14.6 Failure analysis and reverse engineering

Analytical methods are needed not only during fabrication, but also after wafer processing has been completed. When circuits are found malfunctional, either in testing or after field return, the causes must be identified. Hard errors, that is, consistent failures are much easier to locate and to understand than soft errors, that is, the intermittent failures that may take place only under certain operating conditions (for example above certain temperature or frequency). As in wafer-level analysis, non-destructive methods are tried first, and the destructive only afterwards.

In reverse engineering, a chip is 'disassembled' step by step, and the structures, materials and functions are recorded (see Figure 27.5 for IC metallization stripped of all dielectric films). This is practised for example for competitive intelligence or patent infringement examination. Methods like electron beam–induced current (EBIC) can be used to probe electrical functions of a circuit.

2.15 EXERCISES

1. The sheet resistance of a typical aluminium metallization is 0.03 ohm/sq. What is aluminium thickness?
2. Resistance of 200 μm long copper lines was measured to be 40 ohm. From copper deposition process we know that thickness is 300 nm. What is the linewidth?
3. AFM scan area is $1 \times 1\,\mu\mathrm{m}$, which corresponds to 512×512 pixels. What should the AFM-tip radius be so that resolution is tip-limited?
4. Estimate the analytical radius of electron microprobe (EMPA).
5. Can RBS be used to measure dopant profiles?
6. If electron beam is focussed to a 15 nm spot, and at least 100 Auger events (electrons) must be collected to get a signal, what is the detection limit of Auger microprobe?
7. SIMS raw data is ion counts versus sputter time. How can you convert these to concentration versus depth data?
8. What is the acceleration voltage of an atomic resolution TEM?
9. What are the resistivities of bcc-Ta and β-Ta in Figure 2.8?

REFERENCES AND RELATED READINGS

Buchanan, M.: Scaling the gate dielectric: materials, integration and reliability, *IBM J. Res. Dev.*, **43** (1999), 245.

Diebold, A.C.: Materials and failure analysis methods and systems used in the development of and manufacture of silicon integrated circuits, *J. Vac. Sci. Technol.*, **B12** (1994), 2768.

Ohmi, T.: A new paradigm of silicon technology, *Proc. IEEE* (2001), p. 394.

Plummer, J.D. & P.B. Griffin: Material and process limits in silicon VLSI technology, *Proc. IEEE' 89* (March 2001), p. 240.

Runyan, W.R. & T.J. Schaffner: *Semiconductor Measurements and Instrumentation*, McGraw-Hill, 1998.

Schaffner, T.J.: Semiconductor characterization and analytical technology, *Proc. IEEE' 88* (2000), p. 1416.

Schroder, D.K.: *Semiconductor Material and Device Characterization*, 2nd ed., John Wiley & Sons, 1998.

Tong, Q.Y. & U. Gösele: *Semiconductor Wafer Bonding*, John Wiley & Sons, 1999.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

3

Simulation of Microfabrication Processes

Microfabrication processes consist of tens or hundreds of steps that take weeks or months to complete, and therefore the learning cycles can easily become too long. Simulation is one way of shortening the learning cycles. Simulation accuracy is strongly dependent on the details of the process to be simulated, and even a simple simulator can be extremely valuable if it saves enough experimentation time and effort. Simulators can provide meaningful trend data and comparisons between different process options, even though the accuracy might be less than perfect. Simulators can be used to explore possibilities and narrow down options before the experimental work is begun. Simulation can provide information that is not experimentally available or is difficult to measure. Because there is no dopant profiling method with sub-10 nm resolution in both vertical and lateral directions, simulation is the *de facto* method for a two-dimensional dopant distribution analysis.

There are two breeds of process simulators: integrated packages that can be used to simulate the whole fabrication process with many different steps in sequence and dedicated simulators for specific process steps. Dedicated simulators are available for almost all processes, ranging from ion-implantation damage production to lithography defect modelling, to crystal structure prediction of deposited films. Dedicated simulators are more detailed, more accurate and more computation intensive. A basic principles diffusion simulator would start with lattice parameters, interatomic potentials, vacancy production and annihilation rates and atom-defect interactions, and provide diffusion profiles as the output. Integrated packages use simpler models, for instance, macroscopic phenomenological diffusion models based on Fick's equations, but they offer seamless stitching of different process steps into whole processes. Bulk silicon process steps, that is, high-temperature steps that affect dopant distribution inside silicon, epitaxy, diffusion, implantation and oxidation, can be analysed by solving the relevant diffusion equations.

Etching, polishing and deposition produce topography on a wafer. This build-up of topography is difficult to simulate because it involves multiphysics and chemistry – plasmas, fluid dynamics and surface chemical reactions. Film deposition simulators depend on atom arrival angles that are not physical constants like diffusivities but are parameters sensitive to experimental conditions. Etching reactions are complex interactions between the chemical contributions (spontaneous etching, free energy considerations) and physical processes (e.g., ion bombardment enhanced desorption). Topography process simulators are usually semiempirical: some important model parameters are extracted from experiments without fundamental physical validation.

Even though simulation is fast, simulator building is slow and tedious. It is not possible to build simulators for all possible new materials, processes and devices, because the calibration data needs to be available, and it is readily available only for those materials, processes and devices that are widely studied and used. In this sense, the predictive power of process simulation remains poor.

3.1 TYPES OF SIMULATION

Process simulation, device simulation and circuit simulation together are termed TCAD, for technology CAD (Figure 3.1), in contrast to the more established ECAD, electronic simulations, which involve logic and systems simulations. Process simulation deals with physical structures such as atoms and their distributions, device simulation deals with currents and potentials in devices, and circuit simulation is used to study larger circuit blocks. The dopant concentrations produced by a process simulator are used as an input for the device simulator,

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Process simulation
-structures
-dopant profiles
-layer thicknesses
= = > input to device simulation

Device simulation
-electrical, mechanical, thermal, optical behaviour
-current-voltage, force-displacement, potential-flow
= = > input to circuit simulation

Circuit simulation
-output signal and noise
-rise time, speed, delays

Figure 3.1 Levels of simulation

and the device simulator results form the starting material for circuit simulation (Figure 3.1).

Circuit simulation is the most advanced and process simulation is the least developed of the three kinds of simulations. Device simulators for CMOS today are predictive because CMOS device physics is well understood. Of course, continuous scaling to smaller linewidths means that new phenomena must be implemented into process and device simulators regularly.

3.2 1D SIMULATION

A one-dimensional simulator treats matter as layers, and the simulation outputs are layer thicknesses and dopant distributions in the vertical direction (Figure 3.2). One-dimensional simulation has been used since the 1970s when SUPREM from Stanford University emerged. Diffusion, ion implantation, oxidation and epitaxy are treated. Two additional, non-physical process steps are included: film deposition and etching, but these are just geometrical steps, like 'add 500 nm of undoped oxide on silicon', or 'remove the top 50 nm of silicon by etching'. These steps are needed for more realistic models of surfaces and interfaces, but they do not reveal anything about the deposition or etching processes.

Over the years, more layers and more realistic models have been added to 1D simulators, for instance, some simulators can handle the oxidation and doping of polycrystalline silicon. Polycrystalline materials require more inputs than single crystals, for example, grain size and texture, and assumptions of grain boundary diffusion versus bulk diffusion, among others. ICECREM (from Fraunhofer Institute FhG/IIS, Erlangen) is an advanced one-dimensional simulator. It can simulate the following processes:

- epitaxy
- oxidation
- diffusion
- ion implantation
- deposition of undoped oxide films (protective capping layers)
- deposition of doped oxide films (diffusion sources)
- etching (of oxide and silicon).

ICECREM models can account for a number of important real life effects such as high phosphorus concentration in diffusion, implantation through oxide and oxidation enhanced diffusion (OED). These features will be discussed in Chapters 13, 14 and 15. ICECREM output consists of diffusion profiles, oxide thicknesses, sheet resistances and junction depths. Sensitivity analysis can be carried out to study both process-parameter and model-parameter changes.

A typical simulator input file begins with the substrate definition (crystal orientation 100 or 111, doping type and level/resistivity). Grid is defined next: simulation depth is fixed (e.g. 5 μm, and grid spacing is defined (e.g. 0.01 μm). Concentrations that need to be calculated usually range from $10^{15}\,cm^{-3}$ to $10^{21}\,cm^{-3}$. Process steps are then defined in sequence, followed by output commands. Model parameters can be

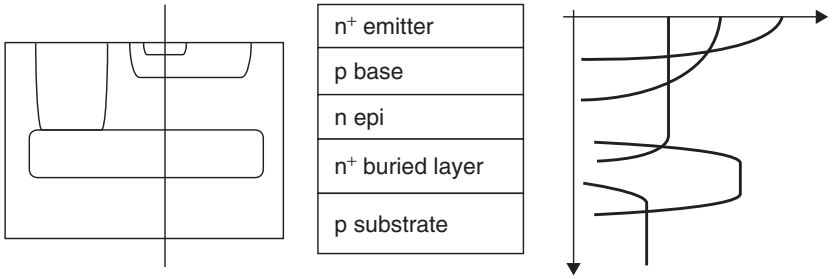


Figure 3.2 Cross section of an npn-bipolar transistor and its 1D simulation model of dopant concentrations along the cut line

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

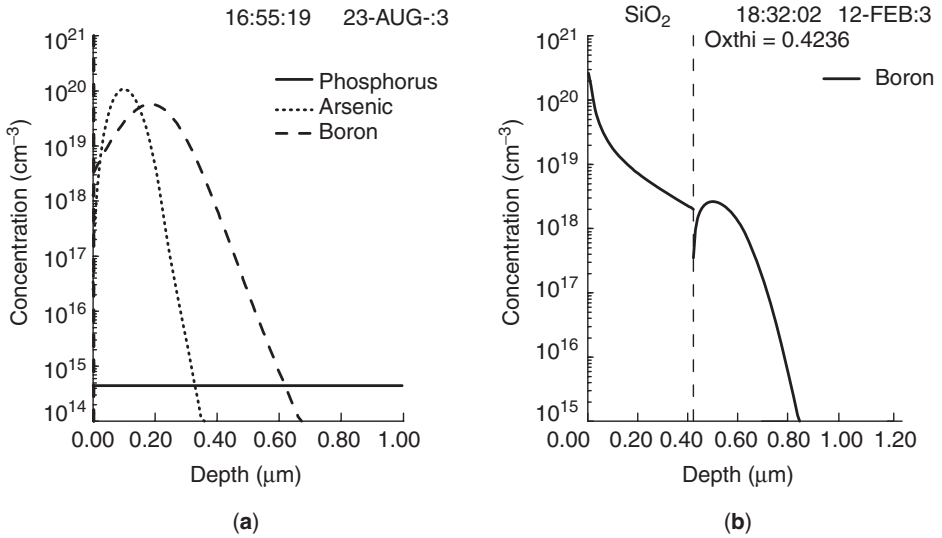


Figure 3.3 (**a**) 1D simulation (ICECREM) of arsenic (150 keV energy) and boron (50 keV) implantation into silicon, dose 10^{15} ions/cm^2 and (**b**) dry oxidation of BF_2^+ implanted silicon (20 keV, 10^{15} ions/cm^2)

modified by the user, but default parameters are good for initial simulations and novice users. Simulation examples in Chapters 6, 13, 14 and 15 are discussed using ICECREM.

1D-simulator output can visualize dopant depth distributions and film thicknesses, as shown in Figure 3.3. There are two important points in the concentration curves: the maximum concentration and its depth, and the junction depth in which the substrate dopant level and the diffused dopant levels match. The junction depths range from tens of nanometres to many micrometres.

3.3 2D SIMULATION

Two-dimensional simulation is indispensable because 1D simulation of more slices cannot predict 2D profiles. This is illustrated in Figure 3.4 for a simple 5 μm linewidth MOS transistor. 1D simulation produces accurate doping profiles and oxide thicknesses along lines A, B and D, but it cannot produce any meaningful results for C (where the implanted dopant spreads laterally under the gate) or E (where oxidation has taken place under a protective nitride layer). The 1D results for A, B and D are valid for 5 μm transistors, but as the device is scaled to smaller linewidths, more and more 2D effects arise, and a 2D simulator will be needed for profiles along B and D as well.

2D-diffusion simulators take into account the oxide and polysilicon structures on top of the silicon, and

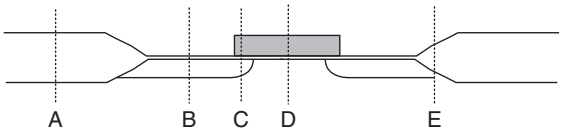


Figure 3.4 Vertical profiles of an MOS transistor: film thicknesses and dopant distributions along lines A, B and D can be simulated with a 1D simulator; but profiles along C and E require 2D simulation

produce dopant profiles that extend, for example, under the gate and masking layer (Figure 3.5). The structures above the silicon surface are usually not simulated, but simply drawn geometries. They are tools to add realism, like the deposition and etching steps in 1D simulators.

Two-dimensional simulators are about cross sections of structures, whereas 1D was only about layers. 2D simulation enables topography simulation. In 1D, it is not possible to study the deposition of films over other films; neither are cross sections relevant. Figure 3.6 shows two different deposition simulations: in both cases, the metal is deposited in a trench, and thickness of the metal on the sidewalls is predicted. Continuum simulators are used in integrated packages, but more and more atomistic simulation is needed. A step-coverage simulator that predicts the metal thickness over a step from the atom arrival angle distribution and surface mobility considerations may be useful, but to see if the crystal structure of the film on the sidewalls is different

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

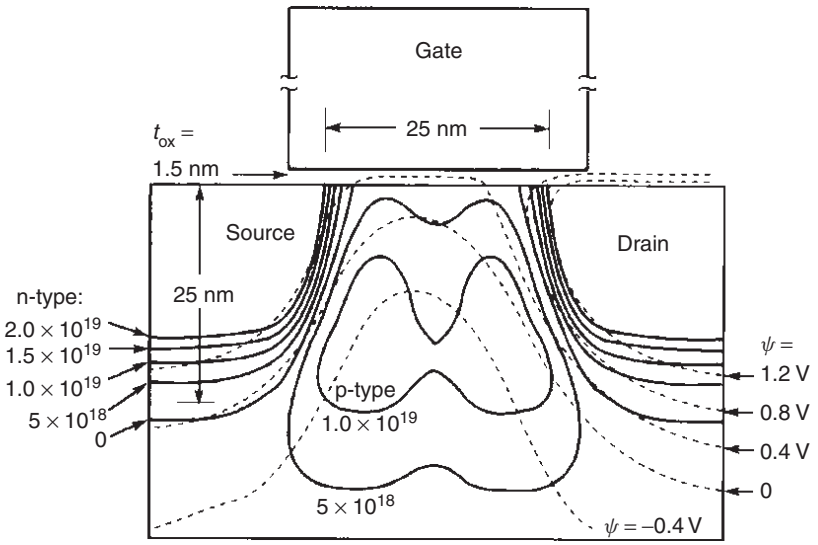


Figure 3.5 2D simulation: dopant concentration profiles of a 25 nm gate length CMOS transistor. Reproduced from Taur, Y. *et al*. (1998), by permission of IEEE

from the horizontal surfaces, we need an atomistic simulator.

2D simulation is computation intensive, and 2D simulators usually have a 1D simulation tool embedded in them, for quick and easy initial 1D tests. Saving on the computational time can be in orders of magnitude. Grid, or simulation mesh, in a 1D simulator, is regular and easy to generate, but in 2D simulators, the mesh generation is much more difficult. In order to reduce the computation time, a dense grid is used where abrupt changes are expected, and a sparse grid where the gradients are not steep. Instead of rectangular grids, triangular grids are often employed.

Optical lithography simulation is a self-contained regime in process simulation. Its main modules are optics, resist photochemistry and development, and its main output is resist profile. This will be discussed in Chapter 10.

3.4 3D SIMULATION

When scaling to smaller and smaller dimensions continues, 3D simulation becomes mandatory. A narrow but long transistor can be simulated by a 2D simulator, but a narrow and short transistor with similar dimensions in both x- and y-directions really needs 3D treatment. Again, complexity and time of simulation increase drastically over the 2D case. If a 1 μm deep layer is simulated in 1D simulator with 10 nm grid spacing, 100 layers need to be calculated. Similar grid size in 2D simulation requires 100×100 squares (10^4), and in 3D it equals 10^6 cubes. Roughly speaking, if 1D simulation takes seconds, 2D takes minutes and 3D, hours.

However, a 10 nm grid is no good for 3D simulation because 3D simulation is used especially for 100 nm devices and alike, and perhaps a 1 nm grid is used. But the question is not only computational; additional physical models need to be developed because more and more atomistic models must be used, and the continuum approximation fails because of the atomic nature of matter. In order to take advantage of 3D-process simulation, 3D-device simulators must be used, just as 2D-process simulators feed into 2D-device simulators. Advanced device simulators must similarly account for the fact that electric current is not a continuous variable, but a stream of charge packets with 1.6×10^{-19} C charge.

Simulation needs to extend from an atomic scale to a reactor scale. On the 1 m scale, simulation is needed to predict gas flows and temperature distributions inside the reactor; on the micrometre scale, simulation is needed to predict doping and deposition inside and on microstructures, and an atomic level simulation is needed for understanding the details of film growth and diffusion. For thin-film deposition, such a simulator would produce a relation between process parameters and film properties. At present, such a multiscale simulation remains a faraway goal.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

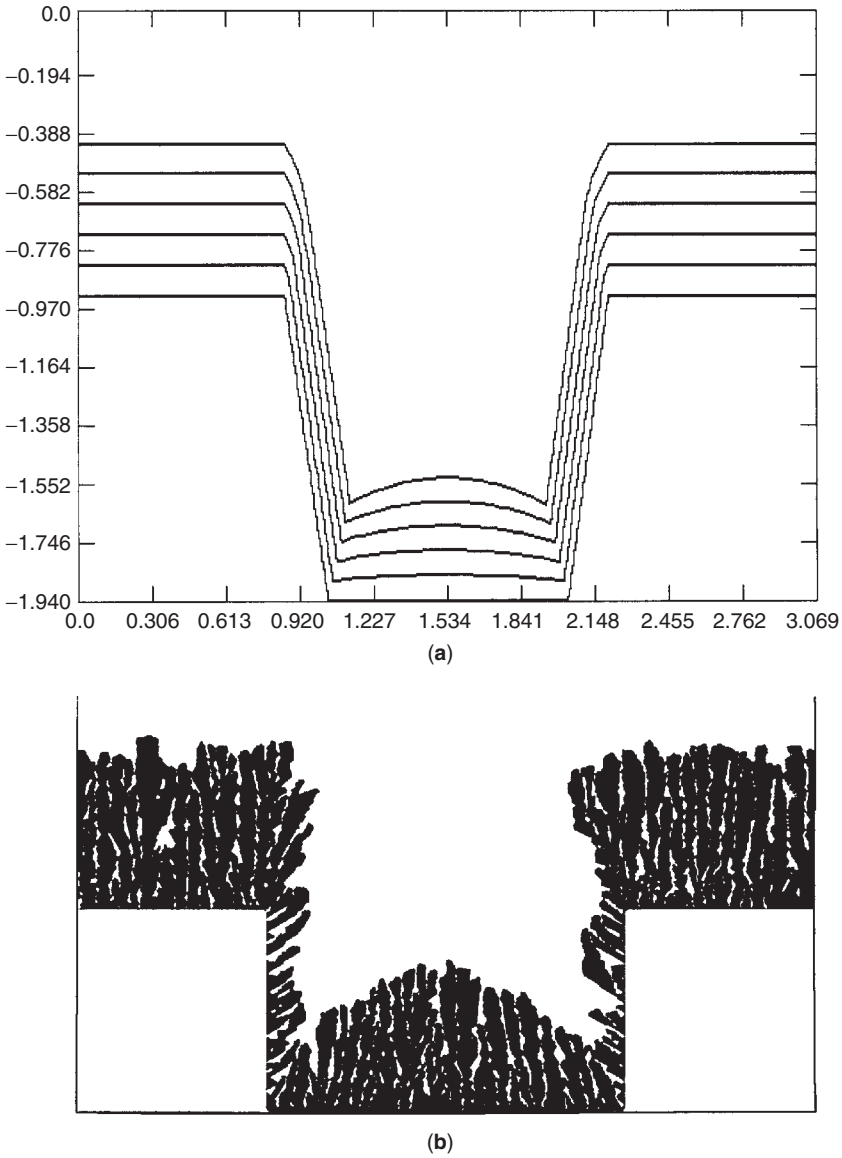


Figure 3.6 Continuum and atomistic metal step-coverage simulation: (**a**) SAMPLE 2D simulation of 0.5 μm thick metal deposition into a 1 μm wide, 1 μm deep trench; only the film thickness is simulated and (**b**) SIMBAD: sputtered tungsten into a trench with prediction of columnar grain structure. Reproduced from Dew, S.K. *et al*. (1991), by permission of AIP

3.5 EXERCISES

1S. What is the difference between the oxidation rates of boron, phosphorus and arsenic doped wafers when all have identical doping levels?

2S. How does the thermal oxide thickness on a phosphorus-doped wafer change with dopant concentration?

3S. What is the energy that phosphorus ions must have to penetrate through 200 nm of oxide?

4S. Compare your simulator with other simulators: how does it reproduce ranges and concentrations for ion implantation of arsenic into silicon? Data from Krusius, P., Process integration for submicron CMOS, *Acta Polytechnica Scandinavica*, **El58** (1987)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

E/(keV)	Dose/(cm^{-2})	Simulator	Range (Å)	Peak concentration (cm^{-3})
40	1.4×10^{13}	TRIM	332	6.0×10^{17}
40	1.4×10^{13}	PREDICT	268	3.8×10^{18}
40	1.4×10^{13}	CUSTOM	270	4.6×10^{18}
90	7.2×10^{14}	TRIM	636	8.6×10^{18}
90	7.2×10^{14}	PREDICT	603	9.9×10^{19}
90	7.2×10^{14}	CUSTOM	530	1.2×10^{20}

5S. Calculate oxide thickness for 10, 100, 1000 and 10 000 m oxidation at 1100 °C.

REFERENCES AND RELATED READINGS

Dew, S.K. *et al*: Modelling bias sputter planarization of metal films using ballistic deposition simulation, *J. Vac. Sci. Technol.*, **A9** (1991), 519–523, fig. 2a.

Ho, C.P. *et al*: VLSI process modelling – SUPREM III, *IEEE TED*, **30** (1983), 1438.

Krusius, P., Process integration for submicron CMOS, *Acta Polytechnica Scandinavica*, **El58** (1987), 1–16.

Law, M.: Process modelling for future technologies, *IBM J. Res. Dev.*, **46** (2002), 339–346.

Lorentz, J. *et al*: Three-dimensional process simulation, *Microelectron. Eng.*, **34** (1996), 85.

Taur, Y. *et al*: 25 nm CMOS design considerations, *IEDM '98* (1998), p. 789.

**Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030**

Part II

Materials

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

4

Silicon

Silicon transistors were first made in 1952, five years after the first germanium-based transistors. The electron mobility in germanium was much higher, and germanium crystal growth was more advanced. However, silicon, with its 1.12 eV bandgap, was better suited to higher operating temperatures, and the reverse currents were also smaller. The real breakthrough came by the end of 1950s when the beneficial role of silicon dioxide was recognized: silicon dioxide provided the passivation of semiconductor surfaces, and it resulted in improved transistor reliability. When it was further noticed that SiO_2 layer could act as a diffusion mask and as isolation for integrated metallization, the way was open for the invention of the integrated circuit. Oxide was a suitable isolation material and aluminium metallization could be patterned on top of the oxide. Neither GaAs nor Ge form stable and water insoluble oxides.

Silicon crystal growth rapidly caught up with germanium, and the steady increase in wafer size has continued up to this day, with 300 mm diameter wafers now in production. For other substrates, smaller sizes are still widely used, and when new materials such as silicon carbide (SiC) are introduced, the crystal growth and the wafering yield are so low that only small ingots and small wafers make sense.

Some 150 million silicon wafers, corresponding to 3 to 4 km^2, are processed annually. The largest proportion of them are 150 mm and 200 mm diameter wafers, ca. 50 million each, with some 20 million wafers of both 100 mm and 125 mm sizes. The latest 300 mm wafers accounted for some 10 million slices in 2003.

4.1 SILICON MATERIAL PROPERTIES

Silicon material properties are an excellent compromise between performance and stability. An energy gap of 1.12 eV makes silicon devices less prone to thermal noise than germanium devices with a 0.67 eV gap. Silicon source gases can be purified to extremely high degrees of purity, meaning that a high resistivity material can be made. Taken together with the high solubility of dopants, up to $10^{21}\,cm^{-3}$ for the common dopants boron, phosphorus and arsenic, this translates to eight orders of magnitude resistivity tailoring opportunities (Figure 4.1). Optical absorption in the visible makes silicon suitable for photodetectors and solar cells, and its transparency in the infrared (above 1.1 μm) is utilized in IR microsystems (Table 4.1).

Silicon is strong: its Young's modulus can be as high as 190 GPa (for <111> orientation). The excellent mechanical properties of silicon have been utilized since the 1960s in micromechanical pressure and force sensors that rely on bending beams and diaphragms. Piezoresistivity detection depends on doped regions for the resistors, and capacitive detection relies on the ability to micromachine shallow air gaps of the order of 1 μm. Both are standard processes in silicon microfabrication.

Stress, σ, and strain (elongation), ε, are correlated via

$$\sigma = \varepsilon E \tag{4.1}$$

with a constant of proportionality E, the Young's modulus. Elongation ε can also be stated as $\Delta L/L$, and stress as force per area, which gives the most familiar expression of Hooke's law: $F/A = E\Delta L/L$. When a piece of material is tensile- stressed, its elongation leads also to a lateral shrinkage of its diameter, $\varepsilon_{\text{lateral}} = \Delta D/D$. Poisson ratio is defined as $\nu = -\varepsilon_{\text{lateral}}/\varepsilon_{\text{tensile}}$. Silicon Poisson ratio, 0.27, in silicon is among the lowest of all solids.

Silicon is as strong as steel, but this fact is disguised by two factors: first, most of us do not have experience with 0.5 mm-thick steel plates, and second, silicon is brittle and the breakage pattern

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

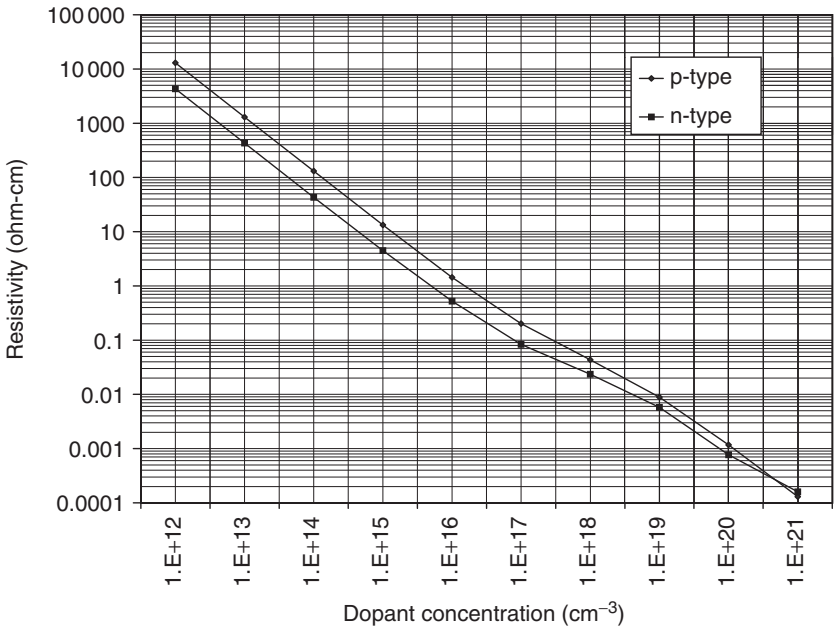


Figure 4.1 Silicon resistivity can be varied over eight orders of magnitude by doping. Data from Hull, R. (1999)

is therefore different from the ductile fracture of multicrystalline steel. Silicon is almost ideally elastic (obeying Hooke's law) up to the yield point, and after that a catastrophic failure takes place. Most metals and oxides obey Hooke's law initially, but then deform plastically before a fracture. The yield strength of silicon is 7 GPa at room temperature; different steel varieties have yield strengths of 2 to 4 GPa while the aluminium yield strength is only 0.17 GPa. Fracture strain for single-crystal silicon is 4%, an exceptionally large value.

4.2 SILICON CRYSTAL GROWTH

4.2.1 Purification of silicon

Silicon-wafer manufacturing is a multistep process that begins with sand purification and ends with final polishing and defect inspection. Silica sand, SiO_2, is reduced by carbon, yielding 98% pure silicon according to the reaction

$$SiO_2 + 2C \longrightarrow Si + 2CO\ (g) \qquad (4.2)$$

This material is known as metallurgical grade silicon (MGS). MGS is converted to gaseous trichlorosilane $SiHCl_3$ (boiling point 31.8 °C) according to the reaction

$$Si + 3HCl \longrightarrow SiHCl_3 + H_2\ (g) \qquad (4.3)$$

The main impurities in MGS (Fe, B, P) react to form $FeCl_3$, BCl_3 and PCl_3/PCl_5. Trichlorosilane gas is purified by distillation, during which $FeCl_3$, and PCl_3/PCl_5 are removed as high boiling point contaminations and BCl_3 as low boiling point contamination, and converted back to solid silicon by the decomposition of $SiHCl_3$ on hot silicon rods by the reaction

$$2SiHCl_3 + 2H_2\ (g) \longrightarrow 2Si\ (s) + 6HCl\ (g) \qquad (4.4)$$

This material is of extremely high purity, and is known as *electronic grade silicon* (EGS). EGS is a polycrystalline material, which is used as a source material in single-crystal growth.

4.2.2 Czochralski crystal growth (CZ)

In CZ-growth, a silica crucible (SiO_2) is filled with undoped electronic grade polysilicon. The dopant is introduced by adding pieces of doped silicon (for low doping concentration) or elemental dopants P, B, Sb or As (for high doping concentration). The crucible is heated in vacuum to ca. 1420 °C to melt the silicon (Figure 4.2). A single-crystalline seed of known crystal

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Table 4.1 Properties of silicon at 300 K

Structural and mechanical		
Atomic weight	28.09	
Atoms, total (cm^{-3})	4.995×10^{22}	
Crystal structure	Diamond (FCC)	
Lattice constant (Å)	5.43	
Density (g/cm^3)	2.33	
Density of surface atoms (cm^{-2})	(100) 6.78×10^{14}	
	(110) 9.59×10^{14}	
	(111) 7.83×10^{14}	
Young's modulus (GPa)	190	(111) Crystal orientation
Yield strength (GPa)	7	
Fracture strain	4%	
Poisson ratio, ν	0.27	
Knoop hardness (kg/mm^2)	850	
Electrical		
Energy gap (eV)	1.12	
Intrinsic carrier concentration (cm^{-3})	1.38×10^{10}	
Intrinsic resistivity (Ω-cm)	2.3×10^5	
Dielectric constant	11.8	
Intrinsic Debye length (nm)	24	
Mobility (drift) (cm^2/Vs)	1500 (electrons)	
	475 (holes)	
Temperature coeff. of resistivity (K^{-1})	0.0017	
Thermal		
Coefficient of thermal expansion ($^\circ C^{-1}$)	2.6×10^{-6}	
Melting point ($^\circ C$)	1414	
Specific heat (J/kg K)	700	
Thermal conductivity (W/m K)	150	
Thermal diffusivity	$0.8\,cm^2/s$	
Optical		
Index of refraction	3.42	$\lambda = 632$ nm
	3.48	$\lambda = 1550$ nm
Energy gap wavelength	1.1 μm	(Transparent at larger wavelengths)
Absorption	$>10^6\,cm^{-1}$	$\lambda = 200$–360 nm
	$10^5\,cm^{-1}$	$\lambda = 420$ nm
	$10^4\,cm^{-1}$	$\lambda = 550$ nm
	$10^3\,cm^{-1}$	$\lambda = 800$ nm
	$<0.01\,cm^{-1}$	$\lambda = 1550$ nm

Source: Data from Hull, R. (1999)

orientation is dipped into the silicon melt. The silicon solidifies into a crystal structure determined by the seed crystal. A thin neck is quickly drawn to suppress the defects that develop because of a large temperature difference between the seed and the melt, and then the pulling rate is lowered. Both the ingot and the crucible are rotated (in opposite directions); ingot rotation is ca. 20 rpm and crucible rotation about 10 rpm.

The ingot diameter is determined by the ingot pull rate. The pulling rate is limited by heat conduction away from the crystallization interface, and therefore large-diameter ingots have lower pulling rates. While a 100 mm diameter ingot can be pulled at 1.4 mm/min, the 200 mm ingot pull rate is 0.8 mm/min. In order to grow low vacancy concentration crystals, pulling rates as low as 0.35 mm/min are employed. Typical pulling time is 30 h, not including heating and cooling, which add another 30 h to the process, for 200 mm ingots.

The ingot length is determined by the yield strength of silicon neck and crucible size. The thin neck is not

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

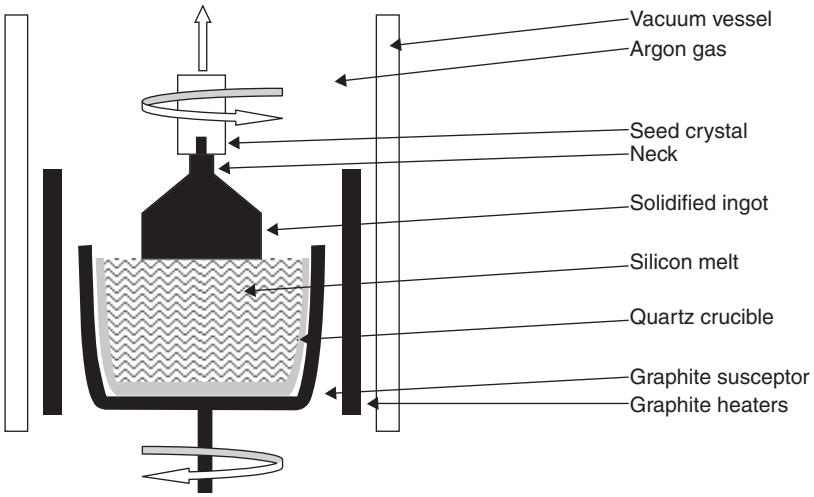


Figure 4.2 Czochralski crystal pulling: silicon (melting point 1414 °C) solidifies as it is pulled up. Pulling speed (∼ mm/min), ingot rotation speed (20 rpm) and crucible counter rotation speed (10 rpm) together determine the ingot diameter

a perfect material as it has defects arising from thermal shock, and torsional forces are also acting on it. Silicon yield strength is significantly lower at high temperatures, but 300 mm ingots can weigh up to 300 kg. Not all EGS can be utilized: ca. 10% of the original polysilicon remains in the crucible. The crucibles cannot be reused; they are extremely expensive disposable objects.

There is an inevitable contamination of the growing crystal from the materials that are essential to the growth set-up: the silica crucible is slightly dissolved during the crystal growth process, and therefore oxygen is always present in CZ-silicon in concentrations of 5 to 20 ppma (according to ASTM standard F121-83). Some of the oxygen evaporates as SiO gas (silicon monoxide) and is transported around the vacuum vessel.

EGS is extremely pure, for instance, boron, phosphorous and iron levels can be as low as 0.01 to 0.02 ppb. However, the crucible is a source of impurities, and for boron, sodium and aluminium, it is the crucible and not the EGS that determines the ingot purity. If synthetic silica is used for the crucibles, much higher purity CZ-ingots can be pulled.

The silica crucible is not mechanically strong enough at ca. 1400 °C temperatures, and a graphite susceptor provides the mechanical strength. The silica crucible reacts with the graphite susceptor according to the equation

$$SiO_2 + 3C \longrightarrow SiC + 2CO$$

This carbon monoxide is the source of carbon, which is always present in CZ-crystals, at concentrations ca. $10^{16}\,cm^{-3}$.

4.2.3 Dopant incorporation

Impurities are incorporated from the melt into the ingot, but different dopants have widely different segregation coefficients. The segregation coefficient is defined as quotient

$$k_o = \text{concentration in solid/concentration in liquid} \tag{4.5}$$

All dopants and metallic impurities are enriched in the melt, and oxygen is perhaps the only material that is incorporated preferentially into the silicon solid phase (see Table 4.2).

Because dopant segregation coefficients are less than unity, excess dopant is needed in the melt, compared with the final ingot. This can be calculated from k_o values easily. As the pulling advances, the melt volume decreases, the dopant concentration in the melt increases and therefore the dopant concentration in the ingot increases along its length. Because the crystal is rotated during growth, the centre- and the edge-boundary layers

Table 4.2 Segregation of dopants and impurities at silicon melt/solid interface

Dopants		Impurities	
Boron	$k_o = 0.8$	Iron	$k_o = 6.4 \times 10^{-6}$
Phosphorus	$k_o = 0.35$	Copper	$k_o = 8 \times 10^{-4}$
Arsenic	$k_o = 0.3$	Nickel	$k_o = 1.3 \times 10^{-4}$
Antimony	$k_o = 0.023$	Gold	$k_o = 2.25 \times 10^{-5}$
Gallium	$k_o = 0.0072$	Oxygen	$k_o = 1.25$

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

will be of different thicknesses, and this leads to radial dopant non-uniformity. There are also stochastic thermal fluctuations in the melt, and these lead to local resistivity variations. Some dopants (As, Sb; and oxygen also) are volatilized from the melt; therefore, concentration along the crystal axis is dependent on the gas flow in the crystal puller.

On the other hand, the concentration of oxygen decreases as the pulling advances. This has to do with the decreased contact area between the melt and the quartz crucible, and also with the flow patterns in the melt and the silica surface temperature. As a consequence, the oxygen concentration decreases along the ingot length. Analog to the mechanisms that cause radial dopant variation, the oxygen incorporation into the ingot also shows radial fluctuations. As a result, it may be that the whole ingot is not within the dopant and oxygen level specifications.

Because molten silicon is electrically conductive, magnetic fields can be used to control the melt behaviour. Magnetic fields reduce local temperature and flow fluctuations, which lead to a more stable melt and consequently to a more uniform growth. The Magnetic Czochralski (MCZ) growth enables a better control of oxygen levels in the crystal. The mechanisms remain to be fully explained, but at least a more uniform melt enables other process parameters, such as argon gas flow, to be varied over a larger range.

4.2.4 Float zone (FZ) crystal growth

If high purity or oxygen-free silicon is needed, float zone (FZ) crystal growth is used. In the FZ-method, a polysilicon ingot is placed on top of a single-crystal seed. The polycrystalline ingot is heated externally by an RF coil, which locally melts the ingot. The coil and the melted zone move upwards, and a single crystal solidifies on top of the seed crystal.

The highest FZ-silicon resistivities are of the order of 20 000 ohm-cm, compared to 100 to 1000 ohm-cm for CZ. Because there is no silica crucible, there is no oxygen, and metal contamination from the crucible is also eliminated. FZ wafers, however, are mechanically weaker than CZ-wafers because oxygen mechanically strengthens silicon. FZ wafers are available only in smaller diameters, 150 mm maximum, with a 200 mm FZ demonstrated but not used in device manufacturing. When doped FZ-silicon is made, dopants are introduced by flushing the melt zone with gaseous dopants such as phosphine (PH_3) or diborane (B_2H_6). High resistivity FZ is often doped via neutron transmutation doping (NTD) according to Equation (4.6)

$$n + {}^{28}\mathrm{Si} \longrightarrow {}^{29}\mathrm{Si} \longrightarrow {}^{29}\mathrm{P} + \mathrm{e}^- \qquad (4.6)$$

A silicon nucleus captures a neutron, and the newly formed nucleus decays by β-decay. This doping method explains why high resistivity silicon (5–20 kohm-cm) is available in n-type.

4.3 SILICON CRYSTAL STRUCTURE

Silicon has a cubic diamond lattice structure (Figure 4.3). The unit cell can be thought of as two interleaved face centred cubic (FCC) lattices with their origins in (0, 0, 0) an d (1/4, 1/4, 1/4). The distance between two atoms is $\sqrt{3}/4a$, and radius $\sqrt{3}/8a$, where a is the unit cell edge length, 5.43095 Å. As shown in Figure 4.3, there are 18 atoms to be considered: 8 at vertices (they are shared between 8 unit cells, and therefore contribute one atom to each unit cell; 6 face atoms are shared between two neighbouring unit cells, and contribute 3 atoms and there are four atoms fully inside the unit cell. The volume fraction of the space filled by silicon atoms is 34%, very low compared to hexagonal close packing, which fills 74% of the space. This open structure of silicon is important for diffusion.

Miller indices define the planes of a crystal. The plane that defines the faces of the cube (see Figure 4.4) intersects axes 1, 2, 3 at (1, ∞, ∞), respectively. The Miller index of a plane is given by the reciprocal of these intersects, that is, (1, 0, 0). The edges that tie planes are designated (1, 1, 0) and the diagonal planes are (1, 1, 1). The crystal structure is of course always the same, but it looks different when viewed from different directions: (100) corresponds to front view; (110) to edge view and (111) to vertex view (Figure 4.5). The set of six equivalent planes (the six faces of the cube) together

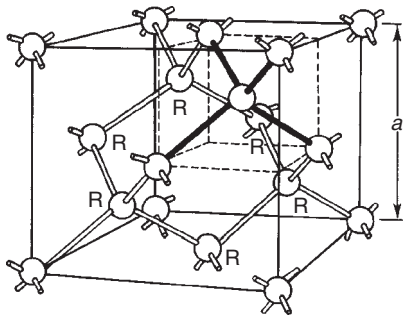


Figure 4.3 Silicon lattice: the unit cell consists of 8 atoms. Reproduced from Jenkins, T. (1995), by permission of Prentice Hall

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

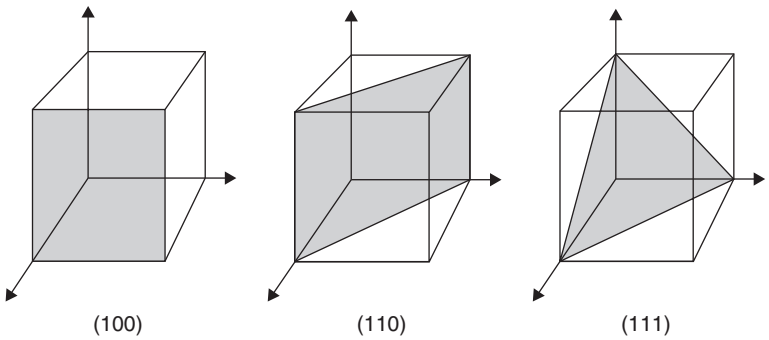


Figure 4.4 Some important silicon crystal planes with their Miller indices

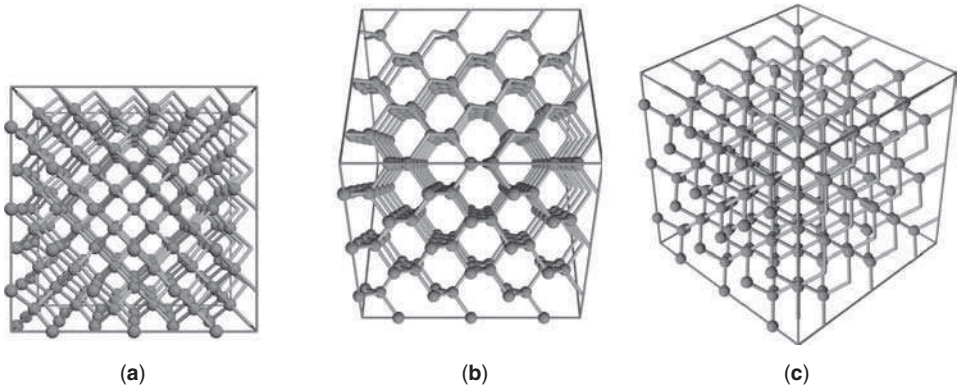


Figure 4.5 Silicon crystal viewed from different angles: (**a**) face view (100); (**b**) edge view (110); (**c**) vertex view (111). Figure courtesy Ville Voipio, Helsinki University of Technology

are designated {100}. There are 12 (1, 1, 0) and 8 (1, 1, 1) planes. Wafers are sometimes cut to other index planes, most notably (311) and (511).

Fourfold symmetry of (100) and sixfold symmetry of (110) and (111) can be seen in Figure 4.5, and it will become apparent in anisotropic wet etching of silicon (to be discussed in Chapter 21).

The angles between the planes can be calculated from the scalar product of the normal vectors

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}||\mathbf{b}| \cos(\mathbf{a}, \mathbf{b}) \tag{4.7}$$

Visual examination shows that (100) and (110) planes meet at 45° and all the other angles can be calculated easily, when the negative unit vectors are accounted for: $\bar{1}10$ is (−1, 1, 0). The angle between (111) and (100) planes is calculated from $1 = \sqrt{3} \cos \alpha$, giving $\alpha = 54.7°$.

In order to get familiar with the silicon crystal structure, the paper fold model shown in Figure 4.6 becomes handy. Copying the model on an overhead transparency and gluing it together will result in a 26-gon, which visualizes the crystal planes nicely. It will be indispensable when crystal-plane dependent etching of silicon will be discussed in Chapters 21 and 28.

Wafers of two crystal orientations are widely used in microfabrication: <100> and <111>. The former is the main material for CMOS and bulk micromechanics; the latter for bipolar transistors, power semiconductor devices and radiation detectors that rely on epitaxial deposition.

4.4 SILICON WAFERING PROCESS

As listed in Table 4.3, silicon ingots are transformed into wafers by a long process which includes mechanical, thermal and chemical treatments and many cleaning and inspection steps.

The silicon-crystal orientation is determined by the seed crystal. After the ingot has cooled down, it is cut to ca. 50 cm stocks, which are measured for crystal orientation by X-ray diffraction. A flat or a notch is

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

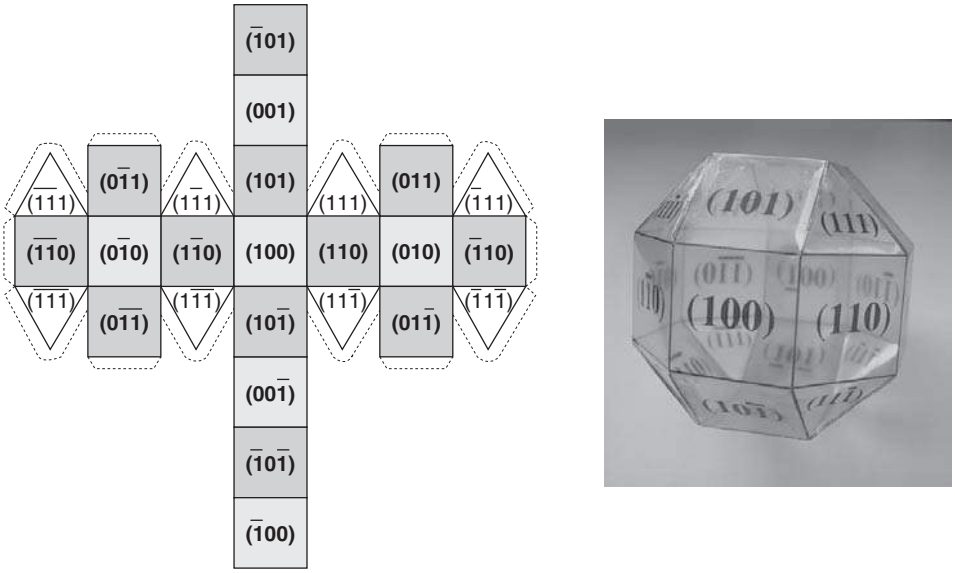


Figure 4.6 Fold-up paper model of silicon crystal planes. (This figure can be copied from Appendix B.) Fold model courtesy of Hiroshi Toshiyoshi, University of Tokyo

Table 4.3 Silicon wafering process

- Ingot crystal orientation by XRD
- Flat grinding
- Sawing ingot into wafers
- Lapping
- Edge smoothing
- Laser scribing
- Etching
- Annealing to destroy thermal donors
- Final polishing
- Inspections

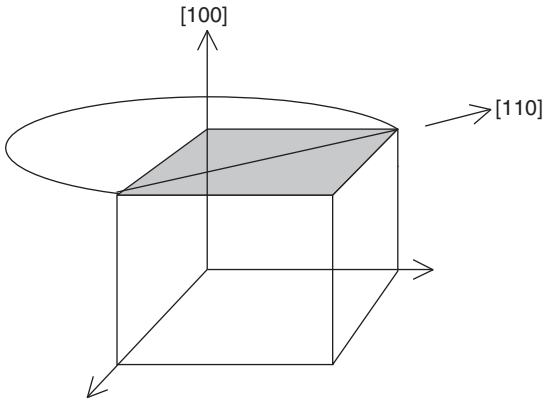


Figure 4.7 A <100> silicon wafer is cut so that one of the (100) planes defines the wafer surface, the vector normal to the surface is in the direction [100] and the flat is along direction [110]

then ground into the ingot to establish orientation. The flat or notch of a <100> wafer is oriented along the [110] direction (Figure 4.7).

The ingot is then sawed to slices. The surface of a <100> wafer is a (100) plane with [100] surface normal vector, usually cut as precisely as practical. <111> wafers are often miscut a few degrees because of epitaxial deposition considerations.

Flat and notches are used by automatic wafer handlers to orient wafers inside the equipment, and devices can be oriented relative to the crystal planes. This latter aspect is especially important in micromechanics in which crystal-plane-dependent anisotropic etching is a major technique. Secondary flats are used to identify the doping type and the orientation of wafers (Figure 4.8).

The next step is lapping: waviness and taper from the sawing are removed by lapping. In lapping, the wafers are rotating between two massive steel plates with alumina slurry. Lapping ensures not only parallelism of wafer surfaces but also equal damage depth. Surface roughness is ca. 0.1 to 0.3 μm after the lapping step. The edges of the wafers are then bevelled in order to prevent the chipping of silicon during wafer handling and to eliminate watermarks during the drying steps.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

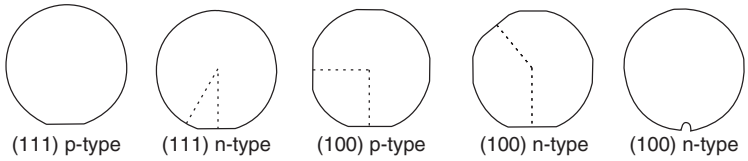


Figure 4.8 Wafer flats and notches for identifying wafer orientation and doping type

Wafer breakage often starts from a crack at the wafer edge, and because silicon is brittle, the crack propagates through the whole wafer. The wafers are marked by laser scribing. This is done early on so that subsequent steps remove the silicon dust generated by marking. Alphanumeric or bar-code marking enable wafer identity tracking during the processing.

Etching is then used to remove the lapping damage: both alkaline (KOH) and acidic ($HF\text{-}HNO_3$) etches can be used. Roughness is reduced somewhat in acid etching, but not in alkaline etching. An annealing step at 600 to 800 °C destroys thermal donors that are charged interstitial oxygen complexes.

Final polishing with 10 nm silica slurry in alkaline solution removes ca. 20 μm of silicon and results in 0.1 to 0.2 nm RMS surface roughness. Silicon is lost in the above-mentioned steps so that ca. half of the original ingot ends up as wafer material. In many power-device and solar-cell applications polishing is not needed because the structures are wide and films are rather thick, therefore, the etched wafer surface quality is enough. This is a significant cost-saving because polishing is an expensive step. On the other hand, in many micro-electro-mechanical system (MEMS) applications, double-side polishing is essential both for double-side lithography and for wafer bonding.

Inspection and cleaning steps constitute a major fraction of all wafering steps. The wafers are measured for mechanical and electric properties. Contactless measurements, for example, capacitance, optical and eddy-current methods, are preferred because contact methods introduce contamination and damage. Wafers are specified for particle cleanliness. Laser light scattering can be used to measure particle size distributions down to 60 nm sizes, but even unaided eye can detect particles larger than ca. 0.3 μm because of their scattering under intense light (e.g., from a slide projector).

Wafers are specified for a number of electrical, mechanical, contamination and other properties as agreed between the wafer manufacturer and chip maker. The specifications in Table 4.4 shows examples of wafer specifications, both for integrated circuits and microelectrical systems. Wafer resistivities and dopant concentrations, and the corresponding short-hand notations are shown in Table 4.5. More discussion on wafer specs will be found in Chapters 24 and 25.

Table 4.4 Specifications for 100 mm wafers, some typical values

	IC	MEMS
Growth method	CZ	CZ
Type/dopant	P/boron	P/boron
Orientation	100	100
Off-orientation	$0.0 \pm 1.0°$	$0.0 \pm 0.2°$
Resistivity	16–24 ohm-cm	1–10 ohm-cm
Diameter	100.0 ± 0.5 mm	100.0 ± 0.5 mm
Thickness	525 ± 25 μm	380 ± 10 μm
Front side	Polished	Polished
Backside	Etched	Polished
Primary flat	<110> ± 1 deg, 32.5 ± 2.5 mm	±0.2°
Oxygen level	13–16 ppma	11–15 ppma
Particles	<20 @ 0.3 μm	<20 @ 0.3 μm

Table 4.5 Resistivity versus dopant concentration

Dopant level	Designation	Dopant concentration (cm^{-3})	Resistivity n/p (ohm-cm)
Very lightly doped	n^{--}, p^{--}	$<10^{14}$	>100/>30
Lightly doped	n^{-}, p^{-}	10^{14}–10^{16}	1–100/0.3–30
Moderately doped	n, p	10^{16}–10^{18}	0.03–1/0.02–0.3
highly doped	n^{+}, p^{+}	10^{18}–10^{19}	0.01–0.03/0.005–0.02
Very highly doped	n^{++}, p^{++}	10^{19}	0.001 < 0.01/0.005

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

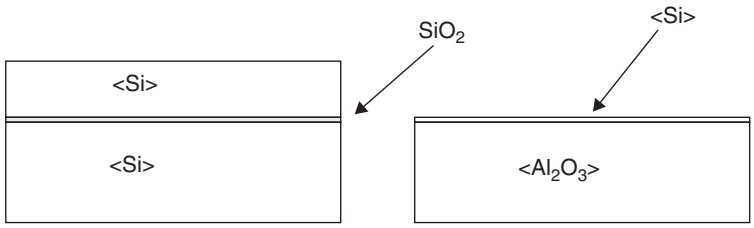


Figure 4.9 Silicon-on-insulator SOI (silicon/oxide/silicon) and SOS (silicon-on-sapphire) wafers

Further processing of the polished wafers leads to more specialized wafers. Epitaxy is a process for growing more silicon on top of a silicon wafer, with the doping level and/or the dopant type independent of the substrate wafer. Bonding of two (or even more) wafers together to create more complex wafers is another further development. Silicon-on-insulator (SOI) wafers can be made by, for example, wafer bonding (Figure 4.9). Silicon-on-sapphire (SOS) wafers rely on epitaxial deposition of silicon on top of a crystalline sapphire (Al_2O_3). It is also possible to create layers inside the wafer for additional functionality. These advanced wafers will be discussed in Chapters 15 (Ion implantation) and 17 (Bonding and layer transfer).

Table 4.6 Sources of non-idealities in silicon wafers

EGS polysilicon	Dopants (B, P) and other impurities (C, metals)
Czochralski growth	Impurities from quartz
	Oxygen from quartz
	Carbon from graphite and SiC
	Vacancies and interstitials
	Precipitates
	Dislocations
Wafering process	Contamination from tools
	Mechanical distortions
Wafer processing	Contamination
	Crystallinity defects
	Precipitation
	Mechanical distortions
	Dislocations

4.5 DEFECTS AND NON-IDEALITIES IN SILICON CRYSTALS

Even though silicon-wafer fabrication results in wafers with extremely well-defined properties, some defects are bound to be found. These defects can be classified according to their origin as grown-in defects and process-induced defects. The former are starting material and crystal-pulling related, and the latter result from the wafering process (at the wafer manufacturer) and from the wafer processing (in the wafer fab) (Table 4.6).

Metallic impurities come from polysilicon, quartz crucible, graphite and other hot parts of the growth system. The segregation coefficients of most metals are very small, and the crystal is purified relative to the melt. Metals are, however, fast diffusers in silicon, and they react with other defects and form clusters. Metals affect electronic devices by creating trapping centres in silicon midgap, reducing minority carrier lifetimes and lowering mobility. Metals can also precipitate at Si/SiO_2 interface and reduce the oxide quality, as will be discussed in Chapter 24. The allowed iron level in silicon wafers is limited to $10^{10}\,cm^{-3}$ (starting material limit) but at the end of an IC precess it can be much higher because fabrication steps introduce more iron.

Point defects are zero-dimensional: vacancies (missing atoms in the lattice), substitutional impurities (foreign atoms at silicon lattice sites) and interstitials (atoms such as oxygen at non-lattice sites) (Figure 4.10). Divacancies and phosphorous-vacancy pairs are also point-like defects. Point defects play an important role in diffusion, which is obvious because solid diffusion requires empty sites for atoms to move in the lattice. Some vacancies are present even at room temperature as a result of thermal equilibrium processes but additional vacancies generated by energetic or high temperature processing play a dominant role in diffusion.

One-dimensional or line defects are called *dislocations*. These come in many varieties, for example, extra half-planes inserted between the regular atomic planes. The order of magnitude of thermally generated stress σ can be gauged by Equation (4.8):

$$\sigma = \alpha E \Delta T \tag{4.8}$$

where strain, $\varepsilon = \alpha \Delta T \alpha$, depends on the silicon coefficient of thermal expansion, Young's modulus E (at

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

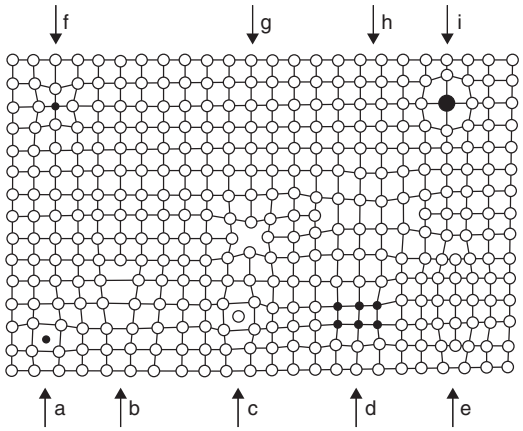


Figure 4.10 Schematic defects. (**a**) Foreign interstitial; (**b**) dislocation; (**c**) self-interstitial; (**d**) precipitate; (**e**) stacking fault (external); (**f**) foreign substitutional; (**g**) vacancy; (**h**) stacking fault (internal); (**i**) foreign substitutional. From Green, M.A. (1995), by permission of University of New South Wales

the temperature in question) and ΔT, temperature difference. The silicon yield strength (a.k.a. critical shear stress) is strongly temperature dependent: at 850 °C it is ca. 50 MPa, at 1000 °C only of the order of 10 MPa, and ca. 1 MPa at 1200 °C. Temperature differences between the wafer centre and the edge can easily lead to thermal stresses above the silicon yield strength. Stresses can be relaxed by slip-line formation.

Area defects include stacking faults, grain boundaries and twin boundaries. Processes that cause volume changes, such as oxidation, are prone to produce defects. Oxidation induced stacking faults (OISF) are a class of such defects.

Bulk defects include voids and precipitates. When the ingot is cooled down, the impurity and the dopant concentration exceed the solid solubility limit (see Figure 14.1 for solubility vs. temperature). Excess dopant or impurity will form precipitates. Oxygen precipitates (O_2P) is one class of such volume defects. Oxygen, which is present in CZ-wafers at 5 to 20 ppma levels, is initially dissolved in interstitials sites, but can precipitate during thermal treatments. Precipitation can take place on the surface or in the bulk. Bulk precipitates act as gettering centres for impurities and are thus beneficial. Carbon atoms act as nucleation sites and centres for oxygen precipitation.

Microvoids are clusters of vacancies formed inside the ingot during crystal pulling. When wafers are cut and polished, these voids end up at wafer surface. A microvoid causes a laser scatterometry signal similar to a particle. Vacancy clusters were therefore classified as particles, and were given the name COP, for Crystal Originated Particles (today, advanced multiangle scatterometry tools can distinguish voids from particles). It was the fact that the number of COPs did not decrease in cleaning (and it could in fact increase!) that lead to a reassessment of their nature. Typical COP sizes are 50 to 200 nm, and they are found in concentrations of 10^4 to 10^6 cm^{-3}.

Haze is defined as light scattering from surface defects, for example, scratches, surface roughness or crystal defects. Haze measurement is by done by scatterometry, and the whole wafer is scanned in haze measurement, in contrast to roughness measurement, which is local area measurement only, for instance, $5 \times 5\,\mu$m area by AFM.

4.6 EXERCISES

1. Calculate an estimate for silicon lattice constant from atomic mass and density.
2. Consider an Olympic swimming pool filled with golf balls and one squash ball. If the golf balls represent silicon atoms, and the squash ball represents a phosphorous atom, what would be the resistivity of a silicon piece with such a doping concentration?
3. Electronic grade polysilicon is available with 0.01 ppb phosphorous concentration. What is the highest ingot resistivity that can be pulled from such a starting material?
4. If 50 kg of ultrapure polysilicon is loaded into a CZ-crystal puller, how much boron should be added if the target doping level of the ingot is 10 ohm-cm?
5. Axial dopant profile along a CZ-ingot can be calculated from

$$C_s = k_0 C_0 (1 - X)^{k_0 - 1}$$

   where C_0 is the initial dopant concentration in the melt, X is the fraction solidified and k_0 is the segregation coefficient. If the wafer-resistivity specifications are 5 to 10 ohm-cm (phosphorus), calculate the fraction of the ingot that yields wafers within this specification.
6. If the neck in a CZ-ingot is 2 mm in diameter, what is the maximum ingot size that can be pulled before the silicon yields catastrophically?
7. If the COP density in the ingot is 10^5 cm^{-3}, what is the COP density on the wafer surface?

REFERENCES AND RELATED READINGS

Borghesi, A. *et al*: Oxygen precipitation in silicon, *J. Appl. Phys.*, **77** (1995), 4169.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Fischer, A. *et al*: Slip-free processing of 300 mm silicon batch wafers, *J. Appl. Phys.*, **87** (2000), 1543.

Green, M.A.: *Silicon Solar Cells*, Centre for Photovoltaic Devices and Systems, NSW, Sydney, 1995.

Hull, R.: *Properties of Crystalline Silicon*, IEE Publishing, 1999.

Jenkins, T.: *Semiconductor Science*, Prentice Hall, 1995.

Müssig, H.-J. *et al*: Can Si(113) wafers be an alternative to Si(001)? *Microelectron. Eng.*, **56** (2001), 195.

Petersen, K.: Silicon as a mechanical material, *Proc. IEEE*, **70** (1982), 420. Reprinted in W. Trimmer (ed.): *Micromechanics and MEMS, Classic and Seminal Papers to 1990*, IEEE Press, 1997, 58–95.

Shimura, F. (ed.): *Semiconductors and Semimetals: Oxygen in Silicon*, Willardson, 1994.

Shimura, F.: *Semiconductor Silicon Crystal Technology*, Academic Press, 1997.

**Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030**

5

Thin-film Materials and Processes

Thin-film processes are needed to make metal wires and to insulate those wires, to make capacitors, resistors, inductors, membranes, mirrors, beams and plates, and to protect those structures against mechanical and chemical damage. Thin films have roles as permanent parts of finished devices, but they are also used intermittently during wafer processing as protective films, sacrificial layers and etch and diffusion masks.

Metallic, semiconducting and insulating films are employed (Table 5.1) in microfabrication. Films are often used, however, not because of their metallic, semiconducting or dielectric properties, but for other features. For example, doped single-crystalline silicon carbide is a semiconductor, but amorphous SiC thin films are insulators for all practical purposes. SiC is frequently used as a structural material in high-temperature/corrosive ambient microdevices because of its excellent mechanical and chemical stability. Similarly, silicon is used not only for its electronic properties but also for its mechanical strength (micromechanics), optical absorption in visible wavelengths (solar cells, photodetectors), low absorption in infrared (waveguides for 1.55 μm optical telecom applications), high Seebeck coefficient (thermoelectric devices) and because of special properties of certain silicon microfabrication processes. Silicon nitride is used for free-standing thin membranes as etch and oxidation mask, as an etch-stop and polish-stop layer and as a passivation material that protects from mechanical and chemical damage.

Note on notations

<Si>	Single-crystal material
c-Si	Single-crystal material
α-Si	Amorphous material
a-Si:H	Amorphous material with imbedded hydrogen (at% usually given)
nc-Si	Nanocrystalline (grain size a few nanometres)
μc-Si	Microcrystalline material (grain size in the range of tens of nanometres)
mc-Si	Multicrystalline (large-grained, polycrystalline, grain size $\gg$ film thickness)
Al-0.5%Cu	Alloy with 0.5% copper
W_2N, Si_3N_4	Stoichiometric compounds
SiN_x, $x \approx 0.8$	Non-stoichiometric compound
W:N	Stuffed material, nitrogen at grain boundaries (non-stoichiometric)
WF_6 (g)	Material in gas phase
W (s)	Material in solid phase
TiW	Exception: TiW is not a compound but pseudoalloy with 30 atom% Ti
$Si/SiO_2/Si_3N_4$	Film stacks are marked with substrate or bottom film on the left

5.1 THIN FILMS VERSUS BULK MATERIALS

In thin films, at least one dimension of the material, the thickness, is small. For narrow lines, two dimensions are small, and for dots all three dimensions are small. This gives rise to prominence of surface effects like surface scattering of electrons, leading to size-dependent resistivity, or at very small dimensions, to quantum effects. The size scale for quantum effects is estimated by Debye lengths, which are of the order of 10 to 100 nm at room temperature.

The density of thin films is often very low compared to bulk materials. Sputtered tungsten films can have a density as low as 12 g/cm^3 compared to the bulk value of 19.5 g/cm^3. Thin films are often porous, which results in long term instability: humidity can be absorbed in the film, and high surface-area porous films oxidize and corrode readily.

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Table 5.1 Materials in microfabrication

	Conducting	Semiconducting	Insulating
Elements	Al, Cu, W, Mo, Ti	Si, Ge	Diamond
Oxides	RuO_2	SnO_2	SiO_2, Al_2O_3, HfO_2
Nitrides	TiN, TaN, W_2N	GaN	Si_3N_4, AlN, BN
Others	$TiSi_2$, $Al_{12}W$	SiC, GaAs, InP	Polymers

Table 5.2 Properties of sputtered molybdenum

Material/thickness	Underlayer	Conditions	Resistivity
Bulk	–	–	5.6 μohm-cm
Thin film, 50 nm	SiO_2	System 1, RT	17 μohm-cm
Thin film, 300 nm	SiO_2	System 1, RT	12 μohm-cm
Thin film, 300 nm	TiW	System 1, RT	9 μohm-cm
Thin film, 300 nm	SiO_2	System 2, RT	15 μohm-cm
Thin film, 300 nm	SiO_2	System 3, 150 °C	9 μohm-cm
Thin film, 300 nm	SiO_2	System 3, 450 °C	8 μohm-cm

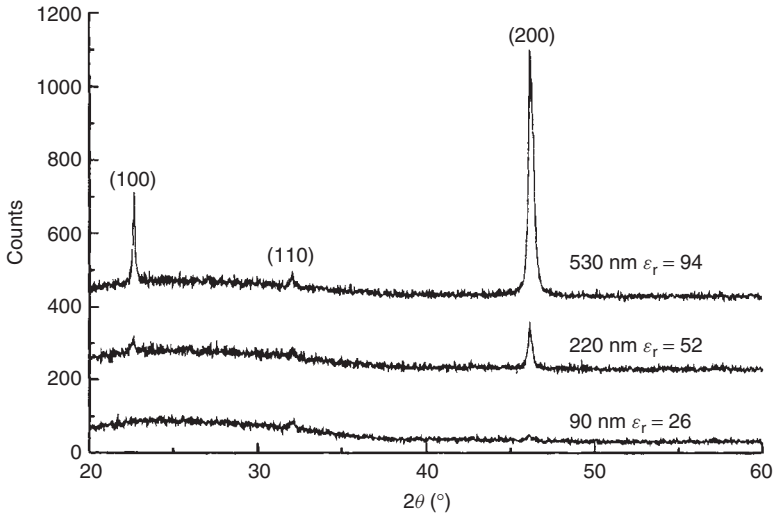


Figure 5.1 $SrTiO_3$ by XRD: thin-film structure and properties are thickness dependent. Reproduced from Vehkamäki, M. *et al*. (2001), by permission of Wiley-VCH

Many thin-film properties, resistivity, coefficient of thermal expansion and refractive index are thickness dependent. Deposition processes have profound effects on all film properties as shown in Table 5.2 for resistivities of sputtered molybdenum films. The films have been deposited in different sputtering systems under slightly different process conditions. In Figure 2.8, tantalum structure and resistivity were seen to depend on underlying layer: tantalum film on tantalum nitride is very different from tantalum film on oxide.

Structure depends on film thickness, and it may be that thick films are polycrystalline even though thinner depositions result in amorphous structure. This is shown in Figure 5.1 for $SrTiO_3$ film. X-ray diffraction (XRD) peaks indicative of crystallinity only appear for thicker films. The dielectric constant ε is also strongly thickness dependent.

Films prepared by different sputtering systems are different, and films prepared by two completely different deposition processes will differ even more. Copper

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

films made by sputtering, evaporation, electroplating or chemical vapour deposition (CVD) can have a factor of 2 differences in resistivity or grain size. When an amorphous film is annealed at high temperature, it will crystallize. But its crystal size and crystal orientation, and surface roughness will be different from a film that was initially polycrystalline, even though the films received identical anneals.

Very thin films are discontinuous and the thickness required for continuous films is process- and material-dependent. One criterion is transparency, which can be calculated from Lambert's law:

$$I = I_o \exp(-\alpha x) = I_o \exp(-4\pi k x/\lambda) \qquad (5.1)$$

With extinction coefficient (k) values 2 to 6 for metal films in the visible range, this translates to ca. 10 to 20 nm as a limit for transparency when a 1/e intensity drop is used as a criterion.

5.2 PHYSICAL VAPOUR DEPOSITION (PVD)

Physical vapour deposition is the dominant method for metallic thin-film deposition. All aluminum films in microfabrication are deposited by PVD, and PVD is used for copper, refractory metals and for metal alloys and compounds like TiW, WN, TiN, $MoSi_2$, ZnO and AlN.

The general idea of PVD is material ejection from a solid target material and transport in vacuum to the substrate surface (Figure 5.2).

Atoms can be ejected from the target by various means.

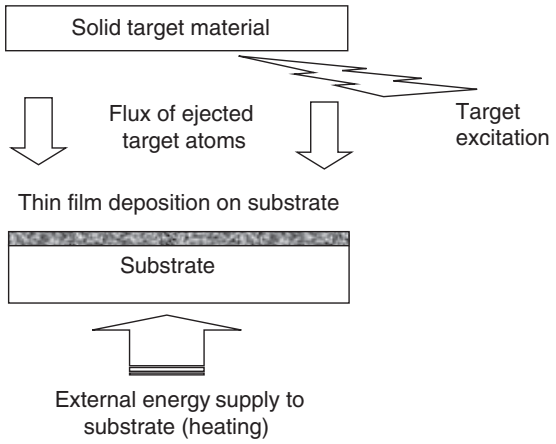


Figure 5.2 The principle of physical vapour deposition in a vacuum system

open source resistive heating → thermal evaporation
electron beam heating → e-beam evaporation
equilibrium source heating → molecular beam epitaxy (MBE)
argon ion bombardment → sputtering
laser beam bombardment → ablation

Shutter blades can be used to prevent deposition on the wafers during unstable flux (e.g., at the start of the deposition or during parameter ramping). Shutter blades enable very accurate and abrupt interfaces to be made, almost at the atomic thickness limit.

5.3 EVAPORATION AND MOLECULAR BEAM EPITAXY

Evaporation of elemental metals is fairly straightforward: heated metals have high vapour pressures and in high vacuum (HV), the evaporated atoms will be transported to the substrate (Figure 5.3). Atoms arrive at thermal speeds, which results in basically room-temperature deposition. Evaporation systems are either high-vacuum (HV) or ultra high–vacuum (UHV) systems, with the best UHV deposition systems with 10^{-11} Torr base pressures, and 10^{-12} Torr oxygen partial pressures.

There are very few parameters in evaporation that can be used to tailor film properties. There is no bombardment in addition to thermalized atoms themselves, which bring very little energy to the surface. Substrate heating is possible, but because of high vacuum requirement, there is the danger of outgassing of impurities from heated system parts.

In high vacuum, the atoms do not experience collisions, and therefore they take a line-of-sight route from source to substrate. Mean free path (MFP) is the measure of collisionless transport, and below ca. 10^{-4} Torr, MFP is larger than the size of a typical deposition chamber (for more discussion on vacuum

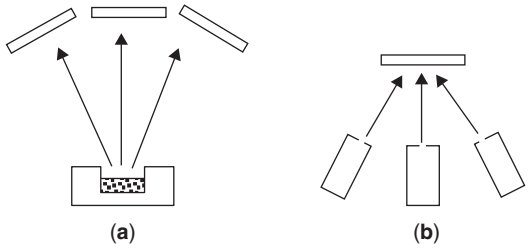


Figure 5.3 (**a**) Evaporation: an atomic beam emanating from an open crucible is transported in high vacuum to the substrate and (**b**) molecular beam system with three Knudsen cells

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

science and technology, refer to Chapter 32). To get uniform film thickness, the substrate direction relative to the beam is important, and substrate rotation is used to ensure uniformity. Uniformity is very much fixed when the chamber geometry is frozen, whereas in gas flow systems such as CVD, uniformity is very much process-dependent.

Low melting-point metals, such as gold and aluminium, can easily be evaporated, but refractory metals require more sophisticated heating methods. Localized heating by an electron beam can vaporize even tungsten (melting point 3660 K), but deposition rates are, however, very low, of the order of angstroms per second. Additionally, X-rays will be generated, which can damage sensitive devices.

It is possible that the molten metal reacts with the crucible because temperatures are very high, even though it is being minimized by use of refractory materials for crucibles: Mo, Ta, W, graphite, BN, SiO_2 and ZrO_2. If a misaligned electron-beam hits the crucible, crucible material will be evaporated and incorporated in the deposited film.

Molecular beam epitaxy (MBE) is a variant of evaporation. Instead of an open crucible, the source material is heated in an equilibrium source known as the Knudsen cell. An atomic beam (in the molecular flow regime, therefore the name MBE) exits the cell through an orifice that is small compared to the source size. Such equilibrium sources are much more stable than open sources, be they heated resistively or by an electron beam.

Alloy evaporation results in a film of a different composition than the source material because of vapour pressure differences of the elements. Compound evaporation is also difficult because most compounds do not evaporate as a molecular species, but are decomposed. Some oxides (e.g., SiO_2, B_2O_3), chalcogenides and halides do evaporate as molecules, and stoichiometric films can be obtained. The use of multiple sources is a standard solution to multi-component films.

Evaporated metal films are usually under tensile stress, in the range of 100 MPa to 1 GPa. Non-metals are found in both tensile and compressive stresses, but the values are smaller than for metals. More discussion on thin-film stresses can be found in Chapter 7.

5.4 SPUTTERING

Sputtering is the most important PVD method. Argon ions (Ar^+) from a glow discharge plasma hit the negatively biased target, slow down by collisions and eject one or more target atoms backwards. The ejected target atoms will be transported to the substrate wafers in vacuum (Figure 5.4). Because sputtering pressures are quite high, 1 to 10 mTorr (three to five orders of magnitude higher than evaporation pressures), sputtered atoms will experience many collisions before reaching the substrate. In a process called *thermalization*, the high-energy sputtered particles (5 eV corresponds to ca. 60 000 K) collide with argon gas ($T = 300$ K), and cool down. Thermalization also occurs to other species present in the plasma, the reflected neutrals (some argon ions are neutralized upon target collision). These neutrals provide energy to the substrate. Thermalization reduces the energy of particles reaching the substrate

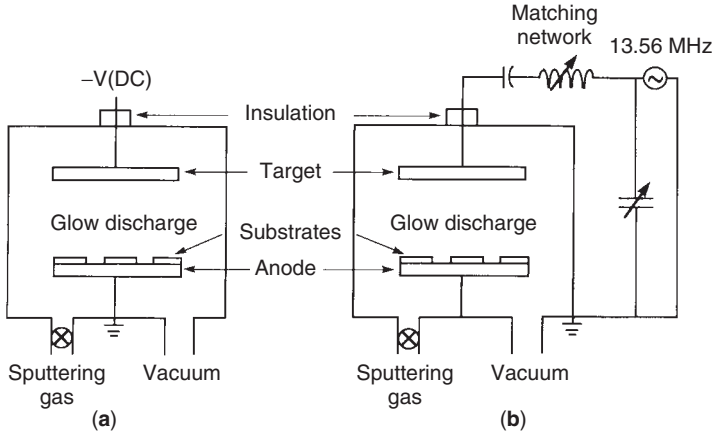


Figure 5.4 Schematic sputtering systems: (**a**) DC and (**b**) RF. Reproduced from Ohring, M. (1992), by permission of Academic Press

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

and it reduces the flux of particles to the substrate. Lower flux means a lower deposition rate, but lower energy leads to less re-sputtering of the film. This re-sputtering can sometimes be very useful, and it will be discussed in the context of bias sputtering in Chapter 32.

In contrast to evaporation, the energy flux to the substrate surface can be substantial. This has both beneficial and detrimental effects: loosely bound atoms (film-forming atoms as well as unwanted impurities) will be knocked out, improving adhesion and making the film denser. But too high energies can cause damage to the film, the substrate and underlying structures (thin oxide breakdown because of high voltages). There will always be some argon trapped in the film but no effect is seen in the first approximation.

Sputtering yield (Y) is a number of target atoms ejected per incident ion. Sputtering yields of metals range from ca. 0.5 (for carbon, silicon and refractory metals Ti, Nb, Ta, W) to 1 to 2 for aluminum and copper to 4 for silver at 1000 eV argon ion energy. Refractory metals have low sputtering yields, which is the fundamental reason for lower deposition rates. In practice, there is another reason that further lowers the deposition rate: refractory metals tend to have higher resistivity and thus lower thermal conductivity, which means that high sputtering powers cannot be applied to refractory sputtering targets. For heavy metals like tungsten and tantalum, sputtering yields are higher with xenon and krypton: these heavy gases transfer energy more efficiently to similar mass target atoms. However, argon is almost exclusively used.

In alloy sputtering, the flux is enriched in the component with higher yield (yields from alloys are even less accurately known than yields from elemental solids; elemental solid yields are used as approximations). The proportion of components in the sputtered flux is (Y_a/Y_b) (X_a/X_b) (X_is are the concentration proportions in target: $X_a + X_b = 1$). Because matter is conserved, the target is enriched in the other component: $(Y_b/Y_a)(X_a/X_b)$. A steady state situation develops and composition remains unchanged.

5.5 CHEMICAL VAPOUR DEPOSITION (CVD)

In chemical vapour deposition (CVD), the source materials are brought in gas phase flow into the vicinity of the substrate, where they decompose and react to deposit film on the substrate. Gaseous by-products are pumped away, as shown schematically in Figure 5.5.

There are various possible CVD reaction types.

pyrolysis	$SiH_4\ (g) \rightarrow Si\ (s) + 2\ H_2\ (g)$
reduction	$SiCl_4\ (g) + 2\ H_2\ (g) \rightarrow Si\ (s) + 4\ HCl\ (g)$
hydrolysis	$SiCl_4\ (g) + 2\ H_2\ (g) + O_2\ (g) \rightarrow SiO_2\ (s) + 4\ HCl\ (g)$
compound formation	$3\ SiH_2Cl_2\ (g) + 4\ NH_3\ (g) \rightarrow Si_3N_4\ (s) + 6\ H_2\ (g) + 6\ HCl\ (g)$

Decomposition of source gases is induced either by temperature (thermal CVD) or by plasma (plasma-enhanced CVD, PECVD). Thermal CVD processes take place in the range 300 to 900 °C (very much source gas dependent), and PECVD processes at ca. 100 to 400 °C, typically at 300 °C (Table 5.3). CVD reaction rates obey Arrhenius behaviour, that is, exponentially temperature-dependent. CVD processes are also complex from the point of view of fluid dynamics.

CVD of silicon on a single crystalline silicon wafer can result in a single-crystalline film. This is termed *epitaxy* and it is an important special case of thin-film deposition. The next chapter is devoted to epitaxial deposition. Most deposition processes lead to amorphous or polycrystalline films.

Silicon dioxide can be deposited by many reactions. Gaseous reactants form a solid film on the wafer and gaseous by-products are pumped away.

$$SiH_4\,(g) + 2N_2O\,(g) \longrightarrow SiO_2\,(s) + 2H_2\,(g) + 2N_2\,(g)$$

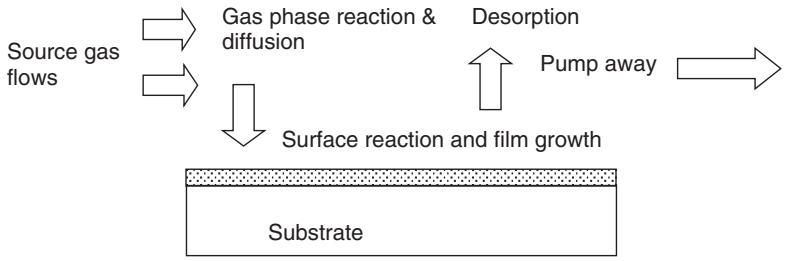


Figure 5.5 CVD process: both gas phase transport and surface chemical reactions are important for film deposition

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Table 5.3 Some widely used CVD processes

Material/method	Source gases	Temperature	Stability
LTO	$SiH_4 + O_2$	425 °C	Densifies
HTO	$SiCl_2H_2 + N_2O$	900 °C	Loses Cl
TEOS	$TEOS + O_2$	700 °C	Stable
PECVD OX	$SiH_4 + N_2O$	300 °C	Loses H
LPCVD poly	SiH_4	620 °C	Grain growth
LPCVD a-Si	SiH_4	570 °C	Crystallizes
LPCVD Si_3N_4	$SiH_2Cl_2 + NH_3$	800 °C	Stable
PECVD SiN_x	$SiH_4 + NH_3$	300 °C	Loses H
CVD-W	$WF_6 + SiH_4$	400 °C	Grain growth

LTO = Low-Temperature Oxide; HTO = High-Temperature Oxide; TEOS = TetraEthylOxySilane, $Si(OC_2H_5)_4$.
The precursor name TEOS has become synonymous with the resulting oxide film; it should be obvious which meaning is used.

The use of N_2O (laughing gas) instead of oxygen is preferred because silane reaction with oxygen is spontaneous and oxide particles are produced everywhere in the system and they float around in the reactor and deposit sporadically on wafers.

CVD is not limited to simple compounds: films can be doped during deposition. CVD oxide can be doped by adding phosphine (PH_3) gas to the source gas flow. Phosphorus doped CVD oxide, also known as *phosphorus doped silica glass* (PSG), is a widely used doped film. Phosphorus oxide is formed by CVD and intermixed with silicon dioxide.

$$4PH_3\ (g) + 5O_2\ (g) \longrightarrow 2P_2O_5\ (s) + 6H_2\ (g)$$

Doped oxide films typically have ca. 5% by weight dopant. Higher doping levels lead to porous, hygroscopic material. Toxicity of PH_3 (and B_2H_6 for BSG) needs to accounted for, but CVD reactors use silane, which is a flammable gas, so the basic designs of CVD reactors are suitable for dangerous gases. Trimethyl phosphite (TMP) and trimethyl borate (TMB) are less toxic alternatives to hydrides.

Phosphorus getters mobile ions like sodium and potassium, and makes PSG a more efficient barrier against the ambient than undoped CVD oxide (which is sometimes known as USG, for undoped silica glass). PSG etch rate is much faster than that of undoped oxide, and PSG is a popular sacrificial layer in micromechanics.

CVD tungsten is deposited in two steps. The silane reduction step deposits a thin nucleation layer over every surface in the system, and high rate blanket deposition with hydrogen reduction is used to achieve the desired total thickness:

$$WF_6\ (g) + SiH_4\ (g) \longrightarrow W\ (s) + 2HF\ (g) + H_2\ (g) + SiF_4\ (g)$$

$$WF_6\ (g) + 3H_2\ (g) \longrightarrow W\ (s) + 6HF\ (g)$$

This process is able to fill holes and trenches and it is very important in multilevel metallization (Chapter 27).

5.5.1 CVD rate and mechanism

The two main differences between PVD and CVD reactions are in flow dynamics and temperature dependence: in PVD, fluid dynamics need not be considered, but CVD processes are flow processes with complex fluid dynamics. In PVD processes, deposition rate depends primarily on target excitation energy. CVD processes are chemical processes, and their rates obey Arrhenius behaviour. The activation energy E_a can be extracted from the Arrhenius formula when the deposition rate has been determined at several temperatures. The magnitude of the activation energy gives hints to possible reaction mechanisms.

Two temperature regimes can be found for most CVD reactions (Figure 5.6): when the temperature is low, the surface reaction rate is low, and there is an overabundance of reactants. The reaction is then in the surface reaction–limited regime. The rate of silicon nitride deposition from SiH_2Cl_2 at 770 °C is ca. 3.3 nm/min. This is compensated by the fact that deposition takes place on up to 100 wafers simultaneously.

When the temperature increases, the surface reaction rate increases exponentially, and above a certain temperature, all source gas molecules react at the surface. The

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

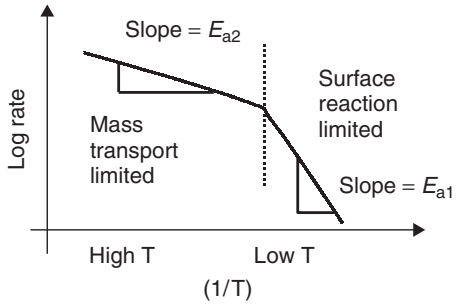


Figure 5.6 Surface reaction–limited versus mass transfer–limited CVD reactions

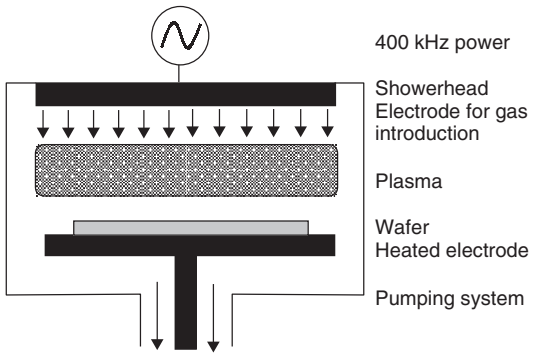


Figure 5.7 Schematic PECVD system

reaction is then in the mass transport–limited regime because the rate is dependent on the supply of a new species to the surface. The fluid dynamics of the reactor then plays a major role in deposition uniformity and rate.

Process temperatures are often severely limited: for instance, after an aluminum–silicon interface has been formed, the maximum allowed temperature is ca. 450 °C to prevent silicon dissolution into aluminum. When aluminum has to be coated by an oxide or nitride layer, plasma activation is usually employed. There is a thermal CVD process for depositing oxide on aluminium (at ca. 425 °C: it is known as (LTO), (for low-temperature oxide, but it has poor reproducibility. Most often plasma activation is employed. Instead of thermal decomposition of the source gases, a glow discharge is utilized. The method is known as PECVD, for plasma-enhanced CVD, and sometimes as PACVD, for plasma-assisted CVD. Much lower temperatures can be used: plasma activation ensures enough reactive species even at low temperatures, typically at ca. 300 °C, but even down to 100 °C (but temperature strongly affects film quality). Whereas typical activation energies for thermal CVD processes are 2 eV (200 kJ/mol), PECVD activation energies are a fraction of that, for example, 0.3 eV for amorphous silicon deposition. PECVD deposition rate is only mildly temperature-dependent.

A simple parallel plate diode reactor for PECVD is shown in Figure 5.7. Wafers are placed on a heated bottom electrode, the source gases are introduced from the top, and pumped away around the bottom electrode. Operating frequency is often 400 kHz, which is slow enough for ions to follow the field, which means that heavy ion bombardment is present. At 13.56 MHz, only the electrons can follow the field, and the ion bombardment effect is reduced.

In thermal CVD, pressure, temperature, flow rate and flow rate ratio are the main variables. In PECVD, we have the additional variable of RF power. In advanced PECVD reactors, RF power can be applied to both electrodes, and the two power sources can supply different frequencies, duty cycles and power levels. The ratio of 13.56 MHz power to kilohertz power is important for film stress tailoring.

Whereas thermal oxide or low-pressure chemical vapor deposition (LPCVD) nitride are really SiO_2 and Si_3N_4, many other (PE)CVD films are non-stoichiometric: plasma nitride SiN_x has, for example, $x = 0.8$. Especially in PECVD, hydrogen is often incorporated into film in considerable amounts, up to 30 atom-%. This can cause device instability later on if hydrogen diffuses into the devices. PECVD can be used to deposit mixed oxides, nitrides and carbides, as well as doped oxides like thermal CVD. Mixture of silane, nitrous oxide and ammonia will result in oxynitride, SiO_xN_y, with varying ratios of nitrogen and oxygen, covering the whole range of compositions (and material properties) between oxide and nitride. Fluorine-doped oxide, SiOF can be deposited, but film instability limits the usable fluorine range to ca. 5%wt, for the same reasons for which phosphorus doping range is limited. Other materials deposited by PECVD include SiO_xC_y and SiC_xN_y, which are used as etch and polish stop layers in multilevel metallizations. Amorphous carbon, a-C:H and related materials resemble diamond in many but not all respects, and they are known as *diamond-like carbon* (DLC). Diamond and SiC can also been deposited by thermal CVD at 700 to 1000 °C, and those materials resemble bulk materials in many respects.

5.6 OTHER DEPOSITION TECHNOLOGIES

Vacuum and reduced pressure deposition methods like PVD and CVD are suitable for films in the thickness range 10 to 1000 nm. This is partly a practical

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

limitation due to deposition rates, which are generally 1 to 100 nm/min. In many cases, thicker films are desired, and PVD or CVD methods quickly become throughput limited. In CVD silicon epitaxy, a 100 μm layer thickness is feasible, even though very expensive. For most polycrystalline and amorphous CVD and PVD films, however, stresses build up to unacceptable levels for thicker films, limiting thicknesses to a few micrometres.

Liquid phase deposition methods include a wide variety of techniques that are unrelated physico-chemically. Compared to PVD and CVD methods, liquid phase methods are extremely simple. A beaker is enough for electroless deposition (with an optional hot plate). Add a current source and an electroplating system is ready. Liquid phase methods are widely used in printed wiring board industry, thin-film head fabrication and in MEMS, and they are being introduced in IC fabrication, for deposition of copper and for inter-metal dielectric layer deposition.

Liquid phase depositions take place at 20 to 100 °C, and film structure and quality are often very different from PVD and CVD films. But as is usual with other deposition technologies, film properties will be strongly influenced by subsequent annealing steps.

Liquid phase deposition methods	Typical applications
- Electroplating/galvanic deposition	Thick conductor layers
	High aspect ratio metallization
- Electroless deposition	Selective metallization
- Spin coating	Photoresists
	Thick polymer layers
	Spin-on-glasses
- Sol–gel	Porous dielectrics
	Thick, complex materials

5.6.1 Electroless deposition

Electroless deposition depends on reduction reaction in an aqueous solution that contains metal salts and a reducing agent. Metal deposition takes place as a result of metal ion reduction. The surface needs to be suitable for electroless deposition and this is achieved by exposing the surface to a catalyst, such as $PdCl_2$. This reducing agent starts the reduction reaction, which then continues locally. Selective deposition is thus possible. Gold, nickel and copper are the usual metals to be deposited by the electroless method. Gold can be deposited from a KOH, KCN, KBH_4 and $KAu(CN)_2$ mixture at rates exceeding 5 μm/min, even though much lower rates are usually used. Temperatures for electroless deposition range from room temperature to 100 °C.

Copper deposition chemistries traditionally use sodium hydroxide in the plating bath, but this has to be eliminated if copper is used in IC metallization. Alternative pH adjustment can be done with TMAH (tetramethyl ammonium hydroxide). Copper sulphate ($CuSO_4$) in formaldehyde (HCHO) and EDTA (ethylene diamine tetraacetic acid) complexing agent are the basic constituents of the bath. Surfactants (polyethylene glygol) and stabilizers (2,2′-dipyridyl) can be added. The reaction is described by

$$CuEDTA^{2-} + 2HCHO + 4OH^- \longrightarrow Cu + H_2 + 2H_2O + 2HCOO^- + EDTA^{4-}$$

The deposition rate is of the order of 100 nm/min. The electroless deposition set-up is extremely simple and no electrical connection needs to be made to the wafers. Selectivity, however, is difficult to maintain. Hydrogen evolution and incorporation into the film is a problem because hydrogen is mobile, and carbon incorporation is another problem. With 2 μohm-cm as the accepted thin-film copper resistivity, electroless deposition can result in much poorer films.

5.6.2 Electroplating/galvanic plating/electrochemical deposition (ECD)

Electroplating takes place on a wafer that is connected as a cathode in metal-ion containing electrolyte solution. The counterelectrode is either passive, like platinum, or made of the metal to be deposited.

Electroplating can be very simple: copper is deposited on the cathode according to the following reduction reaction:

$$Cu^{2+} + 2e^- \longrightarrow Cu\ (s) \quad \text{electrolyte solution: } CuSO_4$$

Gold is plated in a two-step process with the second, the charge transfer reaction, as the rate-limiting step:

$$Au(CN)_2^- \longleftrightarrow AuCN + CN^-$$

$$AuCN + e^- \longrightarrow Au\ (s) + CN^-$$

Electroplating rates vary a lot but are generally in the range of 0.1 to 10 μm/min. Deposited mass is calculated as

$$\text{mass} = \alpha ItM/nF$$

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

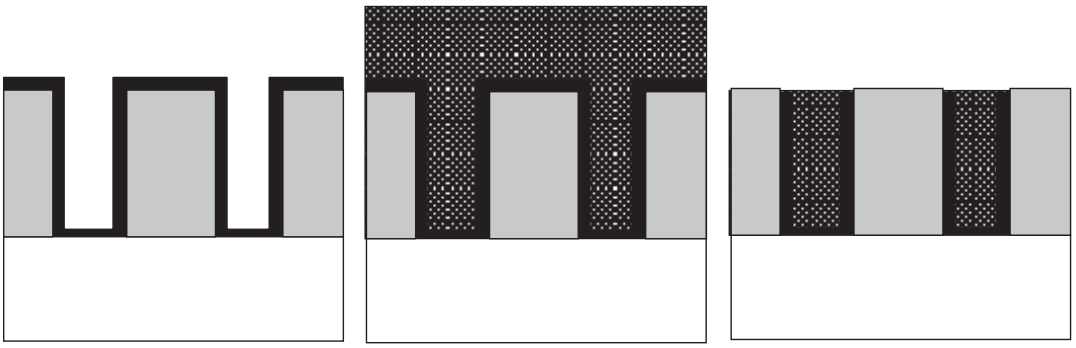


Figure 5.8 Damascene plating: seed layer sputtering; electroplating, polishing

where I is current, t is time, M is molar mass, n is species charge state, α is the deposition efficiency and F is the Faraday constant, 96 500 coulombs.

Noble metals can be deposited at 100% efficiency ($\alpha = 1.00$). In the deposition of less noble metals, hydrogen evolution lowers efficiency, and for some non-metals like phosphorus co-deposition with cobalt (Co:P, 12%, a soft magnetic material), α can be as low as 0.20. Other typical electroplated metals include nickel and iron–nickel (81% Ni, 19% Fe, Permalloy™). Tin–lead (40% lead in eutectic) and indium are plated as solder bumps for chip packaging. Many of the metals used in microfabrication, aluminum, titanium, tungsten, tantalum and niobium, do not have practical electroplating processes.

Three transport processes are active during electrochemical deposition (ECD): diffusion at electrodes due to local depletion of reactant via deposition, migration in the electrolyte and convective transport in the plating bath. The latter is connected to electrochemical cell design, and it is affected by factors such as stirring, heating, recirculation and hydrogen evolution.

Macroscopic current distribution is determined by the plating bath electrode arrangement and wafer and bath conductivity. Electrical contact to the wafer also needs careful consideration. Microscopic (local) current distribution depends on pattern density and pattern shapes. The third scale in ECD is the feature scale: potential gradients inside structures are important especially when high aspect ratio structures are filled.

In practice, the plating solutions are complex mixtures of electrolytes, salts for conductivity control, modifiers for film uniformity and morphology improvement as well as surfactants. Many plating solutions are proprietary. Plating baths are rather aggressive solutions, and photoresist leaching into plating bath or adhesion loss are real concerns for reproducible plating.

Accelerators (brighteners) are additives that modify the number of growth sites. Suppressors are additives for surface diffusion control. Taken together, these additives increase the number of nucleation sites, and keep the size of each nucleation site small, which drives smooth growth. Pulsed plating can also be used in balancing nucleation and grain growth: high overpotential and low surface diffusion favour nucleation, and the opposite conditions favour grain growth.

Damascene plating (Figure 5.8) deposits a film all over the wafer. Polishing is needed to remove excess metal. Metal remains in the grooves and recesses of the wafer, and the wafer surface remains planar. Electroplating can also be done in resist grooves, and more plating applications will be presented in Chapters 23 and 27.

5.6.3 Spin-coating

Spin-coating is a very widely used method for resist spinning and increasingly for other materials as well; for example, spin-on-glasses (SOGs) and thermally stable polymers (known together as spin-on-dielectrics, SODs). It is now a method to deposit films that will remain as structural parts of finished devices.

Spinning is a simple process for viscous materials deposition. Spinners, with typical speeds up to 10 000 rpm, are found in every microfabrication laboratory. The main parameters for film thickness control are viscosity, solvent evaporation rate and spin speed. Spin-coated film thicknesses range from 0.1 μm up to 500 μm, with standard photoresists usually around 1 μm. The coating of thick spin films will discussed in Chapter 10 in connection with thick photoresists.

Dispensing can be in static mode, or slow rotation of ca. 300 rpm can be used (Figure 5.9). Depending

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

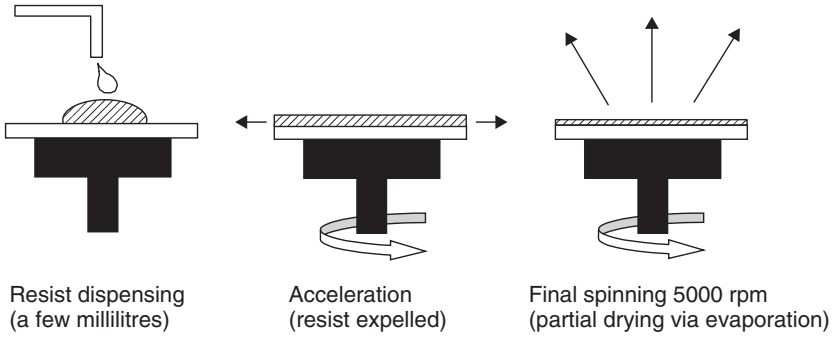


Figure 5.9 Spin-coating process

on the wafer size and desired film thickness, a drop of 1 to 10 ml (cm^3) is dispensed at the wafer centre. Acceleration to ca. 5000 rpm spreads the liquid towards the edges. Half of the solvent can evaporate during the first few seconds, so rapid acceleration is a must because viscosity changes with solvent content, and radially non-uniform thickness will result from viscosity differences. Spin speed can be controlled to ca. ± 1 rpm, and an error of ± 50 rpm will result in 10% thickness differences. Turbulence (both from the spin process itself and from cleanroom airflows) and ambient humidity (which is affected by exhaust from the spinner bowl and the cleanroom environmental control) affect evaporation rate, and consequently, film thickness. Pinhole defects in spin-coated films are thickness-dependent: thinner films are more defective. Pinholes can be caused by particles on the wafers, and also by particles in the dispensed fluid, even though all chemicals in microfabrication have been filtered with submicron filters. Air bubbles formed during dispensing (caused by e.g., an unclean dispense tip) can cause either pinholes or large bubbles, in the millimetre range.

Spin-coated films fill cavities and recesses because they are liquids during spin coating. This is advantageous for gap filling and smoothing, but if uniform thickness over the topography is desired, spinning is not ideal. Room temperature spinning is always accompanied by baking in the range 100 to 250 °C.

5.6.4 Sol–gel

A sol is a colloidal suspension of small (1–1000 nm) particles in a liquid. A gel is 3D solid network that forms in a colloidal liquid. A typical sol–gel process uses metal alkoxides $M–(O–CH_3)_n$ in organic solvents. Alkoxides hydrolyze according to

$$M(OR)_n + xH_2O \longrightarrow M(OH)_n + x\text{ROH}$$

and grow by condensation reaction,

$$(OR)_nM–OH + HO–M(OR)_n \longrightarrow (OR)_nM–O–M(OR)_n + H_2O$$

A great variety of simple methods can be used for sol–gel processing: for example, dipping, spraying and spinning. Compositional variation (by changing alkoxides ratios) is easy. Thickness can be tailored not only by spin speed but also by chemical modifications in the organic side chain R. Film thicknesses of hundreds of micrometres are possible for both glassy SiO-type materials and ceramics like lead–zirconium titanate (PZT).

Drying of gel leads to drastic volume shrinkage (easily by a factor of 10), and the resulting material is known as *xerogel*. Supercritical drying eliminates capillary forces and collapse of the gel, leading to aerogels, which can be 99% void with only 1% solid material. Such a material could be the ultimate dielectric, with a dielectric constant ε close to unity. Application of these materials as structural parts in microdevices will be difficult, but as sacrificial materials they could be easily removable.

5.7 METALLIC THIN FILMS

Metallic thin films have various applications in microfabricated devices.

Conductors: Resistivity is the main consideration: aluminum and copper are main choices for most applications, and gold is often used in RF devices, like inductor coils, to minimize resistive losses. Doped silicon (and polycrystalline silicon) can be used as a conductor, but its resistivity is very high compared to metals.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Contacts to semiconductors: ohmic (metal-like) and Schottky (diode-like) contacts are possible. Aluminum, itself p-type dopant in silicon, makes good ohmic contact to p-type silicon. Platinum silicide is one candidate for silicon Schottky contacts.

Capacitor electrodes: Capacitor electrodes need not be highly conductive. The most important capacitor electrode, the MOSFET gate, is chosen to be polycrystalline silicon because its interface with silicon dioxide is stable, and its lithography and etching properties are good.

Plug fills: When vertical holes need to be filled with a conducting material, CVD tungsten and electrodeposition of copper are employed.

Resistors: Doped semiconductors, metals, metal compounds and alloys can be used as resistors. Heating resistors can be made of almost any material, but precision resistors are difficult to make.

Adhesion layers: Noble metals like gold and platinum do not adhere well to substrates, and therefore thin (10–20 nm thick) 'glue' layers of titanium or chromium are needed.

Barriers: Barriers are needed to prevent unwanted reactions between thin films. Amorphous metal alloys and compounds like tungsten nitride (W:N), titanium–tungsten (TiW), TiN and TaN are the usual materials.

Mechanical materials: Aluminum and nickel are materials for micromechanical free-standing beams and cantilevers, in, for example, micromirrors and resonators. Films such as TiN can be used as mechanical stiffening layers to prevent mechanical changes in the underlying softer films, like aluminum.

Optical materials: Transparent conductors like indium-doped tin oxide (ITO; $In_xSn_yO_2$) are needed in displays and light-emitting devices. In image sensors, metals act as light shields, and in many micro-optical devices, as mirrors. TiN is often deposited on top of aluminum to reduce reflectivity, because lithography is difficult on highly reflecting surface.

Magnetic materials: Nickel and nickel alloys, Ni:Fe, are used in magnetic microactuators. Cores of microtransformers are also made of these materials, which are usually deposited by electroplating.

Catalysts and chemically active layers: Chemical sensors often use films such as palladium and platinum as catalysts.

Electron emitters: Vacuum microemitter tips are often made of molybdenum because of its high melting point and low work function.

Infrared emitters and other IR components: Heated wires emit infrared, and porous metallic films, like aluminum black, act as IR absorbers. Metallic meshes act as IR filters.

Sacrificial layers: Many devices require free-standing structures. These must be fabricated on solid films, which will subsequently be etched away. Copper is often used as a sacrificial material under nickel or gold.

Protective coatings: Sometimes the role of the topmost layer is simply to protect the underlying layers from the ambient: from etching agents or environmental stressors. Nickel and chromium are used as masks for etching.

X-ray components: Masks for X-ray lithography require high atomic mass materials that effectively block X-rays. Tungsten, gold and lead are prime candidates. X-ray mirrors are made by alternating layers of heavy (tungsten, molybdenum) and light materials (carbon or silicon) of X-ray wavelength thicknesses.

The deposition process greatly influences the choice of metals. Not all materials are amenable to all deposition methods, and the resulting film properties (resistivity, phase, texture, adhesion, stress, surface morphology) are closely connected with the details of the deposition process, and may well be idiosyncratic with the equipment. Reproducing results that have been obtained with another piece of equipment can be a nightmare.

5.7.1 Properties of metallic thin films

Low resistivity is required in thin-film form. Thin-film resistivity is often much higher than bulk resistivity. Aluminum, copper and gold thin-film resistivities are close to bulk values; for most others, thin films resistivities are factor of 2 higher. Metals of microfabrication importance are listed below. Resistivities are strongly deposition process–dependent as shown in Table 5.2, and Table 5.4 should be used as a guideline only.

Alloys and compounds TiW, TiN_x and TaN_x have resistivities that are even more strongly deposition process–dependent than simple metals, and the exact composition will also have a profound effect. Resistivities of these metal compounds are usually in the range of 100 to 500 μohm-cm.

Young's moduli are the same order of magnitude for all metals, from 100 GPa for soft metals to 600 MPa for refractory metals. Many metal properties are related to melting point. High melting point equals high bond strength and stable atomic arrangement

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Table 5.4 Properties of metals

Metal	Resistivity (μΩ-cm)	CTE (ppm/°C)	Thermal conductivity (W/cm K)	Melting point (°C)
Al	3	23	2.4	650
Cu	1.7	16	4	1083
Mo	5.6[a]	5	1.4	2610
W	5.6[a]	4.5	1.7	3387
Ta	12[a]	6.5	0.6	3000
Ti	48[a]	8.6	0.2	1660
Co	6.2[a]	12.5	0.7	1500
Ni	6.8[a]	13	0.9	1455
Cr	13[a]	6	0.7	1875
Pt	10[a]	9	0.7	1769
Au	1.7	14	3	1064

[a] Thin-film resistivity is much higher than bulk value: as a rule of thumb, 1.5–2 times the bulk value can be used as an guestimate for thin-film resistivity.

in solid. This correlation is seen in, for example, electromigration resistance.

Electromigration is metal movement with the electron flow. Electrons transfer momentum to metal atoms, which will consequently move and accumulate at the positive end of the conductor and leave voids at the negative end (Figure 5.10). This effect is encountered in aluminum conductors when current densities approach the mega-ampere per square centimetre level, but copper and tungsten tolerate higher current densities. Electromigration will be discussed further in Chapter 24.

5.8 DIELECTRIC THIN FILMS

Dielectric films have, just like metallic films, a plethora of applications in microdevices. The table below classifies dielectric film applications into three categories: structural parts in finished devices, intermittent layers during wafer processing and protective coatings for finished devices. Surprisingly, many films can serve in all these roles.

Active, protective and sacrificial layers during wafer processing

Mask for thermal oxidation	Si_3N_4
Diffusion and ion implantation masks	SiO_2, Si_3N_4
Dopant evaporation barrier	CVD oxide, SiN_x
Etch-stop layer in polymer-based inter-metal stacks	SiN_x
Window definition during selective epitaxial growth	CVD oxide
Etch masks in bulk micromechanics	CVD oxide, Si_3N_4
Dopant sources	PSG, BSG
Spacers in MOS and bipolar transistors	CVD oxide, CVD nitride
Sacrificial layers in surface micromechanics	PSG, resist
Gap fill materials	Oxides, SODs

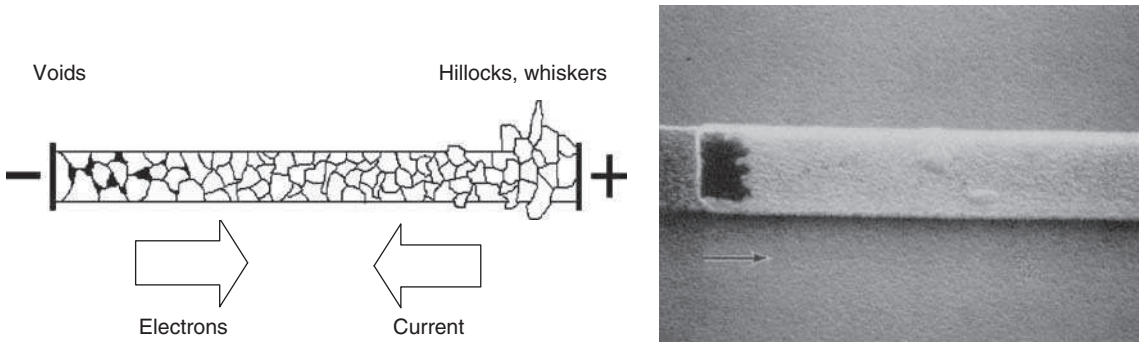


Figure 5.10 Electromigration: atoms are transported from the anode end of a wire towards the cathode with electron wind. Voids are left at the anode end, and hillocks form towards the cathode end: (**a**) schematic. Figure courtesy Antti Lipsanen, VTT; (**b**) SEM micrograph of Al lines (4 μm wide). Reproduced from Hu, C.-K. *et al*. (1993), by permission of American Inst of Physics

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Structural parts of finished devices

Function	Examples
Inter-metal insulation	SiO_2, polymers
Gate oxides in MOS transistors	SiO_2, HfO_2
Capacitor dielectrics	SiO_2, Si_3N_4, Ta_2O_5, $BaSrTiO_3$
Tunnel oxide in EPROMs	SiO_2
Ion barriers	Al_2O_3, Si_3N_4
Tunnel oxides in Josephson junction devices	AlO_x, NbO_x
Dielectric mirrors	CVD oxide, nitride, polysilicon
Micromechanical beams and plates	LPCVD nitride
Antireflective coatings	PECVD SiN_x, SiO_2
Heat sink for lasers and power devices	Diamond
Hydrophobic surfaces	Teflon, diamond
Microfluidic structures	Polymers, oxide, nitride, diamond
Microlenses	Polymers, spin-on glasses

Protective coatings against ambient in final devices

Passivation layer & metal ion barrier	SiO_x, SiO_xN_y
Humidity & scratch protecting barriers	PECVD SiN_x, polyimide
Tribological coating (wear, friction)	Diamond, SiC
Corrosion resistant coatings in harsh environments	Ta_2O_5, SiC

5.9 PROPERTIES OF DIELECTRIC FILMS

Higher deposition temperature usually leads to denser films that are more resistant to etching and polishing and less susceptible to moisture absorption. Thermal oxide etch rate in hydrofluoric acid (HF) is always the same, irrespective of the furnace that was used to grow it. In CVD, and in PECVD in particular, films can have HF etch rates varying enormously depending on the particular type of equipment and process conditions (power, flow rate and ratios, temperature). As a rule of thumb, if thermal SiO_2 etch rate is 100 nm/min, 300 to 1000 nm/min is expected for (PE)CVD oxides. Densification anneal at a high temperature can lower this by a factor of 2.

Films should be free of pinholes, small point-like defects; otherwise they are useless as protective coatings. For plasma-enhanced CVD, <0.1 pinholes/cm^2 is a good value. If the film is less dense than the bulk, it can be either because of porosity or because of pinholes.

5.9.1 Inorganic films

Thermal oxide, SiO_2, is a very high quality dielectric (Table 5.5), but it can only be grown on silicon (single or polycrystalline silicon) and all the other materials on the wafer have to be compatible with ca. 1000 °C oxidizing ambient, which excludes most materials. When silicon dioxide is needed on materials other than silicon, it is done by CVD, either thermal CVD or PECVD.

Thermally grown silicon dioxide is the standard reference material, with its relative permittivity ε_r of ca. 4 (dielectric constant $\varepsilon = \varepsilon_r\varepsilon_0$). In order to minimize capacitances ($C = \varepsilon A/L$) between metal layers, it is preferable to use low dielectric constant films (known as low-k or low-ε materials), many of them polymeric materials, or modified CVD oxides. The topic of dielectric constant will be discussed in connection with multilevel metallization for ICs in Chapter 27.

High dielectric–constant films are required in applications where high capacitance is needed. MOS transistors and DRAM memories are capacitors, and in order to make the capacitors smaller, area has been scaled

Table 5.5 Properties of silicon dioxide and silicon nitride

	SiO_2	Si_3N_4 (LPCVD)
Resistivity (Ω-cm), 25 °C	10^{16}	10^{16}
Density (g/cm^3)	2.2	2.9–3.1
Dielectric constant	3.8–3.9	6–7
Dielectric strength (V/cm)	12×10^6	10×10^6
Thermal expansion coefficient (ppm/ °C)	0.5	1.6
Melting point (°C)	1700	1800
Refractive index	1.46	2.00
Specific heat (J/g °C)	1.0	0.7
Young's modulus (GPa)	87	~300
Yield strength (GPa)	8.4	14
Stress in film on Si (MPa)	200–400 C	1000 T
Thermal conductivity (W/cm K)	0.014	0.19
Etch rate in Buffered HF (nm/min)	100	1

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

down. To keep capacitance constant, capacitor dielectric thickness has been scaled down. This approach cannot be continued indefinitely because of tunnelling currents through thin oxides. High-k dielectrics are a topic in Chapter 25. Thin-film dielectrics have breakdown field in the range of 10^5 to 10^7 V/cm (10–1000 V/μm). This topic is especially important for MOS transistor scaling, with oxide thicknesses in the sub-10 nm range.

5.9.2 Spin-coated inorganic films

Spin-on-dielectrics, SODs, are materials that are spin-coated in liquid state, and cured in a multi-step process to yield solid material. The gap-filling capability of SODs is related to viscosity: low viscosity equals good gap fill, but unfortunately, it is correlated with high shrinkage, too. Spin-on-glasses (SOG) are silicon-containing polymers that can be spun and then cured to produce a silicon dioxide–like glassy material. Numerous commercial formulations for SOGs exist, adjusted for molecular weight, viscosity and final film properties for specific applications. Two basic types of SOG are organic and inorganic SOGs. The inorganic SOGs are silicate-based and the organic are siloxane-based.

Silicate SOGs can be cured to form SiO_2-like layers, which are thermally stable and do not absorb water. They are, however, subject to volume shrinkage during curing, leading to high stresses (~400 MPa). This limits silicate SOGs to thin layers, ca. 100 to 200 nm. Multiple coating/curing cycles can be used to build up thickness, at the cost of quite an increase in the number of process steps.

Addition of phosphorus to SOG introduces changes similar to phosphorus alloying of CVD oxide films. The resulting films are softer and exhibit less shrinkage, and are better in gap filling. However, water absorption increases, which means less stable films.

Organic SOGs based on siloxane (Figure 5.11) do not result in pure SiO_2-like material, but contain carbon after curing. By tailoring the carbon content, the material properties can be modified for lower stress (~150 MPa), and consequently, thicker films. Siloxane films are, however, polymer-like in their thermal stability, and 500 °C is a practical upper limit.

Typical composition of spin-on-glass solution:

siloxane polymer	<20% wt
isopropyl alcohol	20–50%
acetone	10–35%
ethanol	15–20%
1-butanol	Remainder %

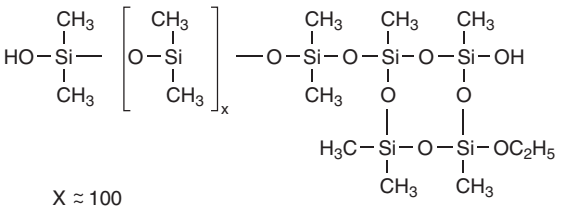


Figure 5.11 Structure of siloxane

Upon curing, the reaction Si–OH + HO–Si → Si–O–Si + H_2O takes place, resulting in a glass-like material. Multi-step curing, first at ca. 100 °C, then at higher temperatures, for example, 175 °C and finally at ca. 400 °C, is required in order to prevent film cracking. Films are prone to cracking because large volume shrinkage of the order of 10% is associated with curing.

5.9.3 Polymer films

Polymeric materials are a different breed from inorganic dielectrics. Historically, no polymeric materials were used as permanent parts of microdevices (but they are used as encapsulation materials), and the reliability and stability of polymeric materials is still inferior to inorganic dielectrics. This is partly inherent, and has to do with porosity that causes, for example, moisture absorption: values below 1% wt are exceptional, with typical values of 1 to 3% wt. It is difficult to achieve etch selectivity between polymers and photoresist, and photoresist stripping remains a problem. Some of these are process development issues that will be solved as polymeric materials mature and experience accumulates.

Polymeric films can replace inorganic films, especially when thick films are needed. Spin coating 10 μm or even 100 μm-thick polymer films is no problem; for inorganic dielectrics, films thicker than a few micrometres are non-standard.

Polymers have thermal limitations: their coefficients of thermal expansion (CTEs) are in the range of 30 to 50 ppm/°C, versus 1 to 20 ppm/°C for elemental metal films and simple inorganic compounds, even though some organic–inorganic hybrid materials have CTEs of 10 to 30 ppm/°C, and decomposition temperatures of 500 °C. The usable temperature range of polymers is limited: photoresist can tolerate ca. 120 °C without degradation, and 350 to 400 °C is the upper limit for most polymers.

Widely used polymer materials in microfabrication include thermally stable aromatic polymers (BCB,

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

benzo-cyclo-butadiene), photopatternable epoxy SU-8, polyimides (some of them photopatternable), fluorinated poly(arylene ethers), fluoropolymer CPFP (cyclised perfluoro polymers like CYTOP™).

PTFE, polytetrafluroethylene (Teflon® is one variety of PTFE) is also used, because of its special surface properties such as superhydrophobicity and extremely low water absorption, <0.10% wt. Note that polymers are sometimes used exactly because of their water absorption: a capacitive humidity sensor measures the change in the dielectric constant due to water absorption in the polymer dielectric. Parylene (poly-para-xylylene) is a versatile material that is strong enough mechanically so that released, free-standing structural parts can be made out of it. Parylene and CYTOP are exceptional polymers because they can tolerate KOH etching. Parylene is deposited by CVD, whereas most other polymers are spin-coated.

Polyimides offer some special properties: some formulations are photopatternable like resists, and form permanent parts in finished devices. Some imides (PI2610) have coefficients of thermal expansion ca. 3 ppm, close to silicon in the plane of the wafer, but ca. 20 ppm/°C perpendicular to the surface. Thermal conductivities of imides are in the range 0.1 to 0.2 W/m K, an order of magnitude higher than that of silicon dioxide, but similar to that of silicon nitride.

Tensile strengths of polymers are in the range of 100 to 400 MPa, and Young's moduli of the order of 1 to 10 GPa, compared with 50 to 500 GPa for inorganic solids and elemental metals. Stresses in polymers are inherently low, <100 MPa, whereas stress minimization in oxides and nitrides is quite a challenge. In addition to normal process variation, polymer properties vary from manufacturer to manufacturer, and the above values are guidelines only.

5.9.4 Measurements for dielectric films

Thickness and refractive index are basic measurements for lossless dielectric films. Optical methods are accurate, quick, non-contact and suitable for both research and manufacturing control applications. Accuracy of measurement is a fraction of a nanometre for both ellipsometry and reflectometry.

Reflectometry assumes a known index of refraction, but measures real thickness by fitting reflections over a wide wavelength range to d-n_f model. Thicknesses from 10 nm to 50 μm can be measured, depending on equipment and algorithm.

Ellipsometry measures thickness and refractive index in a single measurement because both the amplitude and phase of reflected polarized light are measured. For very thin films (<10 nm) optical constants are not really constants, and absolute accuracy of ellipsometry is not very good, but precision is excellent. For thicker films, multiple reflections and interference mean that the solution is periodic, with the period given by Equation 5.2:

$$d = \frac{\lambda}{2}\sqrt{n^2 - \sin^2\phi} \tag{5.2}$$

where ϕ is the angle of the incident laser beam and λ, its wavelength. Measurement at two incident angles (e.g., 50° and 70°) gives additional information, and period matching from the two measurements can give thickness of layers. When film thickness is over 1 μm, ellipsometry becomes difficult.

Ellipsometry needs a fairly large area for measurement, for example, 100 × 100 μm, while reflectometer spots can be as small as a few micrometres, which enables measurement from the structures themselves, without a dedicated test site. The easiest and quickest way to gauge thickness is from interference colours (Tables 5.6 and 5.7). The accuracy of this approach is ca. 10 nm, but the colours repeat at regular intervals, and absolute thickness determination requires additional information.

Table 5.6 Colour chart for Si_3N_4 under tungsten filament illumination

0–20 nm	Silicon
20–40 nm	Brown
40–55 nm	Golden brown
55–73 nm	Red
73–77 nm	Deep blue
77–93 nm	Blue
93–100 nm	Pale blue
100–110 nm	Very pale blue
110–120 nm	Silicon
120–130 nm	Light yellow
130–150 nm	Yellow
150–180 nm	Orange red
180–190 nm	Red
190–210 nm	Dark red
210–230 nm	Blue
230–250 nm	Blue–green
250–280 nm	Light green
280–300 nm	Orange yellow
300–330 nm	Red

Source: Reizman, F. & W. van Gelder: Optical thickness measurement of SiO_2–Si_3N_4 films on silicon, *Solid-State Electron.*, **10** (1967), 625.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Table 5.7 Colour chart for thermal SiO_2 films under daylight fluorescent lighting

Thickness (μm)	Colour	Order
0.05	Tan	
0.07	Brown	
0.10	Dark violet to red–violet	
0.12	Royal blue	
0.15	Light blue to metallic blue	
0.17	Metallic to yellow – green	I
0.20	Light gold or yellow	
0.22	Gold	
0.25	Orange to melon	
0.27	Red–violet	
0.30	Blue to violet–blue	
0.31	Blue	
0.32	Blue to blue–green	
0.34	Light green	
0.35	Green to yellow–green	
0.36	Yellow–green	II
0.37	Green–yellow	
0.39	Yellow	
0.41	Light orange	
0.42	Carnation pink	
0.44	Violet–red	
0.46	Red–violet	
0.47	Violet	
0.48	Violet–blue	
0.49	Blue	
0.50	Blue–green	
0.52	Green (broad)	
0.54	Yellow–green	
0.56	Green–yellow	III
0.57	Yellowish	
0.58	Light orange	
0.60	Carnation pink	
0.63	Violet–red	
0.68	Bluish	
0.72	Blue–green to green	IV
0.77	Yellowish	
0.80	Orange	
0.82	Salmon	
0.85	Dull light red–violet	
0.86	Violet	
0.87	Blue–violet	
0.89	Blue	
0.92	Blue–green	V
0.95	Dull yellow–green	
0.97	Yellow to yellowish	
0.99	Orange	
1.00	Carnation pink	

Source: Pliskin, W. & E. Conrad: Non-destructive determination of thickness and refractive index of transparent films, *IBM J. Res. Dev.*, **1** (1964), 43.

5.10 POLYSILICON

Polysilicon (polycrystalline silicon) is chemical-vapour-deposited by the silane decomposition reaction

$$SiH_4\ (g) \longrightarrow Si\ (s) + 2H_2\ (g)$$

$$630\,^{\circ}C,\ 400\,mTorr\ (rate \approx 10\,nm/min)$$

Undoped polysilicon is not a conductor at all, and in some applications it can be used like an insulator, provided that it is not doped at some later stage. Filling of deep trenches is such an application. Polysilicon can be doped by ion implantation and thermal diffusion processes at ca. 900 to 1000 °C just like single-crystal silicon, but there is the additional possibility of introducing dopants into the feed gas during CVD: B_2H_6 gas for p-type doping and PH_3 for n-type doping.

High doping levels of $10^{21}\,cm^{-3}$ result in polysilicon resistivity of ca. 500 μohm-cm. Electron mobility in polysilicon is an order of magnitude less than in single-crystalline materials, 10 to $50\,cm^2/Vs$. This is doping-dependent, and strongly dependent on deposition and annealing cycles.

Polysilicon deposition can be done either in the truly polycrystalline or in the amorphous (microcrystalline) regime. Grain size of film deposited at 630 °C is 30 to 300 nm, which is similar to linewidths and thicknesses in some applications. For deposition between 580 and 600 °C, grain size decreases and deposition at ca. 570 °C results in amorphous film. This choice affects surface morphology, final grain size after annealing and doping uniformity.

Polysilicon, unlike metals, can be oxidized and it tolerates all process temperatures used in microfabrication; and it can be used as a conductor in spite of its mediocre electrical properties (its grain size, resistivity and stress state will change upon annealing, which may pose problems). Polysilicon interface with thermal oxide is well characterized and polysilicon is the "metal" in MOS transistors. The MOS transistor is a capacitor, and the rather high resistivity of polysilicon is not a major disadvantage.

Polysilicon can be used as a mechanical material just like single-crystal silicon. Its mechanical constants are not unlike those of a single-crystalline material: yield strength 2 to 3 GPa versus ca. 7 GPa; Young's modulus is ca. 160 GPa for both. Thermal conductivity of polysilicon is 0.2 to 0.3 W/cm K, as against 1.57 W/cm K for a single-crystal material, and the coefficients of thermal expansion are identical. The Seebeck coefficient of polysilicon is high

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

(100–400 μV/K), and polysilicon is used in many thermoelectric devices. But CVD offers possibilities for realizing multilayer structures that cannot be made in single-crystal materials. The Fabry–Perot interferometer of Figure 1.8 utilizes two polysilicon layers, and more functionality is built in by leaving some polysilicon area undoped, which effectively results in insulating regions.

5.11.1 Amorphous silicon

PECVD of silicon from silane results in amorphous silicon with a lot of embedded hydrogen. The film is designated a-Si:H and its hydrogen content can be up to 30 atomic-% (and much less in weight %). The film is amorphous because PECVD temperatures are low, in the range of 150 to 350 °C, and the atoms do not have enough energy to find energetically favourable positions but come to rest upon impingement. Amorphous silicon can be deposited on glass, and its biggest industrial application is in the fabrication of thin-film transistors (TFT) for active matrix displays. Electron and hole mobilities in annealed a-Si:H are only ca. 1 to 10 cm^2/V s, which is adequate for switching transistors. *In situ* doping during PECVD is crucial in TFT fabrication because high-temperature doping cannot be done on glass substrates.

Another major application of a-Si:H is in solar cells. Single-crystal silicon has fairly low optical absorption in the visible wavelengths (Table 4.1) but a submicrometre layer of a-Si:H layer can absorb practically all the light impinging on it. Again, glass is a potential substrate, but even cheaper substrates like steel or polymers are being considered.

5.11 SILICIDES

A rather interesting class of conducting thin films is the silicides: compounds of silicon and metal, for example, $TiSi_2$, $CoSi_2$, NiSi, WSi_2 and PtSi. Silicides combine the good properties of silicon, such as high-temperature stability and metal-like resistivity, with the lowest values of ca. 15 μohm-cm for resistivity (Table 5.8).

Silicides are formed by two major methods: CVD and solid-state reaction of metal thin film and silicon. CVD silicides need to be etched like any other films, but the solid state–reacted silicide patterns can be made without silicide etching. The desired pattern is defined in oxide, and metal is deposited. Upon annealing, metal–silicon reaction takes place in those areas where metal and silicon are in contact, but on oxide the metal does not react. The unreacted metal can be etched away to leave silicide and oxide (Figure 5.12).

The silicide is formed under the original surface and the surface of the resulting silicide is approximately at the level of the original silicon surface. This volume expansion/thickness change needs to be accounted for when reacted silicides are made.

Silicide CTEs are typically 15 ppm/°C. Young's moduli for silicides are of the order of 100 GPa. Silicides will be discussed in more detail in Chapter 19.

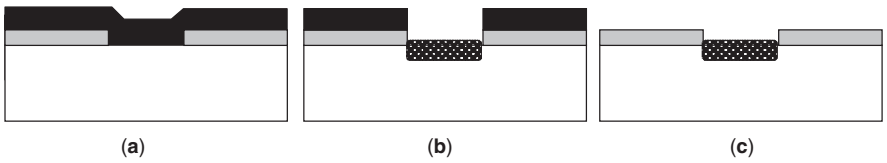


Figure 5.12 Silicide formation by metal–silicon reaction: (**a**) metal sputtering on wafer (**b**) reaction at metal–silicon interface; no reaction on oxide and (**c**) selective etching of unreacted metal leaves silicide

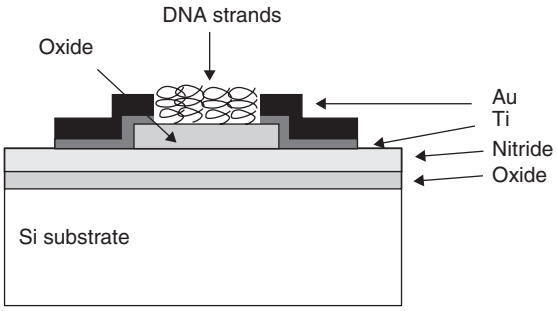
Table 5.8 Silicide properties

Silicide	Resistivity	Formation	Selective metal: silicide etch
$TiSi_2$	15–20 μohm-cm	Ti/Si reaction at ca. 750 °C	$NH_4OH:H_2O_2$
$TiSi_2$	15–20 μohm-cm	CVD $TiCl_4/SiH_2Cl_2/H_2$	–
$CoSi_2$	15–20 μohm-cm	Co/Si reaction at 500 °C	$HCl:H_2O_2$ 3:1
NiSi	15–20 μohm-cm	Ni/Si reaction at 400 °C	HNO_3
WSi_2	30 μohm-cm	CVD WF_6/SiH_2Cl_2 at 400 °C	–
PtSi	30 μohm-cm	Pt/Si reaction	$HCl:HNO_3$ 3:1

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

5.12 EXERCISES

1. Resistor design: How would you fabricate (a) 1 kΩ, (b) 10 kΩ resistors in a process in which minimum linewidth is 3 μm?
2. Polysilicon sheet resistance is 50 Ω/sq. What is polysilicon thickness?
3. The DRAM memory cell is a capacitor. If the cell area is 1 μm^2, with a 4 nm oxide as the capacitor dielectric, and the operating voltage is 2 V, calculate the number of electrons stored in the memory cell.
4. The CVD oxide process is designed to target 500 nm thickness. If the wafers are violet, and the violet changes to pink on wafer edges, what is repeatability and uniformity of this deposition process?
5. If silane (SiH_4) flow in a single-wafer (150 mm) PECVD reactor is 5 sccm (cm^3/min), what is the theoretical maximum deposition rate of amorphous silicon?
6. If 20 nm of nickel reacts with overabundance of silicon, how thick a layer of NiSi will be formed? Densities: Si–2.3 g/cm^3, Ni–8.9 g/cm^3, NiSi–7.2 g/cm^3.
7. $CoSi_2$ is formed by cobalt thin-film reaction with silicon. What is the position of the $CoSi_2$ surface relative to the original silicon surface? Densities: Co–8.9 g/cm^3, $CoSi_2$–5.3 g/cm^3.
8. If ECD current density is 100 mA/cm^2, what will be the nickel deposition rate?
9. Design a process to fabricate a DNA microarray pixel shown below. (Attached gold-labelled DNA strands form electrical contact between gold electrodes). Redrawn from Xue, M. *et al*. (2002).



REFERENCES AND RELATED READINGS

Besser, R.S. *et al*: Chemical etch rate of plasma-enhanced chemical vapor deposited SiO_2 films, *J. Electrochem. Soc.*, **144** (1997), 2859.

Cote, D.R. *et al*: Plasma-assisted chemical vapor deposition of dielectric thin films for ULSI semiconductor circuits, *IBM J. Res. Dev.*, **43**(1–2) (1999), 5.

Elshabini-Riad, A. & F.D. Barlow III: *Thin Film Technology Handbook*, McGraw-Hill, 1998.

Hu, C.-K. *et al*: Electromigration of Al(Cu) two-level structures: effect of Cu kinetics of damage formation, *J. Appl. Phys.*, **74** (1993), 969.

Jiles, D.C. & C.C.H. Lo: The role of new materials in the development of magnetic sensors and actuators, *Sensors Actuators*, **106** (2003), 3; special issue on magnetic sensors and actuators.

Mahan, J.: *Physical Vapor Deposition of Thin Films*, Wiley, 2000.

Ohring, M.: *The Materials Science of Thin Films*, Academic Press, 1992.

Pliskin, W. & E. Conrad: Non-destructive determination of thickness and refractive index of transparent films, *IBM J. Res. Dev.*, **1** (1964), 43.

Reizman, F. & W. van Gelder: Optical thickness measurement of SiO_2-Si_3N_4 films on silicon, *Solid-State Electron.*, **10** (1967), 625.

Ruythooren, W. *et al*: Electrodeposition for the synthesis of microsystems, *J. Micromech. Microeng.*, **10** (2000), 101.

Shacham-Diamand, Y. & V.M. Dubin: Copper electroless deposition technology for ultra-large-scale-integration (ULSI) metallization, *Microelectron. Eng.*, **33** (1997), 47.

Smith, D.L.: *Thin-film Deposition: Principles and Practise*, McGraw-Hill, 1995.

Srikar, V.T. & S.M. Spearing: Materials selection in micromechanical design, *J. MEMS*, **12** (2003), 3.

Vehkamäki, M. *et al*: Atomic Layer Deposition of $SrTiO_3$, *Chem. Vapor Deposit.*, **7** (2001), 75.

Xue, M. *et al*: A self-assembled conductive device for direct DNA identification in integrated microarray based system, *IEDM 2000* (2002), p. 207.

IBM J. Res. Dev., **42**(5) (1998); special issue on electrochemical microfabrication.

**Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030**

6

Epitaxy

Epitaxial deposition is a very special case of thin-film deposition. Epitaxy means the growth of a single crystalline layer on top of a single crystalline substrate. The growing layer registers the crystalline information from the layer below. In order to do so properly, the crystal lattices of the two layers must be closely matching. Because crystal information is 'transmitted' across the substrate–film interface, surface quality of the starting wafers is of paramount importance. Defects, be they native oxide, crystal defects (dislocations, stacking faults) or metal impurities, can destroy epitaxial growth. Epitaxy is a delicate process, and high quality epitaxial films are difficult to make. Epitaxy can fail partially and result in a defective single crystalline material, or it can fail completely, and result in a polycrystalline film. Whether the defective material is usable for devices depends on the density and location of those defects: if defects are confined to the substrate–epi interface and the epilayer is mostly defect-free, the material is usable; but this depends on the device operating principle, and engineering judgement is needed to decide on acceptable defect levels.

Epitaxy has nothing to do in particular with silicon or semiconductors: epitaxy is a phenomenon that is seen in many classes of solids. However, semiconductor-on-semiconductor epitaxy, both Si/Si and $GaAs/Al_xGa_{1-x}As$, has been, and remains, the most voluminous industrial application of epitaxial deposition. Insulators like calcium fluoride (CaF_2) and yttrium oxide (Y_2O_3) can be grown epitaxially on silicon, and so can cobalt silicide ($CoSi_2$). Epitaxial silicon can be grown on sapphire (crystalline aluminum oxide, Al_2O_3) and epitaxial cerium oxide, CeO_2, can be grown on silicon, and epitaxial YBCO superconductor can be grown on CeO_2.

In solid phase epitaxy (SPE), the film registers the crystalline structure from the underlying single-crystalline substrate. Amorphous films can thus be converted to epitaxial films by annealing. Of course, all the limitations of clean surfaces, matching lattice and so on still apply. Epitaxy from liquid phase (LPE) is also possible: both saturated solutions and melts can be used as sources for epitaxial growth. LPE was the dominant technology in the early days of III-V semiconductor laser and LED fabrication, but it has largely been superseded by gas-phase and vacuum systems.

In homoepitaxy, the substrate and the growing film are the same material. Silicon epitaxy on silicon enables freedom in doping level and doping type tailoring. Epitaxial wafers account for some 20% of all wafers sold. A lightly doped epitaxial p-type layer (10 ohm-cm) can be grown on a heavily p-doped substrate wafer (0.2 ohm-cm). This is the material for advanced microprocessors and other high-performance logic circuits. n-Silicon on p-substrate is used in many micromechanical devices because of electrochemical etch stop. The number and thickness of layers is practically unlimited: in IGBT (Insulated Gate Bipolar Transistor) power transistors a moderately doped n-layer is grown first, followed by a thicker lightly doped layer. In semiconductor laser structures, there can be hundreds of epitaxial layers. Another benefit of epitaxy is the absence of oxygen and carbon, which are always present in CZ-silicon. Uniformity of epitaxial layers is good, for both thickness and resistivity, and if very tight resistivity specification is needed, epitaxial wafers override bulk silicon wafers.

Hardware for epitaxial deposition is varied: in principle, almost any deposition system can be used for epitaxial deposition under some conditions but there are a couple of established technologies for epitaxial deposition. CVD epitaxy of silicon with $SiH_{4-x}Cl_x$ ($0 \leq x \leq 4$) source gases is the standard method. In the compound semiconductor field, MOCVD (Metal Organic CVD; also known as MOVPE for Vapour Phase

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Epitaxy) and MBE, molecular beam epitaxy, are the two main epitaxy techniques.

The term *epi-poly* is used in micromechanics. It is self-contradictory: epitaxial films are single crystalline, and poly means polycrystalline. What is meant is that a CVD epireactor has been used to deposit a thick layer of silicon, using epi growth conditions (temperatures around 1100 °C), but growth is on an amorphous substrate, for example, SiO_2, resulting in a polycrystalline film. Standard polysilicon deposition in an LPCVD reactor at 630 °C is a very slow process, ~10 nm/min; whereas epitaxial growth rates are of the order of 1 μm/min, a factor of 100 higher. Typical epi-poly thicknesses are 10 to 20 μm, compared with 0.1 to 2 μm typical of LPCVD polysilicon, which is used as a CMOS gate and surface micromechanics structural layer.

6.1 HETEROEPITAXY

Epitaxy on dissimilar materials is termed *heteroepitaxy*, with examples such as AlAs on GaAs, GaN on SiC or SiGe on Si. The $Al_xGa_{1-x}As$ system is favourable because lattice constants of all GaAs and AlAs differ by

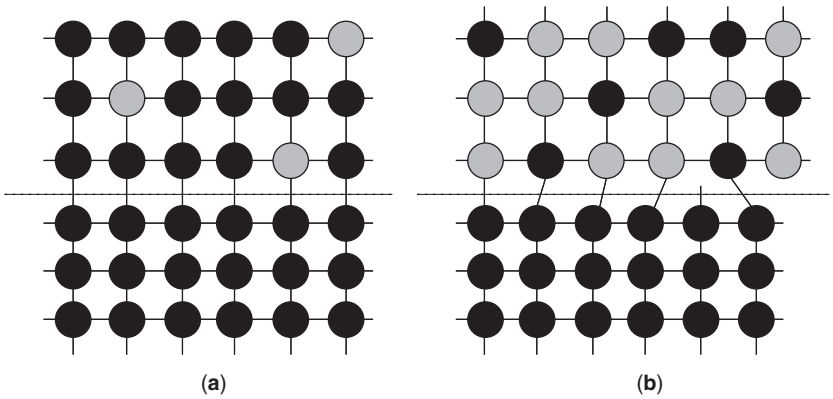


Figure 6.1 $Si_{(1-x)}Ge_x$ alloy grown on silicon (Si black, Ge gray): (**a**) strained (pseudomorphic) epitaxial SiGe layer with lattice constant matching silicon lattice constant parallel to the surface, but relaxed in the perpendicular direction; (**b**) large lattice constant difference leads to misfit dislocations

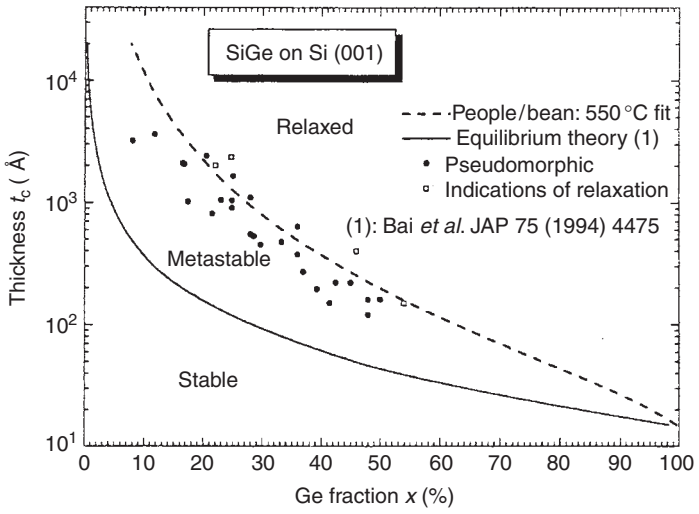


Figure 6.2 In the stable region, the SiGe film on silicon is so thin that it conforms to the silicon lattice; above critical thickness, it relaxes via misfit dislocations. From Herzog, H.-J. *et al*. (2000), by permission of Elsevier

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

less than 0.2%, and superlattices of AlAs/GaAs/AlAs type can be grown easily, with periods down to atomic layer thickness, equipment limitations allowing.

Heteroepitaxy for silicon materials is difficult because no good lattice matching materials can be found. The most important application is the growth of $Si_{(1-x)}Ge_x$ on silicon. The lattice constant of silicon is 5.43 Å and that of germanium is 5.66 Å. The lattice constant of SiGe alloys is described fairly well as a linear combination of silicon and germanium lattice constants by

$$a_{Si_{(1-x)}Ge_x} = (1-x)a_{Si} + xa_{Ge} \qquad (6.1)$$

There exists a critical thickness t_c (which depends on lattice constant and therefore germanium fraction) below which mismatch can be accommodated by elastic deformation, as shown in Figure 6.1(a). The relation tying epitaxial thickness and germanium fraction (and therefore lattice constant) is shown in Figure 6.2. Above t_c, the lattice relaxes via misfit dislocations, and the crystalline quality may become useless for device applications.

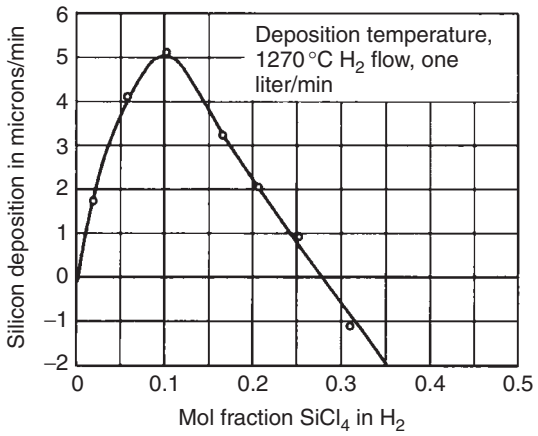


Figure 6.3 Epitaxial growth rate as a function of $SiCl_4/H_2$ flow ratio. Typical growth condition is 1 μm/min, $SiCl_4/H_2$ (1%/99%). Above ca. 2 to 3 μm/min the resulting film is polycrystalline, not epitaxial. From ref. Theurer, H. (1961), by permission of Electrochemical Society Inc.

6.2 CVD HOMOEPITAXY OF SILICON

As an example of homoepitaxy, CVD silicon epitaxy is described. The reactor is heated to ca. 1200 °C under hydrogen flow, which reduces native oxide.

$$SiO_2\ (s) + H_2\ (g) \longleftrightarrow SiO\ (\text{vapour}) + H_2O\ (\text{vapour}) \qquad (1150\text{–}1200\,^\circ\text{C}) \qquad (6.2)$$

Growth commences when silane gases of the type SiH_xCl_{4-x} $(0 \le x \le 4)$ are introduced into the reactor.

$$SiH_4\ (g) \longrightarrow Si\ (s) + 2H_2\ (g), \qquad T = 1000\,^\circ\text{C} \qquad (6.3)$$

$$SiCl_4\ (g) + 2H_2\ (g) \longleftrightarrow Si\ (s) + 4HCl\ (g), \qquad T = 1250\,^\circ\text{C} \qquad (6.4)$$

The latter reaction is reversible, and cleaning is possible with HCl when the reaction proceeds from right to left, that is, hydrogen chloride etching of silicon. Excessive etching should be avoided because surface roughness tends to increase in etching. Silicon tetrachloride can also be used as a silicon etchant.

$$SiCl_4\ (g) + Si\ (s) \longrightarrow 2SiCl_2\ (g) \qquad (6.5)$$

This reaction can be prevented when the $SiCl_4$ fraction is limited below 27% (see Figure 6.3), but much more dilute silanes are usually used, with 99% hydrogen typical.

The $SiCl_4$ process temperature is, however, very high and undesirable dopant diffusion takes place during epitaxy. Low temperature, and therefore minimal diffusion, is an important consideration when sharp interfaces must be made. SiH_4 reaction is better in this respect, but due to lower temperature, the rate is lower. Trichlorosilane (TCS), $SiHCl_3$, and dichlorosilane (DCS), SiH_2Cl_2 are good compromises between deposition rate and operating temperature (see Equation (4.3)).

$$SiH_2Cl_2\ (g) \longleftrightarrow Si\ (s) + 2HCl\ (g) \quad T = 1150\,^\circ\text{C} \qquad (6.6)$$

Typical epitaxial growth rates are 1 to 5 μm/min. They depend on the silane gas chosen, on temperature and on flows. Epi reactions are subject to general CVD reaction rate laws discussed in Chapter 5 (see, for instance, Figure 5.6). Growth rate can be increased by operating at higher temperature but above certain limits, gas phase nucleation or some other mechanisms lead to polycrystalline rather than epitaxial deposits. At lower temperatures, surface reactions may be too slow for epitaxial arrangements to take place, and polycrystalline films result.

Epitaxial layer growth is assumed to proceed at surface kinks and steps (Figure 6.4). These are energetically favourable nucleation sites, compared to flat open areas. Perfectly flat surfaces offer inherently fewer points for atoms to position themselves, and growth is therefore

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

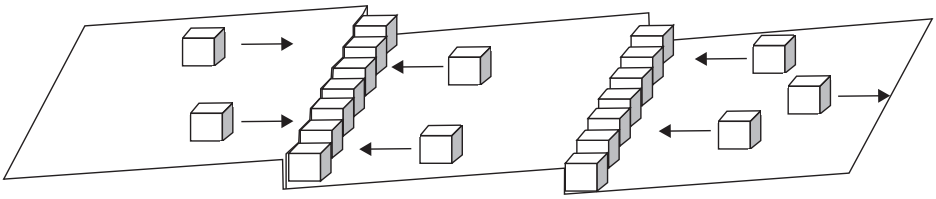


Figure 6.4 Terrace step kink (TSK) growth model of epitaxy: growth proceeds at kinks, and atoms on flat surface diffuse to energetically favourable positions at kinks. Wafer miscut creates terraced structure

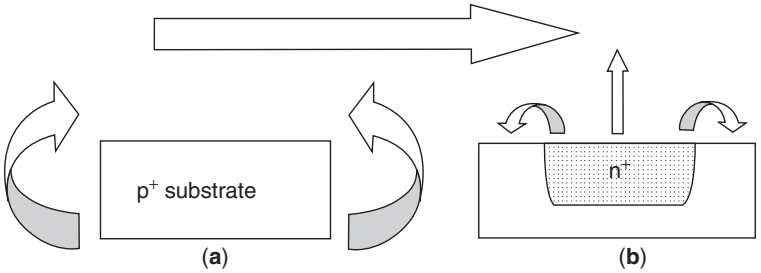


Figure 6.5 Autodoping: dopants evaporated from heavily doped substrate add to intentionally added dopant (substrate autodoping); dopants from heavily doped regions influence doping locally (lateral autodoping)

difficult. It can be aided by miscut wafers: instead of slicing the ingot perfectly, for example, a 3° misorientation is used (typical of <111> material). Atomic steps so created act as nucleation sites for epitaxy.

6.2.1 Doping of epilayers

Epitaxial layer doping level and dopant type can be chosen independent of the substrate. Gaseous dopants, PH_3, B_2H_6 and AsH_3, are added to the source gas flow, enabling doping during epitaxial growth. Dopant concentration can be varied over 7 orders of magnitude ($10^{13}-10^{20}\,cm^{-3}$). In many applications, several epilayers with different doping levels and/or types are grown sequentially, or in graded structures where composition or doping level changes in minor steps, for example, from Si to $Si_{0.7}Ge_{0.3}$ in tens of increments of germanium concentration.

Epitaxial growth need not be the first process step: doped silicon is also single-crystalline silicon and epitaxy on it works just as well. In bipolar transistor fabrication, a buried layer formation by diffusion is the first step (see Figure 3.2), followed by epitaxial deposition of a lightly doped layer on top of a heavily doped buried layer. Base and emitter diffusions will then be done in this lightly doped epitaxial layer. More discussion on epitaxy on structured wafers can be found in Chapter 26.

Because of the high temperatures involved, dopant diffusion will inevitably take place during epitaxy. If the epilayer doping level is lower than that of the substrate, the epilayer will be doped from the substrate through two different mechanisms: (1) solid-state diffusion across the substrate–epi layer interface and (2) dopant atom outdiffusion from the substrate into gas stream and subsequent vapour phase doping, known as *autodoping* (Figure 6.5). Autodoping depends on the volatility of dopants, with antimony (Sb) being the best (the lowest vapour pressure) and arsenic and boron having somewhat higher, and phosphorus the highest vapour pressure. Autodoping comes both from the substrate itself, and also from any doped regions that have been made in steps preceding epitaxy.

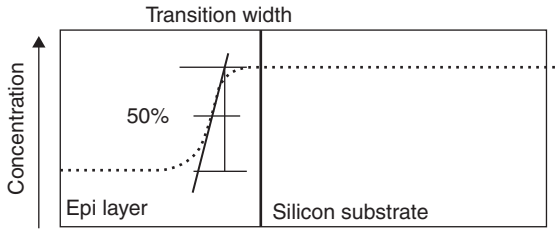


Figure 6.6 Transition width at substrate–epi interface. Lightly doped epitaxial layer on heavily doped substrate

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

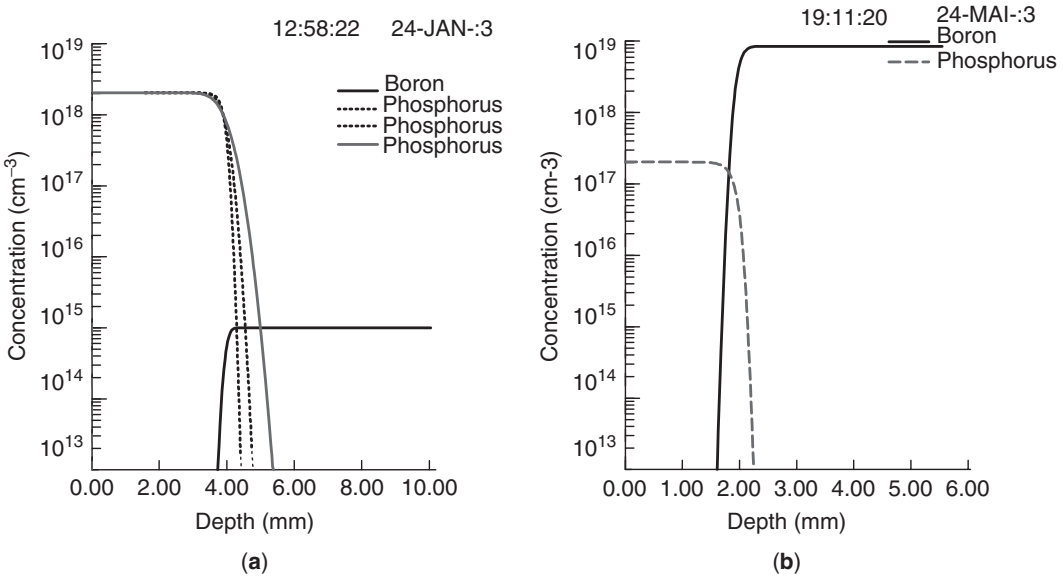


Figure 6.7 (**a**) ICECREM simulation of epitaxial interface sharpness: three different growth temperatures (1050 °C, 1100 °C, 1150 °C) have been used to grow a nominally 4 μm thick phosphorous doped epilayer on boron doped substrate. Low temperature leads to sharper interface; (**b**) lightly phosphorus doped epi on heavily boron-doped substrate

6.2.2 Measurement of epitaxial deposition

Three measurements must be carried out on epitaxial wafers: thickness, resistivity and surface quality. Surface quality is assessed first and foremost by optical inspection: pyramids, mounds and hillocks scatter light, which can be detected by optical methods. Nomarski interference contrast microscope detects surface height differences and infrared depolarization reveals stresses. Laser scattering measures particles and microroughness. Optical methods are fast, and 100% of wafers are inspected.

Thickness of epilayers can be measured by Fourier transform infrared (FTIR) spectroscopy: constructive and destructive interference from reflections at the surface and at the substrate–epi interface are detected. FTIR requires, however, a highly doped substrate (resistivity below 0.025 ohm-cm). On resistive substrates, spreading resistance profiling (SRP) is used. SRP requires sample bevelling, that is, it is sample-destructive. One wafer in 25 or one in 100 is measured by SRP. SRP can also measure multilayer structures. Transition width measurement is done by SRP or SIMS, and it is done, for example, once for 1000 wafers.

SRP also measures resistivity, but simpler and faster methods are used for routine measurements. Resistivity is measured by the mercury probe capacitance–voltage method (Hg-CV-method) for p/p and n/n structures and by the four-point probe method for n/p and p/n structures. In both methods, a metal contact is made on silicon, even though liquid mercury-drop contact is much more benign than tungsten-needle contact of 4PP. Wafers are not usable after metal probes. Non-contact measurements would be much in need, but most are rather cumbersome and require special conditions to be fulfilled.

6.3 SIMULATION OF EPITAXY

Epitaxy simulators currently used in process integration studies are not physically based. A true physical simulator would use temperature, flow rate and surface reaction rate constants as inputs, and it would reproduce growth rate and dopant distribution as the outputs. Instead, epitaxy simulators are really hybrids between film deposition and diffusion simulators: deposition rate and temperature are given, and the dopant profile is calculated from diffusion constants at the relevant temperature.

The inputs for the epitaxy simulator are the following:

- dopant type of wafer
- growth rate and time
- growth temperature
- dopant type and concentration in the flow.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

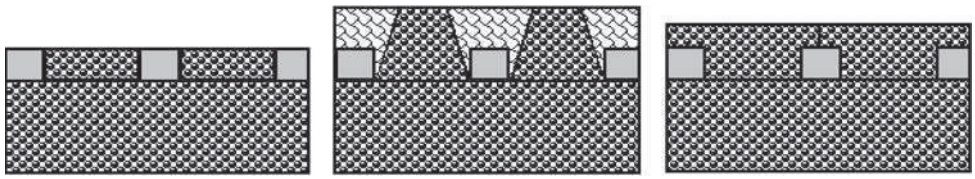


Figure 6.8 (**a**) Selective epitaxy: no deposition on oxide; (**b**) blanket deposition: epitaxy on single-crystalline substrate, polycrystalline on oxide; (**c**) epitaxial lateral overgrowth (ELO): merging of epitaxial film fronts over oxide

Such a semiempirical simulator can predict the dopant profile across the substrate–epi interface, taking into account both outdiffusion from the substrate and diffusion from the epilayer into the substrate.

Some rough guides to gas-phase dopant concentration and the resulting epilayer doping are given below:

Dopant in gas phase	Dopant in epitaxial film
10^{-10} bar	10^{15} cm^{-3}
10^{-8} bar	10^{17} cm^{-3}
10^{-6} bar	10^{19} cm^{-3}

Note that phosphorus and boron incorporation into growing silicon is very strong: its concentration in the film is much higher than its gas-phase concentration. Arsenic incorporation into the epitaxial film is somewhat more pronounced.

Simulation of epitaxial deposition by ICECREM is shown in Figure 6.7. In the simulation shown in Figure 6.7, the same deposition rate, 0.2 μm/min, has been used for all temperatures. This is a limitation in epitaxy simulation: rates are temperature-dependent, but they have to be manually given; they do not follow from first principles.

6.4 ADVANCED APPLICATIONS OF EPITAXY

If there are both oxide and single-crystal silicon areas on the wafer, growth will be epitaxial on silicon, and polycrystalline on the oxide (Figure 6.8). In selective epitaxial growth (SEG), the film grows only in those areas where single-crystal silicon is present; elsewhere, growth is suppressed. Selective epitaxy can be done many times over, as long as high-quality seed is available. Masking materials have to be compatible with the process steps in question: silicon dioxide and silicon nitride are the obvious candidates.

Epitaxial growth requires crystal orientation information from the substrate, but once this information is registered, epitaxial growth can continue over amorphous or polycrystalline material. Epitaxial lateral overgrowth (ELO) technique incorporates patterned seed areas, oxide isolation and lateral overgrowth. One of the main problems in ELO is the point where the two growth fronts merge: defect density can be very high.

Crystallization of amorphous material can be used to obtain epitaxial films. Chemical vapour–deposited α-Si on sapphire single-crystal wafer can be turned into a single-crystalline film under suitable annealing conditions. Defect densities vary enormously for different heteroepitaxial and re-crystallization schemes; while sometimes defective epitaxy or partial re-crystallization can be beneficial for device operation, defects will hinder all device functions at other times.

6.5 EXERCISES

1. What are the resistivities of the substrates and epilayers in Figure 6.7?
2. Can a laboratory scale with 0.1 mg resolution be used for epilayer thickness measurements?
3. Growth rates as a function of temperature are given below for SiH_4 epitaxy. If deposition takes place at 1000 °C, is it in mass-transfer or surface reaction–limited regime?

700	750	800	850	900	950	1000	1050	1100°
0.04	0.09	0.2	0.4	0.5	0.6	0.7	0.75	0.8 μm/min

4S. For an n^+/n^- structure (substrate 10^{18} cm^{-3}, epi 10^{15} cm^{-3}), calculate the transition width as a function of epitaxy temperature for a 4 μm thick epilayer.

5S. Initial wafer doping level is 10^{15} cm^{-3} phosphorus. Epilayer is boron-doped with 10^{17} cm^{-3} concentration. Calculate junction depth as a function of growth temperature.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

6S. If pnp-bipolar transistors are made, the buried layer has to be p-type. Calculate boron updiffusion for different epitaxy conditions when the buried layer doping is $10^{18}\,cm^{-3}$ and epilayer doping is $10^{15}\,cm^{-3}$.

REFERENCES AND RELATED READINGS

Baliga, J.B.: *Epitaxial Silicon Technology*, Academic Press, 1986.

Crippa, D., D.R. Rode & M. Masi: Silicon epitaxy, *Semiconductors and Semimetals*, Vol. 72, Academic Press, 2001.

Herzog, H.-J. *et al*: SiGe-based FETs: buffer issues and device results, *Thin Solid Films*, **380** (2000), 36.

Meyerson, B.S.: UHV/CVD growth of Si and Si:Ge alloys: chemistry, physics, and device applications, *Proc. IEEE* '80 (October 1992), p. 1592.

Ohmi, T. *et al*: Formation of device-grade epitaxial silicon films at extremely low temperatures by low-energy bias sputtering, *J. Appl. Phys.*, **66** (1989), 4756.

Theurer, H.: Epitaxial silicon films by the hydrogen reduction of $SiCl_4$, *J. Electrochem. Soc.*, **108** (1961), 649.

Wu, Y.H. *et al*: The effect of native oxide on epitaxial SiGe from deposited amorphous Ge on Si, *Appl. Phys. Lett.*, **74** (1999), 528.

**Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030**

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

7

Thin-film Growth and Structure

In this chapter, we deal with deposition processes and the resulting film structures. Interface stability and sharpness, grain size, texture, stress and other film properties are dependent on film deposition processes, but they depend on preceding and subsequent process steps too. Structures already made on the wafer set various limitations on the processing conditions. Now, we will also consider deposition on non-planar surfaces, which introduces new considerations.

7.1 GENERAL FEATURES OF THIN-FILM PROCESSES

The general features of thin-film deposition processes are visualized in Figure 7.1. Thin-film deposition involves thermal physics, fluid dynamics, plasma physics, gas-phase chemistry, surface chemistry, solid-state physics and materials science. We must deal with source materials (sputtering targets, precursor chemicals, electrolyte compositions), we must address the transport of source material to the substrate (in high vacuum, low vacuum, atmospheric pressure or liquid), and we have to understand surface processes (adsorption, reaction, desorption, ion-bombardment induced effects). Characterization of films entails dozens of techniques ranging from optical to nuclear, electrical to mechanical. This multidisciplinarity leads to a great number of phenomena and models that must be taken into account, both in experimental work and in simulation.

There are a few basic methods of source excitation and their different configurations. Thermal activation can be either resistive, photothermal or electron beam–induced, and laser or ion beams can be used. Plasma sources range from simple DC-diodes to microwave, helical and inductive configurations. In the liquid phase, the choices are less numerous, and electrochemical and chemical potential differences are the main driving forces.

Transport of material from the source to the wafer can be directional or diffuse. With directional deposition reactor geometry, the wafer position and the structures on the wafer determine the flux that can be easily modelled. Evaporation and molecular beam epitaxy (MBE) are examples of directional, line-of-sight deposition systems. With diffuse transport, the arrival of the depositing specie is usually difficult to model, as in mass-transport limited regime of chemical vapour deposition (CVD).

Film deposition on the substrate surface is a sum of many factors. In the first approximation, the deposition is independent of the substrate (this distinguishes the deposition from growth processes such as thermal oxidation and epitaxy, which are intimately coupled with the substrate). But the surfaces do interact with the deposition processes via available chemical bonds, contamination and crystallography. An important parameter is the sticking coefficient, or the probability that an impinging particle will remain on the surface. A high-sticking coefficient means that the particle will come to rest at the point of impingement, and a low-sticking coefficient means that only the energetically favourable attached specie will stick, and the others will desorb. Sticking coefficients range from 0.001 to 1, and they are generally lower for CVD processes than for physical vapour deposition (PVD).

Even if no annealing is done immediately after film deposition, the films will experience thermal treatments during subsequent processing. Thermal loads from these treatments can be considerable, and they affect many film properties, such as grain size, resistivity and stress. Film surfaces and interfaces will be modified during these anneal steps by diffusion, dissolution or chemical reactions.

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

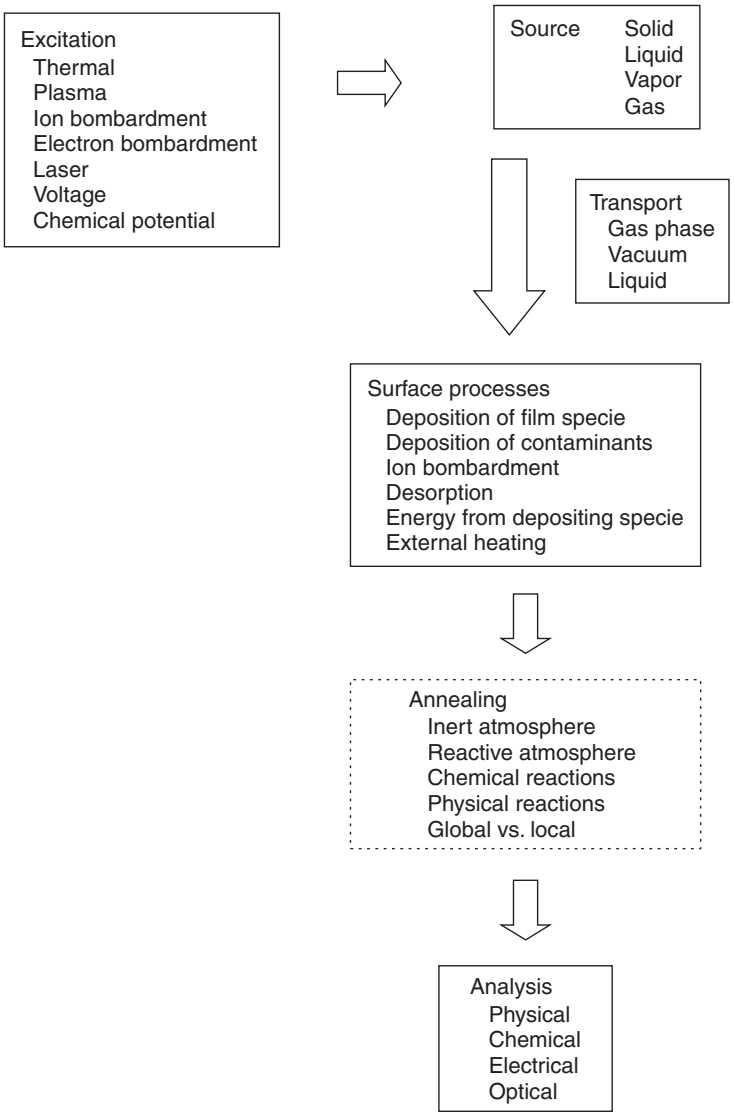


Figure 7.1 General features of thin film deposition processes

7.2 PVD-FILM GROWTH AND STRUCTURE

Atoms impinging on a surface attach to the surface either with chemical bonds ($\approx$1 eV; chemisorption) or by short-range van der Waals forces ($\approx$0.3–0.4 eV; physisorption).

These adatoms are able to move because of their own initial energy or by substrate-supplied energy or because they receive energy from the impinging particles.

There are two main modes of film growth: 2D and 3D (Figure 7.2). Two-dimensional growth, also called *layer-by-layer* growth, is the preferred mode. It is encountered in many epitaxial depositions. Three-dimensional growth is also known as *island growth*. Island growth is common when metals are deposited on insulators where the bonds between film atoms are stronger than the bonds between film atoms and the substrate. A third mode, called Stranski–Krastanov, is a mixture of 2D- and 3D-modes. Understanding of growth mechanisms is elusive and it is difficult to predict which growth mode would take place.

If we measure the early stages of thin-film growth by surface-sensitive techniques, for example, Auger

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

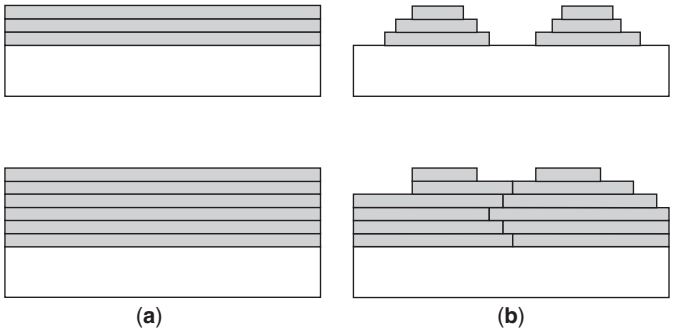


Figure 7.2 Thin-film growth modes: (**a**) 2D (layer-by-layer) and (**b**) 3D (island) growth. Early stage and coalescence

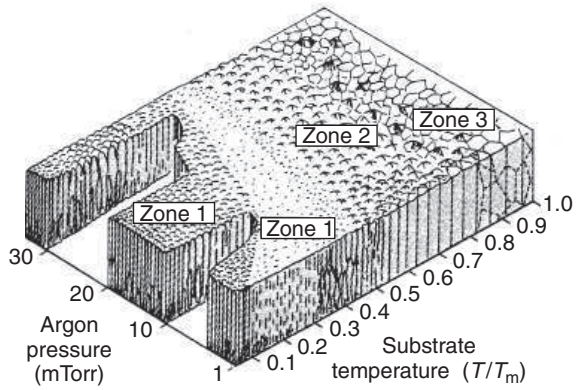


Figure 7.3 A zone model of sputtered thin-film microstructure. Reproduced from Thornton, J.A. (1986), by permission of American Inst of Physics

electron spectroscopy or X-ray photoelectron spectroscopy (XPS) (which probe 1 or 2 nm deep), we can distinguish the mechanisms: in 2D-growth mode; the signal from the substrate quickly dies out because the whole surface becomes covered by the deposited layer. In 3D-mode, the substrate signal slowly decreases as the proportion of open substrate area is diminished.

In the initial stages of 3D-growth, numerous small nuclei are formed on the surface. This is a transformation from vapour phase to solid phase. These small nuclei are mobile, and they grow by merging with other nuclei, but they can also incorporate atoms from the vapour phase. Some of the impinging atoms re-evaporate immediately and do not contribute to growth, and some small nuclei also re-evaporate. The nuclei grow in size to become islands, but remain separate, and more nuclei can form on the area between the islands. Coalescence is driven by surface energy (and surface area) minimization, like the droplet movement on a surface. Islands merge eventually to form a continuous layer. For PVD metal films this happens at ca. 10 to 20 nm thickness (100–200 atomic layers). Films thinner than this are optically transparent but they can be electrically conductive (percolated). Such films have applications as permeable electrodes in gas sensors and as top metals in optical devices.

Zone models of PVD explain the structure of thin films (Figure 7.3). The first question is which materials will form amorphous films and which will result in (poly) crystalline films. Silicon and other covalently bonding materials often end up as amorphous films, and many compounds and metal alloys with dissimilar-sized atoms similarly result in amorphous films. Elemental metal deposition usually results in polycrystalline films. The crystallinity of the sputtered films is determined by complex interactions between the substrate (its chemical and structural features and temperature) and the growing film. In the zone-model, pressure and temperature are the main variables to explain film

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

microstructure (temperatures are normalized to melting point temperatures, T/T_m, in K). Zone 1 is small-grained and porous. Zone 2 has larger columnar grains and Zone 3 exhibits still larger grains. The intermediate region is termed *Zone-T* (for transition).

Z1 is the region where the low momentum of the impinging specie is combined with slow chemical processes due to low temperature: the film atoms come to rest almost immediately and do not move. This leads to a porous structure with columnar grains (see Figure 3.6 for simulated columnar-grain structure). Such a structure is under moderate tensile stress. The voids between the grains are nanometre-sized, which leads to measurable density reduction and poor stability because of the absorption of moisture and oxygen. Impurities such as oxygen can change the intrinsic stress from tensile to compressive and complicate the simple model described above.

At lower pressure, ion bombardment induces densification of the film, and the film stress is highly tensile. A further increase in ion bombardment (at lower pressure or higher sputtering power) leads to the disappearance of voids and conversion to compressive stress. Higher temperature leads to enhanced surface diffusion that can be calculated from Equation 7.1:

$$x^2 = \sqrt{4Dt} \tag{7.1}$$

where $D = D_0 \exp(-6.5T_m/T)$ and surface diffusion constant D_0 is of the order of $10^{-7}\,m^2/s$ and t is the time it takes to deposit the next atomic layer. For atoms to diffuse distances similar to void sizes (~nanometre), Equation 7.1 can be used to estimate temperatures where transition from Z1 to Zone T takes place.

Z2 occurs at $T/T_m > 0.3$, so the surface diffusion is significant. The grains grow larger, and the defects are eliminated. Z3 occurs at $T/T_m > 0.5$, and the diffusion process is very fast. Elimination of the voids enhances diffusion. The films are annealed during deposition. The grains are more isotropic and the films 'lose memory' of the deposition-process details.

The final grain size is determined by subsequent annealing steps. The sputtered aluminium grain size is ca. 0.5 μm, similar to a typical film thickness. In 3 μm lines, there are always many grains across the line, but in 0.5 μm lines, the situation changes dramatically: there are practically no three-grain boundaries and the grains are end-to-end, known as *bamboo structure*. All processes that depend on grain boundaries, such as diffusion and electromigration, are strongly affected.

Film structure can change not only continuously as described above but also abruptly. Tantalum films sputtered under different conditions can end up in either body centred cubic (bcc) structure or as tetragonal β-Ta. Resistivity of bcc-Ta is ca. 20 μohm-cm with temperature coefficient of resistivity (TCR) 3800 ppm/°C. Values for β-Ta are ca. 160 μohm-cm and 178 ppm/°C, respectively (see Figure 2.8 for another tantalum deposition experiment). In Chapter 19, $TiSi_2$ phase transformation upon annealing will be discussed.

Grains in polycrystalline films can have any crystal orientation, but in practice, films are often strongly textured: the distribution of grain orientations are along one or two main crystal planes. For example, aluminium films usually have a (111) texture, that is, (111) planes are parallel to the wafer surface. For undoped LPCVD, polysilicon (110)-orientation crystals dominate, but for *in situ* phosphorus doped poly (311) is the dominant orientation.

The texture is established during deposition, and it is not much affected by subsequent annealing steps below (2/3) T_m even though the grain size is. Texture inheritance is common: subsequent films easily acquire the same texture as the underlying film. Thin seed layers can therefore be used to modify the thick layers. This is true for CVD and electrodeposition too.

7.2.1 Characterization of PVD films

PVD films, especially sputter-deposited films, can be modified by a number of parameters. System configuration and geometry come to play via target-substrate distance, base pressure/gas phase impurities and power coupling scheme/bias voltage; and process parameters such as pressure and power affect the momentum of the impinging atoms and ions, and substrate temperature is important for desorption, diffusion and reactions.

Collimated sputtering is a technique in which a mechanical grid is placed between the anode and the cathode, and off-angle atoms do not contribute to the flux arriving at the wafer, but are deposited on the collimator walls. Collimated sputtering is better in filling the bottoms of holes and trenches. In Table 7.1, a collimated system is compared with a conventional system, and analysed for an extensive range of film parameters. These characterization measurements relate to R&D phase, and in manufacturing sheet resistance will be used for quick monitoring.

Electrical characterization described in Chapter 2 and above has been DC, but circuits that operate at gigahertz frequencies must be measured at proper frequencies. The same applies to dielectric films too.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Table 7.1 Sputtered titanium nitride (TiN) film characterization: collimated vs. standard

Film property	Analytical technique	Collimated TiN	Standard TiN
Thickness (nm)	RBS (density = 4.94 g/cm^{-3})	81 nm	161 nm
	TEM cross section	82 nm	178 nm
Sheet resistance	Four-point probe	13.7 ohm/sq	7.4 ohm/sq
Rs uniformity	Four-point probe	3.3%	5%
Resistivity (μohm-cm)	Rs by four-point probe, Thickness by TEM	112	132
Density	Thickness by TEM & RBS, Density by RBS	4.88 g/cm^{-3} 93% of bulk	4.47 g/cm^{-3} 86% of bulk
Stoichiometry (Ti/N)	RBS	1.31	1.00
Phase	Glancing angle XRD	TiN (38–1420)	TiN (38–1420)
(JCPDS card #)	Electron diffraction	TiN (38–1420)	TiN (38–1420)
Preferred orientation	$\theta - 2\theta$ XRD Electron diffraction	(220)	(220)
Net stress Gpa	Wafer curvature	2.7 (tensile)	3.1 (tensile)
Grain structure	Cross-section TEM	Columnar	Columnar
	Plane view TEM	2D equiaxial	2D equiaxial
Average grain size	TEM	19.2 nm	18.3 nm
Average roughness	AFM	0.43 nm	1.23 nm
Min/max roughness		8 nm	18.7 nm
Specular reflection	Scanning UV	248 nm: 142%	145%
(% of Si reference)		365 nm: 55%	95%
		440 nm: 57%	123%
Impurities	Auger	O < 1%	O < 1%
(atom %)		C < 0.5%	C < 0.5%

Source: Wang, S.-Q. & J. Schlueter: Film property comparison of Ti/TiN deposited by collimated and uncollimated physical vapor deposition techniques, *J. Vac. Sci. Technol.*, **B14**(3) (1996), 1837.

7.3 CVD-FILM GROWTH AND STRUCTURE

CVD reactions have much lower sticking coefficients than PVD reactions. CVD processes are diffusive processes, whereas PVD processes are line-of-sight processes (in the first approximation). This means that deposition around corners, and even under overhang structures, is possible in CVD but impossible in PVD. CVD temperatures are high compared to PVD processes, which means that the adatoms have high surface mobilities, which also enhances step coverage.

The main parameters in CVD processes are flow rates, flow-rate ratio of reactants, temperature and pressure. In PECVD, RF power plays an important role. In Figure 7.4, PECVD silicon grain sizes are recorded as a function of $SiH_4/(SiH_4 + H_2)$ flow ratio. High-frequency (70 MHz) PECVD was employed, and glass wafers were used as substrates at 225 °C. Keeping all other deposition parameters constant, a change in the gas ratio has resulted in enormous grain-size and surface-roughness variation. In LPCVD, polysilicon deposition using SiH_4 as a source gas, a similar grain-size variation can be seen as a function of temperature: at 630 °C large grains (of the order of 100 nm) are formed, below 600 °C the grain size is reduced and at 570 °C the film is amorphous.

CVD films can be either amorphous, polycrystalline or single crystalline (epitaxial) as deposited. Epitaxial films remain single crystalline during annealing; polycrystalline films experience grain growth and even phase transitions. Amorphous films either stay amorphous or crystallize. Silicon dioxide and aluminium oxide are exceptional amorphous films because they remain amorphous throughout typical microfabrication temperatures. Pictured below are Al_2O_3 and $SrTiO_3$ films: aluminium oxide is amorphous and strontium titanate is polycrystalline (Figure 7.5).

Dielectric films have a number of measurements different from metallic films. One special feature is the use of etch rate as a quality criterion. With dielectrics, thermal SiO_2 acts as a reference film that can always be used to eliminate etchant concentration or temperature effects. Boron nitride is a new material that has been

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

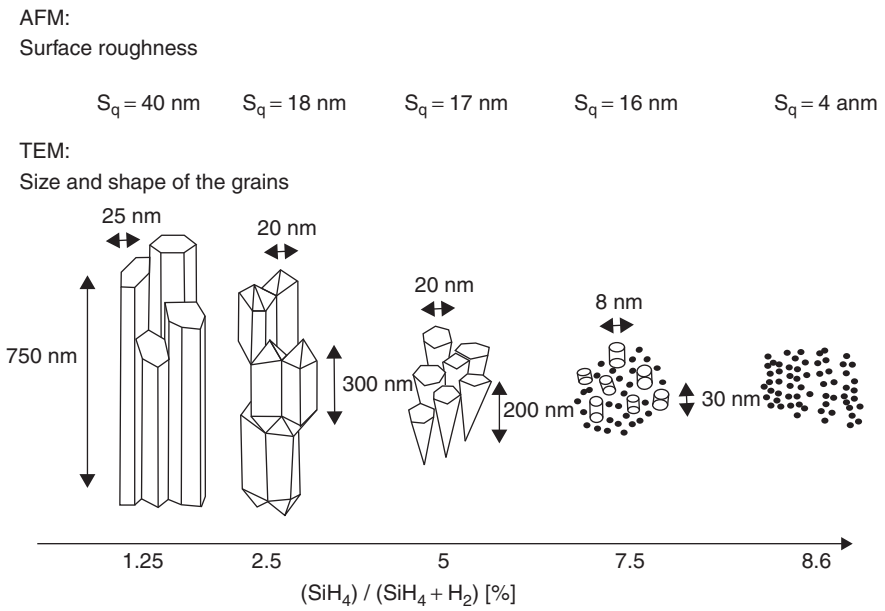
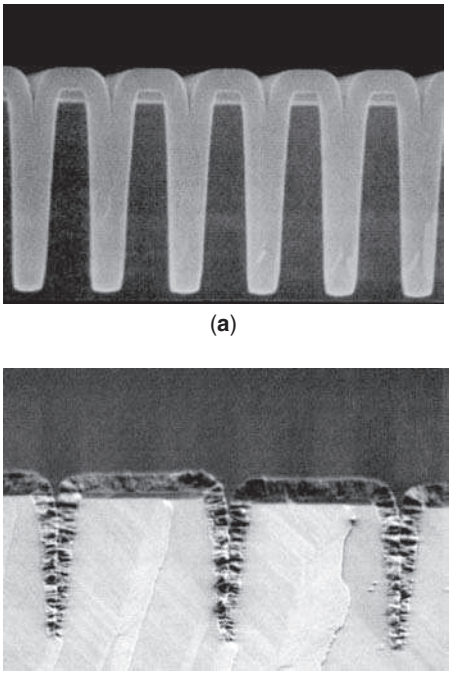


Figure 7.4 Microstructure evolutions of silicon films deposited by PECVD. Grain-size measurement by transmission electron microscope (TEM); surface roughness by atomic force microscope (AFM). Reproduced from Vallat–Sauvain, E. *et al*. (2000), by permission of AIP



(**a**)

(**b**)

Figure 7.5 SEM micrographs of thin-film structure: (**a**) amorphous aluminium oxide. From Ritala, M. *et al*. (1999), by permission of Wiley-VCH and (**b**) polycrystalline strontium titanate. Reproduced from Vehkamäki, M. *et al*. (2001), by permission of Wiley-VCH

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Table 7.2a PECVD conditions

Gases	B_2H_6 (1%)/NH_3	$B_3N_3H_6/N_2$
Flow rates	1800 sccm/120 sccm	100 sccm/200 sccm
RF power	500 W	200 W
Pressure	660 Pa (=5 Torr)	400 Pa (=3 Torr)
Temperature	400 °C susceptor	300 °C susceptor
Deposition rate	300 nm/min	370 nm/min

Table 7.2b Film properties

Uniformity	<5% (3σ)	3% (3σ)
Refractive index	1.746	1.732
Stress	−400 MPa	−150 Mpa
Etch rate in RIE	62 nm/min	28 nm/min
Etch rate H_3PO_4 167 °C	1–11 nm/min	–
Etch rate BHF	0.5 nm/min	<1 nm/min
B/N ratio	1.02	1.02
Hydrogen content	<8 at%	<8 at%
Density	1.89 g/cm^3	1.904 g/cm^3
Structure	Amorphous	Amorphous
Step coverage	60% (1 × 1 μm)	80% (0.5 × 0.5 μm)
Optical bandgap	4.7 eV	4.9 eV
Dielectric constant	3.8–5.7	3.8–5.7
Breakdown potential	6–7 MV/cm	6–8 MV/cm

Source: Cote, D.R. *et al*: Low-temperature CVD processes and dielectrics, *IBM J. Res. Dev.*, **39** (1995), 437

studied because of its potential as an insulator in multilevel metallization: it has lower dielectric constant than nitride (3.8–6 vs. 6–7) and low etch and polish rates (Table 7.2). It is not used in volume manufacturing.

Many of the measurements listed above are often laborious, and in production control, ellipsometric or reflectometric thickness and refractive index measurements would probably be used.

7.4 SURFACES AND INTERFACES

Surface roughness of thin films varies considerably. In general, high-temperature deposition results in smoother films. Epitaxial films are of course very smooth, but many amorphous films can also be extremely smooth. There is a strong correlation between surface smoothness and volume homogeneity: thermal oxide, amorphous silicon (recall Figure 7.4) and TEOS oxide are both smooth and homogeneous, whereas doped polysilicon and silicides are rough and inhomogeneous. Volume inhomogeneity makes the measurement of thin-film properties difficult. It is usual then to treat the film as if it was a stack of many layers, each with slightly different properties, for example, interfacial mixed layer, bulk of film and surface layers modelled as three materials each with materials constants of their own.

Thermodynamics gives hints for interface stability. The change in Gibbs free energy $\Delta G = G_{\text{products}} - G_{\text{reactants}}$ is positive for a stable pair of materials. For the reaction

$$\text{Ti} + \text{SiO}_2 \longrightarrow \text{TiO}_2 + \text{Si} \qquad (7.2)$$

the change in Gibbs free energy is $\Delta G = G_{\text{TiO}_2} - G_{\text{SiO}_2} = (160 - 165)\,\text{kcal} = -5\,\text{kcal}$, indicative that the reaction can proceed as written. Thermodynamics, however, is about initial and final states, and not about rates: some thermodynamically favourable processes are so slow that no effects are seen during device lifetime. But if thermodynamics forbids a reaction, it cannot proceed: the change in Gibbs free energy for

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

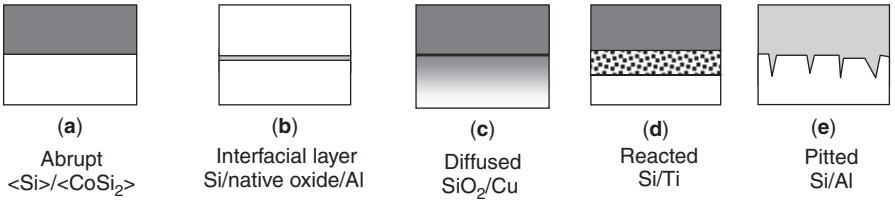


Figure 7.6 Possible interface structures: (**a**) abrupt; (**b**) interfacial layer; (**c**) diffused; (**d**) reacted and (**e**) pitted

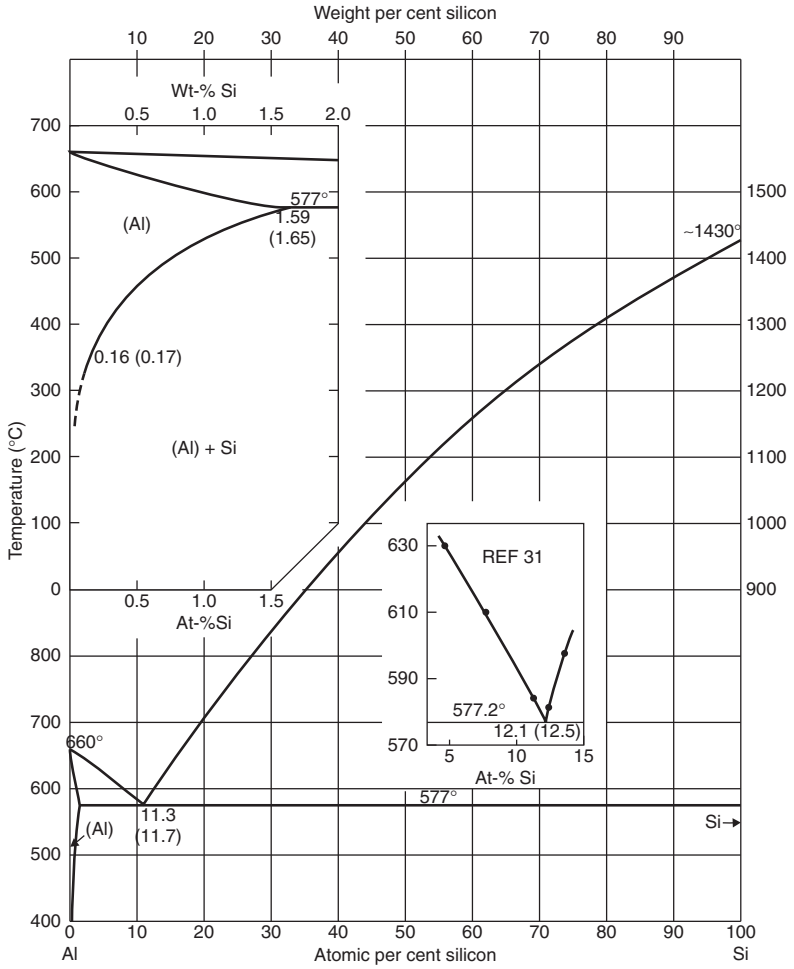


Figure 7.7 Aluminium/silicon phase diagram. Reproduced from Hansen, M. & K. Anderko (1958), by permission of McGraw Hill

cobalt/silicon dioxide reaction is positive, and cobalt does not reduce the oxide. This means that titanium silicide and cobalt silicide formation reactions are very different from interfacial oxide point of view.

Interface types also vary significantly. Abrupt interfaces (Figure 7.6(a)) are not the only idealizations: they are encountered in epitaxy; but other methods, CVD, PVD and electrochemical deposition, also produce almost ideally sharp interfaces. Native oxides are almost universally encountered on interfaces (Figure 7.6(b)); however, in many cases, those ca. 1 nm films do not destroy the device functionality.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

The case of silicon dioxide/copper (Figure 7.6(c)) shows copper diffusion into the oxide. The silicon/titanium pair will react and form silicide (Figure 7.6(d)). Many metals do form silicides, copper silicides form at very low temperatures, 200 to 300 °C, nickel, cobalt and titanium at successively higher temperatures, and W, Mo and Ta will also form silicides; not all of them, simple $MeSi_x$ compounds but complex mixtures of various silicides, for example, Me_2Si_5, Me_2Si_3, $MeSi_2$, MeSi. Aluminium reacts with tungsten and titanium to form $Al_{12}W$ and Al_3Ti, respectively.

Aluminium does not form a silicide. Annealing at 425 °C will dissolve native oxide, ensuring good electrical contact. However, too much annealing will lead to pitting: silicon is soluble in aluminium (as shown in Al-Si phase diagram, Figure 7.7), and open volume is left behind as the silicon atoms migrate into aluminium. Aluminium, on the other hand, will diffuse to fill in the space left by silicon dissolution. This leads to the case depicted in Figure 7.6(e). These aluminium spikes can be micrometres deep, and extend beyond the pn-junction. To prevent junction spiking, aluminium can be alloyed with silicon: a silicon concentration of 0.5% (wt%) will saturate aluminium at 425 °C, and 1% Si will prevent silicon dissolution at 500 °C. The other, more general solution is to implement a diffusion barrier.

7.5 ADHESION LAYERS AND BARRIERS

Adhesion is a major issue in thin-film technology. As a rule of thumb, poor adhesion is the norm, and only special attention will lead to good adhesion. Some materials have poor adhesion due to their chemical nature: noble metals are noble because they do not react, and therefore they do not form bonds across the substrate interface. Adhesion is also related to surface cleanliness: residues or dirt from the previous step will almost inevitably lead to poor adhesion. Deposition process variables do play a role: in sputtering, energetic ions and atoms will kick off loosely bound atoms, but in evaporation, there is no inherent removal of weakly bonded atoms.

Adhesion layers are additional films with the role of adhesion improvement, and, in the first approximation, have no effect on the device structure or operation. The thickness of the adhesion layer is in the range of 10 nm because volume properties are of no interest, but only its surface properties. The adhesion layer and the structural film are deposited immediately after each other in the same vacuum chamber: freshly formed adhesion-layer surface ensures cleanliness and thus eliminates one main factor of poor adhesion. Adhesion-layer films are selected on the basis of their bond-forming abilities: titanium and chromium are the two most widely used materials. Typical pairs of adhesion layer/noble metal include Ti/Pt, Ti/Au and Cr/Au. Adhesion layers are also useful for near-noble refractory metals like tungsten.

Barriers are additional layers between two materials. Their role is to prevent reactions between adjacent layers, be it diffusion, chemical reaction or any other type of unwanted interaction. Many aspects of barriers are similar to adhesion layers: barriers are not needed for device operation as such, but their presence either makes the fabrication process more robust, or the resulting device more stable. Barriers are thin, like adhesion layers, with 10 to 100 nm as typical barrier thickness.

Total barriers must prevent all fluxes through them: atom diffusion and charge carrier transport. In the case of metallization, the current has to flow through the barrier, but atom movements must be prevented. Metallic barriers have relatively loose requirements for resistivity (the distance is <100 nm only). Most barrier materials have resistivities around 100 to 500 μohm-cm, one-to-two orders of magnitude higher than the conductors. While resistivity is not a problem, contact resistivity must be low, and barrier height considerations may exclude some materials.

The first barriers to be implemented were 100 nm thick TiW films between aluminium and silicon to prevent Al-Si junction spiking. TiW grain size is ca. 100 nm: if sputtered in argon, grain boundaries offer fast diffusion paths, and pure TiW is not a very effective barrier. But deposition in poor vacuum led to the incorporation of oxygen and nitrogen, which passivated grain boundaries. When the mechanism was elucidated, reactive sputtering of TiW in $Ar + N_2$ atmosphere was adopted. Reactive sputtering leads to 10 nm grain size and nitrogen at grain boundaries, both of which lead to improved barrier performance. Amorphous films would be preferable as barriers, and a-WN has been one candidate. Copper metallization needs barriers not only between copper and silicon, but also between copper and silicon dioxide because copper diffuses into oxide. Tantalum and tantalum compounds such as TaN are used. Silicon nitride can be used as a dielectric barrier between copper and oxide because it is stable in contact with both silicon and copper.

When active devices are made on glass (or on steel), such as thin-film transistors, the substrate has to be isolated from the silicon devices. Barriers like silicon dioxide (both CVD oxide and spin-on-glass (SOG)) as well as Al_2O_3 have been used.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

7.5.1 Measurement of adhesion layers and barriers

The first adhesion test is tape-pull test: adhesive tape (standard office tape is commonly used) is attached to the thin film and pulled off. If the film peels off with the tape, it has failed the adhesion test. More advanced tests use a quantifiable pull force.

Adhesion layer and diffusion-barrier stability can be checked by electrical and physical measurements. Sheet-resistance increase is a quick and simple measurement. Copper resistivity is very low, 1.7 μohm-cm, and when the barrier fails, the copper can react with the silicon underneath, bringing about a resistance increase because copper silicides CuSi and Cu_3Si are high-resistivity materials. They can be identified by X-ray diffraction, but the resistance increase is indicative of silicide formation. Pn-junction diode leakage is another quick electrical measurement.

Auger-depth profiling is the standard physical measurement. Auger measurement is slow and sample destroying, but it can be done on a blanket wafer without any sample preparation. Usually the as-deposited sample is compared with the annealed sample(s), and barrier failure is evidenced by intermixing of metal and silicon across the barrier. Accumulation of material at the interfaces, and atom distributions across the film are helpful in understanding the reactions behind the barrier failure.

Note that the Auger analysis shown in Figure 7.8 does not indicate TiO_2 formation even though the coexistence of titanium and oxygen might suggest it: Auger is about atoms and not about compounds. XRD could show TiO_2 formation by the appearance of diffraction peaks identified as arising from TiO_2.

7.6 MULTILAYER FILMS

Performance of simple elemental or compound films, with or without barrier or adhesion layers, is often not enough, and multilayer films are introduced to offer improvement. Early integrated circuits used aluminium for metallization. In order to improve interface stability, Al-Si (1%) was adopted, and later TiW diffusion barrier was added and Al-Si was replaced by Al-Si-Cu for improved electromigration resistance. For many generations, (0.8 − 0.5 − 0.35 − 0.25 μm) IC metallization was done with a Ti/TiN/Al/TiN film stack. Titanium acts as an adhesion promoter, TiN as a diffusion barrier, Al as a current-carrying film and the top TiN has the dual role of mechanical stiffening of the structure and reflectivity reduction. Metallization reliability has been greatly improved by the adoption of such multilayer metallization schemes, but a price has been paid elsewhere: the etching of such multilayer structures is difficult.

Periodic multilayers have been fabricated for various purposes: Si/Mo and W/C and similar light element/heavy element structures are designed for X-ray optics. Periodicities are of the order of nanometres (≈ X-ray wavelength). Multilayer structure of AlN/TiN with ca. 10 nm periodicity has been found to have excellent tribological properties, for instance, hardness in excess of its constituent materials. ZrO_2/HfO_2 multilayers have been used in order to improve leakage currents in the deposited capacitor dielectrics. These polycrystalline multilayers have been termed *nanolaminates*.

Minimum thickness/minimum period of the multilayer structures depends on the growth process

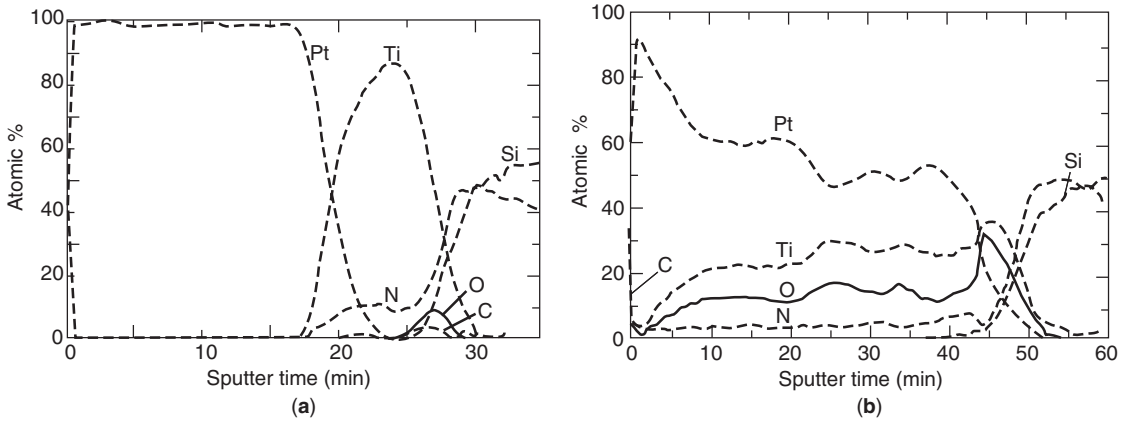


Figure 7.8 Auger depth profile of Pt/Ti/SiN_x/Si structure: (**a**) as deposited and (**b**) oxygen annealed at 600 °C: the interdiffusion of films is almost complete. Oxygen and carbon accumulation on the surface in the as-deposited sample indicate cleaning problems. Reproduced from Kang, U. *et al.* (1999), by permission of Institute of Pure and Applied Physics

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

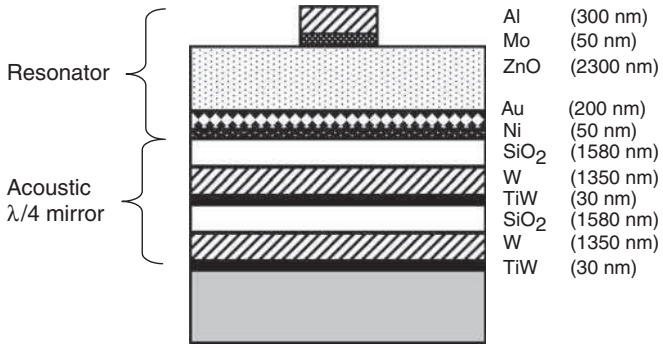


Figure 7.9 Bulk acoustic resonator structure on a glass wafer: a piezoelectric ZnO resonator is sandwiched between gold and aluminium electrodes. TiW, Ni and Mo are thin adhesion promotion layers. W and SiO_2 form λ/4 acoustic wavelength filters. Adapted from VTT Microelectronics annual research review 2001

characteristics and also on the sharpness of interfaces. For epitaxial growth, atomic layer structures are possible; for example, delta-doping layer is a single atomic layer of dopant between two semiconductors. Interface abruptness depends on the reactor-operating principle: if growth is dependent on the gas flow in the reactor, minimum thickness is determined by the gas residence time in the reactor (discussed in Chapter 32), which can be fractions of seconds or tens of seconds. Flow systems, such as CVD, are thus not suitable for very thin layers. Beam systems, evaporation, sputtering and molecular beam epitaxy MBE with shutters enable subsecond turn-off and turn-on of the deposition. When multilayer structures are so thin that quantum effects arise, they are termed *superlattices*.

Dielectric mirrors with λ/4 layer thicknesses for high reflectance surfaces involve multiple dielectric layers. Undoped polysilicon, oxide and nitride are the usual films. For visible wavelengths, layer thicknesses around 100 nm are typical. Similar λ/4 structures are used in thin-film bulk acoustic resonators (TFBAR): multilayers of W:SiO_2, with thicknesses ca. 1.5 μm, act as acoustic mirrors (Figure 7.9).

In PECVD deposition, oxynitride films of composition SiO_xN_y can be easily made. By tailoring the composition, the refractive index can be tailored from 1.46 to 2, full range between oxide and nitride indices (Figure 7.10). By sandwiching the SiON film between two lower refractive index films, it acts as a waveguide. Doping of oxide by phosphorus (PSG) or germanium can also be used to tailor the refractive index, but only over a limited range before the other film properties change too much.

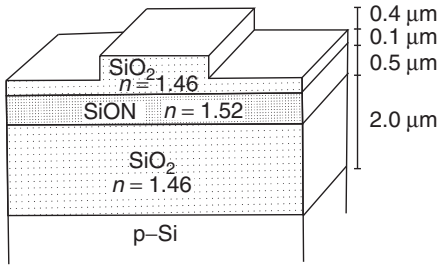


Figure 7.10 Refractive index $SiO_2/SiO_xN_y/SiO_2$ waveguide: n_f 1.46/1.52/1.46. Reproduced from Hilleringmann, U. & K. Goser (1995), by permission of IEEE

7.7 STRESSES

Thin films are under either compressive or tensile stresses when deposited on the wafers. Stresses consist of extrinsic stresses, caused by thermal expansion mismatch between the film and the substrate, and of intrinsic stresses that depend on the film microstructure and the deposition process.

Extrinsic stresses can be estimated from thermal expansion coefficient differences:

$$\sigma = E_f(\alpha_f - \alpha_s) \times \Delta T/(1 - \nu) \qquad (7.3)$$

(by convention, negative stresses are compressive)

where E_f = Young's modulus of the film
ν = Poisson ratio of the film
α = coefficient of thermal expansion
ΔT = temperature difference.

In the first approximation, the temperature difference is the difference between the deposition and measurement

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

temperatures, but the situation is really much more complex because stress relaxation can occur during high-temperature deposition.

The coefficient of thermal expansion (CTE) of silicon is $2.6 \times 10^{-6}/^{\circ}\mathrm{C}$ (around room temperature). The only other materials used in microfabrication that have smaller coefficients are silicon dioxide, silicon nitride and diamond which have CTEs $0.5 \times 10^{-6}/^{\circ}\mathrm{C}$, $2.4 \times 10^{-6}/^{\circ}\mathrm{C}$ and $1.1 \times 10^{-6}/^{\circ}\mathrm{C}$, respectively. Oxide, nitride and diamond, are therefore the only materials that can develop compressive extrinsic stresses over silicon substrates. Aluminium CTE is 23 ppm, which is fairly high, tungsten CTE is 4 ppm and polymers have CTE values in the range of 30 to 100 ppm.

Intrinsic stresses are caused by many mechanisms that are not fully understood. Deposited polycrystalline films are not at their energy minimum. An exceptionally low deposition temperature means that the arriving atoms do not have enough energy to find energetically favourable positions, and the film builds up without relaxation. Voids and incorporated foreign atoms contribute to intrinsic stresses. Bombardment during deposition has a pronounced effect on many film properties, including stresses, because the bombardment pinches off loosely bound atoms, resulting in a more uniform, less stressed film. Too high bombardment, on the other hand, implants atoms into the film in a non-equilibrium way, and compressive stresses build up. Crystallization and phase transitions, and other processes that lead to volume changes, such as outgassing, lead to stress changes.

Evaporated metal films are usually under tensile stresses. Sputtered films can be under tensile or compressive stresses. Sputtering, with ion bombardment during deposition, is a much more complex process than evaporation, and stress tailoring can be achieved by:

- bias power
- argon pressure
- sputtering gas mass
- temperature
- deposition rate.

Sputtered film stress can be tailored by the deposition pressure: films are usually under compressive stress if deposited at low pressure (ca. 0.1 Pa in a magnetron sputtering system) but turn to tensile stress as the deposition pressure is raised (to ca. 1 Pa) (Figure 7.11). This crossover pressure increases with the atomic mass. However, this is not a universal solution, because pressure affects not only the film stress but also many other properties such as deposition rate and film density.

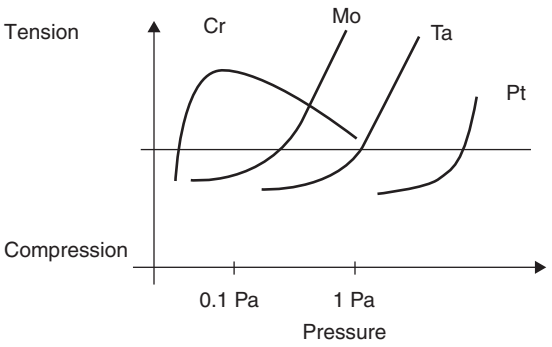


Figure 7.11 Sputtering pressure and film stress. Atomic masses: Cr 52, Mo 96, Ta 181, Pt 195. Redrawn after Ohring, M. (1992), by permission of Academic Press

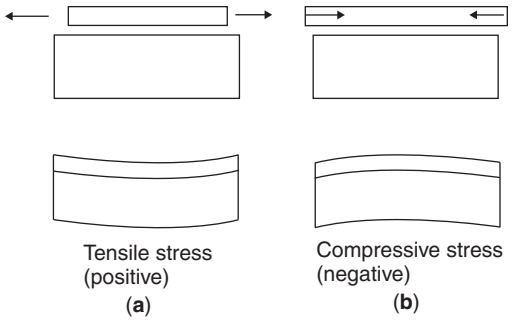


Figure 7.12 Thin-film stresses: a film that must be elongated to fit a wafer is under tensile stress (positive) and a film that is compressed to fit a wafer, is under compressive (negative) stress

Stresses in thin films cause wafer curvature, as shown in Figure 7.12. Imagine a free film attached to a massive wafer and forcefit to the wafer size. Next, imagine, stress relaxation through the wafer curvature. A film under tensile stress will result in a concave shape, while a compressively stressed film will end up with a convex profile.

Figure 7.12 gives a macroscopic depiction of stresses, but the same reasoning works on the atomic level as well: germanium lattice constant is 4.2% larger than that of silicon, therefore germanium and silicon–germanium films on silicon are compressively stressed, and silicon films on SiGe are under tensile stress.

Stress at room temperature is a sum of intrinsic and extrinsic stresses. Since extrinsic stresses are usually tensile (with the exception of oxide, nitride and diamond), and total stresses can be close to zero, this means that intrinsic stresses from the deposition process are compressive. This is often the case.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Wafers are ca. 1000 times thicker than films, and because all solids have similar elastic constants, wafer stresses and strains are ca. 1000 times less than those of thin films. Thin-film stresses are of the order of 10 to 1000 MPa (1000 MPa = 10^{10} dyn/cm^2).

Annealing temperature can be used to tailor stresses: a long-time, low-temperature anneal of fine-grained LPCVD silicon (deposited at 580 °C) will result in a slightly compressively stressed film, while high-temperature anneal will result in tensile stress (Figure 7.13).

Bimetal thermometer is a classic example of a thermal expansion coefficient mismatch. Bimorph structures can be used as sensors and actuators in microsystems, but the initial shape has to be known. Shown in Figure 7.14 are SiO_2/Al and SiO_2/Ti cantilevers, which are bent because of stresses in the structures, without external sensing or actuation force. In a single material cantilever (e.g.,

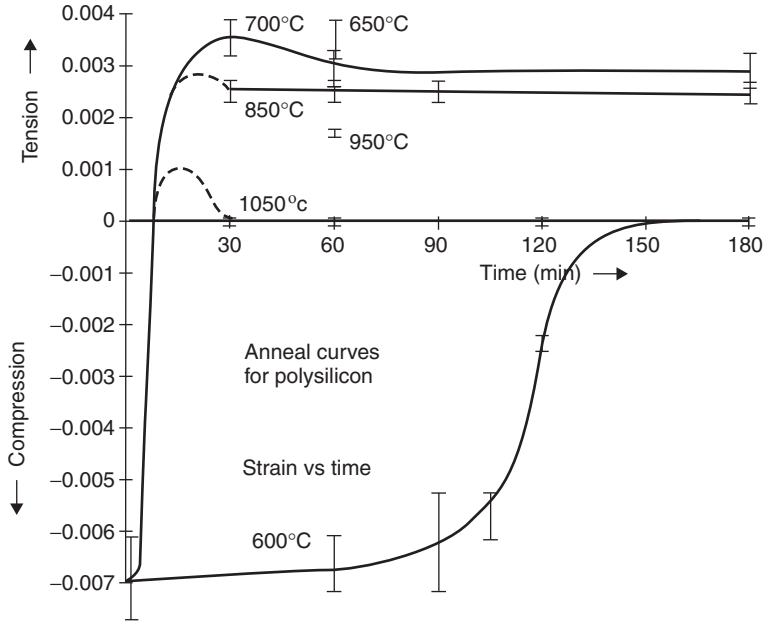
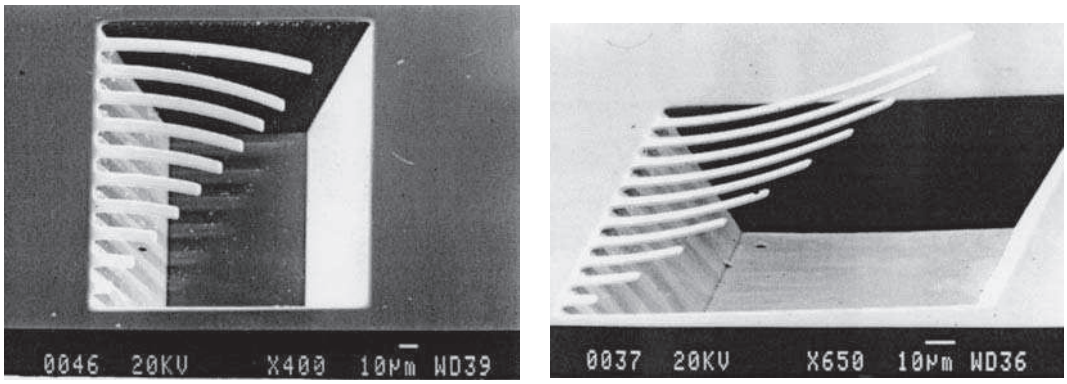


Figure 7.13 Different anneal processes for 580 °C deposited polysilicon. Reproduced from Guckel, H. (1988), by permission of IEEE



(**a**) (**b**)

Figure 7.14 (**a**) Compressive stress in SiO_2/Al cantilevers causes downward bending and (**b**) tensile stress in SiO_2/Ti cantilevers leads to upward bending. Reproduced from Fang, W. & C.-Y. Lo (2000), by permission of Elsevier

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

LPCVD polysilicon), the stress gradients can lead to similar bending.

7.7.1 Stress measurement

Thin-film stresses are usually measured by wafer-curvature measurements: the curvature needs to be measured both with the film and without the film (either before the deposition; or after etching away the film) because wafer bows of 30 μm are typical, and they would lead to 100% errors in stress values easily. Optical techniques or scanning probes can be used for curvature measurement.

Film stress is given by the Stoney formula:

$$\sigma = (E_s t_s^2 / 6 t_f (1 - \nu)) \times ((1/R) - (1/R_0)) \qquad (7.4)$$

t_s = substrate thickness
ν = Poisson ratio of the substrate (0.27 for silicon)
t_f = film thickness
R = radius of curvature for the substrate + film system (negative for convex)
R_0 = radius of curvature for substrate without film.

Stresses can also be measured by Bragg–Brentano X-ray diffraction. Lattice spacing d_f in the direction normal to the surface is measured and compared to a relaxed film lattice spacing d_r. Strain is calculated as $\varepsilon_{33} = (d_f - d_r)/d_r$ and stress as $\sigma_{11} = -(E_f \times \varepsilon_{33})/2\nu_f$. Note that there is a fundamental and practical difference compared with the Stoney formula: in Bragg–Brentano we need to know the thin-film elastic constants E_f, ν_f, whereas in the Stoney formula, only the film thickness needs to be known, but elastic constants of the substrate are needed, and these are generally well known. Bragg–Brentano is used for epitaxial films, in which film elastic constants are well understood and known.

7.8 THIN FILMS OVER TOPOGRAPHY: STEP COVERAGE

Deposition on a patterned substrate introduces new considerations as the film must go over steps. Both film thickness and structure will be different on horizontal and vertical surfaces, especially in sputtering and PECVD, where particle bombardment during the deposition is present. A basic explanation for different step coverage is the angle for the arriving atoms. On horizontal free surfaces, it is 180°, in convex corners it is 270° and in the bottom concave corners it is only 90°, as depicted in Figure 7.15. This leads to cusping, or the most pronounced deposition at the step corners.

High-temperature CVD processes like TEOS and HTO, and LPCVD processes of nitride and polysilicon

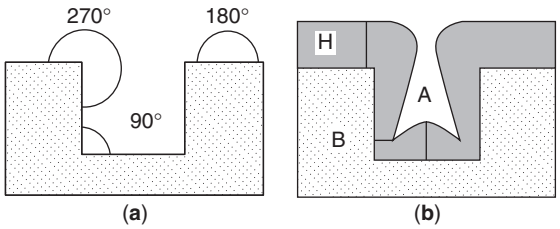


Figure 7.15 (**a**) Arrival angles of depositing specie at different positions and (**b**) step coverage: B/H; bottom coverage: A/H

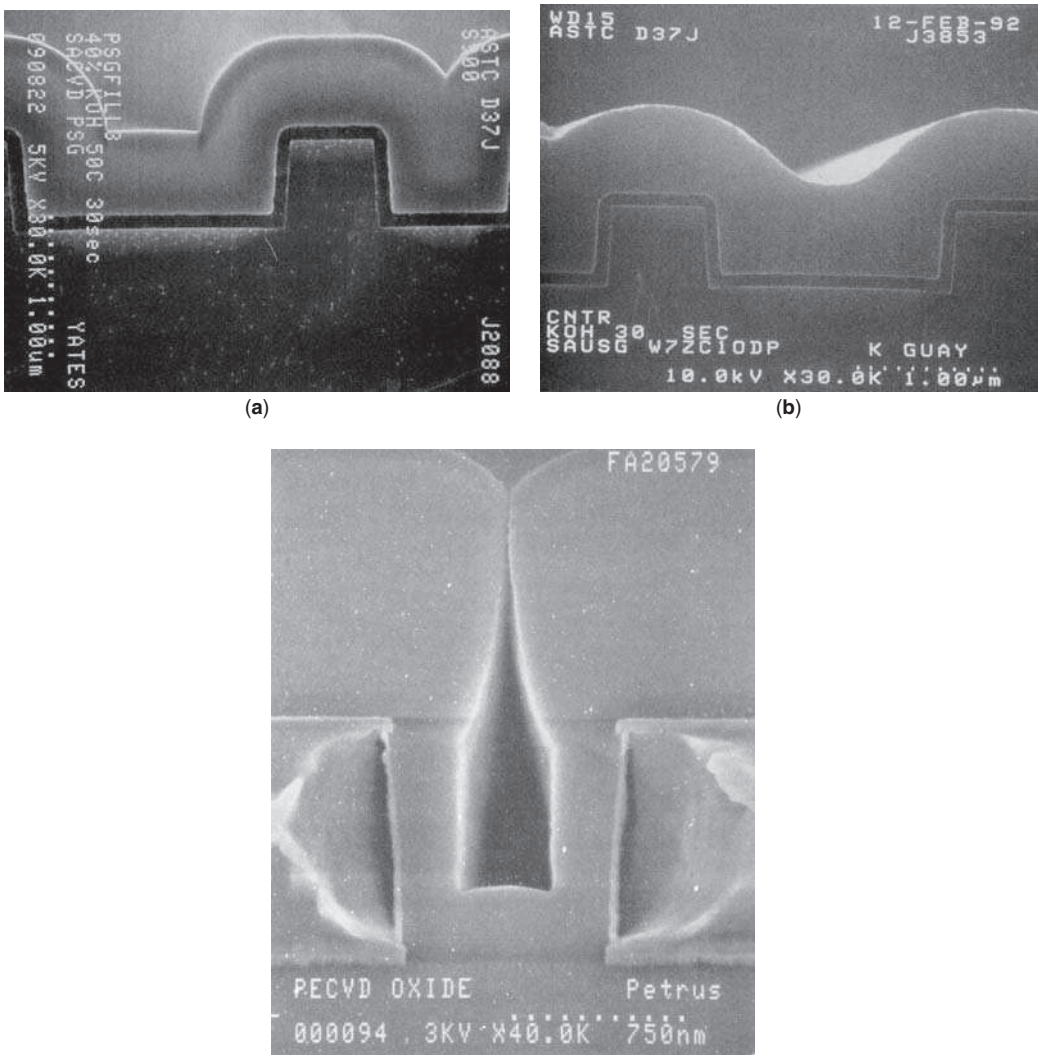
and CVD-tungsten have a nearly perfect conformal deposition, that is, both step coverage and bottom coverage are 100%. This comes from fast surface diffusion at relatively high deposition temperatures, and from low-sticking coefficient, which means that weakly bound specie do not contribute to film growth. Spin films have a flow-like profile, which means that they cover small gaps and spaces well, but on large areas (both recesses and mesas) the film thickness saturates to a constant value.

Step coverage in evaporation is very poor. Sputtering and PECVD form the middle ground: the step coverage is strongly deposition-condition dependent (see Figure 3.6 for simulated sputter-deposited profiles). In PECVD, source gases, flow ratios, RF power, temperature, pressure and phosphorus doping can affect the step coverage (Figure 7.16). Conformal deposition is no guarantee that film quality on the sidewalls is equal to that of planar areas: etch rates of sidewall oxide films can be significantly faster compared to planar reference areas. Measurement of sidewall film etch rate requires destructive cross-sectional imaging, but planar area measurements cannot be trusted.

Gap filling is important for both yield (in fabrication) and reliability (in the field): if voids are left between the structures, these can act as traps for residues and sites for absorption of moisture (Figure 7.17). Voids can remain closed during some process steps without any adverse effects, but the following etch or polish steps can open them up unexpectedly, leading to problems.

Step coverage is a strong function of the aspect ratio. It has to be remembered that aspect ratio is a dynamic variable: a contact hole that is initially 1:1 turns into a 2:1 aspect ratio hole as the metal deposition proceeds, and just before closure, aspect ratio approaches infinity. Figures 7.5 (a) and (b) and 7.16 (a) and (b) show excellent gap filling. Step coverage is usually no major problem for low-aspect ratio structures, say <0.5:1, but at 1:1 and higher-aspect ratios, the step coverage rapidly deteriorates. It is important to remember that on real

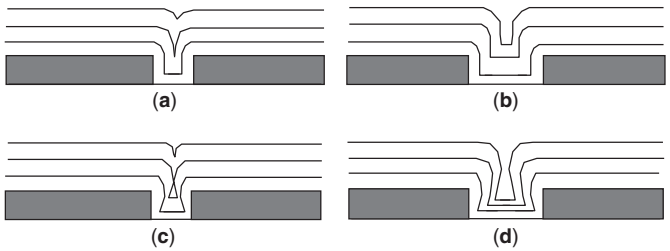
Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030



(**a**) (**b**)

(**c**)

Figure 7.16 Step coverage in different CVD processes: (**a**) phosphorus doped CVD oxide with conformal (100%) step coverage, (**b**) undoped CVD oxide with flow-like profiles and (**c**) PECVD oxide from silane/nitrous oxide reaction leads to a void formation. Reproduced from Cote, D.R. *et al*. (1995), by permission of IBM



(**a**) (**b**)

(**c**) (**d**)

Figure 7.17 (**a**) Gap filling with conformal step coverage. (**b**) Conformal deposition of a larger gap with the same process does not lead to gap filling but the original step height remains. (**c**) Void and (**d**) cusp are formed when step coverage is maximum at the step corner

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

microdevices, there are always structures of various shapes and variable spacings, and the film deposition over all these spaces needs to be considered. It is far too simple to consider one size only.

Good step coverage in metallization is essential for reliability. Even though the metal film will be continuous even with, say, 10% step coverage, current density will increase dramatically at the thinnest point, causing a major reliability problem.

7.9 SIMULATION OF DEPOSITION

Topography simulation (for deposition, etching and polishing) works on fluxes and surface processes: at each grid point, the incoming flux (from the fluid phase) and surface-reaction probability are evaluated (with a return flux of reaction products in the case of etching/polishing, or non-sticking specie in the case of deposition) to calculate the new surface height. In principle, the generation of the incoming specie could be simulated (for instance, ion and radical production in plasma) but this is usually not integrated into a topography simulator; rather, it is a part of a reactor simulator. New surface points are calculated and those points are connected to represent the surface. Accuracy is increased by calculating new points between existing points when they are far apart; and similarly, by eliminating points that become close to each other.

Deposition models define atom arrival angles, and various models are available in most simulators: fully directional, hemispherical, conical, etc. Etch models include isotropic and anisotropic models, and user definable mixtures of the two. Model selection is very much an empirical question, and the predictive power of topography simulation is diminished by this semiempirical tailoring of model parameters.

Input for a typical topography simulation includes

- the surface topography already made
- the material to be deposited
- the deposition model (angular distribution of depositing specie)
- thickness/rate and time.

Adjustable parameters include surface diffusivity, which determines how much lateral movement the impinging specie is allowed before it is 'frozen' in the growing film.

Topography simulator SAMPLE 2D, developed at University of California, Berkeley, has been used to obtain the profiles shown in Figure 7.18. Hemispherical deposition model is an approximization of sputter deposition. Trench dimensions have been varied to see the effect of the aspect ratio on step coverage. In the 1:1 aspect-ratio trench step, the coverage is ca. 15%, but in the 2:1 aspect-ratio trench, the coverage is only a meagre 5%. Slightly sloped profile in the 2:1 trench leads to ca. 10% step coverage.

Note that step coverage over isolated lines is always the same irrespective of the line aspect ratio: step coverage depends on the atom arrival angles and, by definition, the isolated lines have a large unobstructed space next to them, and, therefore, will result in identical step coverage.

Monte Carlo (MC) and molecular dynamics (MD) simulations offer more realism, for example, the prediction of step coverage based on relaxation (Figure 7.19). Calculations can be speeded up by treating matter as 100 Å cluster spheres instead of individual atoms. Clusters, and thus the atoms, come to rest at stable positions, for example when touching three other spheres. The arrival of new material and the rearrangement of already deposited films can be simulated simultaneously. Temperature and sticking coefficient are used as parameters for surface mobility.

2D simulation can overestimate the bottom coverage by 40%, compared to 3D. This is intuitively easy to understand because 2D simulation treats the recesses as infinitely long trenches, with very large acceptance angles along the trenches, whereas 3D simulation takes into account the real acceptance angle.

7.9.1 Scales in simulation

The fundamental simplification of many topography/thin-film simulators is the fact that surface-controlled reactions are assumed. On a microscopic scale this is true: material is being added to or removed from a surface, but on a macroscale this is a gross simplification. Etching and deposition processes can be either surface-reaction limited or transport-process limited. The transport of reactants from gas flow to surface (as in a CVD reactor) or the removal of reaction products by convection (like removal of hydrogen bubbles that result from silicon etching) can be more critical to etching or deposition than the surface processes. Whether it is the surface reaction or the transport mechanism that determines the reaction rate has to be studied for each process. If the reaction is transport limited, then the simulation should be able to model fluid dynamics at the reactor scale, in addition to the surface processes at the micrometre scale.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

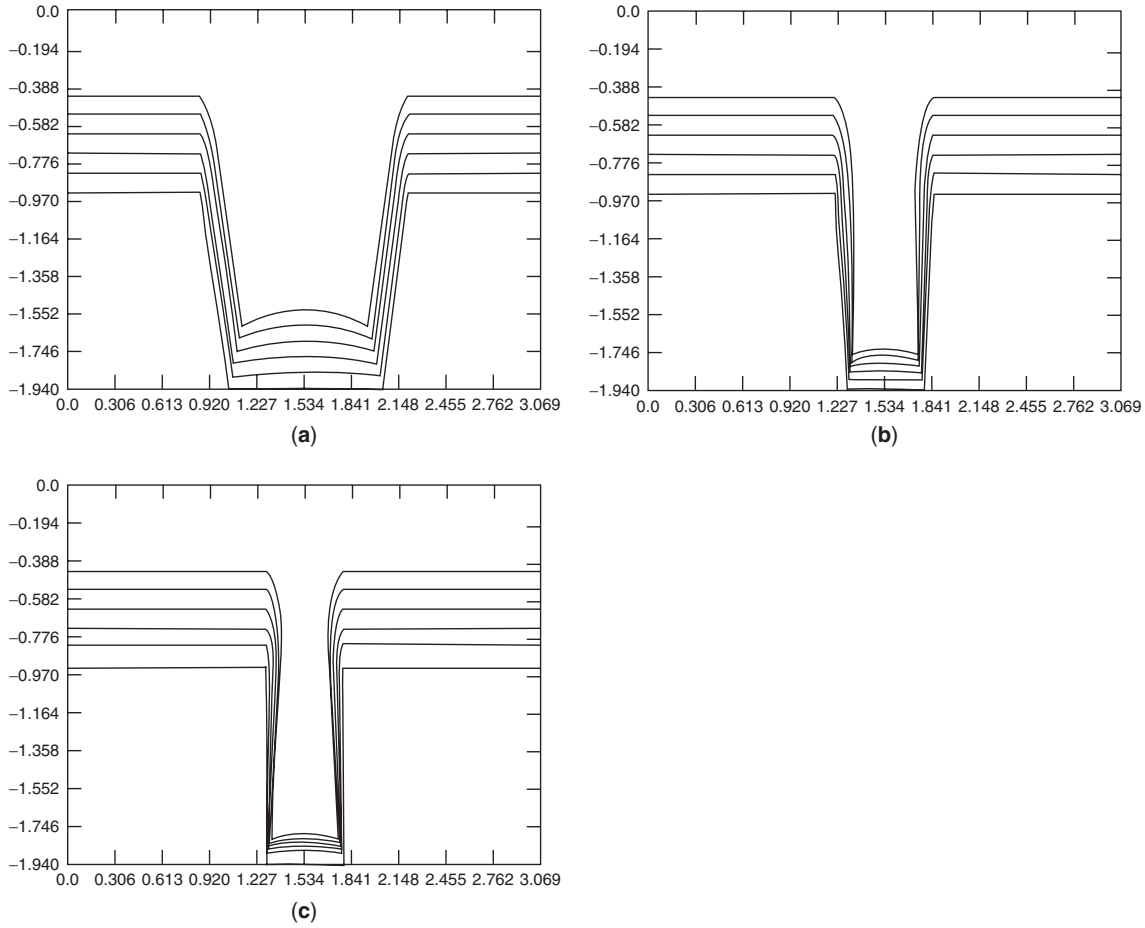


Figure 7.18 Simulation of deposition step coverage with SAMPLE 2D. Hemispherical deposition model corresponds to sputtering. Trench widths are 1 μm and 0.5 μm, depths 1 μm. Wall angle either 90° or ca. 81°. Film thickness is 0.5 μm in all cases

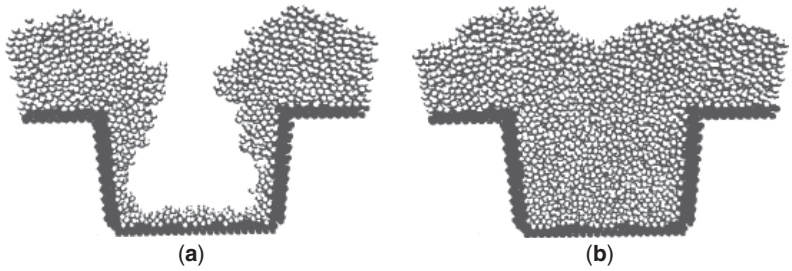
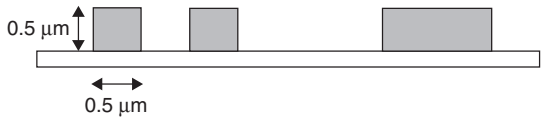


Figure 7.19 3D Monte Carlo simulation of aluminium deposition into a contact hole: (**a**) high-rate deposition and (**b**) low-rate deposition. Both depositions are at the same temperature. The simulation is 3D, but only a cut through the contact hole centreline is shown. Reproduced from Baumann, H.F. & G.H. Gilmer (1995), by permission of IEEE

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

7.10 EXERCISES

1. The speed of sound in ZnO is 5700 m/s. What is the intended operating frequency for the TFBAR shown in Figure 7.9?
2. Calculate the wafer bow that a thin film of 100 nm thickness and 100 MPa stress induces on a 675 μm-thick, 150 mm diameter silicon wafer. Also calculate the same for a 100 nm-thick film of 500 MPa stress on a 380 μm-thick, 100 mm-diameter wafer?
3. A periodic lattice of W and C is used as a λ/4 X-ray mirror. What are the layer thicknesses that should be used for 100 eV X-rays?
4. Oxygen is soluble into titanium up to 34 atomic%. What will be the thickness of a silicon dioxide film that can be dissolved by a 50 nm-thick titanium film? Titanium density is 4.5 g/cm^3, silicon dioxide density is 2.3 g/cm^3.
5. What is the step coverage in Figures 7.15(b), 7.16(c), and 7.19(a)?
6. Draw the deposited film profile over a given topography for the six different cases listed below:
   (a) Sputtered aluminium, 300 nm thick
   (b) CVD TEOS 0.3 μm thick
   (c) Electroplating 0.5 μm copper
   (d) PECVD oxide 0.2 μm thick
   (e) Evaporated aluminium, 100 nm thick
   (f) SOG application, 300 nm thick.



7. $TiAl_3$ is formed in the reaction between aluminium and titanium films. What will happen to the volume of the metal line? Al: 2.7 g/cm^3; Ti 4.5 g/cm^3; $TiAl_3$ 3.35 g/cm^3.

REFERENCES AND RELATED READINGS

Baumann, H.F. & G.H. Gilmer: 3D modelling of sputter and reflow processes for interconnect metals, *IEDM 1995*, p. 89.

Chou, B.C.S. *et al*: Fabrication of low-stress dielectric thin-film for microsensor applications, *IEEE EDL*, **18** (1997), 599.

Cote, D.R. *et al*: Low-temperature CVD processes and dielectrics, *IBM J. Res. Dev.*, **39** (1995), 437.

Fang, W. & C.-Y. Lo: On the thermal expansion coefficients of thin films, *Sensors Actuators*, **84** (2000), 310.

Guckel, H. *et al*: Fine-grained polysilicon films with build-in tensile strain, *IEEE TED*, **35** (1988), 800.

Hansen, M. & K. Anderko: *Constitution of Binary Alloys*, 2nd ed., McGraw-Hill, 1958.

Hilleringmann, U. & K. Goser: Optoelectronic system integration on silicon: waveguides, photodetectors, and VLSI CMOS circuits on one chip, *IEEE TED*, **42** (1995), 841.

Kang, U. *et al*: Pt/Ti thin film adhesion on SiN_x/Si substrates, *Jpn. J. Appl. Phys.*, **38** (1999), 4147.

Laurila, T. *et al*: Failure mechanism of Ta diffusion barrier between Cu and Si, *J. Appl. Phys.*, **88** (2000), 3377.

Murarka, S.P.: *Metallization, Theory and Practice for VLSI and ULSI*, Butterworth-Heinemann, 1993.

Ohring, M.: *The Materials Science of Thin Films*, Academic Press, 1992.

Raaijmakers, I.J. *et al*: Microstructure and barrier properties of reactively sputtered Ti-W nitride, *J. Electron. Mater.*, **19** (1990), 1221.

Ritala, M. *et al*: Perfectly conformal TiN and Al_2O_3 film deposited by atomic layer deposition, *Chem. Vapor Deposit.*, **5** (1999), 7.

Rossnagel, S.M. *et al*: Thin, high atomic weight refractory film deposition for diffusion barrier, adhesion layer and seed layer applications, *J. Vac. Sci. Technol.*, **B 14** (1996), 1819.

Smith, D.L.: *Thin-film Deposition*, McGraw-Hill, 1995.

Thornton, J.A.: The microstructure of sputter-deposited coatings, *J. Vac. Sci. Technol.*, **A4**(6) (1986), 3059.

Vallat-Sauvain, E. *et al*: Evolution of microstructure in microcrystalline silicon prepared by very high frequency glow-discharge using hydrogen dilution, *J. Appl. Phys.*, **87** (2000), 3137.

Vehkamäki, M. *et al*: Atomic layer deposition of $SrTiO_3$, *Chem. Vapor Deposit.*, **7** (2001), 75.

Wang, S.-Q. & J. Schlueter: Film property comparison of Ti/TiN deposited by collimated and uncollimated physical vapor deposition techniques, *J. Vac. Sci. Technol.*, **B14**(3) (1996), 1837.

Wang, S.-Q. *et al*: Step coverage comparison of Ti/TiN deposited by collimated and uncollimated physical vapor deposition techniques, *J. Vac. Sci. Technol.*, **B14**(3) (1996), 1846.

Wang, Y.Y. *et al*: Synthesis and characterization of highly textured polycrystalline AlN/TiN superlattice coatings, *J. Vac. Sci. Technol.*, **A16** (1998), 3341.

Xu, Y.P. *et al*: A study of sputter deposited silicon films, *J. Electron. Mater.*, **21** (1992), 373.

**Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030**

Part III

Basic Processes

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

8

Pattern Generation

A pattern generation tool transcribes the circuit design data into a physical structure. It must be able to expose single pixels and expose them fairly fast, since designs can consist of millions of pixels. The first pattern generators were optomechanical shutter systems with a flash bulb. Aperture blades were sized and positioned, followed by the exposing flash. After mechanical movement of the wafer, the aperture sizing operation and flashing was repeated, with operating frequency of ca. 1 Hz. This method was employed in the early era of microfabrication when linewidths were above 10 μm.

The most precise way of delineating structures is by drawing a single feature with a focused beam of electrons, ions or photons. This is faster than the mechanical aperture method but still very slow. It has three main applications:

1. Direct writing for ultimate resolution.
2. Direct writing in research and small series production.
3. Writing photomasks for optical lithography.

Beam writing is several orders of magnitude slower than optical lithography with photomasks but it offers ultimate resolution, down to ca. 10 nm compared with 100 nm for the best optical lithography tools. It is also flexible because designs can be changed immediately by rewriting the code. Optical lithography (recall Figure 1.3) is the mainstay of microlithography, but the photomask cost increases rapidly as linewidths are scaled down, and photomask writing and inspection time can be considerable. Electron beam writing is an option for R&D or pilot production, but equipment for electron beam lithography is complex and sensitive and it requires a lot of servicing and maintenance for an ultimate resolution and reasonable uptime.

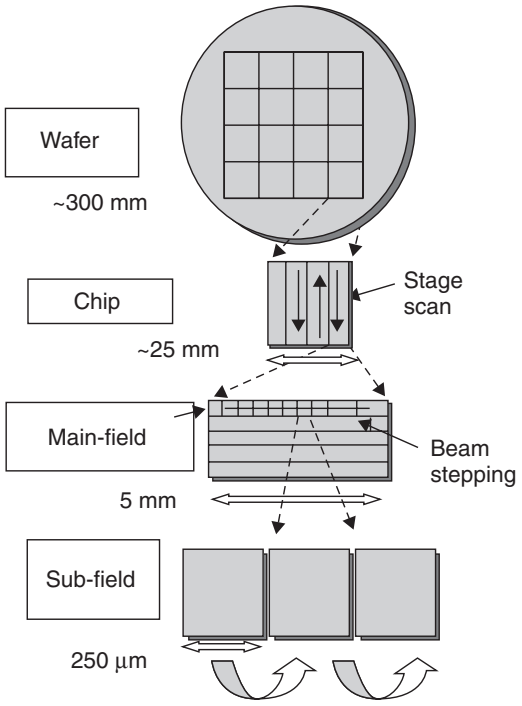


Figure 8.1 Electron beam lithography system: subfield is electrically scanned, and other movements are introduced to write larger areas. Reproduced from Yamaguchi, T. (2000), by permission of American Inst of Physics

8.1 BEAM WRITING STRATEGIES

Electron and laser beam systems are the standard tools for pattern generation. They combine high resolution and flexible data management. The simplest writing strategy is termed *raster scan*: it uses a single Gaussian beam and divides the pattern to be drawn into small rectangles and makes an 'exposure-no-exposure' decision for each

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

rectangle. Vector scanning enables skipping of empty (non-exposed) spaces, making the system much faster, at the expense of system complexity. Variable shaped beam is another improvement over raster scan: when larger than minimum pixel size structures are drawn, writing speed is enhanced dramatically.

Electron beam (and laser beam) writing area is very small: ca. $250 \times 250\,\mu m$ area, that is, the area that can be scanned electromagnetically (e-beam) or acousto-optically (laser beam). If an area larger than $250 \times 250\,\mu m$ needs to be drawn, additional movements must be introduced (Figure 8.1). The stage scan is a mechanical movement, controlled by an interferometer. Pattern placement in different subfields is thus a sum of two rather different mechanisms.

8.1.1 Alignment

Alignment is a major criterion in all lithography techniques. In Electron beam lithography (EBL), alignment relies on electron scattering from alignment marks. It can be done in two basic ways. Global alignment uses marks placed on wafer edges. This is fast if ultimate accuracy is not necessary. Chip-alignment uses alignment marks at each chip location. The accuracy can be further increased if alignment marks are visited regularly during writing, rather than just at the beginning of writing. Processing usually begins with a zero layer lithography: only alignment marks are exposed on the zero layer and etched into the wafer, for example, $1\,\mu m$ deep, $10\,\mu m$ wide and $100\,\mu m$ long. These may deteriorate as more layers are deposited and etched, but their global nature makes them better than a sequential layer-to-layer alignment scheme.

8.2 ELECTRON BEAM PHYSICS

Electrons are light mass objects, and when they hit resist with high energy (10–50 kV typical), they scatter forward (recall Figure 2.12). Even though the beam spot on resist top surface is very small, scattering broadens the beam inside the resist and the resist is exposed on a larger area than the beam spot. Forward scattering is not, however, the major component of resist exposure: most of the resist exposure comes from secondary electrons that have been created when the beam slows down. These 2 to 50 eV electrons have a range of a few nanometres in resist.

Beam spots in the 5 nm range are available. This is not limited by the wavelength of electrons ($\lambda = 8$ pm for 25 kV) but rather by electron source size and electron optics aberrations and diffraction for highly collimated beams. Interactions in solid further limit minimum size: effective beam diameter is given by

$$d_{\text{eff}}\ (\text{nm}) = 0.9\ (t/V)^{1.5} \tag{8.1}$$

resist thickness t is in nm and voltage in kV.

Some electrons experience backscattering (large angle scattering) with ca. micrometre ranges. Exposure dose thus depends on the neighbouring structures. This is known as the *proximity effect*. The proximity effect can be combated by biasing structures smaller or larger so that the final pattern is of desired size and shape.

8.3 PHOTOMASK FABRICATION

Instead of direct writing of millions of pixels on a wafer, beam writers can be used to write photomasks for optical lithography. The simplest photomasks are just laser-printed overhead transparencies: they are suitable for structures in the size range of hundreds of micrometres and for simple demos, for example, in a student lab. The printed circuit board industry uses more advanced laser plotters and polyester transparency films, with minimum lines of ca. 30 to $50\,\mu m$. Polymer-based masks suffer from wear and tear and from dimensional instability.

Photomasks proper are glass plates with chromium (ca. 100 nm thick) on them. Soda lime glass is used for larger linewidths ($>3\,\mu m$) and quartz is the material of choice for micron and submicron work. Optical lithography with photomasks is the dominant patterning technology because optical exposure is fast: illumination through a photomask exposes up to 10^{10} pixels in a one second exposure. But the original mask pattern that optical lithography so efficiently reproduces must be written slowly feature by feature. The enormous throughput difference warrants making the mask plates, which can be costly: a set of 15 plates (corresponding to $1\,\mu m$ CMOS process) costs 15 000 USD; and a set of 25 plates for $0.25\,\mu m$ CMOS costs ten times more.

Writing time for a mask plate can be limited by several factors, which depend on the pixel size, total area, resist sensitivity and electronic and mechanical scan speeds

$$\tau_1 = AS/I \tag{8.2}$$

where A is area, S is the exposure dose, I is beam current.

Exposed pixel size, d, affects writing time via

$$\tau_2 = A/fd^2 \tag{8.3}$$

where f is the beam incrementing rate (up to 500 MHz).

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Electronic scan time and wafer stage mechanical movement time must be considered for a complete system

$$\tau_3 = A/Lv \tag{8.4}$$

where L is the electronic scan length and v is stage speed.

The time to write for a $10\,\text{cm} \times 10\,\text{cm}$ area is approximately one hour, as the calculations below show. Typical resist sensitivities vary between 1 to $10\,\mu\text{C/cm}^2$ $100\,\mu\text{C/cm}^2$ is usual for high-resolution resist, poly methyl methacrylate (PMMA) and beam currents range from 1 to 250 nA (or even less for modified SEMs that are used as e-beam writers), which gives τ_1 of the order of 400 to 40 000 s for 250 nA depending on resist sensitivity. Write time τ_2 is, for example, 10 000 s (0.1 μm pixel, 100 MHz). Assuming 250 μm electronic scan length and 1 cm/s stage speed, τ_3 writing time corresponds to 4000 s. Depending on resist selection, either τ_1 or τ_2 gives the limiting write time. If highly sensitive resist is chosen, then pixel size sets the limit.

Photomasks with chrome-on-glass also go by the name binary masks, because there is either a transmission or a blockade of light, but nothing else. In phase-shift masks, PSM, the phase of the light is manipulated while traversing the mask. PSMs will be discussed in Chapter 38.

If the mask is mostly covered by chrome, with only a small percentage of open area, it is said to be a dark field (DF) mask; if it is mostly transparent, with only small percentage of chrome, it is designated a light field (LF) mask, also known as *bright field* (BF) mask.

Process flow for mask fabrication

1. mask blank preparation deposition of chrome on quartz; resist application;
2. pattern writing e-beam or laser; slow writing of elementary shapes;
3. pattern processing resist development chrome etching (wet etching) resist stripping;
4. metrology CD (critical dimension) control;
5. inspection for pattern integrity defects (in chrome) pattern fidelity (shape and position);
6. cleaning particle removal, soft error reduction;
7. repair focused ion beam etching and/or deposition;
8. final defect inspection.

Adapted from Skinner, J.G. *et al*.

Optical lithography can be done with reduction optical systems (to be discussed in the next chapter), which means that the patterns on the mask are larger than final structures on the wafer. This is a great relief for mask makers: 1 μm final size on a wafer corresponds to 5 μm on the mask when 5X reduction optics is used.

8.4 PHOTOMASKS AS TOOLS

Photomasks are tools for process and device engineers (Figure 8.3). The process engineer wants to see the resolution of the optical lithography process, and this is checked by linewidth test structures. Process robustness is tested by structures that span a range of values around the baseline process. For example, if the design linewidth is 3 μm, test structures may span the range 1 to 10 μm. The same applies for spaces between the lines. Linewidth is dependent on the immediate neighborhood, and therefore test structures should include lines of different kinds: isolated, nested, dense, sparse, and so forth (Figure 8.2).

The device engineer designs different geometries of devices: for example, square and octagonal inductor coils, or straight and meandering resistors (Figure 8.3). For transistor parameter extraction, a set of test transistors with dimensions of, for example, 2, 3, 5, 10, 20 and 50 μm are used.



Figure 8.2 Test structure for lithography and etching: the central line is surrounded by dark field and light field areas, and it is found as an isolated line as well as an array line. In the ideal case linewidth should be independent of its neighbourhood

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

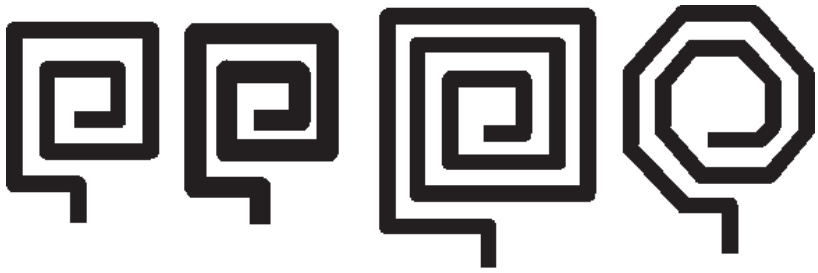


Figure 8.3 Test structures for inductor coils: the process engineer is interested in different linewidths and spacings; the device engineer wants to test different coil shapes and see the effect of the number of coil turns

Writing shapes other than rectangles can be difficult for mask makers. Photomasks are written by machines designed to do XY-orthogonal structures. The CAD programs for IC design support drawing on XY-grid, and even data conversion from design program to mask writer program can be difficult for non-rectilinear shapes. Photomasks are, however, not necessarily XY-symmetric. For instance, stitching of subfields can be made as small as 6 nm in X-direction, but not in Y-direction, because the former depends on beam scanning, but the latter on the mechanical stage movement. Smoothly curving lines needed in integrated optics are difficult, and circles and arbitrary angles pose difficulties, too. Edge definition of structures other than XY-lines can, of course, be increased by using smaller writing grid, or double exposure, both of which increase writing time considerably.

8.5 PHOTOMASK INSPECTION, DEFECTS AND REPAIR

Photomask fabrication requires, in addition to a scanning beam equipment, a repertoire of inspection and repair equipment. Three basic control measurements for masks are linewidth, position and defects. Linewidth is a local measurement, over a test structure pattern. With linewidths in the micrometre range, measurement should be able to discern ca. 10 nm. Pattern position is a global measurement and it is usually fixed to a mask writing tool, controlled by a stage interferometer, and measured to ca. 10 nm accuracy over 10 cm mask plate size.

Defects on the mask are fatal because they will be reproduced on the wafers. Defects can be classified into two broad categories of hard defects and soft defects. Soft defects are mainly particles or resist residues that can be cleaned away. Hard defects are permanent spots or scratches in chrome or in quartz.

Two basic inspection strategies are used: optical inspection combined with a comparison to a known perfect mask plate (known as *die-to-die*) or a comparison between design data and the finished mask plate (*die-to-data*). There are usually hundreds of identical chips on a photomask plate and if they have been independently drawn, it would be statistically improbable that they would have defects at the same locations. This could be the case, however, if there is a systematic error in the data, for example, structures that are beyond the capability of the mask writer system (e.g., too narrow lines have been designed, or too narrow spaces between the lines).

When defects are detected on a mask plate, it is often financially attractive to repair them rather than to write a new plate. Defects come in many guises, but from a repair point of view there are two grand classes of defects:

- missing chrome
- extra chrome.

The former requires the deposition of a layer that will prevent light transmission. Usually, a metallic layer is deposited, for example, tungsten. The latter defect type requires the removal of extra chrome. Both can be accomplished with focused ion beam (FIB) techniques but the real difficulty lies in guiding the FIB to a detected defect site.

Geometric/topological classification of defects (see Figure 8.4):

- protrusion (extra chrome attached to a feature)
- intrusion (partial loss of chrome in a feature)
- bridge (chrome connecting two features)
- necking (discontinuity in a line)
- pinhole (hole in a chrome)
- pin spot (extra chrome on a light field area).

From the yield and reliability point of view not all defects are equal. Defect must be understood as a

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030



Figure 8.4 Mask defects: defects smaller than the feature size will affect final dimensions and, therefore, current density, electric field and other device parameters. Redrawn after Skinner, J.G. *et al*., by permission of SPIE

very broad term: anything that prints on the wafer or changes critical dimension by more than 10% is counted as a defect. This can be a light transmission error, a pattern error, a stochastic scratch or an undulating line edge.

Defect size is important: not all defects are able to destroy the functionality of the chip. As a rule of thumb, defects greater than one-third the minimum linewidth are prospective 'killer defects'. Mask buyer can specify defects and accept plates with some defects that have been classified as non-fatal.

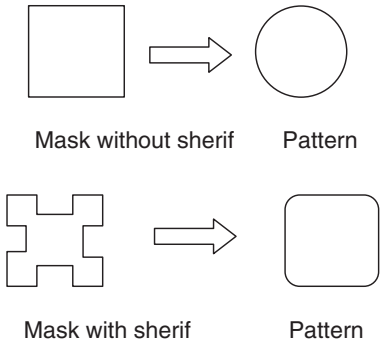
Optical defects not related to written patterns include the following:

- transmission variability in glass (LF areas)
- transmission variability in chrome (DF areas).

Transmission defects are subtle, and even if detected, it is not straightforward to repair them. Phase-shift mask making is very expensive partly because of difficulties in inspection and repair or transmission defects.

8.6 EXERCISES

1. How deep will (a) 10 keV e-beam penetrate into silicon and (b) 50 keV beam into quartz?
2. What is the smallest possible feature size that can be written with a 50 keV electron beam?
3. What is the photomask writing time for a gigabit circuit with 1 000 000 000 contact holes, when the incrementing rate is 500 MHz and mask plate area 8 cm × 8 cm? The photomask is 4X the final size.
4. What process and materials parameters do you need to know in order to estimate the electron beam heating of a mask plate and resist during EBL? How does beam-induced heating affect linewidth control?
5. Use a laser printer to make simple line/space test structures with 600 dpi and 1200 dpi resolutions, and check by microscope for linewidths, line edge roughness and reproducibility.
6. How is the electron beam system throughput affected if 5X masks are drawn, instead of 1X masks?
7. Sherifs are proximity correction structures at the corners of lines: sherifs result in a more rectangular final shape compared with a simple rectangular initial shape. If the sherif size is half the feature size, calculate how the e-beam writing time is affected!



REFERENCES AND RELATED READINGS

Allen, P.C.: Laser scanning for semiconductor mask pattern generation, *Proc. IEEE'90* (October 2002), p. 1653.

McCord, M.A. & M.J. Rooks: Electron beam lithography, in P. Rai-Choudhury (ed.): *Handbook of Microlithography, Micromachining and Microfabrication*, Vol. 1, p. 139.

Pugh, G. *et al*: Impact of high resolution lithography on IC mask design, *Custom Integrated Circuits Conference* IEEE (1998), p. 149.

Skinner, J.G. *et al*: Photomask fabrication procedures and limitations, in P. Rai-Choudhury (ed.): *Handbook of Microlithography, Micromachining and Microfabrication*, Vol. 1, p. 377.

Yamaguchi, T.: EB stepper – a high throughput electron projection lithography system, *Jpn. J. Appl. Phys.*, **39** (2000), 6897.

Conference series "Photomask" organized by SPIE and BACUS is organized annually.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

9

Optical Lithography

Lithography work flow consists of the following major steps when viewed from the point of view of the wafer:

1. Photosensitive film (photoresist) application
2. Alignment of mask and wafer
3. Exposure of the photoresist
4. Development of patterns.

The alternative view is that of information flow; this will be discussed in Chapter 10 in conjunction with lithography simulation.

Optical lithography is basically photography. The original image to be transferred, the photomask, which corresponds to the negative in photography, is set in a mask-aligner/exposure tool. It is aligned to the photoresist-coated wafer, and exposed by UV radiation (Figure 9.1). Exposure changes photoresist solubility, which enables selective removal of resist in the development step. In positive resists, the exposed areas become more soluble in the developer, and in negative resists, the exposed parts become insoluble.

This resist pattern can be used as an etch mask. Photoresist is removed after etching. The patterning process continues with new doping and deposition steps, and new lithographic steps. Layers have to be aligned to each other, as in multiple exposure photography. Overlay of successive layers is a critical factor in lithography, not only in resolution.

There are three rather different elements in the optical lithography process:

- *Optics*: radiation generation, propagation, focusing, diffraction, interference;
- *Chemistry*: photochemical reactions in the resist, development;
- *Mechanics*: mask-to-wafer alignment.

We will discuss lithography first from a tool point of view, and then from a pattern point of view: the shape and size of patterns that can be printed on the wafer.

9.1 LITHOGRAPHY TOOLS (ALIGNMENT AND EXPOSURE)

The simplest lithographic technique is contact lithography: the photomask and the resist-covered wafer are brought into intimate contact, and exposed. The resolution is determined by mask dimensions and diffraction at mask edges. Extremely small patterns can be made in theory but making photomasks with submicron features is prohibitively expensive. Damage to mask is frequent when the mask and the wafer are brought into contact, which makes contact printing not very production worthy.

Proximity lithography is a modification of contact lithography: a small gap, for example, 3 to 50 μm is left between the mask and the wafer. The wavefront traversing the mask is diffracted by the mask patterns, and Fresnel diffraction formulae have to be used to estimate resolution. Both contact and proximity lithography are done in one and the same machine: the gap between the mask and the wafer is an adjustable parameter, with values from zero up (Figure 9.2). Contact/proximity lithography systems are 1X: the image is the same size as the original. The role of optical system I (Figure 9.1) is then to provide uniform illumination. Optical system II does not exist.

In projection optical systems, the optical system II of Figure 9.1 is the key element: it provides an image of the mask on the wafer. Reduction optics can be used, and this is a great improvement over 1X systems. With 5X reduction projection optics, the original photomask features can be made rather large, for example, 1 μm for 0.2 μm final feature size. Fraunhofer far-field diffraction governs the optics of projection systems.

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

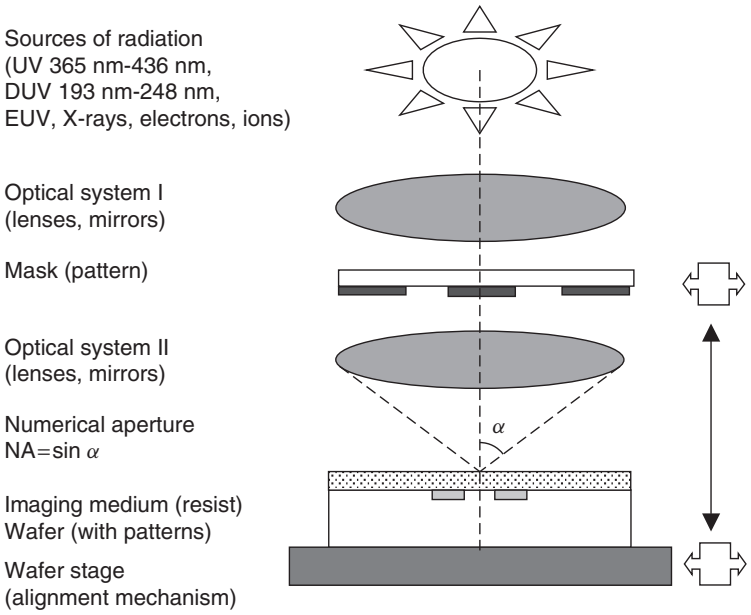


Figure 9.1 Optical lithography: alignment and optical exposure of photosensitive resist film. Note that mask image reduction can be done in projection optical system

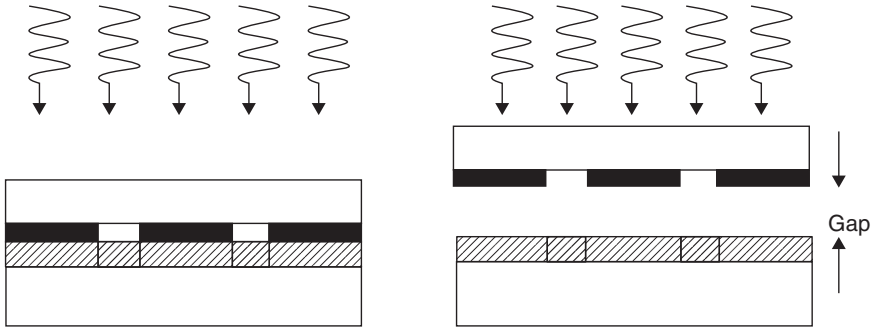


Figure 9.2 Contact and proximity lithography. Proximity gap is typically 3 to 50 μm

Projection optics is often used for chipwise exposure: one chip is exposed, and the wafer is moved to a new position, and another chip is exposed. This approach is termed *step-and-repeat*, and the systems are known as *steppers*. It is certainly slower than full wafer exposure (at the introduction of step-and-repeat, throughput was ca. 30 WPH (wafers per hour), compared with 100 WPH of 1X projection optical systems), but several advantages are apparent. First of all it is much easier to make optical systems for, say, 20×20 mm exposure fields than for 150 mm, let alone for 200 mm or 300 mm wafers. Second, alignment can be done for each chip individually. Third, experimentation is easy: for example, all chips can be exposed differently (Figure 9.3), in order to find the optimum exposure dose and focus conditions, and to check process robustness. It is possible to change reticle between exposures, and have many different chips on one wafer in any proportion. Inclusion of test chips is thus flexible. Step-and-repeat photomasks are called *reticles*, and sometimes the word 'mask' is reserved for 1X full wafer masks only.

Step-and-repeat was an existing technique in the photomask industry: the original chip pattern was written on a mask blank and the final 1X full wafer mask with hundreds of identical chips was made by

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

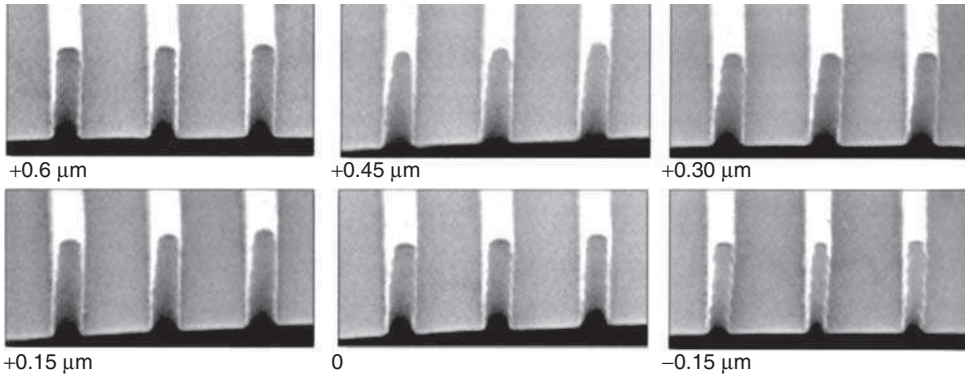


Figure 9.3 0.20 μm lines printed in 0.7 μm-thick resist by 248 nm exposure. Different focus depths have been tried. Reproduced from Peterson, B. *et al*. (1996), by permission of ICG Publishing Ltd, London

copying the original pattern many times over to another mask blank.

Step-and-scan is an alternative high-resolution optical approach. In step-and-scan the reticle and the wafer move in unison, and the exposing radiation enters through a narrow slit. 4X-reduction scanners are widely employed in manufacture of advanced CMOS chips.

In projection optical system, the reticle is not in physical contact with the wafer, which greatly improves mask lifetime. During 1X contact/proximity period, mask makers had big business making new working copies of existing designs on a regular basis. Photoresist debris can of course be cleaned from the mask, but frequent cleaning itself is a danger to the mask: chrome adhesion loss, chrome etching, scratches and mechanical damage in handling or electrostatic charging from spray nozzles used in cleaning are potentially damaging.

Soft defects: particles, chrome-etch residues, resist flakes, and so on, can be removed by cleaning once detected. One way to battle soft defects is pellicle: a protective transparent film is attached above the reticle immediately after mask inspection. Airborne particles will settle on the pellicle film, which is ca. 100 μm above the chrome pattern. This eliminates particle defects because they will be out of focus during lithography. This approach is of course not applicable in contact or proximity lithography.

5X reduction makes mask-making much easier. Errors in both resist image and the etched chrome image on the mask are reduced, leading to tighter linewidth tolerances on the wafer (Table 9.1). Mask writer placement error is also reduced, improving overlay between two layers. The more complicated optics of reduction systems (in contact printing there is no imaging optics) introduce some distortion but this is a minor price to be paid.

Table 9.1 1X and 5X lithography systems compared

Linewidth variability	1X	5X
Resist image on mask	8%	1.6%
Chrome image on mask	8%	1.6%
Resist image on wafer	10%	10%
Etched image on wafer	10%	10%
Residual sum of squares RSS	18.1%	14.3%
Overlay variability	**1X**	**5X**
Mask writer placement	72 nm	14.4 nm
Wafer alignment error	50 nm	50 nm
Stepper table error	30 nm	30 nm
Lens distortion	15 nm	30 nm
Residual sum of squares RSS	94 nm	68 nm

Source: Rai-Choudhury, P. (1997).

9.2 RESOLUTION

9.2.1 Contact/proximity printing

Making closely spaced narrow lines is the main challenge in microlithography; not the making of individual narrow lines. An individual narrow line can be made even accidentally by for example overexposure (but line shape will be far from ideal). Resolution, or the ability to separate two patterns, is then the criterion for patterning accuracy (Figure 9.4). Proximity lithography minimum resolvable period $2b_{min}$ is calculated from

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

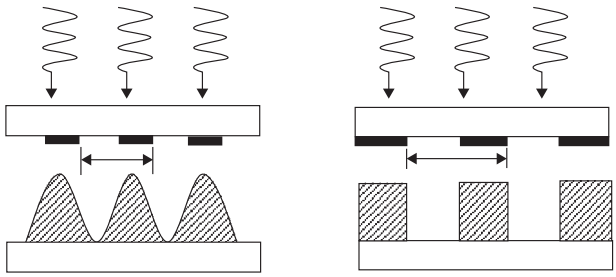


Figure 9.4 Resist profiles and resolution: (**a**) microlithographic resolution is not enough to produce useful resist patterns (even though optically the structures are clearly resolved) and (**b**) for larger lines and spaces, proper resist profiles can be produced. Positive resist: exposed parts are dissolved in development

Fresnel diffraction and approximated by

$$2b_{\min} = 3\sqrt{\frac{\lambda}{n} \times \left(g + \frac{d}{2}\right)} \tag{9.1}$$

Typical values for these parameters are

λ	Wavelength of exposing radiation	$\lambda = 436\,\text{nm}$, mercury lamp g-line
g	Gap between mask and photoresist	$g \approx 0 - 50\,\mu\text{m}$
d	Resist thickness	$d \approx 1\,\mu\text{m}$
n	Resist refractive index	$n \approx 1.6$

Perfectly vertical resist walls (90°) are difficult to make. Positive resists usually have a slightly positive slope, 85° to 89°, negative resists have similar retrograde profile. This is a natural consequence of exposure light intensity through the mask.

In MEMS and thin film head fabrication, resists can be 10 to 100 μm thick, or even thicker. The resolution formula 9.2 is valid in the interval

$$\lambda < \text{gap} < L^2/\lambda \tag{9.2}$$

where L is the linewidth.

X-ray lithography is proximity lithography, but with much smaller wavelength: $\lambda \approx 1\,\text{nm}$ is used, and therefore much smaller lines can be printed. X-ray lithography can also expose thick resists (100–1000 μm) quickly because synchrotron light sources provide intense X-ray beams. Because of good collimation, vertical resist sidewalls will result, enabling resist height to width ratios above 100:1.

9.2.2 Resolution: projection optical systems

Resolution of projection optical system is approximated by Rayleigh relations:

$$\text{resolution} = k_1\lambda/NA \tag{9.3}$$

$$\text{depth of focus} = k_2\lambda/NA^2 = \pm\lambda/(2NA^2) \tag{9.4}$$

NA is the numerical aperture of the system (Figure 9.1) and λ is the exposure wavelength. Rayleigh criterions are optical, whereas we are interested in microlithographic resolution that intricately involves masks and resists. These are incorporated into the parameters k_1 and k_2. Using $k = 1$ criterion for 0.15 *NA* system at 436 nm wavelength (corresponding to 1980's stepper) ca. 3 μm resolution is possible. Over the years, optics designs have pushed *NA*s higher, up to 0.8, and shorter wavelengths (365 nm, 248 nm, 193 nm) have been employed.

Parameters, k_1 and k_2, were long considered constants, but recently they have been aggressively scaled down. This requires much higher degree of control of all aspects of the lithographic system: resist uniformity and mask quality have to be improved; and for further dowscaling of k_1, Optical Proximity Correction must be employed, and later on Phase Shift Masks must be introduced. Assuming $k_1 = 1$, 0.6 NA exposure tool with 248 nm wavelenght is capable of 400 nm resolution, but it has production resolution of 300 nm which corresponds to $k_1 = 0.7$, and it is capable of 200 nm in a research laboratory, which means that $k_1 = 0.5$. Lithography scaling is driven exclusively by CMOS. Most microfabrication industries do not share the tools and techniques of deep submicron CMOS lithography.

9.3 BASIC PATTERN SHAPES

There are four basic shapes that have to be patterned: line, trench, hole and dot. An opaque chromium line on a mask will end up as a line on the wafer if positive resist is used, but as a trench in the case of negative resist (Figure 9.5). A transparent opening in chromium will result in a trench with positive mask, and in a line with negative resist. Masks of Figures 9.5(a) and (b) are thus interchangeable if resist polarity is switched.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

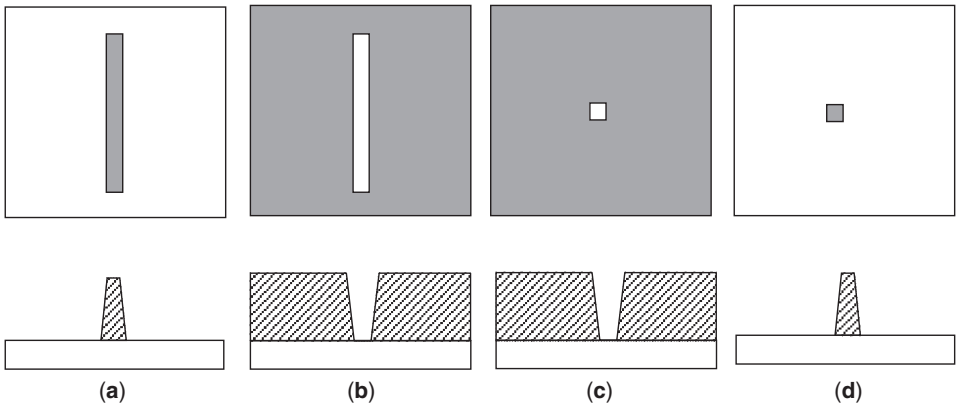


Figure 9.5 Basic pattern shapes and their positive resist profiles (**a**) line (LF); (**b**) trench (DF); (**c**) hole (DF) and (**d**) dot (LF)

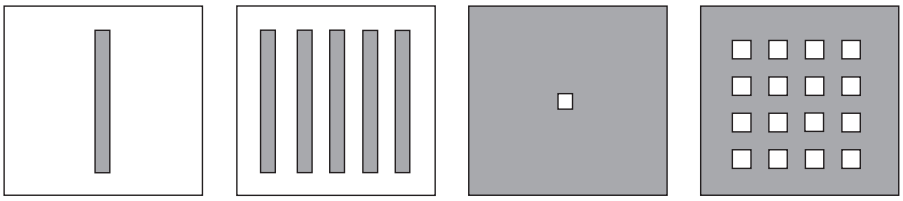


Figure 9.6 Isolated vs. array features

Patterns come in two basic varieties: isolated and array (Figure 9.6). Lithography for these is different, and the ultimate lithographic resolution is also shape dependent. For example, stray light is a major issue for a light field structures, whereas in dark field patterns, it is not so much of an issue.

Isolated lines can be made fairly easily in any desired width. But resolution, that is, the ability to print two lines close to each other is what determines the device-packing density on the wafer. Microlithographic resolution, line plus space, is called pitch.

In CMOS circuits, the minimum linewidth is usually that of polysilicon gate, which is an isolated line. Contact hole and trench minimum linewidths are usually slightly larger (e.g. by 10%); isolated dots may have a minimum size 20 to 50% larger. Resolution is not usually divided equally between line and space: 0.8 μm resolution can mean 0.35 μm wide polygate with 0.45 μm space.

9.4 ALIGNMENT AND OVERLAY

Because microdevices are built-up layer-by-layer, overlay of successive layers relative to previous layers is a paramount performance criterion of optical lithography align/exposure tool. Overlay refers to general pattern placement, and alignment refers to the specific spots on the wafer, the alignment marks (a.k.a. alignment keys or targets) that are used for the alignment procedure. Because alignment is limited to specific structures (usually on the wafer or chip edge), it is not a full guarantee of overlay elsewhere. Overlay is affected by lens aberrations, wafer chuck irregularities (equipment related problems), mask pattern misplacement (mask fabrication problems) or distortions on the wafer itself, such as warpage or site flatness. We will, however, use the term alignment as a general term for layer-to-layer registration because it is an easy operational concept. The term "mask aligner" nicely underlies the importance of alignment. As a rule of thumb, alignment of 1X systems is ca. one-third of the minimum linewidth. A contact/proximity aligner that can print 3 μm minimum lines is typically capable of 1 μm registration between levels. A 5X projection stepper with 0.5 μm minimum linewidth can align to ca. 0.1 μm.

Alignment needs to be evaluated over long time: device fabrication processes take weeks or even months. For example, temperature differences between different exposures will affect alignment because of thermal expansion of the wafer, the wafer stage and the

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

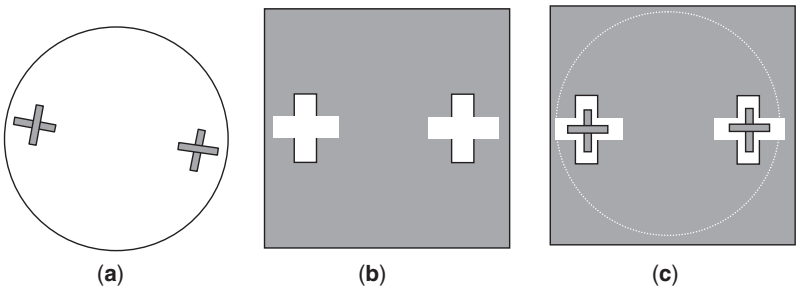


Figure 9.7 Alignment operation: (**a**) wafer with alignment marks; (**b**) photomask with alignment marks and (**c**) after linear translation and rotation of the wafer the alignment marks on wafer and mask coincide

photomask. The lenses in the optical path of the exposure tool are subject to constant UV flood, and they too need to be thermally stabilized.

Alignment needs to be discussed from two rather different points of view:

1. *Equipment view*: This is an optomechanical problem of finding alignment marks on the mask and on the wafer, and manipulating them to coincide.
2. *Device design view*: This is a design issue and it depends on overlaps and spacings that structures need for the device to operate, for instance metallization has to overlap contacts.

Alignment could be done using the devices themselves, but this is impractical because of micrometre dimensions and multiple identical structures. Therefore separate alignment marks are used. Alignment marks are much larger than device features because they exist only for alignment, and have nothing to do with resolution. Alignment is usually done on a wafer level, with two alignment marks as far from each other as possible, to increase theta (rotational) resolution (Figure 9.7).

Alignment sequence determines which layers are aligned to each other. Layers are not necessarily aligned sequentially to a preceding layer, but to some important previous layer. A contact hole is aligned to a resistor, but the metal layer can be aligned either to the contact hole, to make sure that the whole contact hole is covered, but it can also be aligned to the resistor; after all, the metal has to make contact with the resistor. These issues will be dealt with in Chapter 24.

9.4.1 Lithography metrology

Lithography produces test structures of itself. Test structures must include resolution structures with the same dimensions as the devices themselves, but also smaller and larger structures so that process robustness and linearity can be checked. Optical microscopy and scanning electron microscopy (SEM) are standard methods. Even when linewidths are below optical microscopy resolution, it is useful as an initial check: for instance, resist adhesion loss, delamination and other gross errors can be seen. Linewidth control is usually accepted as $\pm 10\%$ of design value. Linewidth measurements by stylus/AFM or SEM form the basis of lithography process control. Resist thickness has a profound effect on linewidth, as will be discussed in the next chapter.

9.5 EXERCISES

1. What is the best possible resolution in optical contact lithography?
2. What is the diffraction limited resolution of 10 nm X-ray photons?
3. 100 mm diameter silicon wafer has 1 μm lines fabricated on it. The photomask is made of soda lime glass with a coefficient of thermal expansion (CTE) of 10 ppm ($10 \times 10^{-6}/^{\circ}\mathrm{C}$). How accurately must the temperature in the patterning process be controlled in order to keep distortions from thermal expansion over 100 mm wafer below 0.3 μm? Silicon CTE is $2.5 \times 10^{-6}/^{\circ}\mathrm{C}$.
4. Make a graphical presentation of projection lithography resolution versus depth of focus!
5. A 50 μm thick resist must be used in an electroplating process. What is the minimum feature size that can be used?

REFERENCES AND RELATED READINGS

Helbert, J.N.: *Handbook of VLSI Micro lithography*, Noyes Publications, 2001.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Moreau, W.: *Semiconductor Micro lithography*, Plenum Press, 1988.

Peterson, B. *et al*: Approaches ro reducing edge roughness and substrate poisoning of ESCAP photoresists, *Semicond. Fabtech.*, **8** (1996), 183.

Rai-Choudhury, P.: (ed.): *Handbook of Micro lithography, Micromachining and Microfabrication*, Vol. 1, SPIE, 1997.

Schneider, C. *et al*: Automated photolithography critical dimension controls in a complex, mixed technology, manufacturing fab, *Advanced Semiconductor Manufacturing Conference* (2001) IEEE/SEMI, p. 33.

Shaw, J.M. *et al*: Negative photoresists for optical lithography, *IBM J. Res. Dev.*, **41** (1997), 81.

Microlithography World magazine: http://sst.pennnet.com/home.cfm

Giesecke+Devrient GmbH, Giesecke+Devrient Currency Technology GmbH, and Papierfabrik Louisenthal GmbH. - Ex. 1030

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

10

Lithographic Patterns

We will now discuss photoresists. Resist chemistry and resist working principles will be covered. In Chapter 9, we treated resists as if they were digital on/off materials that either react under exposure or do not; now we are dealing with more realistic cases: resists have exposure threshold energy, finite contrast and finite selectivity in developers. Resists are also optical materials and they are part of an optical system with reflections, interference and absorption. All these aspects become more pronounced when resists go over topography; patterning on a planar surface is fairly straightforward. Simulation of lithography will also be presented.

10.1 RESIST APPLICATION

The lithography process starts by a surface preparation step like almost all microfabrication processes. In order to remove moisture, the wafers are baked. The next step, wafer priming, also known as *adhesion promotion*, ensures known surface conditions. Hexamethyl disilazane vapour (HMDS, $(H_3C)_3$–Si–NH–Si–$(CH_3)_3$) is applied at reduced pressure to form a monomolecular layer on the wafer surface, making the wafer hydrophobic, which prevents moisture condensation. This is especially important for materials like metals, polysilicon and PSG, because resist adhesion to these materials is poor. Adhesion promotion is also a guarantee against cleanroom humidity variations and an equalizer for wafers with different storage times.

Spin coating is the standard resist application method (recall Figure 5.9). A few millilitres of resist is applied on a static or a slowly rotating wafer. Acceleration to ca. 5000 rpm spreads the resist over the wafer, leaving a very uniform layer. The remaining solvent evaporates during soft bake, for example, 90 °C, 30 min in an oven or 90 °C, 60 s on a hot plate.

Spin speed can be used to tailor resist thickness over one decade, for example, 0.5 to 5 μm, but beyond that a new resist formulation with different solid content must be used. Viscosity is dependent on resist solid content (which can vary from 20–80%) and temperature. The solvent evaporation rate depends on ambient environment, and a closed spinner bowl with saturated solvent vapour and adjustable exhaust can be used to control evaporation.

On a planar surface, a 5 nm thickness variation across the wafer is standard for a 1 μm thick resist. Spin processing over severe topography is difficult: liquid-like film will fill grooves and crevasses, and a highly non-uniform resist thickness results (Figure 10.1). This is a problem for textured solar cells (Figure 1.6) or deep-etched MEMS structures (Figure 1.10). On the other hand, this planarizing effect is sometimes used to advantage.

There are three more resist coating technologies: electrochemical coating, spray coating and casting. Electrochemical coating requires special resist formulations, spray is applicable to thin resists. Casting is suitable for thick resists only. These techniques are especially suited to applications in which resist coverage is needed over severe topography, where spin coating is notoriously bad.

Thin resists are preferred for better resolution; but thinner resists are prone to particle defects, and pinhole density rapidly increases when resist thickness is scaled down. Spin-bowl cleaning is also a major particulate control issue: frequent cleaning prevents layer growth, and thus flaking of residual film from the walls. Even monolayer resists have been used in research applications. They can be used as etch masks for shallow etchings in the 10 nm range, or as electrodeposition masks, but clearly are not general purpose resists. Monolayer resists are not spin coated: self-assembled monolayers (SAMs) and Langmuir–Blodgett techniques are employed.

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

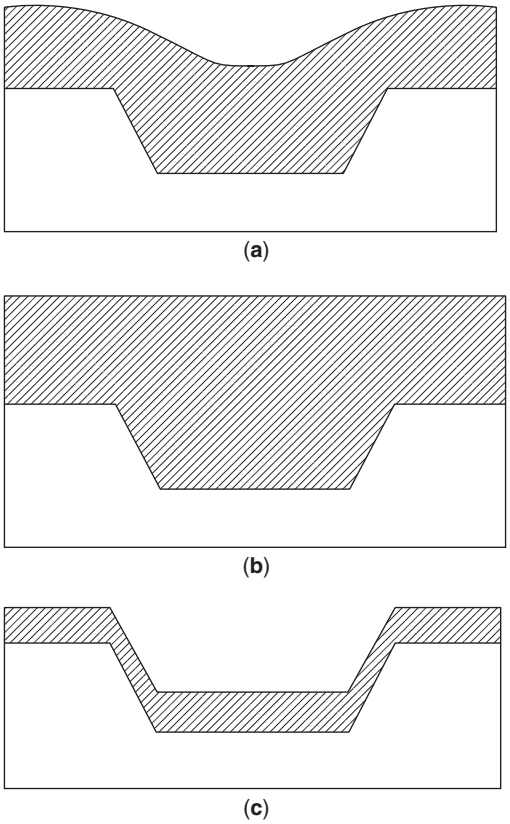


Figure 10.1 Resist over topography (**a**) spin-coated; (**b**) cast and (**c**) electrodeposited or aerosol spray coated

10.1.1 Thick resists

'Thick' can mean very different thicknesses to different people. For IC people, 5 μm is already thick; 5 times the standard thickness. In MEMS and thin film head (TFH) fabrication for magnetic recording, 'thick' can be anything from 5 to 200 μm, and in X-ray lithography, 'thick' extends to the millimetre range.

Thick-resist (and spin-on-glass) processing has a few extra factors that need attention, compared to standard resists. Rapid solvent evaporation has to be prevented because rapid and large shrinkage leads to defective and non-uniform films. One solution is a closed spinner bowl that creates a saturated solvent–vapour atmosphere. This buys extra time to ensure uniform resist spreading before viscosity increases so much that flow is stopped. The solvent evaporates during final spinning to some extent, but for thick resists, it is advantageous to perform an additional slow spinning step in the end, to further dry the resist. Thick resists are very sensitive to levelling, and if the film is not dry, it will flow on an uneven surface after spin coating. It is also possible to apply a thick resist by multiple coatings of thinner layers. Soft baking for solvent removal must be done after each application.

10.1.2 Edge bead

Spin-film definition at the wafer edge is often poor: the resist always flows over the edge, but the film at the edge is discontinuous or non-uniform. Some film is easily transported to the back of the wafer, which may cause contamination in subsequent process steps. Drying during spinning increases viscosity at the edges, which causes accumulation of material on the rim of the wafer. This is known as *edge bead.*

Edge bead removal (EBR) is a process in which a directed solvent jet etches the resist away from the wafer edges. This does not diminish the number of usable chips because the edge chips are usually non-functional anyway. The opposite of EBR is sometimes used in MEMS: in order to prevent edge chipping during long wet etching, edges are protected by extra resist.

10.2 RESIST CHEMISTRY

Resists have three main components:

- base resin, which determines the mechanical and thermal properties;
- photoactive compound (PAC), which determines sensitivity to radiation;
- solvent, which controls viscosity.

The most common base resin for positive resists is phenolic Novolak, which is soluble in alkaline developers. Diazonapthoquinine (DNQ), a photoactive compound, acts as an inhibitor; and the unexposed resist is therefore non-soluble in developer. Upon exposure, DNQ decomposes and releases carboxylic acid, which makes the exposed resist soluble (Figure 10.2).

The calculation of exposure uses the normalized concentration $M(x, t)$ of the remaining inhibitor: it describes the fraction of inhibitor left after exposure at a certain time in a certain position inside the resist. The optical absorption α in the photoresist is described by

$$\alpha = AM(x, t) + B \tag{10.1}$$

where A is the exposure-dependent and B, the exposure-independent absorption. A and B are known as *Dill parameters*, and their values for novolak resists are in

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

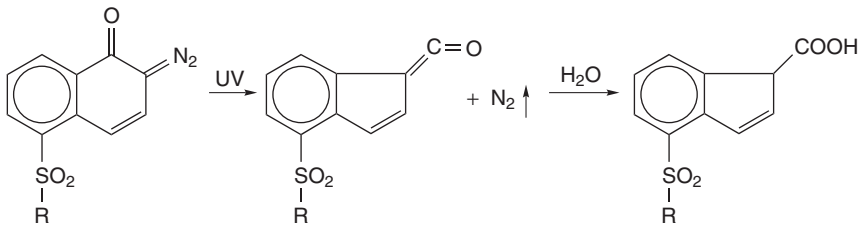


Figure 10.2 Diazonapthoquinine (DNQ)-novolak-resist reaction upon UV exposure. The photoactive compound reacts to form carboxylic acid, which is soluble in the developer. Reproduced from Neureuther, A.R. & C.A. Mack, by permission of Int Soc for Optical Engineering

the range 0.4 to $1\,\mu m^{-1}$ for A and 0.01 to $0.1\,\mu m^{-1}$ for B. The decrease of inhibitor concentration depends not only on the light intensity $I(x,t)$, but also on sensitivity to exposing radiation C, and of course, on inhibitor concentration M. Time-dependent inhibitor concentration is given by

$$\partial M/\partial t = -I(x,t)M(x,t)C \qquad (10.2)$$

The sensitivity parameter C is also known as *Dill C* and its value for novolak resists is of the order of $0.01\,cm^2/mJ$. A, B and C are, of course, wavelength-dependent. Analytical solutions to resist exposure are very difficult and simulation is extensively used.

Resist sensitivity can be tailored for different wavelengths (or for electrons, ions or X-rays; the name *photoresist* is used in non-optical lithographies as well). Sensitivity is important for productivity. With typical exposure energies of the order of 100 to $500\,mJ/cm^2$ for DNQ positive resists, exposure times for standard $1\,\mu m$ thick resists are of the order of 1 s with 500 W lamps. In the first approximation, a $10\,\mu m$ resist needs 10 s exposure, and a $100\,\mu m$ thick resist requires 100 s (development time, which is ca. 1 min for a $1\,\mu m$ resist, must also be multiplied by thickness ratio).

Deep-UV (DUV, 248/193 nm) resists with chemical amplification (CA) are more sensitive. The first DUV lamps had too low intensities for practical throughputs and this problem led to the development of high-sensitivity chemically amplified resists in the 1980s. CA resist works in two steps: photoacid generator (PAG) molecules decompose upon photon impact and these decomposition products catalyse more PAG decomposition so that a single photon can lead to 1000 decomposition reactions. In the second step, in post-exposure bake, the photoreaction products diffuse (nanometres or a few tens of nanometres) and react, and the reaction products are responsible for the solubility difference between exposed and unexposed resist.

Because the reaction is catalytic, the exposure dose is very small and the system throughput is high. CA resists need only 10 to $50\,mJ/cm^2$ exposure doses, one-tenth of that for novolak resists. However, the very fact that the reaction is catalytic poses a danger: if the reaction is quenched, and multiplication stops, the resist is not exposed. This can happen because of airborne contaminants that react with the resist. Ammonia is one prime culprit, and ammonia cannot be completely eliminated from cleanroom air because it is such an essential component of cleaning baths, and ammonia is released by HMDS priming process. The two-step nature makes lithography time-sensitive. Lithographic performance is a sum of illumination and post-exposure bake, and the two steps need to be done sequentially without time delays.

Negative resists can become insoluble because of molecular weight increase due to polymerization. The resist becomes cross-linked either via free-radical or acid-catalysed polymerization. Alternatively, chemical reactions in the resist can generate photoproducts that bring about solubility differences. The cross-linking feature that makes negative resists stable also makes photoresist removal difficult, an obvious dilemma.

Negative resists were the original resists in microfabrication, but in the 1970s positive resists overtook them. Negative resists have, however, a larger market than positive resists, owing to their predominance in the printed circuit board industry where low cost and high sensitivity are combined with fairly large linewidths. Negative resist developers are solvents, and some solvent diffuses into the resist, causing swelling and loss of linewidth control. Positive resists are developed in weak alkaline solutions that are easier and safer to handle. New negative resists have been introduced over the years, and today, resolution is not anymore the determining factor in the positive/negative choice. For thick resists ($>20\,\mu m$), negative tone is

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

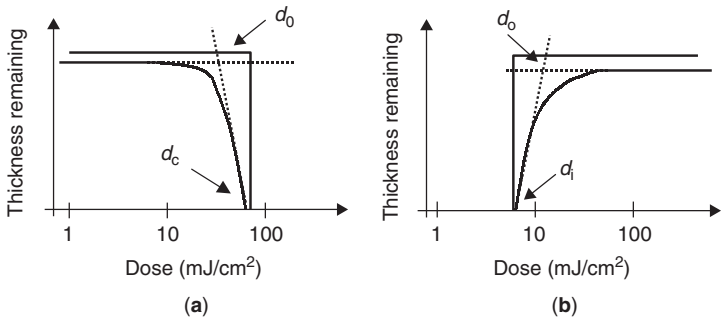


Figure 10.3 Resist contrast plots on thickness–exposure dose axes for infinite contrast resist and real resists (**a**) positive resist and (**b**) negative resist

preferred because high absorption in positive resists limits exposure depth.

10.2.1 Contrast

Photoresist contrast is important for both resolution and profile. A sigmoid (non-linear) response function is essential for patternability. Optical wavefronts after mask are not ideal square waves but rather attenuated sine waves, and linear response as a function of exposure dose is rather useless because the photoresist patterns are smoothly curving bumps, and not clearly defined rectangular shapes.

Contrast is calculated for positive and negative resists as

$$\gamma_p = (\log(d_c/d_0))^{-1} \qquad \gamma_n = (\log(d_o/d_i))^{-1} \quad (10.3)$$

where d_c is the dose to clear all resist and d_0 is extrapolated dose at the kink of the contrast curve, and for negative resists, d_o and d_i are defined analogously (Figure 10.3). Typical contrasts are 2 to 5 for novolak-based positive resists, and 5 to 10 for DUV resists.

10.3 THIN FILM OPTICS IN RESISTS

A photoresist is a part of an optical system involving the illumination light source, the lenses and the photomask, and we have to also include the substrate, because light reaching through the resist to the substrate will be reflected back, and it contributes to pattern formation (Figure 10.4).

Photoresist thickness determines the optical path length for the incoming and outgoing rays. Constructive and destructive interference inside the photoresist lead to intensity variation in the vertical direction through the resist. This is seen as standing wave patterns in the developed resist. In the extreme case, the parts that receive least light (in positive resist) will not be developed by a developer that has high selectivity between exposed and unexposed parts (high-contrast developer). Post-exposure bake, which enhances diffusion of photoproducts, will make the standing wave effect smaller.

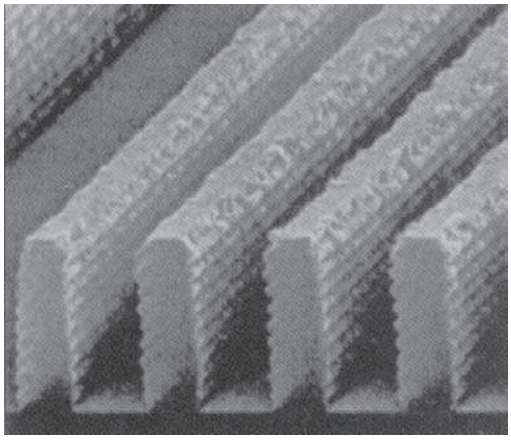


Figure 10.4 Reflections at the air–resist and resist–substrate interface result in interference pattern of standing waves. Reproduced from Peterson, B. *et al.* (1996), by permission of Henley Publishing

Thin-film interference in the resist leads to thickness-dependent exposure doses. Depending on the resist thickness, the total dose needed to expose the resist changes. If destructive interference takes place in the top surface of the resist, almost all the illumination energy is absorbed in the resist, whereas in the case of constructive interference at the top surface, only half the energy stays inside the resist. Maxima and minima alternate at $\lambda/(4n)$ intervals; for example, for the exposure of a resist of refractive index 1.64 to light of wavelength $\lambda = 365$ nm, this interval is 56 nm. On a planar surface, this problem can easily be solved by better control of the photoresist

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

spinning process, but on a structured surface there is no general solution to the variable resist thickness problem (Figure 10.5).

Swing ratio is a measure of the variation introduced by thin film–optical effects. It is determined as exposure dose variation (max–min) divided by mean value. It can be defined similarly for linewidth. It is analogous to a lossy Fabry–Perot interferometer, and swing rate can modelled as

$$S = 4\mathrm{e}^{(-\alpha D)}\sqrt{(R_1 R_2)} \qquad (10.4)$$

where R_1 is the reflectivity at the air–resist interface;
R_2 is the reflectivity at the resist–substrate interface;
α is the resist absorption coefficient;
D is the resist thickness.

Obviously, there are four ways to minimize the swing ratio. One strategy is to minimize R_1, which translates to a top antireflective coating (TAR). Light traversing TAR twice will interfere destructively and minimize reflections if the TAR thickness matches the $\lambda/4n$ condition. The TAR refractive index is given by $n_{\mathrm{TAR}} = (n_{\mathrm{resist}} \times n_{\mathrm{air}})^{1/2}$. With resist n's typically around 1.65, the TAR refractive index should be ca. 1.3. The TAR thickness would then be ca. 70 nm.

Photoresist-like spinning is a popular method for coating the TAR, and the material is very much photoresist-like (non-absorbing, however), and it will be removed by the developer. Added process complexity is small. The TAR is insensitive to the substrate material, and therefore, this is a fairly general method to reduce reflections and swing. If, however, the TAR is deposited over steps in a way similar to the resist, the TAR thickness will be variable, and its effectiveness reduced.

Reduction of R_2 involves bottom antireflective coatings, BARCs. BARCs work by index matching just as TARs but also by absorption: absorbed light will not re-enter the resist. BARC thicknesses are not unlike those of TARs, but the materials and processes are. BARCs must tolerate developers, because if they did not, they would undercut the resist patterns. BARCs are therefore patterned by dry-etching. Spin-on polymer-based BARCs do exist, but inorganic BARCs that will be left as permanent parts of the finished devices are also used. Titanium nitride, TiN, is a BARC for aluminum lithography, but it is deposited in the same process as the aluminum, not in conjunction with resist processing. Oxides and nitrides can also be used as BARCs. It is difficult to remove them selectively, and most often, they too remain as parts of finished devices. Inorganic BARCs can act as hard masks for etching: the resist is used as mask for BARC etching, and BARC is then used as a mask for film etching.

Absorption strategy involves resist tailoring. Standard αs are around 0.2 to 1 $\mu\mathrm{m}^{-1}$. Adding dyes to increase α to, for example, 2 $\mu\mathrm{m}^{-1}$ means that all radiation will be absorbed in the top resist layer, and the bottom part will not be exposed. So, there is an optimum between swing ratio reduction and resist profile. Top-surface imaging (TSI), which will be discussed shortly, overcomes the absorption dilemma by using very thin resists, which are not sensitive to profile variation like standard resists.

The fourth possibility, resist thickness increase, is at odds with resolution: if we wish to print narrow lines, thinner resists are better. Scaling to smaller linewidths with this strategy is therefore not an option at all.

10.3.1 Lithography over steps

Viscous flow of photoresist over steps leads inevitably to uneven resist thickness, and linewidth change at step edges (Figure 10.5). Because spin-coating results in variable resist thickness over steps, linewidth will be dependent on the underlying steps via resist thickness changes.

On non-planar surfaces, the effect of structures from previous steps causes some problems. Reflections from

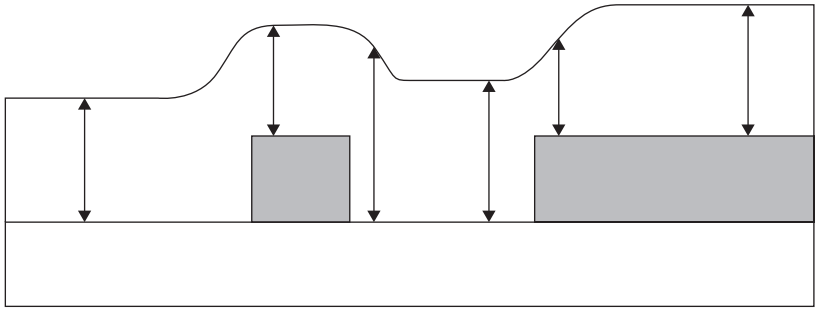


Figure 10.5 Resist thickness variation over topographic features

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

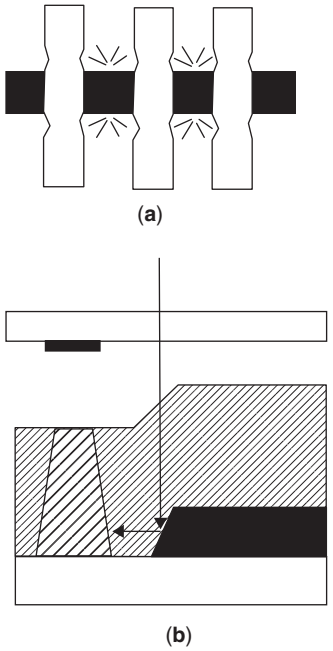


Figure 10.6 Reflective notching. (**a**) Top view of distorted resist lines and (**b**) cross-sectional view shows how the underlying metal line reflects incoming light into resist sidewall

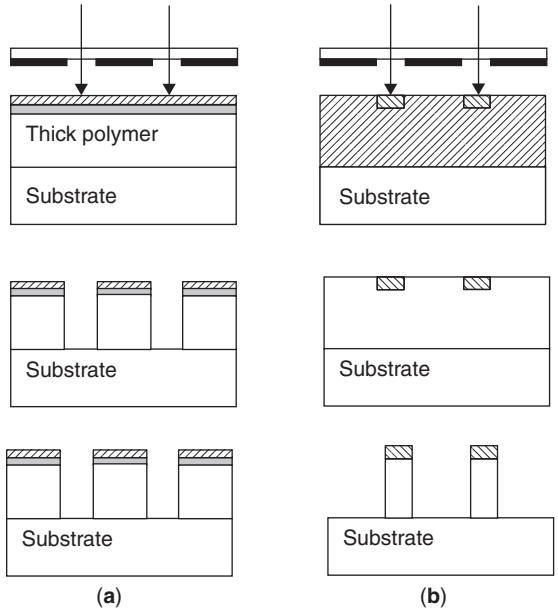


Figure 10.7 Multilayer resist and top-surface imaging. (**a**) Tri-layer resist process: exposure of thin top resist; etching of thin hard mask; etching of thick resist and (**b**) top-surface imaging process: exposure; silylation; plasma etching

underlying metal lines can cause resist exposure in unwanted places. This is called *reflective notching* (Figure 10.6).

10.4 EXTENDING OPTICAL LITHOGRAPHY

10.4.1 Top-surface imaging and multilayer resists

Top-surface imaging (TSI) and multilayer resists (MLR) offer true improvements in resolution, and therefore, device-packing density. Both bilayer and tri-layer resists have been tried. TSI and MLR rely on the fact that high resolution is easier to achieve in a thin imaging layer.

In MLR, a thick planarizing layer is applied first, followed by a hard mask layer of glass-like material (e.g., spin-on-glass). A very thin imaging layer is then applied (Figure 10.7). MLR eliminates focus depth effects if the planarizing resist works well. After developing the thin top imaging resist, plasma etching is used to pattern the hard mask, which then acts as a mask for dry development (oxygen plasma etching) of the thick planarizing layer.

Top-surface imaging uses a dyed resist for maximum absorption in the thin top layer. The exposed areas will be treated chemically: a silylation reaction takes place in the exposed regions, and a plasma-tolerant Si–O compound is formed. This Si–O compound acts as a hard mask for the dry development process, much like the deposited hard mask in the multilevel resist process.

Both MLR and TSI suffer from process complexity, and have not been practised as much as early estimates gave reason to believe. Performance of optical lithography has been improved by a multitude of evolutionary steps in lens design, thinner resists, improved process control and by adoption of planarization, which relieves depth-of-focus problems.

10.4.2 Resist trimming of light field structures

Because the price of optical lithography tools is increasing rapidly, there is a need for cheap alternative tools and/or methods. Two simple techniques for tweaking the optical lithography process for smaller dimensions are presented. Neither method can improve resolution but can be used to print narrow isolated lines and trenches.

Minimum resist line is first produced by optical lithography, and the isotropic plasma etching of

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

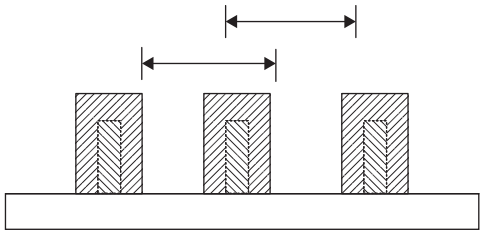


Figure 10.8 Resist trimming: resist lines made narrower by isotropic etching of the resist in oxygen plasma. Resolution (line + space) remains constant

photoresist is then performed (Figure 10.8). Resist line gets narrower and thinner. This method is most suitable when reasonably narrow lines can be used as starting point. Lines of 1.0 μm original width and thickness can be narrowed down to 0.2 μm; a 0.4 μm horizontal narrowing from both sides. Resist thickness after thinning is 0.6 μm because isotropic thinning was employed. This is a useful approach for studying simple structures, such as individual lines of scaled-down dimensions. Small MOSFETs of ca. 20 nm gate lengths have been made by resist trimming by using a 200 nm initial linewidth. But line plus space remains intact, and no more devices can be made to fit on a wafer.

10.4.3 Chemical shrink of dark field structures

The resist thinning method does not work for dark field patterns: any loss of linewidth will result in wider structures. A poor man's method of small DF structures is based on resist flow: resist will flow when heated above glass-transition temperature. This flow will, under favourable conditions, make holes and trenches smaller in a controlled fashion. This method has been successfully used in contact hole scaling studies.

A more advanced version for making narrow dark field patterns consists of patterning, overcoating, baking and rinsing (Figure 10.9). The overcoating material reacts with the resist during baking, and forms a non-soluble layer on the sidewalls of the contact hole, making the hole smaller (should there be photoresist residue at the bottom, it would block the contact hole). 0.25 μm contact holes have been reduced to 0.10 μm with this method.

10.5 LITHOGRAPHY SIMULATION

The lithographic pattern formation starts with the designer's layout file, which is turned into a physical mask plate in a mask shop. This mask is inserted into the exposure tool, where it modifies the illumination from the light source. After complex photochemistry steps in the photoresist, development creates patterns in the resist (Figure 10.10). This information flow has many points where errors can occur, and where dimensions are not accurately transferred. Some of these are data errors related to formats used in drawing and mask writing, and some are physical, and related to both mask writing and exposure resolution, and to etching tolerances.

It should be noted that the mask writing process has a similar information flow and similar error sources: the mask writer has finite resolution, the photoresist used in mask writing is similar to resists used in optical lithography, and chrome etching has its non-idealities just like any other etching process.

Lithography simulation is a self-contained speciality within simulation. It is partly physical simulation (optical modelling) and partly semiempirical simulation like etch simulation (development modelling).

Lithography simulators have three basic functions as shown in Figure 10.11. The first module is optical modelling, the second is photochemical, time-dependent, diffusion modelling and the third module is an etch simulator specifically developed for resists (Figure 10.11). Development of a novolak resist in an alkaline developer is an etching reaction, and it uses models similar to etching, but because its application field is very specific,

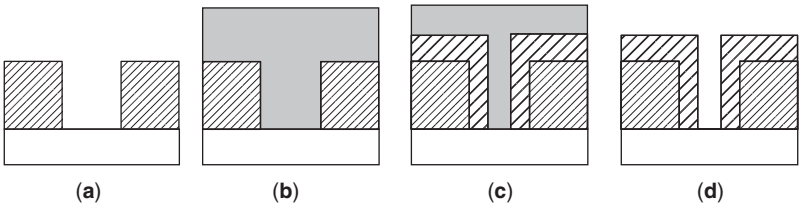


Figure 10.9 Chemical shrink technology for contact hole narrowing: (**a**) minimum contact hole exposed by optical lithography; (**b**) polymer deposition; (**c**) curing and (**d**) washing away the unreacted polymer. Redrawn from Ishibashi, T. *et al*. (2000), by permission of Institute of Pure and Applied Physics

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

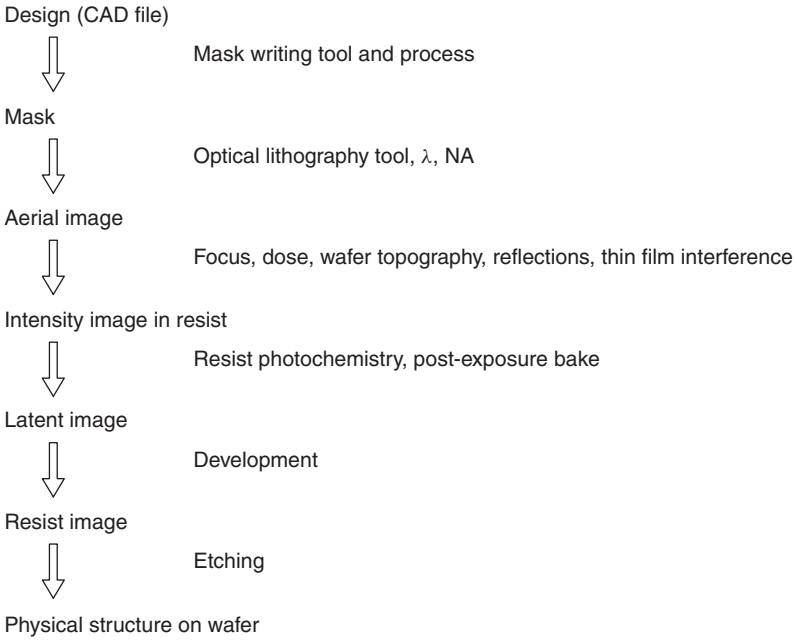


Figure 10.10 Lithography information flow. Adapted from Brunner, T. (1997), by permission of IEEE

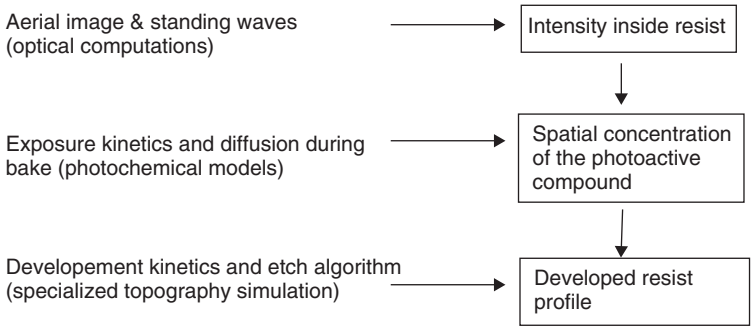


Figure 10.11 Modules of lithography simulation. Redrawn after Neureuther, A.R. & C.A. Mack (1997), by permission of SPIE

higher accuracy is possible. These steps have been modelled with good success even though an understanding of many basic mechanisms in resist exposure and development is yet to be uncovered.

SAMPLE 2D simulator contains optical lithography models. Lithography simulation input parameters include light source data like wavelength, exposure dose, numerical aperture and coherence; resist thickness and Dill parameters A, B and C; wafer and resist refractive indices and development rate parameters. SAMPLE can predict resist profiles with standing waves (Figure 10.12).

10.6 LITHOGRAPHY PRACTICE

After lithography, various processes are possible, and all of them exhibit rather different requirements for resists in terms of optimum thickness and profile, chemical stability, thermal and mechanical specifications, and so on (Figure 10.13). Resists face a serious scaling trade-off: thickness has to be scaled down for better resolution, but etch resistance and implant-blocking capability cannot be sacrificed; and thin resists are also more prone to pinholes. New resist chemistries based on aromatic and fluoropolymers are being developed. After

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

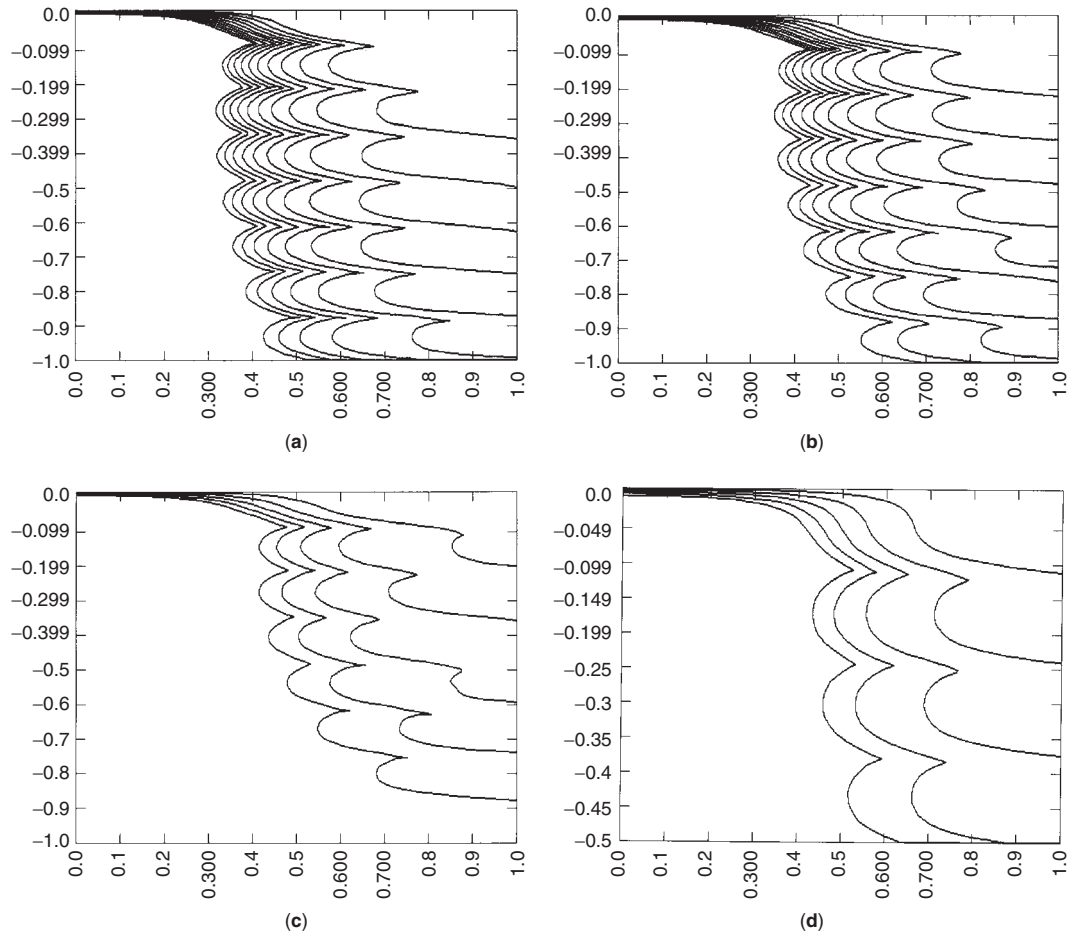


Figure 10.12 SAMPLE 2D simulation of resist exposure and development: nominal linewidth is 1.0 μm (only the right hand side is shown because the structure is symmetric). (**a**) exposure dose 100 mJ/cm^2, development time 65 s; (**b**) 80 mJ/cm^2 dose, 75 s development leads to sloped profile and (**c**) dose 70 mJ/cm^2, development 70 s, leads to incomplete development. In (**d**), conditions are identical to (**c**) but resist thickness is only 0.5 μm

etching, implantation or deposition, the resist has to be easily removed. This is obviously at odds with adhesion and stability.

Each of the steps following lithography has its special features and requirements:

Wet etching

- resist adhesion is important, resist may peel off;
- resist will not tolerate hot, strong acidic or alkaline etch solutions.

Plasma etching

- resist will be etched in plasma, its size and shape will change;
- resist will be damaged by plasma (both bombardment and thermal effects);
- removal of damaged resist is difficult.

Deposition

- plating solutions are often chemically aggressive.

Ion implantation

- resist thickness of 1 μm will stop B, P, As and Sb ions with <200 keV energy;
- beam current heats resist, cooling or current limitation are needed;
- resist carbonizes under heavy doses (>10^{15} cm^{-2}), difficult to remove.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

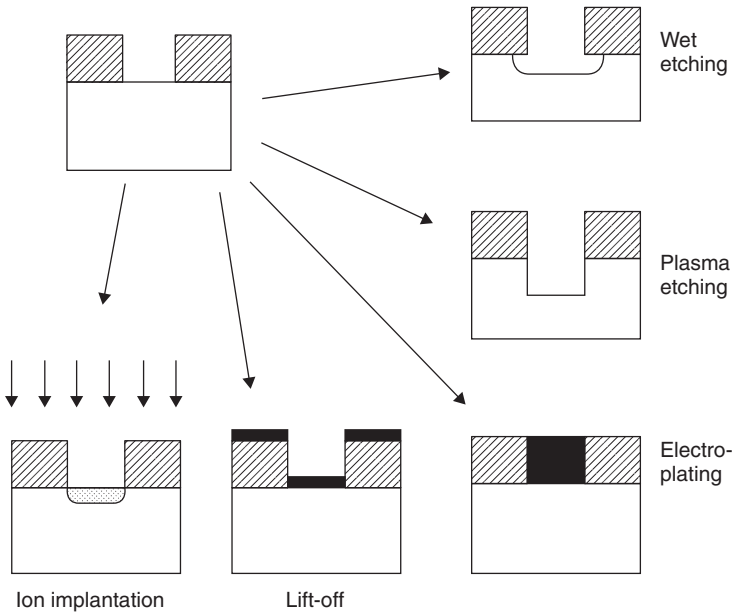


Figure 10.13 Processing after lithography puts varying demands on resists

Lift-off

- thickness of the film needs to be less than resist thickness;
- resist sidewall profile preferably retrograde;
- deposition process $T < 120\,^{\circ}\mathrm{C}$ because of resist thermal limitation.

10.7 PHOTORESIST STRIPPING/ASHING

After the photoresist has served its role as a protective layer, it must be removed. There are a number of methods to accomplish this (Table 10.1). The choice depends on the particular process step, the materials present on the wafer, resist nature and established laboratory practice (which may be determined by historical precedence, environmental concerns or other idiosyncratic factors). Oxygen plasma is a universal method, and the liquid phase methods are more or less specific to certain applications.

Sulphuric acid is a strong oxidant, and therefore an effective resist remover; however, it cannot be used if the wafer is metallized because the acid will etch metals too. Acetone is a fairly mild remover, and it cannot be used if the resist has been damaged or transformed by plasma or ion bombardment. Oxygen plasma alone will often suffice, but it is common practice to use two-step resist stripping: plasma (dry) removal followed by wet removal.

Table 10.1 Photoresist stripping

Techniques	Mechanism
Oxygen plasma	Oxidation in vacuum
Ozone discharge	Oxidation under atmospheric pressure
Acetone	Dissolution in liquid
Ozonized water	Bond breaking and dissolution
Sulphuric acid	Oxidation in liquid
Organic amines	Oxidation and dissolution in liquid
H_2O_2	Oxidation in liquid

The cost structure of photoresist stripping varies with the methods: in plasma or ozone ashing, equipment purchase cost is a major issue but oxygen bulk gas is cheap; in wet stripping (e.g., H_2SO_4) the cost of chemicals is important because large volumes are used (and disposed of). Some organic amine strippers are very expensive and can only be used for a few hours; the cost is dominated by material cost.

Ultrapure ozonized water, UPW-O_3, (*in situ* generation of 10–100 ppm ozone in DI-water) is potentially a major cost-reduction invention in stripping. Strip rates of 150 nm/min can be achieved, and utilization of ozone is very efficient even though the simple chemical reaction might suggest otherwise:

$$CH_2 + 3O_3 \longrightarrow CO_2 + H_2O + 3O_2 \qquad (10.5)$$

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

CH_2 can be used as a model molecule for photoresist. This calculation shows that 10.3 grams of ozone is needed to remove 1 gram of resist, for example, a batch of 25 wafers (200 mm) would need ca. 10 to 100 kg of ozonized water. But fortunately, much less is needed; ozone breaks up longer molecules, and the smaller molecules are water soluble.

10.8 EXERCISES

1. What fraction of resist ends up on the wafer in spin coating?
2. Estimate the contrasts of resists in Figure 10.3.
3. How much resolution can be gained by adopting TSI?
4. By how much will the swing ratio be reduced if a top antireflection coating can reduce air/resist reflections by 20%? By how much will the swing ratio be reduced if the absorbance increases from 0.5 to $1\,\mu m^{-1}$?
5. Calculate some good and bad resist thicknesses for novolak resist at 365 nm exposure.
6. What is the linewidth in Figure 10.4?
7. If a wafer with 350 μm thick resist is baked on a hot plate that is 0.1° off-horizontal, what will be the resist non-uniformity due to gravitational flow?

REFERENCES AND RELATED READINGS

Ausschnitt, C.P. *et al*: Advanced DUV photolithography in a pilot line environment, *IBM J. Res. Dev.*, **41** (1997), 21.

Bruce, J.A. *et al*: Characterization of linewidth variation for single- and multiple-layer resist systems, *IEEE TED*, **34** (1987), 2428.

Brunner, T.: Pushing the limits of lithography for IC production, IEDM 1997, p. 9.

Hartney, M.A. *et al*: Oxygen plasma etching for resist stripping and multilayer lithography, *J. Vac. Sci. Technol.*, **B7** (1989), 1.

Heschel, M. & S. Bouwstra: Conformal coating by photoresist of sharp corners of anisotropically etched through-holes in silicon, *Sensors Actuators* **A70** (1998), 75.

Holmes, S.J. *et al*: Manufacturing with DUV lithography, *IBM J. Res. Dev.* **41** (1997), 7.

Ishibashi, T. *et al*: Advanced microlithography process with chemical shrink technology, *Jpn. J. Appl. Phys.*, **40** (2000), 419.

Loechel, B.: Thick-layer resists for surface micromachining, *J. Micromech. Microeng.*, **10** (2000), 108.

Neureuther, A.R. & C.A. Mack: Optical lithography modeling, in P. Rai-Choudhury (ed.): *Handbook of Microlithography, Micromachining and Microfabrication*, SPIE.

Peterson, B. *et al*: Approaches ro reducing edge roughness and substrate poisoning of ESCAP photoresists, *Semicond. Fabtech.*, **8** (1996), 183.

Rai-Choudhury, P.: (ed.): *Handbook of Microlithography, Micromachining and Microfabrication*, Vol. 1, SPIE 1997.

Satou, I. *et al*: Progress in top surface imaging process, *Jpn. J. Appl. Phys.*, **39** (2000), 6966–6971.

Usujima, A. *et al*: Generation mechanism of photoresist residue after ashing, *J. Electrochem. Soc.*, **141** (1994), 2487.

IBM J. Res. Dev., **41**(1/2) (1997), special issue on optical lithography.

Conference series "Advances in Resist Technology and Processing" by SPIE is organized annually.

**Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030**

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

11

Etching

The pattern transfer process consists of two steps: lithographic resist patterning and the subsequent etching of the underlying material. The resist pattern can always be removed if found faulty on inspection, but once the pattern has been transferred on to solid material by etching, rework is much more difficult, and often impossible.

Etching is often divided into two classes, wet etching and plasma etching. Wet etching equipment consists of a heated quartz bath ($10 000), and plasma-etch equipment is a vacuum chamber with an RF-generator and a gas system (costing up to millions of dollars).

The basic reactions in etching are as follows:

Wet etching

$$\text{solid} + \text{liquid etchant} \longrightarrow \text{soluble products}$$

$$\text{Si (s)} + 2\text{OH}^- + 2\text{H}_2\text{O} \longrightarrow \text{Si(OH)}_2(\text{O}^-)_2\ \text{(aq)} + 2\text{H}_2\ \text{(g)} \qquad (11.1)$$

Plasma etching

$$\text{solid} + \text{gaseous etchant} \longrightarrow \text{volatile products}$$

$$\text{SiO}_2\ \text{(s)} + \text{CF}_4\ \text{(g)} \longrightarrow \text{SiF}_4\ \text{(g)} + \text{CO}_2\ \text{(g)} \qquad (11.2)$$

There are three steps that must take place for etching to proceed:

- transport of etchants to surface;
- surface reaction;
- removal of product species.

If etching does not take place, any of the three steps could be causing the problem: transport could be prevented or reduced by, for instance, a thick boundary layer; a native oxide or residues from the previous steps could retard or prevent etching; or the products may not be volatile or soluble enough, and they redeposit on the wafer. Gas bubbles formed according to Equation 11.1 can protect the surface from further etching.

Etch rates are typically 100 to 1000 nm/min, for both wet and plasma processes. The lower limit comes from manufacturing economics, and the upper limit from resist degradation, thermal runout and damage considerations. Silicon etching is exceptional: rates up to 20 μm/min are available in both wet etching ($HF:HNO_3$) and in plasma etching (DRIE) in SF_6/C_4F_8.

There are materials that cannot be wet etched, for example, SiC, GaN, TiC and diamond. These materials, can, however, be plasma etched. Some materials cannot be etched even by plasmas because no suitable source gas/volatile product combination exists. In that case, purely physical etching, known as *ion milling* or *ion beam etching* (IBE), can be used: argon ion bombardment will erode any material. Many solid-state laser garnets and magnetic materials (of the type $Gd_3Ga_5O_{12}$, gadolinium gallium garnet) are etched by ion milling. It is, however, difficult to find suitable non-eroding masking materials: if anything can be etched by argon bombardment, this applies to masking materials as well. Typical ion milling rates are 10–100 nm/min, an order of magnitude less than in plasma etching.

Note on terminology

The term *dry etching*, as opposed to wet etching, is often used as a synonym for plasma etching, but there are dry methods that do not involve plasma, for example XeF_2 gas etching. Plasma etching, in the older literature, can also mean a specific type of etch reactor, the parallel plate plasma reactor, in which the wafer is placed on the grounded electrode. The opposite of the plasma etcher is the RIE reactor (reactive ion etching), with the wafer on the powered electrode. Today, both plasma etching and RIE are used as general terms and not as reactor descriptions.

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

11.1 WET ETCHING

Wet etching mechanisms fall into two major categories:

metal etching:

$$\text{electron transfer} \quad \text{Me (s)} \longrightarrow \text{Me}^{n+}\ \text{(aq)} + \text{ne}^-$$

insulator etching:

$$\text{acid–base reaction} \quad \text{SiO}_2 + 6\text{HF} \longrightarrow \text{H}_2\text{SiOF}_6\ \text{(aq)} + 2\text{H}_2\text{O}$$

The rate limiting steps in etching are similar to those encountered in CVD (Chapter 5):

1. The surface reaction is slow, and it determines the rate.
2. The surface reaction is fast, and rate is determined by etchant availability (transport of reactant by diffusion and convection).

Surface reaction–limited processes exhibit activation energies of 30 to 90 kJ/mol. The rate increases with increasing etchant concentration and it is insensitive to stirring. Crystal planes can etch differently in surface reaction–limited etching. Aluminum etching in H_3PO_4 is surface reaction–limited: Al_2O_3 dissolution is the rate-determining step, with 54 kJ/mol activation energy.

Transport-controlled reactions are characterized by activation energies of 4 to 25 kJ/mol. Their rate increases with agitation and stirring because more reactant is being brought to the vicinity of the surface. Furthermore, all crystal planes etch at the same rate, which is natural because the reaction is not surface-limited. Silicon etching in a HF:HNO_3 mixture is limited by HF diffusion through the product layer. The activation energy is 17 kJ/mol.

11.1.1 Wet etching tools

Wet processing comes in three major variants: tank (bath), spray tool and single-wafer processor. The tank is, for example, a quartz vessel with heating and temperature control. It is filled with water and chemicals and the wafers are immersed in liquid for the required time, and then transferred to similar tanks for rinsing. Spray tools handle a cassette (or cassettes) but instead of immersion, liquid is sprayed from stationary nozzles on rotating wafer cassette(s). After the first spraying, the process continues with either another chemical or DI-water spray and nitrogen drying in the same vessel. Fresh mixing of chemicals and lower liquid volumes are spray tool advantages over tanks. Single-wafer tools are akin to photoresist spinners, and in a sense, they are spray tools too. However, processing acts on the wafer topside only.

The heating of wet process tanks uniformly is no easy task, because highly reactive and corrosive chemicals are used at high temperatures (e.g., 180 °C boiling nitric acid to etch nitride, or 120 °C peroxo sulphuric acid for cleaning, known as *Piranha*). The materials of the tanks and heaters must be compatible with the process: in chemical, thermal and mechanical respects. Teflon® and quartz are often used in the most demanding applications, but both are expensive materials and difficult to machine. Polypropylene is used for less critical applications, while stainless steel is the material for solvent tanks.

Temperature uniformity depends on stirring and convective heat transfer. This is not trivial because stirring can affect the etch process in other ways too: it can enhance reactant supply, reaction product removal or heat removal from an exothermic reaction.

Heating will result in higher etch rates, but there are practical limitations: resist (or other masking material)

Table 11.1 Wet etchants for photoresist masked etching

SiO_2	NH_4F:HF (7:1) BHF, 35 °C
SiO_2	NH_4F:CH_3COOH:$C_2H_6O_2$ (ethylene glycol):H_2O (14:32:4:50)
poly-Si	HF:HNO_3:H_2O (6:10:40)
Al	H_3PO_4:HNO_3:H_2O (80:4:16), water can be changed to acetic acid
Mo	H_3PO_4:HNO_3:H_2O (80:4:16)
W, TiW	H_2O_2:H_2O (1:1)
Cr	$Ce(NH_4)NO_3$: HNO_3:H_2O (1:1:1)
Cu	HNO_3:H_2O (1:1)
Ni	HNO_3:CH_3COOH:H_2SO_4 (5:5:2)
Ti	HF:H_2O_2
Au	KI:I_2:H_2O; KCN:H_2O

Table 11.2 Wet etchants for other applications

SiO_2, PSG	HF (49%) sacrificial layer removal (>1 μm/min)
SiO_2	DHF, dilute HF, usually 1%, for removing native oxide (ca. 10 nm/min)
<Si>	KOH (10–50%) anisotropic crystal plane-dependent etch
Nitride	H_3PO_4 boiling at 160 – 180 °C, CVD oxide mask
Si	HNO_3:HF:CH_3COOH various compositions, rate > 10 μm/min possible
Pt, Au	HNO_3:HCl (1:3) 'aqua regia'

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

may not tolerate higher temperatures, or the etch may evaporate. Changing concentration can either increase or decrease etch rate: silicon etch rate increases from 0 to 20% KOH concentration, and decreases for higher concentrations.

The oxide etch rate goes down linearly with decreasing HF concentration. However, the aluminium etch rate goes up when HF concentration decreases: 49% HF etches aluminium 38 nm/min, but HF:H_2O (1:10) results in 320 nm/min rate. This is because water has an active role in aluminium surface oxidation. Buffering agents and other additives can dramatically change etch rates, as shown in Table 11.3.

Wet etching is an indispensable tool in defect analysis: microstructural defects like stacking faults and pinholes can be made visible by wet etching. Sirtl, Secco, Wright, Dash and Sailor are etchants for delineating defects.

11.1.2 Etching profiles

The isotropic etching front proceeds as a spherical wave from all points open to the etchant (Figure 11.1). Because the etch profile is rounded, isotropic etching cannot be used to make fine features (Figure 11.2). Undercutting is similar to vertical etched depth. For a thin-film thickness of 500 nm, undercutting is also 500 nm, and etch bias, that is, the difference in etched feature size to mask size, is 1000 nm.

The isotropic profile is the most commonly encountered etch profile. Most wet etchants result in an isotropic profile, and it is also encountered in plasma and dry etching. Dry etching of silicon with XeF_2 gas, without plasma, results in isotropic profiles. Similarly, HF-vapour etching of oxide is isotropic dry etching. In plasma etching, the degree of isotropy can be controlled by the etching parameters, from fully isotropic to fully anisotropic (which may not be easy).

Undercutting can be compensated by making the initial mask feature larger than the desired width, for light field structures and vice versa for dark field structures. This approach works quite well for isolated structures, but in dense arrays its utility is compromised.

Wet etching profiles are seldom perfectly isotropic, and both deep slopes and gently sloping sidewall profiles are possible. The main parameters affecting the slope are the same as those governing the other main features of etching: etchant concentration and temperature. Silicon

Table 11.3 HF-based wet etch rates (nm/min) for selected materials at room temperature

Etchant	Material					
	SiO_2	TEOS	PSG	Si_3N_4	Al	Mo
HF (49%)	1763	3969	4778	15	38	0.15
NH_4F:HF (7:1) (BHF)	133	107	1024	1	3	0.5
HF:H_2O 1:10	48	157	922	1.5	320	0.15
NH_4F:HF:glycerine 4:1:2	89	186	1375	0.8	1	0.3

Source: Kim, B.-H. *et al.* (1999).

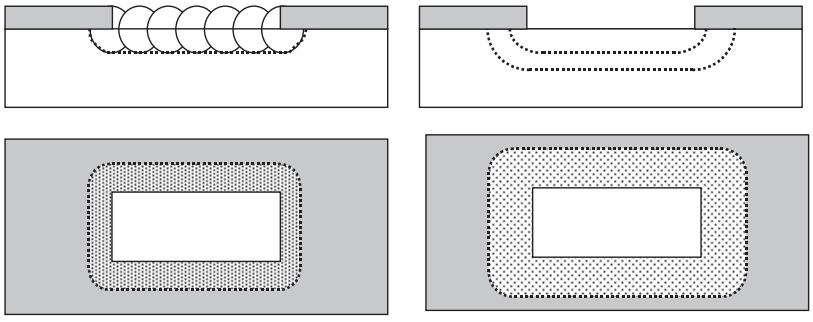


Figure 11.1 Cross-sectional and top views of isotropic (spherical wave front) etching at two stages of the process. Mask shown in gray; the dotted portion shows the mask that has been undercut

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

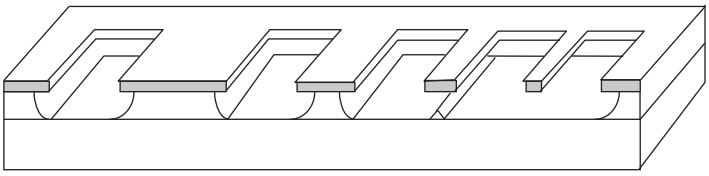


Figure 11.2 Undercutting in isotropic etching: wide lines are narrowed but narrow lines are completely undercut and released

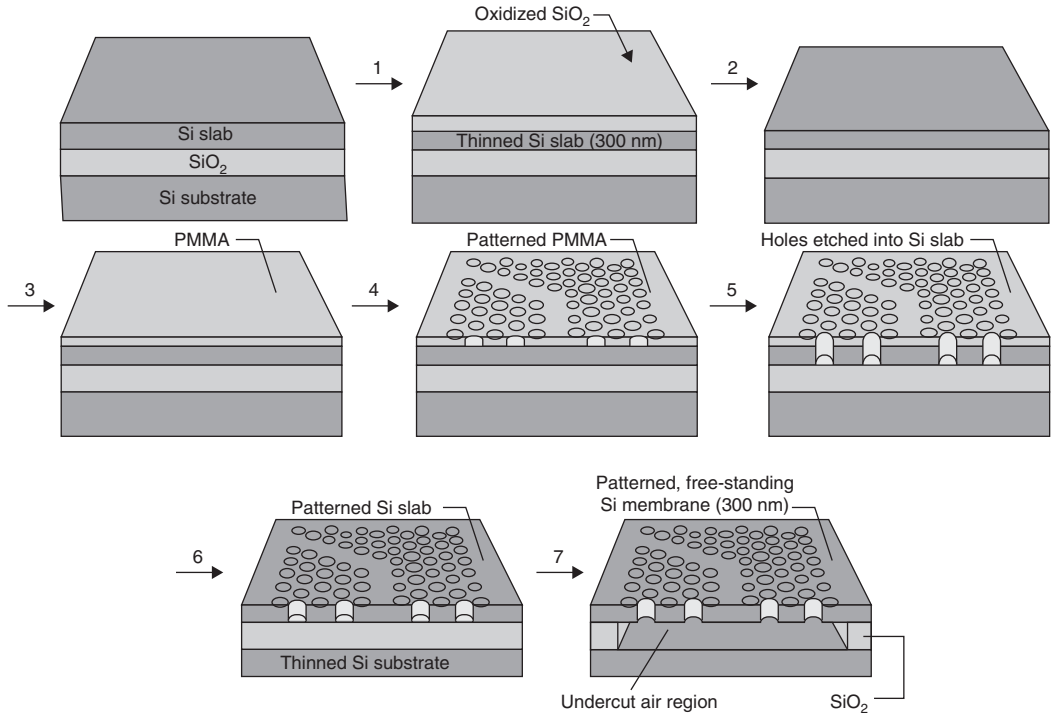


Figure 11.3 Photonic crystal fabrication on a SOI wafer: plasma etching defines release holes, and SiO_2 is isotropically etched under silicon membrane. Reproduced from Loncar, M. *et al*. (2000), by permission of American Inst of Physics

dioxide etching in buffered HF (BHF) can produce steep slopes at 7:1 NH_4F:HF ratio at 25 °C, but 30:1 ratio at 55 °C leads to a gentle slope. Gentle slopes may be desirable for step coverage in subsequent deposition steps. When multi-layer films are etched, profile control is even more difficult than with simple films. In the best case, a single etch step can etch both films.

Undercutting is sometimes desirable and even necessary. Free-standing structures, beams, cantilevers and membranes are made by releasing them by isotropic etching, as shown in Figure 11.3 for a photonic crystal. Free-standing structural layer fabrication demands isotropic undercut etching (wet or dry). The topic will be discussed in more detail in Chapter 22. In reverse engineering and failure analysis, thin films are removed selectively by isotropic etching (wet or dry) to reveal the wanted structures, layer by layer.

Wet etching processes are easy in theory but difficult in practice:

1. Reaction products may affect the etching reaction, for example, hydrogen evolves when silicon is etched by hydroxide (KOH, for instance), and this hydrogen can prevent the etchant from reaching the surface.
2. Etching reaction produces substances that catalyse the reaction, for example, NO in HF-HNO_3-based silicon etching or silicon in EDP (ethylene diamine pyrocathecol) etching of silicon.
3. Etching reaction is sensitive to stirring/convective mass and heat transfer.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

4. Etching reaction is exothermic and temperature rises during etching (for these reactions, stirring decreases the etch rate because it decreases temperature).
5. Evaporation leads to concentration changes during etching.

11.1.3 Etching with a hard mask

In wet etching the resist is usually not consumed by the etchant, and the gravest danger is adhesion loss. This is dependent on priming, feature size, resist thickness and the chemical character of the resist. Generally, thicker resists are mechanically more stable. Interface stability is important for the etched profile because the etchant can easily propagate along the film/resist interface.

Photoresists are materials that combine photoactivity and mechanical/thermal/chemical stability, and, obviously, photoactivity is the property that cannot be sacrificed. In order to find optimum materials as etch/plating/implant masks, the concept of *hard mask* has been devised. The mask material is etched with photoresist masking, the photoresist is then stripped and the etch/plating/implant process is performed using the hard mask only. The hard mask material can be optimized to suit the application, irrespective of the photoresist.

The wet etchant for Si_3N_4 is boiling concentrated phosphoric acid (H_3PO_4) at 180 °C. The photoresist cannot tolerate such etching conditions. Instead, oxide is used as an etch mask: CVD oxide is deposited on top of nitride, and the oxide is patterned by the photoresist and HF-etched. After resist stripping, the oxide acts as a mask for nitride etching (Figure 11.4).

When CF_4-plasma was found to etch nitride, people were willing to invest in plasma etching even though it was immature technology and not very production worthy, just because the alternative was definitely difficult.

In silicon etching in KOH, silicon dioxide or silicon nitride hard masks are standard materials. When glass wafers (or thick oxides) are etched, nickel, chromium, polysilicon and amorphous silicon are suitable masking materials for concentrated HF (49%). Silicon carbide (PECVD SiC), tantalum pentoxide (Ta_2O_5) and aluminium nitride (AlN) are excellent hard masks for many wet and dry etching processes. Aluminum nitride, however, is easily etched by alkaline solutions such as KOH or even dilute NaOH photoresist developer. This fact can sometimes make processing much faster and easier compared to other hard masks, which are very stable materials (which is why they were chosen in the first place).

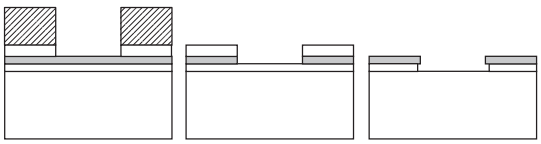


Figure 11.4 Wet etching an oxide/nitride stack: CVD oxide hard mask is etched by HF with resist mask; nitride in etched by H_3PO_4, and oxide (both bottom oxide and mask oxide) are etched by HF

11.2 ELECTROCHEMICAL ETCHING

Silicon is not etched in HF. If, however, silicon is made an anode in an electrochemical etching set-up, etch rates of ca. 1 μm/min are observed. Depending on current density, silicon can be etched in two rather different modes: pore formation and electropolishing. In pore formation, etching proceeds vertically downwards, leaving a silicon 'skeleton' with up to 80% empty space. Electropolishing resembles wet etching, in the sense that the whole surface is being etched.

The electrochemical etch set-up is shown in Figure 11.5. Hydrofluoric acid, with or without ethanol and/or water is used as an electrolyte. Platinum is the standard cathode. Both electropolishing and pore formation take place in the anodic regime.

The reactions that take place in HF-electrolyte are:

$$Si + 6HF \longrightarrow H_2SiF_6 + H_2 + 2H^+ + 2e^-$$

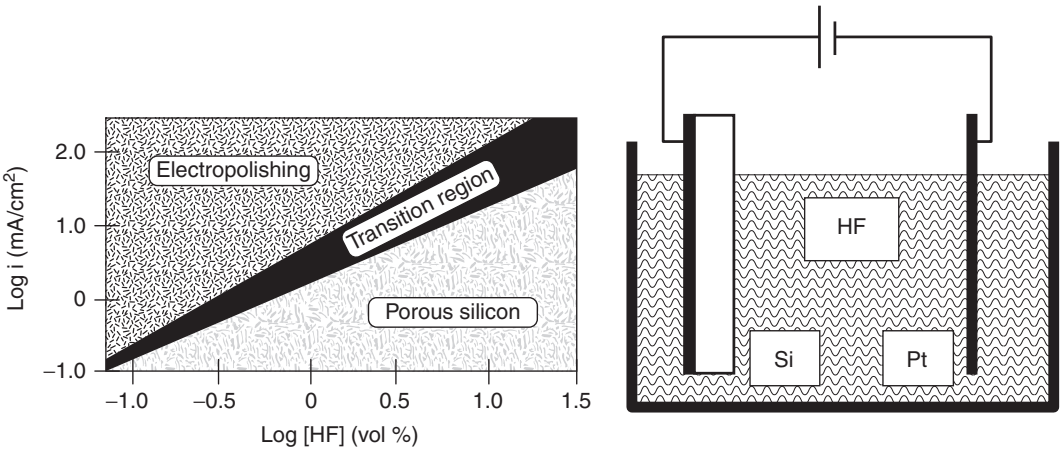
(pore formation at low current density)

$$Si + 6HF \longrightarrow H_2SiF_6 + 4H^+ + 4e^-$$

(electropolishing at high current density)

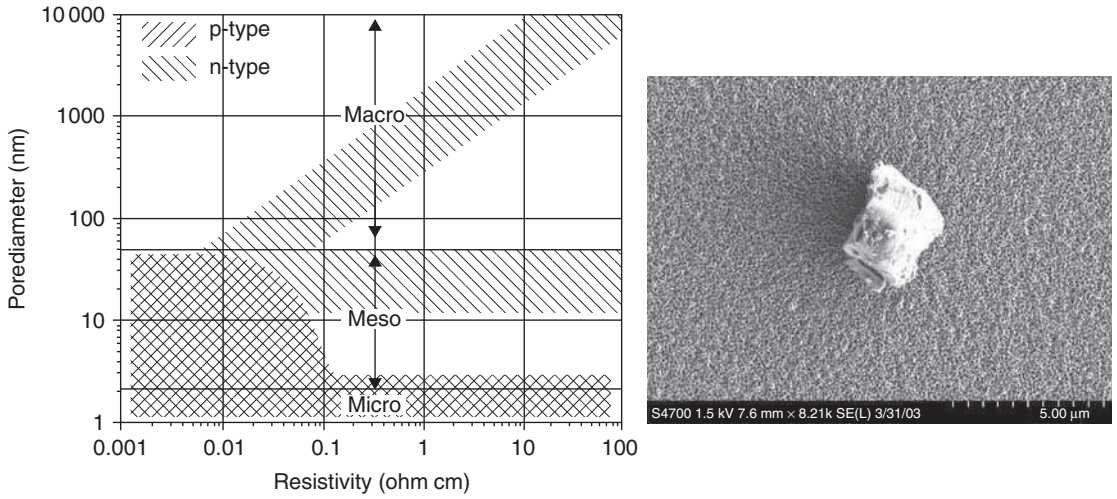
Pore formation starts at the wafer surface from a defect or an intentional initial pit. Electronic holes from the bulk silicon are transported to the surface, and they react at the defect or pit. Further etching occurs at the newly formed pore tips, because they attract more holes due to higher electric field strength, and the process leads to a uniform porous layer depth as the holes are consumed by the growing tips and other surfaces are depleted of holes. This etching mode takes place under low hole concentration and it is limited by hole diffusion, and not by mass transfer in the electrolyte cell. If hole density increases, some holes reach the surface and react there, leading to surface smoothing. This is the electropolishing regime, in which ionic transfer from the electrolyte plays a role.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030



(**a**) (**b**)

Figure 11.5 (**a**) Regimes of silicon anodic etching in HF: porous silicon formation and electropolishing. Reproduced from Collins, S.D. (1997), by permission of Electrochemical Society Inc; (**b**) Electrochemical etching set-up



(**a**) (**b**)

Figure 11.6 (**a**) Pore size ranges of electrochemically etched silicon: macroporous, mesoporous and microporous regimes. Reproduced from Lehmann, V. (1995), by permission of IEEE; (**b**) 50 nm pore size (with a micron particle). SEM micrograph courtesy Eero Haimi, Helsinki University of Technology

Illumination contributes to hole concentration in n-silicon (but not in p-type Si) and a very wide range of pore sizes from 0.2 to 20 μm can be etched by varying electrolyte concentration, current density and illumination (Figure 11.6). As a rule of thumb, pore diameter in micrometres is half the resistivity in ohm-cm: for 1 μm pores, 2 ohm-cm n-silicon is suitable. For small pores, low resistivity is needed; for large pores, high resistivity material has to be used. If pore formation starts from an unobstructed surface, a random pore array results. If initial pits are prepared by lithography and etching, pores can be arranged at will.

There are a couple of drawbacks in electrochemical etching (and deposition): electrical contact has to be made to the wafer backside, and this contact has to tolerate the etchant. Concentrated HF (49%) is often

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

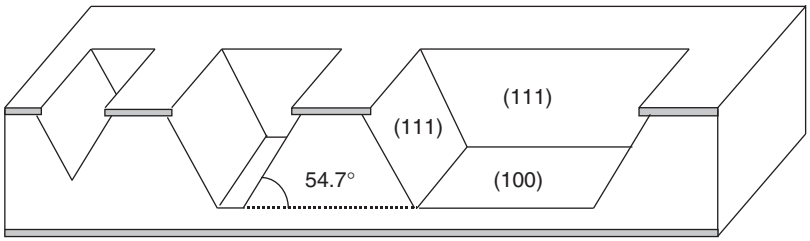


Figure 11.7 Anisotropic wet-etched profiles in <100> wafer. The sloped sidewalls are the slow-etching (111) planes; the horizontal planes are (100). Etching will terminate if the slow-etching (111) planes meet

employed, which seriously limits the choice of metals. Alternatively, a wafer holder can be used to protect the wafer backside, and any metal is good. However, such a holder takes up area on the wafer front, reducing the number of usable chips.

Porous silicon is single-crystalline silicon, even though it is a sponge-like network rather than true solid. Epitaxial deposition on porous silicon is possible, and other thin films can be deposited too. Depending on deposition process step coverage, pores will either be filled or buried by thin film material. Conformal CVD into macroporous grooves is no different from CVD into etched grooves of similar dimensions. Porous silicon presents a curious case in which etch selectivity can be obtained between silicon and silicon: porous silicon etching proceeds rapidly because the sidewalls between the pores can be as small as a few nanometres, whereas solid silicon is attacked from the top surface only. Etch rate ratio can be as high as 100 000:1. This selectivity, together with lithographic patterning and pore-size tailoring (by doping type and level), leads to interesting sacrificial layer techniques in which porous silicon is etched away underneath solid silicon. This will be dealt with in Chapter 22.

11.3 ANISOTROPIC WET ETCHING

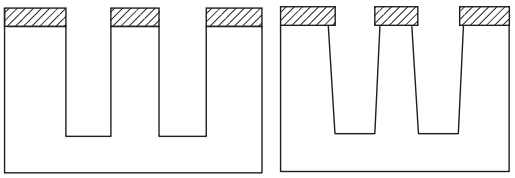
Isotropy, or homogeneity of space in all directions, is sometimes useful as we can neglect directions. Wet etching with its spherical-wave etch fronts is such a process. Anisotropic processes are spatially directional, but there are two completely different usages of the term anisotropic etching: anisotropic wet etching and anisotropic plasma etching.

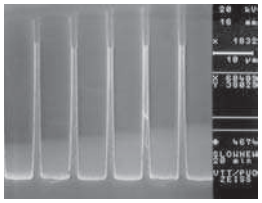
Potassium hydroxide, KOH, and tetramethyl ammonium hydroxide, TMAH, are the common anisotropic wet etchants for silicon. In KOH etching, the rates of different crystal planes can differ by a factor of 200. Silicon (100) crystal planes are fast etching, whereas (111) planes are slow etching. This results in structures bound by the (111) planes (Figure 11.7). The variety of shapes that can be made is astonishingly large, as will be seen in Chapters 21 and 28.

11.4 PLASMA ETCHING

Anisotropic plasma etching is synonymous with vertical or near vertical sidewalls. Anisotropy results from directional ion bombardment in the plasma reactor. Vertical walls and highly accurate reproduction of photoresist dimensions translate to closely spaced structures (Figure 11.8). High packing density of devices is possible by anisotropic plasma etching.

When etch bias becomes significant relative to linewidth, wet etching faces serious problems. In IC fabrication, this led to adoption of plasma etching at ca. 3 μm linewidths. With anisotropy, that is, vertical sidewalls, undercut compensation schemes became unnecessary, and all the resolving power of lithography





(c)

Figure 11.8 Plasma-etched anisotropic profiles (**a**) ideal vertical; (**b**) practical vertical with a slight undercut of the mask and sloped sidewall and (**c**) SEM micrograph of RIE profile

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

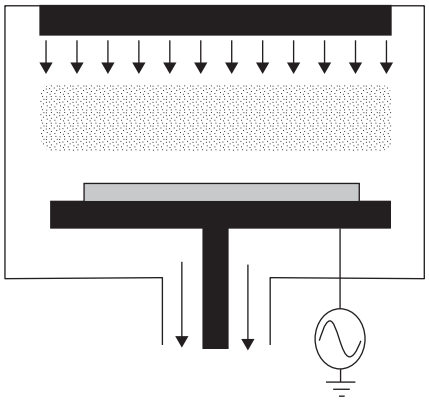


Figure 11.9 Plasma etching system (RIE, Reactive Ion Etcher): gases are introduced through the top electrode, wafers are on the powered bottom electrode

tools could be used to increase device-packing density. Plasma etching has been an indispensable tool since the early 1980s, and it has always been able to etch, with high precision, those structures that lithography has been able to print in photoresist.

Plasma etching is done in a vacuum chamber by reactive gases excited by RF-fields (Figure 11.9). Both the excited and ionized species are important for plasma etching. Excited molecules like CF_4^* are very reactive, and ionic species like CF_3^+ are accelerated by the RF field, and they impart energy directionally to the surface. Plasma etching is thus a combination of chemical (reactive) and physical (bombardment) processes.

11.4.1 Plasma etch chemistries

In a plasma discharge, a number of different mechanisms for gas-phase reactions are operative. Discharge generates both ions and excited neutrals, and both are important for etching.

Ionization	$e^- + Ar \longrightarrow Ar^+ + 2e^-$
Excitation	$e^- + O_2 \longrightarrow O_2^* + e^-$
Dissociation	$e^- + SF_6 \longrightarrow e^- + SF_5^* + F^*$

The most abundant species in the plasma reactor is the source gas. Etch reaction products are the next most abundant, and they may represent a few or 10% of all moieties. Excited neutrals may be present at a few percent, but ions are just a very minor component, 1 in 100 000. They are, however, often important for the mechanism.

Table 11.4 Typical etch gases

Fluorine	Chlorine	Bromine	Stabilizers	Scavengers/ others
CF_4	Cl_2	HBr	He	O_2
SF_6	BCl_3		Ar	
CHF_3	$SiCl_4$		N_2	
NF_3	$CHCl_3$			
C_2F_6				
C_4F_8				
XeF_2				

Plasma etching is based on reaction product volatility. Silicon is easily etched by halogens (Table 11.4): both fluorides (SiF_4), chlorides ($SiCl_4$), and bromides ($SiBr_4$) of silicon are volatile at room temperature, at millitorr pressures. No ion bombardment is needed for etching if the reactions are thermodynamically favoured and the role of ion bombardment is to induce directionality. Silicon nitride (Si_3N_4) is etched by fluorine, producing SiF_4 and NF_3. Aluminum is spontaneously etched by Cl_2, but the surface of aluminium is always protected by native aluminum oxide, and aluminium etching can only commence after this oxide has been removed. Ion bombardment is essential for native oxide removal.

11.4.2 Plasma etch mechanisms

Chemical bonds need to be broken for etching to take place. Bond energies, therefore, give indications of possible etching reactions (Table 11.5). Reactions that lead to bonds stronger than the Si–Si bond will etch silicon; and if the products have stronger bonds than Si–O, silicon dioxide will be etched. These simple predictions are experimentally confirmed: fluorine, chlorine and bromium will etch silicon because silicon–halogen bonds are stronger than silicon–silicon bonds. Only Si–F bond is stronger than Si–O bond and therefore only fluorine is predicted to etch oxide. However, because of ion bombardment, oxide is slightly etched in chlorine and bromine plasmas also, but to a much lesser extent than in fluorine plasmas.

In practice, the volatility of reaction products (i.e., high vapour pressure) is used as a criterion for etchant selection. Boiling points of reaction products

Table 11.5 Bond energies (kJ/mol)

C–O	1080	Si–F	550
Si–O	470	Si–Cl	403
Si–Si	227	Si–Br	370

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Table 11.6 Etch product boiling points (T_{bp}, °C)

SiF_4	−90	$SiCl_4$	−70	CO_2	−56
NF_3	−206	$AlCl_3$	190	PH_3	−133
WF_6	2.5	$GaCl_3$	78	AsH_3	−116
WOF_4	110	$TiCl_4$	−25		
TaF_5	96.8	$WOCl_4$	211	$SiBr_2$	5.4
MoF_6	17.5	WCl_6	275		
$MoOF_4$	98	$InCl_2$	235		
NbF_5	72	$MoCl_5$	194		
	$PtCl_4$	370d			
	PbCl4	−15			
	$Cr(CO)_6$	110d			

Note: d – decomposition

Table 11.7 Non-etchable reaction products (T_{bp}, °C)

$CuCl_2$	620	TiF_4	>400
CuF_2	950d	PbF_2	855
$CrCl_2$	824	CrF_2	1100
AlF_3	1290s	TiF_3	1200

Note: d – decomposition; s – sublimation

(Table 11.6 and 11.7) can be used to estimate volatility, but tabulated values of boiling points are usually for a pressure of 1 atm, not for reduced pressures. Reaction products like WOF_4 (from CF_4 and O_2 etching of tungsten) and $AlCl_3$ (Cl_2 etching of aluminium) have boiling points around 200 °C, and they are volatile enough for practical etching, but AlF_3 or CrF_2 have boiling points ca. 1000 °C and, therefore, fluorine is not a suitable etchant for these materials (Table 11.7). Ion bombardment enhances removal of material, and it can be used to drive reactions that might otherwise not be suitable for etching. Such reactions are, however, prone to residues.

Bombardment supplies energy to horizontal surfaces. These surfaces experience ion-induced desorption, ion-induced damage and ion-activated chemical reactions. Sometimes etchant gases (together with resist erosion products) form films on the sidewalls, and these films prevent etching laterally. Sidewalls do not experience ion bombardment, and, therefore, film formation and etching reactions are different from horizontal surfaces (Figure 11.10). Low-pressure operation usually favours anisotropy because bombardment is more directional, but it requires either a bigger pump or reduced flow rate, in which case the rate is lower (Figure 11.10).

Deep silicon etch processes (also known as DeepRIE, or DRIE) utilize both effects. In the Bosch process (named after the company that developed it), SF_6 and

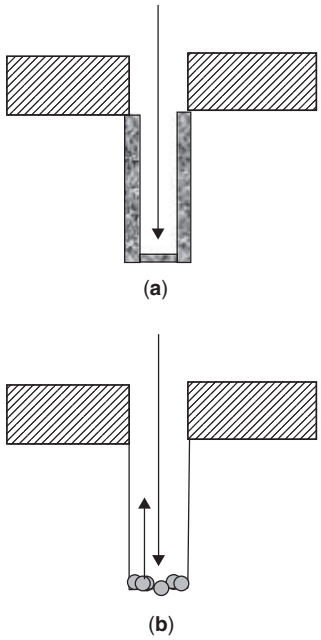


Figure 11.10 Mechanisms of anisotropy in plasma etching (**a**) sidewall passivation: ion bombardment preferentially removes passivation film from horizontal surfaces only and (**b**) suppression of spontaneous chemical reactions by cryogenic cooling; only ion-enhanced reactions can proceed

C_4F_8 gases are pulsed: a C_4F_8 pulse deposits a protecting polymer film all over the structure. SF_6 etching removes the polymer film from the trench bottom by ion-assisted etching, but the sidewalls do not experience ion bombardment, and they remain protected (but are slightly etched by the chemical component). The next pulse deposits a new protective film and then another SF_6 pulse is fed into the reactor. The pulsed operation leads to an undulating sidewall (see Figure 20.9), which introduces difficulties in some applications. In cryogenic deep etching, continuous SF_6/O_2 flow is used and etching proceeds vertically because lateral etching is suppressed by low temperature (−120 °C) and the $SiO_xC_yF_z$ residue film also protects the sidewalls.

Exact plasma etch mechanisms remain unknown in many cases. It has been shown that damaged single-crystal tungsten is etched much faster than the perfect crystal. Silicon etch rate has been shown to be synergistic with both ion bombardment and chemical components: etching with argon ion bombardment or with XeF_2 gas alone results in a very low etch rate, whereas simultaneous Ar^+ /XeF_2 process etches silicon 1 to 2 orders of magnitude faster.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

In plasma etch simulation, plasma physics provides ion and neutral energies, diffusion models are needed for fluxes of particles impinging on the surface, and then the surface reactions need to be understood. There can be competing reactions at every stage: SF_6 molecules are ionized in plasma, but F^- ions can react with oxygen in the plasma, which decreases active fluorine concentration; CHF_3 acts not only as a fluorine source, but also as a source of $(CF_2)_n$ polymer, which will deposit on the wafer. Simple model systems such as argon bombardment of fluorinated silicon surfaces have been simulated but predictive first principles plasma etch simulators remain to be developed.

11.5 CHARACTERIZATION OF ETCH PROCESSES

11.5.1 Linewidth and profile

Linewidth is also known as CD, for critical dimension, in the IC industry. Linewidth measurement checks deviation from design values. A deviation of 10% is acceptable for digital devices, but this error budget has to be divided between lithography and etching.

The sidewall profile of the finished feature has important implications for subsequent process steps: step coverage of the next deposition process depends on it. The profile can be measured with top view optical or SEM measurements, but destructive cross-sectional SEM pictures are considered the ultimate profiles.

Linewidth can be measured by scanning over the line either with a mechanical stylus or with a laser or electron beam. Line edges are seldom abrupt, and judgement must be used to locate the line edge properly. Real lines do not have perfectly vertical sidewalls, but sloped or even retrograde walls, with edge roughness that can be a significant fraction of the linewidth for narrow lines (Figure 11.11). Multiple scans must be made to average over edge roughness. Substrate and film roughness add noise to stylus measurements, and for soft materials, stylus penetration can be a problem. Linewidth can also be measured electrically, as was discussed in Chapter 2.

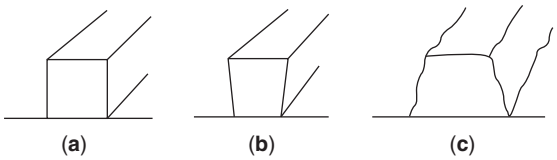


Figure 11.11 Line profiles (**a**) ideal vertical wall; (**b**) retrograde wall and (**c**) positively sloped wall with rough edge

11.5.2 Selectivity

Selectivity is a measure of etch rate ratios (ERR). Selectivity can be defined between film and substrate and between film and photoresist or other masking materials. Selectivities range from 1:1 to 100:1 in typical plasma etching processes. Resist selectivities range from 1:1 to 10:1 in plasma etching (with 100:1 possible). In wet etching, resist selectivity is often good, but resist adhesion loss and peel-off are severe limitations.

Etch stop is the term used for etching processes, in which the selectivity is so high that etching essentially stops when the underlying material is reached. This will be discussed more in the Chapter 21, because it has important applications in bulk micromechanics. When polymeric films are etched, selectivity and photoresist stripping are problematic: resist is polymeric material too and selectivity between two similar materials is difficult to achieve. PECVD oxide or nitride layers, can be used to cap polymer layers.

11.6 ETCH PROCESSES FOR COMMON MATERIALS

11.6.1 Silicon

Fluorine, chlorine and bromine processes are standard for silicon etching, resulting in reaction products SiF_4, $SiCl_4$ and $SiBr_4$, respectively. Fluorine processes are safer to use, but seldom fully anisotropic. Chlorine processes result in vertical sidewalls inherently, and the same applies to bromine processes. These two gases are, however, are highly toxic, and the equipment for Cl_2 or HBr etching must be equipped with a loadlock. Loadlocks complicate system operation but simultaneously improve repeatability since the reaction chamber is not exposed to room air and humidity.

SF_6- and CF_4-based processes have typically 10 to 40% oxygen added to them. Oxygen has several roles: it reacts with SF_n and CF_n fragments, and keeps fluorine concentration high by preventing fluorine recombination with the fragments. Oxygen etches resist, and contributes to sidewall film formation by oxidation and via its effect on resist consumption.

11.6.2 Silicon dioxide

Silicon dioxide etching is driven by ion bombardment. Isotropic plasma etching of oxide is, therefore, difficult, but high-enough radical concentration will result in reasonable isotropic etch rates. Any fluorine-containing gas can be used as an etchant for oxide, CF_4 or SF_6,

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

for example. However, both gases etch silicon too, and they are suitable for non-selective etching only.

CHF_3 is used as oxide etch gas when selectivity against silicon is required. It provides fluorine and carbon for etching (SiF_4, CO_2 etch products), and CF_2^* radicals, which are polymer precursors. Polymerization takes place on silicon surfaces, whereas on oxide surface $(CF_2)_n$ polymerization does not take place due to oxygen supply: ion bombardment–induced reactions on oxide result in CO_2 formation.

11.6.3 Silicon nitride

Nitride etching has aspects of both silicon and oxide etching. SF_6- and CF_4-based processes etch nitride fast, but isotropically and without selectivity against silicon. They are, however, selective against oxide with selectivities of ca. 2:1. CHF_3-based processes, on the other hand, etch nitride and provide selectivity against silicon. In fact, CHF_3-oxide etch processes usually perform well as nitride etch processes, and result in anisotropic profiles unlike SF_6- and CF_4-based processes.

11.6.4 Aluminum

Aluminum has native oxide, Al_2O_3, which is very difficult to etch. Chlorine (Cl_2) and chlorine-containing gases are used, with $AlCl_3$ as the main etch product. Multi-step etching is needed to etch aluminium: in the first 10 s, high power is used to sputter native Al_2O_3 away, power is then reduced to etch the bulk of aluminium. Aluminum is spontaneously etched in Cl_2, and a polymerizing agent is needed to passivate sidewalls for anisotropic profile; $CHCl_3$ and CH_4 are often used. In some low-pressure reactors, Cl_2/BCl_3 gases without polymer-forming gases will result in clean, anisotropic profiles. Nitrogen or argon is often added to stabilize the plasma and to improve photoresist selectivity.

11.6.5 Copper

Copper is not plasma-etched in current microfabrication processes. It is a difficult material to etch because neither fluorides (CuF_2), nor chlorides ($CuCl_2$), are volatile at room temperature. Increased temperature will help, but even at 100 to 200 °C, the rate is low and the photoresist is severely attacked. Organic etch gases have been tried with modest success. The first step is the oxidation of copper, followed by volatile compound formation. $Cu(hfac)_2$ (hfac – hexafluoroacetylacetonate) etching reaction proceeds according to

$$CuO + 2Hhfac \longrightarrow Cu(hfac)_2 + H_2O$$

The reaction products must be stable enough so that they can be transported away. Decomposition would result in redeposition residues and non-uniform etching.

If aluminium is alloyed with copper (to improve electromigration resistance), aluminium etching will be difficult for the same reason. Al-0.5%Cu is still fairly easy to etch but Al-4%Cu leaves residues of copper chlorides, which are difficult to remove.

11.6.6 Refractory metals and silicides

Tungsten etching is similar to silicon in many respects. In fluorine plasmas, the reaction product is WF_6; in oxygen–halogen plasmas, it is WOF_4 or $WOCl_4$. Tungsten hexafluoride has a boiling point of 17 °C and isotropic etching profile easily results. Oxyfluorides and oxychlorides are less volatile and ion bombardment is needed to remove them completely, which translates to better anisotropy. Molybdenum, too, is etched by both chlorine and fluorine plasmas, with or without oxygen. For titanium etching, chlorine etching is preferred, but fluorine etching is possible; and for TiW (30 at % Ti), SF_6 is a typical choice. Tantalum and niobium are etched similarly. Silicides WSi_2, $MoSi_2$ and $TaSi_2$ are etched in processes that resemble silicon and/or respective metal etching.

11.7 ETCH TIME AND SPACERS

Etch time seems like a simple concept: film thickness divided by etch rate. A slight overetch is required because there are uncertainties in both etch rate and in film thickness, which typically vary by, say 5%. However, when the films to be etched run over topography, the situation changes dramatically.

If film deposition is conformal, film thickness at the edge of a step will be the sum of the film thickness and step height. If anisotropic etching is stopped at the end point calculated from planar film thickness, a residue equal to original step height remains at the edge (Figure 11.12).

Long overetch will eventually remove this residue but this makes high demands on etch selectivity between the two materials. Sometimes it is desirable to leave this residue in place, and utilize it in the fabrication process. It is then termed *spacer*. Spacers have various applications, which will be discussed in

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

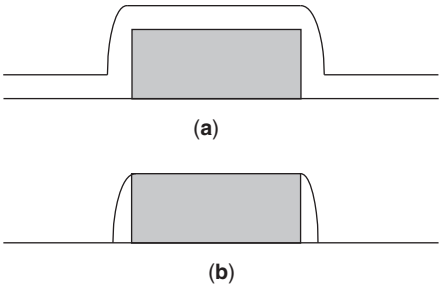


Figure 11.12 Spacer formation (**a**) conformal deposition over a step and (**b**) anisotropic etching to end point. For complete removal of top film, thickness to be etched is the sum of step height and top film thicknesses

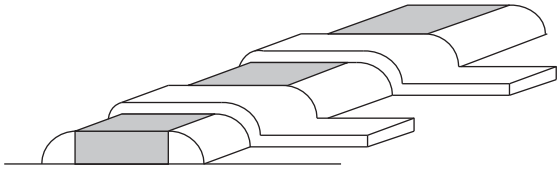


Figure 11.13 Etching to end point leaves spacers, which, if conductive, short neighbouring lines. If spacers are dielectric, they can form a permanent part of the device

Chapters 19, 25 and 26. Note that it is essential for spacer formation that etching is anisotropic; in isotropic etching, sideways etching would remove the material at the step edge.

If the bottom film is a conductor and top film is a dielectric, the spacer can be left in place. However, if the bottom film is a dielectric and the top film is conductive, then all the conductor lines etched in the top film will be electrically connected with each other through the conductive spacer at step edge (Figure 11.13).

11.8 COMPARISON OF WET ETCHING, ANISOTROPIC WET ETCHING AND PLASMA ETCHING

In many applications, the choice of wet versus plasma etching is a question of convenience: certain equipment or etch bath is available or some suitable masking material is handy. When sloped etch profiles are required, or when undercutting is needed, isotropic etching must be used. Isotropic wet etching of silicon can be done at fairly high rates – microns per minute or even tens of microns per minute. Through-wafer etching is done either by anisotropic wet etching or by DRIE. The ink jet example of Figure 11.14 shows how different etch techniques are utilized in one device: manifold etching is done by TMAH anisotropic wet etching,

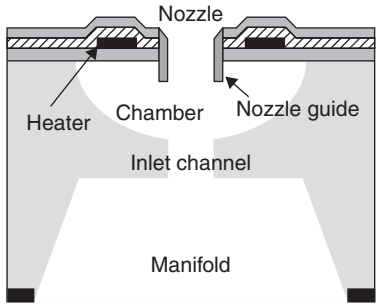


Figure 11.14 Ink jet etching features: isotropically wet etched chamber, DRIE inlet channel, anisotropic TMAH manifold etch, anisotropic nozzle guide (spacer) etch. Reproduced from Shin, S.J. *et al*. (2003), by permission of IEEE

critical inlet channel is defined by DRIE, chamber geometry is made hemispherical by isotropic wet etching and anisotropic plasma etching is needed in making the nozzle guides, which are similar to spacers from the fabrication point of view.

11.9 EXERCISES

1. What would you use as plasma etch gases and etch masks for etching the following materials:
   - diamond
   - SiC
   - GaN
   - GaAs
   - $PbZrTiO_3$
   - BCB (benzocyclo butadiene polymer)?
2. Polysilicon etched depth in chlorine plasma is given in the table below. Determine the etch rate.

Time (s)	Depth (nm)
20	50
40	185
60	325
80	455

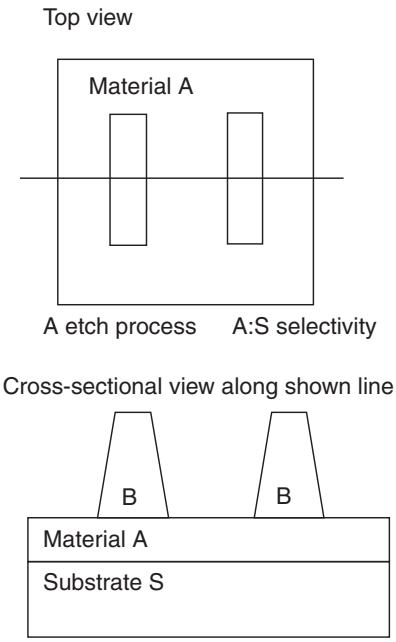
3. What is the activation energy of the etching of <100> silicon in 20% TMAH?

Temperature (°C)	Rate (μm/hr)
60	29
70	36
80	62
90	87

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

4. How much underlying oxide is lost when a tungsten film of 500 nm thickness is etched from a sample that has 300 nm steps on it? Tungsten: oxide selectivity is 10:1.
5. Etch rate could basically be measured easily by weighing the sample before and after etching, and translating that into the rate by taking the area into account. What resolution scale is needed to determine rates for:
   - tungsten etching, 500 nm thickness
   - silicon etching, 20 nm thickness.

   Densities: W – 19.5 g/cm^3, Si – 2.65 g/cm^3
6. How can the porosity of porous silicon be measured by weighing?
7. What is the resistivity of the p-type wafer shown in Figure 11.6(b)?
8. Draw cross-sectional figures of the shown structure under the following etch conditions, for two etch times: right at etch end point; and after 50% overetch.



Profile	A:S selectivity
anisotropic	∞
anisotropic	5:1
anisotropic	1:1
isotropic	∞
isotropic	5:1
isotropic	1:1

9. How much dimensional error does chromium wet etching introduce to (a) 1X photomasks and (b) 5X reticles?

REFERENCES AND RELATED READINGS

Bell, F.H. & O. Joubert: Polysilicon gate etching in high density plasmas, *J. Vac. Sci. Technol.*, **B14** (1996), 3473.

Bien, D.C.S. *et al*: Characterization of masking materials for deep glass etching, *J. Micromech. Microeng.*, **13** (2003), S34.

Collins, S.D.: Etch stop techniques for micromachining, *J. Electrochem. Soc.*, **144** (1997), 2242.

Hsiao, R.: Fabrication of magnetic recording heads and dry etching of head materials, *IBM J. Res. Dev.*, **43** (1999), 89.

Kim, B.-H. *et al*: MEMS fabrication of high aspect ratio track-following microactuator for hard disk drive using silicon on insulator, *Proc. IEEE MEMS '99*, (1999), 53.

Lehmann, V.: Porous silicon – a new material for MEMS, *Proc. IEEE MEMS* (1995), p. 1.

Loncar, M. *et al*: Waveguiding in planar photonic crystals, *Appl. Phys. Lett.*, **77** (2000), 1937.

Moreau, W.: *Semiconductor Microlithography*, Plenum Press, 1988.

Oehrlein, G.S. & J.F. Rembetski: Plasma-based dry etching techniques in the silicon integrated circuit technology, *IBM J. Res. Dev.*, **36** (1992), 140.

Schroder, D.K.: *Semiconductor Material and Device Characterization*, 2nd ed., John Wiley & Sons, (1998), pp. 582–584 defect etching.

Shin, S.J. *et al*: Firing frequency improvement of back shooting ink-jet printhead by thermal management, *Transducers'03* (2003), p. 380.

Walker, P. & W.H. Tarn: (eds.): *Handbook of Metal Etchants*, CRC Press, 1991.

Williams, K.R. & R.S. Muller: Etch rates for micromachining processes – Part I, *J. MEMS*, **5** (1996), 256–269.

Williams, K.R., Gupta, K. & M. Wasilik: Etch rates for micromachining processing – Part II, *J. MEMS.*, **12** (2003), 761.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

12

Wafer Cleaning and Surface Preparation

Microfabrication takes place under highly controlled conditions: all materials for cleanroom construction, processing equipment and wafer-handling tools are carefully selected to minimize particle, molecular or ionic contamination. Water, gases and chemicals are purified of contaminants and filtered of particles. These are, however, passive precautions, and active wafer cleaning must be undertaken before practically every major process step. Wafer-cleaning steps can account for up to 30% of all process steps.

Wafer cleaning is about contamination control, but it is also about leaving the surface in a known and controlled condition. This means damage removal, surface termination (hydrophobicity/hydrophilicity control) and prevention of unwanted adsorption. Therefore, many people prefer to call this activity *surface preparation*.

The main sources of contamination are the fabrication processes themselves. Air cleanliness in an advanced cleanroom is so good that airborne particles are not the main contamination source anymore, but airborne gaseous contaminants need careful attention. The human contribution has also been reduced significantly with correct gowning and working procedures or by factory automation. These matters are dealt in more detail in Chapter 35.

The purity of starting materials is important: liquid chemicals for advanced IC processes come with 1 or 0.1 ppb (parts per billion) impurity specifications. Sputtering target purities are, for example, 99.999%. Similar '5Nine' purities are typical for many process gases, but some applications need 99.99999% (7N) purity. Water purity is measured by resistivity: typical requirement is 18 Mohm-cm. This de-ionized water (DIW) is also known as UPW, for ultra pure water.

Because of device-size downscaling, contamination becomes even more critical. Finer patterns demand control of finer particles, and ultra-thin gate oxides necessitate low metal contamination levels for good integrity (low interface trap density, low oxide charge, and small leakage current).

12.1 CONTAMINATION FORMS

Contamination comes in various forms, which have different sources, effects on device and cleaning methods. The main classes of contamination are

- particles
- metals
- organics
- volatile inorganic contamination
- native oxide
- microroughness.

Particle-size monitoring is becoming a problem in advanced integrated circuits; in 130 nm processes, particles greater than 65 nm are monitored. A few decades ago, particles of the size 1/10 of minimum linewidth could be detected (with reasonable throughput), and more recently, particle detection at one-third of minimum linewidth was the norm. As scaling continues, it may be that monitored particle size will be identical to minimum linewidth. Particles are also a major concern in wafer bonding (Chapter 17), irrespective of linewidth.

Metal contamination cannot be avoided as long as machine parts are made of metals; so, metal contamination has to be controlled by cleaning. Metal contamination on the surface can spread into the silicon bulk, and dissolved metals and metal precipitates in the bulk act as recombination centres for charge carriers. Precipitates at silicon/oxide interface or in the critical areas of the device are detrimental because they affect diffusion profiles via their effect on crystal defects. If metals segregate into the oxide during oxidation, they can prevent, retard or degrade oxide film growth, and result in poor-quality oxides.

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Organics can cause increased contact resistance or abnormal film growth. This often comes through their prevention of the cleaning process. When wafers are ramped to high-temperature processes in an oxygen-containing atmosphere (e.g., 1% O_2 in N_2), organic contamination will usually be volatilized, but ramping in an inert atmosphere (N_2 or Ar) can cause carbon inclusions in the growing films or silicon carbide formation.

A model molecule for surface organics is trimethyl siloxane TMS, which is the reaction product of priming agent HDMS. The by-product of TMS decomposition is ammonia, which can contaminate chemically amplified DUV resists.

$$2\text{Si–OH} + (\text{CH}_3)_3\text{Si–NH–Si}(\text{CH}_3)_3 \longrightarrow 2\text{Si–O–Si}(\text{CH}_3)_3 + \text{NH}_3$$

Native oxide films grow readily on silicon. Growth is not instantaneous, however, and proper surface finishing can protect the surfaces for extended periods of time. Hydrofluoric acid cleaning ('HF-last') leaves the surface hydrophobic with H-termination (Figure 12.1). In normal cleanroom air, 42% RH and 1.2% H_2O concentration, a 0.5 nm native oxide film will grow in a few hours, but in dry air, native oxide formation is greatly reduced. Native oxide formation depends on the wafer type too: <111> wafers and heavily doped wafers oxidize faster.

Native oxides degrade contacts, cause crystallinity defects in epitaxial growth, prevent solid-state reactions and contribute to gate oxide integrity degradation because native oxide film quality is not uniform like that of thermally grown or CVD oxides. HF-last cleaning step is typical for silicon epitaxy – dilute HF (1:100) is used to remove oxide just prior to epitaxy.

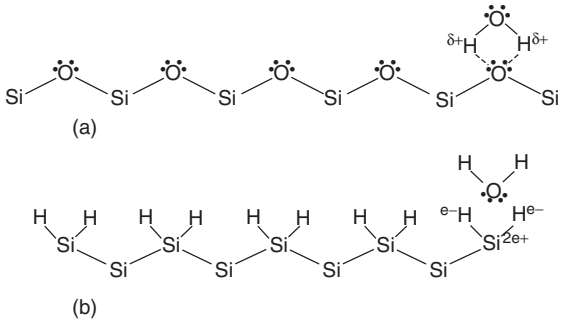


Figure 12.1 Silicon surface after cleaning: (**a**) hydrophilic surface after ammonia peroxide cleaning attracts water and (**b**) hydrophobic surface after HF cleaning repels water. Source: T. Hattori (ed.) (1998)

Measurement of native oxides can be done by spectroscopic ellipsometry, but not without difficulties. The optical constants of nanometre films are not identical to thicker films and they need to be calibrated against other methods. XPS signal strengths (Si–Si bonds and Si–O bonds give signals at slightly different energies) can be used.

Contact angle is used to characterize surface hydrophilicity/hydrophobicity. Hydrophilic surfaces have small contact angles, and water spreads evenly on hydrophilic wafers (Figure 12.2). Ammonia peroxide cleaning is the standard procedure for making hydrophilic surface finish. On hydrophobic surfaces, water forms distinct droplets. HF-last cleaning results in hydrophobic surfaces (contact angle $>90^\circ$). Water sometimes remains on the wafer after rinsing, resulting in watermarks during drying. These can be minimized by tailoring the contact angle to either high or low values. Superhydrophobic surfaces, with contact angles $>150^\circ$ can be made by deposition of fluoropolymers like Teflon®.

Microroughness can be classified as contamination because it has effects similar to other sources of contamination. Wafers come from manufacturers with 0.1 nm RMS surface roughness. Many of the cleaning processes rely on etching mechanisms and lead to increased surface roughness. Cleaning solution composition and time have to be optimized with respect to both cleaning

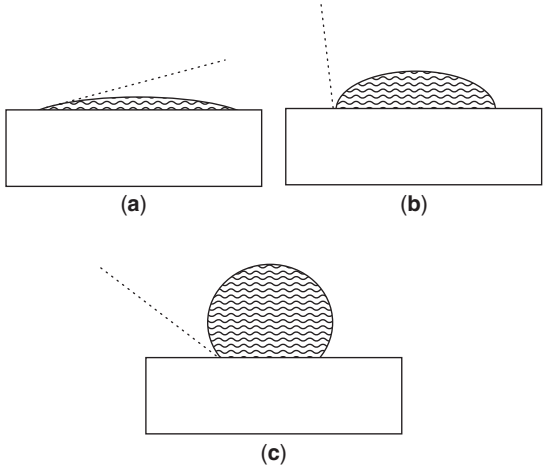


Figure 12.2 Contact angles of water droplets on wafer: (**a**) hydrophilic surface after ammonia-peroxide cleaning, 20°; (**b**) hydrophobic surface after HF cleaning, ca. 95° and (**c**) superhydrophobic surface, 150°. (Copyright Springer)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

efficiency and roughness increase. Decomposition of cleaning solutions and impurities can also catalyse surface reactions leading to increased roughness.

12.2 WET CLEANING

Acid, base and solvent wet cleanings are the main methods of cleaning. Dry cleaning by, for example, vapours and plasmas offers some advantages that will be discussed in Chapter 34. Wet cleaning is simple, it has high throughput and it cleans both the front and the back of the wafer simultaneously (see Figure 12.3). Wet benches are reliable tools, but chemical consumption can be high. There are two main approaches: either using rather concentrated chemicals for cleaning many batches before changing the chemicals or using dilute chemicals and changing them after each and every batch.

From the end of the 1960s till the early 1990s, wet cleaning relied on a few proven methods, which were, however, never studied in detail, and whose working mechanisms were unknown. In the 1990s, a vast amount of work was done in uncovering the mechanisms of contamination and contamination removal.

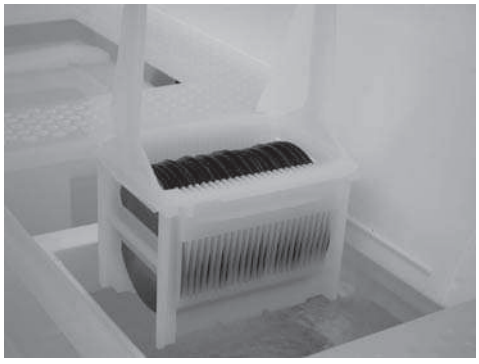


Figure 12.3 A wafer cassette with 25 wafers of 100 mm diameter is being lowered into a cleaning bath. Photo courtesy Paula Heikkilä, Helsinki University of Technology

The standard clean, known as the *RCA-clean* (invented at RCA Laboratories), consists of a sequence of different wet cleans. They are each effective in

Table 12.1 Wet-cleaning solutions: typical compositions and conditions

Name/alias	Chemical composition	Temperature/time
RCA-1 SC-1, standard clean; aka APM; ammonia peroxide mixture	NH_4OH:H_2O_2:H_2O (1:1:5)	50–80 °C, 10–20 min
RCA-2 SC-2; standard clean-2; aka HPM, hydrogen chloride-peroxide mixture	HCl:H_2O_2:H_2O (1:1:6)	50–80 °C, 10–20 min
SPM Sulphuric peroxide mixture, aka Piranha	H_2SO_4:H_2O_2 (4:1)	120 °C, 10–20 min
DHF (dilute HF) Standard chemicals come in the following concentrations:	HF:H_2O (1:20 – 500)	Room temperature, 1 min
HCl	37%	
H_2SO_4	96%	
H_2O_2	30%	
NH_4OH	29%	
HF	49%	

Bath life: If the bath is used for more than one batch before changing, chemical concentration is monitored, and, for example, ammonia evaporation or peroxide decomposition can be compensated by 'spiking', that is, refreshing the bath with an injection of fresh chemicals.
Disposal: HF requires a separate disposal system because its health effects are different from other mineral acids, which may all be collected in the same container. Sometimes, acids that contain heavy metals must be collected separately (e.g., titanium or cobalt containing salicide etchants).

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

removing different types of contamination. Table 12.1 lists the main wet-cleaning solutions commonly in use. Cleaning is always closely connected with both preceding and following process steps, and therefore cleaning strategies in different labs and wafer fabs can be very different in respect to cleaning bath chemistry, bath sequence, concentration, time and temperature. For instance, instead of the standard ammonia peroxide clean in 1:1:5 NH_4OH:H_2O_2:H_2O ratios, some users prefer 1:4:100, and even though all users do employ the ammonia peroxide step in pre-oxidation cleaning, additional HCl:H_2O_2, HF and H_2SO_4:H_2O_2 cleans are combined in variegated ways.

Chemical consumption in wet benches is a major environmental concern. With larger wafer sizes, larger tanks have to be used, with increasing volumes of expensive high-purity liquids, which are dangerous to handle, and which have to be disposed under controlled conditions. Full fabrication process of a 200 mm IC wafer consumes a cubic metre of ultrapure water, and tens of kilograms of liquid chemicals are required. Hundreds of litres of acid waste are produced. Rinse water can be recycled, and acid recovery and reuse are also common practices.

12.3 PARTICLE CONTAMINATION

Particle contamination is dangerous in lithography, but lithography is rather insensitive to metal ion contamination. Deposition processes are sensitive to small particles that can 'grow' in size during conformal deposition such as CVD when the film encapsulates the particle. This may eliminate the particle as an electrical contaminant, but lithography- and topography-forming steps will be aware of it.

Table 12.2 Sources of particles

– Chemical reactions in deposition and etching
– Moving parts in tools: robot arms, valves, doors
– Static parts: wafer holders, cassettes, o-rings
– Vacuum: pumping, venting, condensation
– Gases, chemicals, water

Fabrication processes themselves are major sources of particles. Listed in Table 12.2 are some materials and mechanisms that contribute to particle contamination.

In liquid, both the wafer surface and the particles acquire surface charge. These charges lead to either attractive or repulsive forces between particles and surfaces. Surface charge is characterized by zeta potential. It is independent of particle size but it depends on the electrolyte pH: in acidic conditions (low pH) the zeta potential is positive, and in alkaline solution it tends to be negative, as shown in Figure 12.4. Like charges repel each other and opposite charges attract each other. Acidic cleans, such as HF, which result in positive zeta potential for most particles and negative zeta potential for silicon surface, are therefore prone to particle adhesion, whereas alkaline cleaning baths, like ammonia peroxide, are less susceptible to particle adhesion.

12.3.1 Particle removal in wet cleaning

The two main mechanisms for wet cleaning are

1. dissolution/decomposition
2. etching.

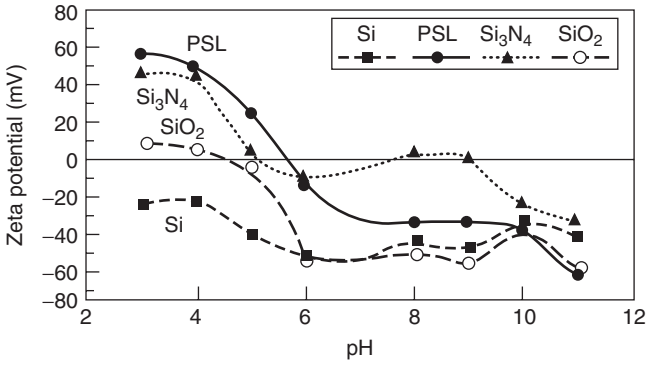


Figure 12.4 Zeta potential: pH influences particle adhesion and removal (PSL polystyrene latex). Source: T. Hattori (ed.) (1998)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

They have a very important distinction for surface roughness – etching processes tend to make surfaces rougher.

Ammonia peroxide solution works by oxidizing the silicon surface, and subsequently etching the oxide away.

$2H_2O_2 \longrightarrow 2HO_2^- + 2H^+$	peroxide disproportionation
$Si + 2HO_2^- \longrightarrow SiO_2 + 2OH^-$	silicon oxidation
- - - - - - - - - - - - - - - - - -	
$Si + 2H_2O_2 \longrightarrow SiO_2 + 2H_2O$	total reaction for oxidation
$SiO_2 + OH^- \longrightarrow HSiO_3^-$ (aq)	oxide etching (cf. Si etch in KOH)

Silicon etch rate in ammonia peroxide is ca. 0.1 to 0.5 nm/min (depending on concentration) and a typical clean removes ca. 1.5 nm of silicon. This leads to undercutting and removal of the particles.

Particle-removal efficiencies of different ammonia concentrations of RCA-1 are shown in Figure 12.5. In the first approximation, cleaning efficiency depends on the removed silicon depth, but more detailed analysis hints at reduced removal efficiency in dilute solutions. Megasonic agitation is widely used to enhance particle removal.

Ammonia peroxide cleaning results in oxidized surface, which is beneficial because it protects the silicon surface. For instance, during ramping wafers to high temperatures, volatile contamination will be removed before the thin oxide is baked away.

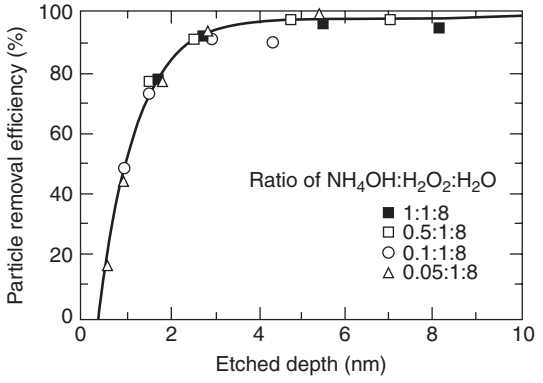


Figure 12.5 Etching as a method for particle removal: ca. 4 nm undercut etch is enough to remove most particles. Ammonia dilution is used as a parameter. Source: T. Hattori (ed.) (1998)

12.3.2 Wafer particle measurements

Particle measurements on wafers down to 60 nm size range can be performed by laser scattering equipment. A laser illuminates the wafer surface, and forward-scattered (Mie-scattering) light is measured. Scattering events can be caused by all irregularities on wafer: vacancy clusters (COPs) are pits, and they, too, scatter light. On very clean wafers COPs can account for 90% of 'particles'. Various optical designs (tilted incident laser beam, variable detector angle, measurement of both reflected and scattered signals) can be used to distinguish the nature of the scattering sources.

Scatterometric particle sizes are calibrated against contamination standards that have polystyrene latex spheres (PSL) of certified sizes on them. These PSL are nearly spherical, have tight size distribution and have a known refractive index of ca. 1.6. The number of particles is better calibrated against etched features with known light-scattering properties and known positions on the wafer. Such standards can be cleaned and reused, whereas contamination standards cannot.

Because real particles are not spheres with known optical constants, particle sizes cannot strictly be measured by light scattering (as witnessed by the fact that equipment from different manufacturers, and even different models from the same manufacturer do not give the same particle sizes). Latex sphere equivalent (LSE) size should be reported. Mirror-polished unpatterned wafers are good for basic studies, but real wafers present a number of problems. Because forward-scattered light is reflected by the wafer before reaching the detector, thin films on the wafer must be taken into account. On oxide, particle calibration needs to be done for each film thickness. On metallized wafers, surface roughness leads to decreased signal-to-noise ratio, and therefore small particles cannot be detected. Correlating a scattering event to a physical particle is usually difficult, even though scatterometry produces a map of the wafer. If particles can be seen in SEM, chemical identification is possible by either EMPA or EDX analysis. This can be important for particle source identification.

On patterned wafers, the situation becomes even more difficult. Pattern recognition software can be used to remove regular patterns from stochastic particle signals, but detection limit and equipment throughput are sacrificed.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

12.4 ORGANIC CONTAMINATION

There are many sources of organic contamination in the cleanroom. Table 12.3 below lists some of the most usual ones.

12.4.1 Organics removal

Sulphuric acid peroxide mixture (SPM) removes organics by oxidizing decomposition. This is however, a slow method, and other mechanisms are at work. Bond breakage and subsequent formation of smaller molecular mass fragments that are more soluble can explain fast organics removal. SPM cleaning leaves difficult-to-remove sulphur residues, and RCA-1 step is often carried out immediately after SPM to turn sulphides into soluble sulphates.

Oxidation of wafer surface by peroxide and the subsequent removal of this thin oxide by HF is shown in Figure 12.6. Organic films can prevent oxidation by peroxide for some time, which leads to unequal oxide thickness, and, after HF etching, to increased surface roughness. Extended cleaning would remove organics and lead to uniform oxide thickness and consequently no roughness increase.

Table 12.3 Sources of organic contamination

– Liquid chemicals and vapours used in fabrication processes: HMDS, isopropyl alcohol (IPA), acetone
– Gases, for example according to reaction $n\mathrm{CF}_4 \rightarrow (\mathrm{CF}_2)_n + 2n\mathrm{F}^*$
– Organic films (resist, spin-on polymers)
– Wafer holders and boxes
– Vacuum systems: pump oils, o-rings
– Cleanroom materials: sealants
– Intake air

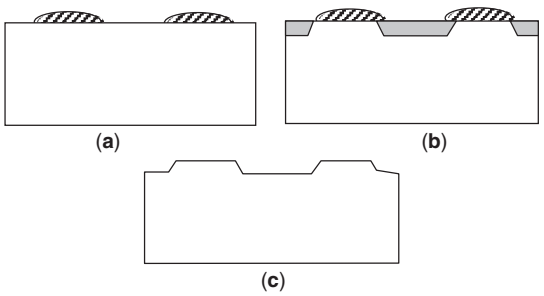


Figure 12.6 Organics removal: (**a**) organic residue on surface; (**b**) residue retards oxidation in H_2O_2 and (**c**) oxide removal in HF results in increased surface roughness. (Based on Hattori/Realize Inc.)

Because sulphuric acid constitutes an environmental concern and a safety hazard, other candidates have been sought for organics removal. Ozonated DI-water with 10 to 100 ppm ozone has proven to be very effective for some organic contamination. Furthermore, it is a room temperature process, versus 120 °C SPM. The ultimate cleaning method for organic contamination is thermal oxidation: no organic compound can tolerate 1000 °C in oxygen atmosphere. This provides a reference surface for analytical methods, but of course it is not a practical cleaning process.

12.4.2 Measurement of organic contamination

Organic contamination can be conveniently measured by FTIR (Fourier transform infrared spectroscopy), which identifies not only elements but also chemical bonds, as shown in Figure 12.7. FTIR can be operated in attenuated total reflection mode (ATR-FTIR) to improve sensitivity. XPS is very surface sensitive, and it can also identify chemical bonds, which is often important in understanding the origin of the contamination.

Molecular surface contamination can be measured by thermal desorption spectroscopy (TDS). TDS consists of a furnace connected to a mass spectrometer, and desorption of contaminants is monitored as a function of the furnace temperature. Silicon surface condition has also been clarified by TDS: at 340 °C, water desorbs, at 400 °C, hydrogen-terminated silicon surface undergoes reaction $\mathrm{SiH}_2 \rightarrow \mathrm{SiH} + \frac{1}{2}\mathrm{H}_2$ and at 500 °C $\mathrm{SiH} \rightarrow \mathrm{Si} + \frac{1}{2}\mathrm{H}_2$. Baking can therefore be used as an *in situ* surface-cleaning method.

12.5 METAL CONTAMINATION

There are numerous sources of metals, even though alternative materials like silicon, Teflon®, SiC and quartz are extensively used in making process equipment and wafer-handling tools. Table 12.4 lists some common sources of unwanted metals.

Table 12.4 Sources of metal contamination

– Tool materials (shutter blades, collimators, chucks)
– System components (pipes, valves)
– Wafer handling (tweezers, robot arms, wafer holders)
– Impurities in chemicals (buffered HF, BHF, is a known source of copper)
– Chemicals themselves (some photoresist developers are NaOH)
– Human contribution (sodium from sweat, heavy metals from cosmetics)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

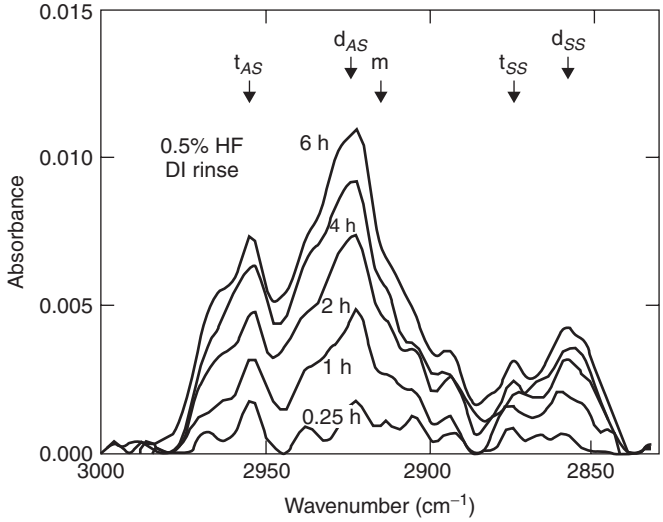


Figure 12.7 Infrared spectroscopy shows how organic contamination builds up over 6 h on an HF-rinsed wafer, evidenced by increased absorbance due to CH(m), CH_2(d) and CH_3(t) bonds. Reproduced from E. Grannemann (1994), by permission of AIP

12.5.1 Device effects of metal contamination

Metal contaminants degrade performance of electronic devices in various ways, depending on their chemical and physical nature, that is, reactivity with silicon and silicon dioxide and diffusion. Harmfulness of metal atoms depends on where they end up on the wafer: metals and metal precipitates in active areas lead to serious yield problems, while metals trapped in the bulk of the wafer are relatively harmless. Deep-level impurities act as majority carrier traps. Recombination velocity has its maximum when deep-level energy is in the middle of the forbidden gap, and therefore Zn, Cu, Au and Fe are especially harmful impurities, as shown in Figure 12.8.

MOS transistors can fail via various metal-induced mechanisms; for instance, junction leakage, oxide dielectric strength failure or threshold voltage shift.

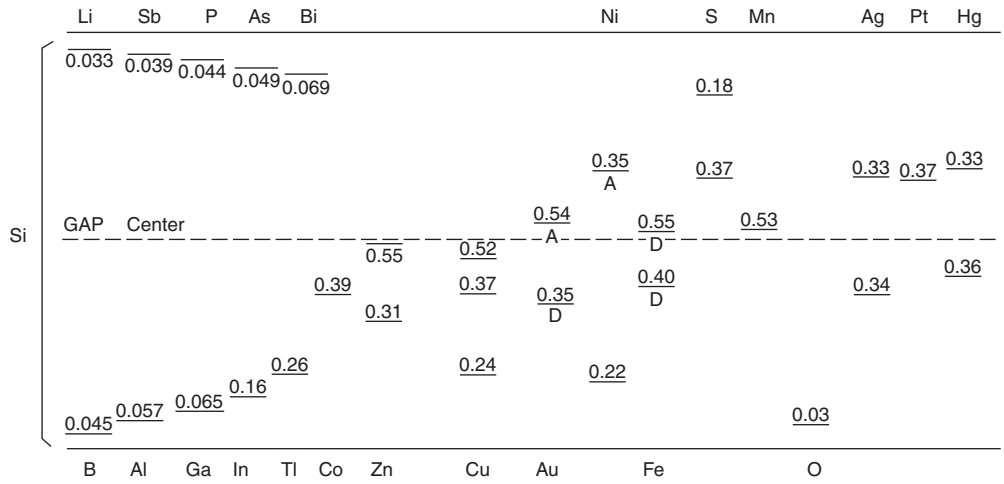


Figure 12.8 Ionization energies of impurities in silicon. Reproduced from S.M. Sze & J.C. Irvin (1964), by permission of Pergamon

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Segregation of contaminants between Si and SiO_2 has a major impact on the effects of metallic contamination: during thermal oxidation, Al, Ca, Cr and Mg are incorporated into the oxide and contribute to oxide quality problems, whereas Fe, Cu and Ni diffuse in silicon bulk.

Non-electronic devices are less sensitive to metal contamination, but metals cannot be completely ignored: metal contamination causes stacking faults in oxidation, and metals can catalyse peroxide decomposition, which leads to reduced particle-cleaning efficiency in RCA-1.

12.5.2 Metal removal

Acidic solutions $HCl–H_2O_2$ and $H_2SO_4–H_2O_2$ are the main methods for metal removal. Dilute HF, which removes a thin oxide layer, will additionally remove some metallic contaminants. Ammonia solutions (RCA-1) can also form complexes with metals and remove Cu and Ni.

The cleaning efficiencies of $HCl–H_2O_2$ and HF are very different, though. Both can reduce Fe and Ni levels below detection limit, but HF is much more effective in removing Al, and $HCl–H_2O_2$ in removing Cu. Dilution of HF needs to be specified because various workers use different concentrations. For aluminium removal, 0.1% DHF (by weight) is enough, but below that the removal efficiency rapidly deteriorates. HCl concentration in $HCl–H_2O_2$ has to be at least 5% for it to remove iron.

The wet chemicals themselves contain metallic impurities, and at the 10 ppb level their deposition on wafer surface is of some concern. For example, iron at 1 ppb level in RCA-1 solution results in a surface concentration of 10^{12} atoms/cm^2. Metal removal after RCA-1 has to be performed. The use of higher-purity chemicals helps to reduce the need in the first place, but it cannot be relied upon as the sole method because of statistical effects, both in manufacturing and in use (if RCA-1 bath is used several times, contamination from previous batches remains in the solution). RCA-1 must be accompanied by a cleaning step that removes metals efficiently. However, both HF- and HCl-based solutions lead to increased particle counts.

Newer cleaning solutions include $HF:H_2O_2$, which has both oxidizing and metal-removal capabilities. It can be used at room temperature versus 70 °C, which is typical of RCA-cleans. $HF:H_2O_2$ seems to increase surface roughness, so cleaning time needs to be optimized.

12.5.3 Measurement of metallic contamination

Metal contamination surface concentrations range from 10^{10} to 10^{14} atoms/cm^2, depending on technology generation, contamination-control strategies and particular process steps. Total reflection X-ray fluorescence (TXRF) uses a grazing incident angle to probe the wafer surface to nanometre depth. It is most sensitive for medium-mass atoms, and less sensitive towards both ends of the mass range. Detection limit of TXRF is ca. 10^9 atoms/cm^2. TXRF is a non-destructive method that can be used on whole wafers.

In vapour-phase decomposition (VPD) and wafer surface analysis (WSA) methods, surface impurities are first collected in oxide (native oxide or chemical oxide), which is then decomposed by HF and collected in a droplet. This concentrate is analysed by the graphite furnace atomic absorption spectroscopy method (GFAAS) or by the inductive coupled plasma-mass spectrometer (ICP-MS), which can have sensitivities as low as 10^8 cm^{-2}.

Metallic contaminants can be measured by their effects on charge carriers. Minority carrier lifetime will be degraded by contamination. Surface photovoltage SPV and microwave photoconductivity decay (μPC) methods provide this information.

12.6 RINSING AND DRYING

Rinsing in DI-water and drying must be considered as essential parts of any cleaning process. As a general strategy, we should keep the wafer wet all along the cleaning process and reduce the number of times when wafers are drawn from liquid to air. When drying is required, there are a number of methods available: spinning, nitrogen blowing, vapour drying, lamp drying, vacuum drying, and dry wafers can also emerge from slow removal from hot DI-water. Spinning techniques are prone to charging and particle adherence, which are inherent in high-speed spinning equipment. Various isopropyl alcohol (IPA) drying methods rely on low surface tension and good wettability of IPA. In Marangoni drying, the wafer is drawn from water into IPA-nitrogen atmosphere, and water is pulled back, leaving a dry surface. IPA drying methods must be considered for chemical consumption, hot vapours and solvent accumulation.

12.7 PHYSICAL CLEANING

Three methods of physical removal of particles are widely used:

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

– brush scrubbing
– jet scrubbing
– ultrasonic/megasonic.

In brush scrubbing, nylon or PVA brushes physically touch the wafer and brush away the particles. This is effective especially when lots of particles or large particles have been deposited on the wafer. Therefore, brush scrubbing is often done after wafer scribing or polishing steps.

In jet scrubbing, high-pressure water is sprayed on the wafer. The removal mechanism is similar to brush scrubbing but no physical contact with the wafer is needed. Increasing pressure improves cleaning efficiency, but electrostatic charging can damage thin films.

In sonic cleaning, shock waves supply localized sound energy that helps in particle removal. Ultrasonic agitation (20–40 kHz) is also beneficial in wet removal of photoresist. However, cavitation may damage the wafers. Above 1 MHz, this is not an issue, and the method is termed 'megasonics'. Megasonic agitation improves particle removal even for very small particles, <100 nm size.

12.8 EXERCISES

1. Translate surface iron contamination of $10^{10}\,\mathrm{cm}^{-2}$ into a number of monolayers!
2. If there is one monolayer coverage of organic contamination on the wafer, how much is that counted as carbon atoms/cm^2?
3. Area of an NMOS transistor with 1 μm minimum linewidth is about the same as that of a red blood cell, $5 \times 8\,\mu\mathrm{m}$. The source/drain areas are doped to very high concentration, but the number of dopant atoms is only 10^9 because of small area. What concentration will result if the blood cell decomposes on the transistor, releasing its phosphorus atoms and doping the silicon?
4. Calculate the daily (24 h) chemical and DI-water consumption for an SPM-DIW-rinse-RCA1-DIWrinse-DHF-DIW rinse-RCA2-DIWrinse1-DIWrinse2 cleaning cycle when a tank for 25 wafers of 200 mm diameter is used. Assume a 4 h changing interval for RCA-cleans and 24 h bath life for SPM and DHF.
5. What happens to particle contamination in (a) wet etching and (b) plasma etching?
6. If we had an Olympic swimming pool full of UPW, how many droplets of sweat can be dissolved before Na^+ and Cl^- exceed the specification level of 0.1 ppb?

REFERENCES AND RELATED READINGS

E. Grannemann: Film interface control in integrated processing systems, *J. Vac. Sci. Technol.*, **12** (1994), 2741.

T. Hattori (ed.): *Ultraclean Surface Processing of Silicon Wafers*, Springer (1998).

W. Kern: The evolution of silicon wafer cleaning technology, *J. Electrochem. Soc.*, **137** (1990), 1887.

W. Kern (ed.): *Handbook of Semiconductor Wafer Cleaning Technology*, Noyes Publications (1993).

H. Kitajima & Y. Shiramizu: Requirements for contamination control in the gigabit era, *IEEE TSM*, **10** (1997), 267.

S. Middleman & A.K. Hochberg: *Process Engineering Analysis in Semiconductor Device Fabrication*, McGraw-Hill (1993).

T. Ohmi, *et al.*: Dependence of thin-oxide film quality on surface microroughness, *IEEE TED*, **39** (1992), 537.

H. Okorn-Schmidt: Characterization of silicon surface preparation processes for advanced gate dielectrics, *IBM J. Res. Dev.*, **43** (1999), 351.

D.K. Schroder: *Semiconductor Material and Device Characterization*, 2nd ed., John Wiley & Sons (1998).

S.M. Sze & J.C. Irvin: Resistivity, mobility and impurity levels in GaAs, Ge and Si at 300°K, *Solid-State Electron.*, **11** (1964), 599.

F. Zhang, *et al.*: The removal of deformed submicron particles from silicon wafers by spin rinse and megasonics, *J. Electron. Mater.*, **29** (2000), 199.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

13

Thermal Oxidation

Silicon dioxide, SiO_2, is probably a more important material in silicon technology than silicon itself: while GaAs and Ge have higher electron mobilities than silicon, and enable potentially faster devices; they do not have native oxides that protect their surfaces, and neither do stable, thick oxides exist. Silicon dioxide has functions as capacitor dielectric and isolation material, in which case the oxide forms a part of the finished device. But oxides are used intermittently many times during silicon processing as a masking material for diffusion or etching, and as a cleaning method to reclaim perfect silicon surface.

13.1 OXIDATION PROCESS

Silicon is easily oxidized: a native oxide of nanometre thickness grows on the silicon surface in a couple of hours or days, depending on surface conditions, and similar thin oxides form easily in oxygen plasma or in oxidizing wet treatment. These oxides are, however, limited in their thickness and they are not stoichiometric SiO_2. Deposited CVD oxides are used in some applications where low temperatures are absolutely necessary, but superior silicon dioxides are grown in 800 to 1200 °C, Figure 13.1. Two basic schemes are used: wet (aka. steam) and dry oxidation.

Wet oxidation: $\mathrm{Si\ (s) + 2H_2O\ (g) \longrightarrow SiO_2\ (s) + 2H_2\ (g)}$

Dry oxidation: $\mathrm{Si\ (s) + O_2\ (g) \longrightarrow SiO_2\ (s)}$

Thermal oxidation is a slow process: dry oxidation at 900 °C for 1 h produces ca. 20 nm thick oxide and wet oxidation for 1 h produces ca. 170 nm. Exact values are dependent on silicon crystal orientation: oxidation rate of <111> is somewhat higher than that of <100> silicon; highly doped silicon oxidizes faster than lightly doped material, and the higher the oxygen pressure, the higher the rate.

Thin oxides, such as CMOS gate oxides, Flash memory tunnel oxides and dynamic random access memory (DRAM) capacitor oxides are of the order of 1 to 20 nm. These oxides are grown in dry oxygen at 850 to 950 °C. Thin oxides also have many auxiliary and sacrificial roles: a thin oxide under nitride relieves stresses caused by the nitride film. Thicker oxides are used for device isolation and as masking layers for ion implantation, diffusion and etching steps. They are usually 100 to 1000 nm thick, and grown by wet oxidation.

13.2 DEAL–GROVE OXIDATION MODEL

A model for oxide growth has been put forth by Deal and Grove. It is a phenomenological macroscopic model that does not assume anything about the atomistic mechanisms of oxidation. Oxygen diffusion through the growing oxide and chemical reaction at the silicon/oxide interface are modelled with the classical Fick diffusion equation and chemical rate equation (Figure 13.2).

Oxidation is modelled as if the boundaries were stationary (which is a reasonable assumption because oxidation is slow). The diffusion equation for oxygen is

$$0 = D(\mathrm{d}^2C/\mathrm{d}z^2) \tag{13.1}$$

where C is the oxygen molar concentration (in units mol/m^3), subject to the boundary conditions

$$C = C_\mathrm{s} \quad z = 0 \tag{13.2}$$

at the SiO_2 surface and

$$-D(\mathrm{d}C/\mathrm{d}z) = R \quad z = Z \tag{13.3}$$

at the SiO_2/Si interface, where R is the reaction rate at the interface (in units mol/m^2 s).

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

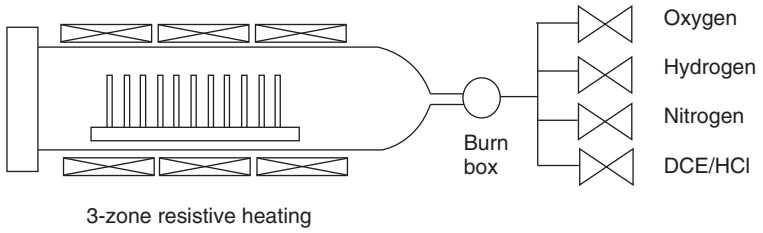


Figure 13.1 Horizontal oxidation furnace: wafers are vertically loaded in quartz boats

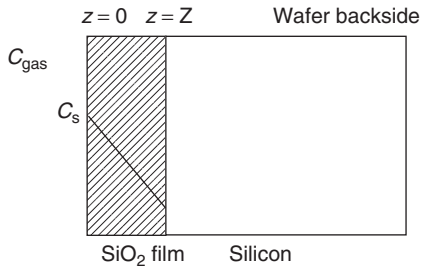


Figure 13.2 Model of thermal oxidation: oxygen diffuses through SiO_2 film and reacts at the SiO_2/Si interface. Concentration of oxygen inside oxide decreases linearly

The latter equation specifies that all oxygen reaching the interface will react there to form oxide: there will be no build-up of unreacted oxygen inside oxide or silicon.

For a reaction like Si (s) + O_2 (g) → SiO_2 (s), the rate is assumed to be first order, that is, $R = kC$, directly related to concentration of reactive species, C, and characterized by a rate constant k. We can then rewrite the second boundary condition as

$$-D(\mathrm{d}C/\mathrm{d}z) = kC \quad \text{at } z = Z \tag{13.4}$$

A solution that satisfies these conditions is

$$C = C_s - (kC_s/(kZ + D)) \tag{13.5}$$

Rate (at the interface $z = Z$) is then

$$R = kC(Z) = kDC_s/(kZ + D) \tag{13.6}$$

To calculate thickness growth rate, we must convert molar concentration to volume through density:

$$RM_{SiO_2} = \rho_{SiO_2}(\mathrm{d}Z/\mathrm{d}t) \tag{13.7}$$

where the molar volume of SiO_2 is $\upsilon = M_{SiO_2}/\rho_{SiO_2}$ (60 g/mol/2.2 g/cm^3 = 27.3 cm^3/mol).

When we solve for $Z(t)$ from the rate equation, we get

$$\mathrm{d}Z/\mathrm{d}t = (kDC_s\upsilon)/(kZ + D) \text{ subject to } Z = 0 \text{ at } t = 0 \tag{13.8}$$

This leads to the oxide thickness equation:

$$t = Z/(KC_s\upsilon) + Z^2/(2DC_s\upsilon) \tag{13.9}$$

When thin oxides are considered, we can ignore the second term, and rate is then simply

$$Z = kC_s t \tag{13.10}$$

or growth is linear in time and related to the rate constant k.

For thick oxides, we can ignore the first term, and we get

$$Z = \sqrt{2DC_s\upsilon t} \tag{13.11}$$

or growth is parabolic, related to diffusion length $\sqrt{Dt}$.

The Deal–Grove model thus predicts linear oxidation rate initially, followed by a parabolic behaviour for thicker oxides, Figure 13.3. The linear regime covers only the initial stages of oxidation with some success. The model works much better for thick oxides, and theory and experiment agree that doubling oxide thickness requires quadrupling oxidation time in the parabolic regime (this can be used as a quick estimate for oxidation time once one process is known and fixed).

Dry oxidation is slower than wet oxidation (Figure 13.4) even though diffusion of oxygen molecules through silicon dioxide is faster than diffusion of water molecules. But water solubility in silicon dioxide is 4 orders of magnitude larger that oxygen solubility, and therefore, concentration of the oxidant in oxide is much greater.

13.2.1 Oxidation of other materials

Very few materials can tolerate oxidizing ambients at ca. 1000 °C. No metal can withstand such conditions. Silicon and silicon-containing compounds are really exceptional in this respect.

Polysilicon oxidation presents a number of complications compared to single-crystal oxidation. The polysilicon surface is not smooth like a single-crystal surface

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

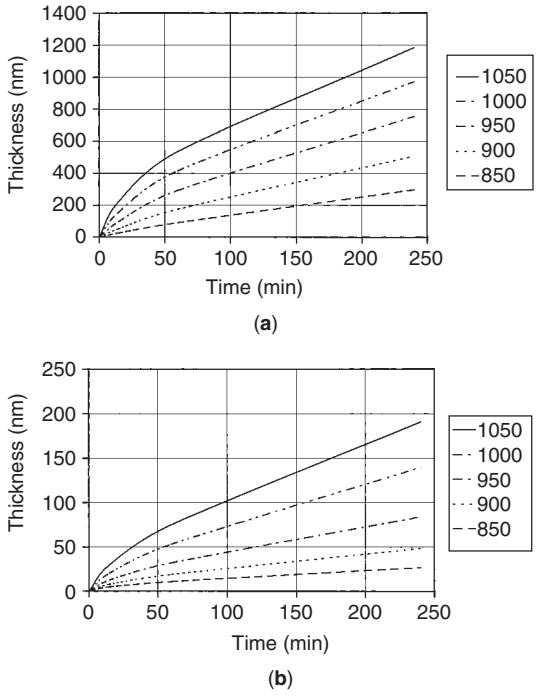


Figure 13.3 Oxidation of <100> silicon at temperatures between 850 and 1050 °C: wet and dry

and the oxide quality will be inferior to oxides grown on smooth surfaces. Polysilicon consists of grains of many orientations, which have different oxidation rates. Polysilicon texture is most often (110) and the oxidation rate of undoped poly falls between (100) and (111) rates. In polycrystalline materials, there are two different diffusion paths: through the bulk, and along grain boundaries. Because grains grow during oxidation, this introduces complications in the analysis. In doped polysilicon, dopants precipitate at grain boundaries. Boron doping leads to minor rate enhancement and phosphorus-doping to clearly increased oxidation rate via increased vacancy concentration, just as in the case of the single-crystal material.

Silicides will generally oxidize to form SiO_2, with the exception of $TiSi_2$, which will turn into TiO_2. Tungsten polycide gates (WSi_2/poly) can be processed similarly to polysilicon. Making the silicide silicon-rich, $WSi_{2.2}$, will ensure proper oxidation. Silicon carbide, SiC, can be oxidized to produce SiO_2 with standard silicon oxidation processes but the rate is very low compared to silicon oxidation.

13.3 OXIDE STRUCTURE

Thermally grown silicon dioxide is glassy, and exhibits only short-range order, in contrast to quartz, which is crystalline SiO_2. The basic unit of silica structure is SiO_4 (Figure 13.5).

In a perfect arrangement, such as crystalline quartz, all oxygen atoms bond to two silicon atoms (oxygen has valence 2, silicon has valence 4) but in thermal oxide some bonds are not made, leaving unbonded charged oxygen atoms, making the oxide less stable than quartz. This is also reflected in their properties: quartz density is 2.65 g/cm^3, silicon oxide density 2.2 g/cm^3; Young's modulus is 107 GPa for quartz and 87 GPa for oxide.

When dopant atoms are incorporated into silicon dioxide network, they can take either substitutional or

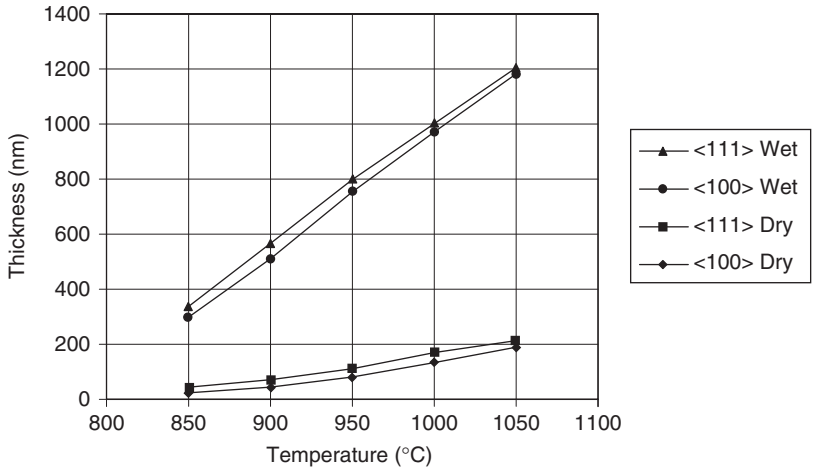


Figure 13.4 Difference between <100> and <111> silicon oxidation (constant oxidation time 240 min)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

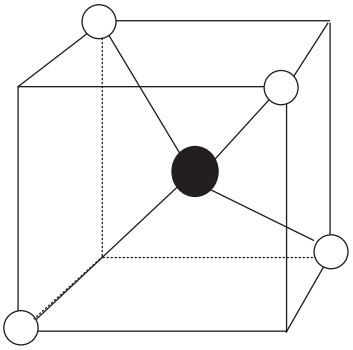


Figure 13.5 Basic structure of silica: a silicon atom tetrahedrally bonds to four oxygen atoms

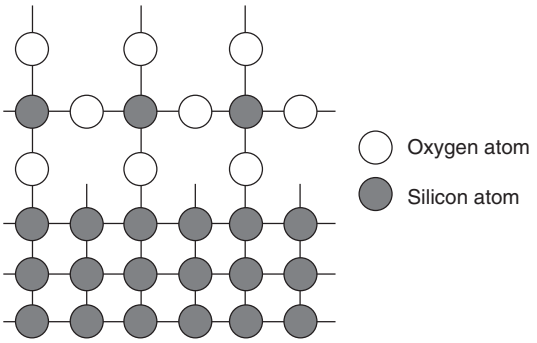


Figure 13.6 The structure of silicon–silicon dioxide interface: some silicon atoms have dangling bonds

interstitial positions. Boron and phosphorus can take the position of a silicon atom in the network and form oxides themselves (B_2O_3, P_2O_5), hence the name *network formers*. However, due to their electrical properties, they affect oxide differently. Phosphorus, a group V element, will donate an extra electron to a non-bridging oxygen and stabilize the oxide, whereas boron with one electron missing makes oxide less stable. Sodium, potassium and lead are interstitial network modifiers that bond to one silicon atom only and do not form glasses themselves.

When silicon and oxygen react to form SiO_2, silicon is consumed: for an SiO_2 layer of thickness D, silicon thickness consumed is $0.45D$ as can be calculated from molar volumes:

Density of Si	Molar mass	Molar volume
2.3 g/cm^3	28 g/mol	12.17 cm^3/mol
Density of SiO_2	Molar mass	Molar volume
2.2 g/cm^3	60 g/mol	27.27 cm^3/mol

The original surface is somewhat below the oxide mid-point. This volume change leads to restrictions in the oxidation of structured surfaces, because stresses can become excessively large in the corners of the structures. On the other hand, the fact that oxidation consumes silicon can be used as a cleaning method: thin oxide is grown and immediately removed by hydrofluoric acid (HF) etching, to reveal a perfect silicon surface.

Another consequence of volume change is that oxide and silicon cannot fully fill the space at the interface. Some atoms do not have their full valence, but have dangling bonds (Figure 13.6). These bonds act as traps for charge carriers.

Thermal oxidation is often complemented by a post-oxidation anneal (POA) in nitrogen. This step densifies the film and anneals out some defects. It of course adds to thermal load, and has to be considered when doping profiles are fine-tuned. Hydrogen anneal is often used to passivate dangling bonds: hydrogen attaches to the free valence of the silicon, and eliminates further charge trapping. However, high electric fields can easily accelerate electrons to such energies that hydrogen atoms are released during device operation.

Oxide thickness is usually measured by optical methods: either by ellipsometry or reflectometry. Thermal oxides can be grown with very tight specifications, for a 10 nm thick oxide, uniformity is 1%, that is, equal to one atomic diameter. For thermal oxides, refractive index value $n = 1.46$ is usually used, but for very thin oxides this is not valid. A quick and easy way to gauge oxide thickness is by its colour; Table 5.7 shows oxide colours.

Various electrical measurements are also used: breakdown voltage is one of many. High-quality silicon dioxide can sustain 10 MV/cm, even 12 MV/cm, while poly-oxides have 5 MV/cm breakdown fields. Oxide defects and electrical quality are closely connected; this topic will be discussed further in Chapter 24.

13.4 SIMULATION OF OXIDATION

Oxidation simulation, together with diffusion simulation, is the backbone of all process integration simulators. Thermal oxidation is well understood, and can be accurately modelled. However, the atomistic mechanisms of thin oxides (and early stages of oxidation in general) are still under intensive study.

Oxidation simulation requires as input:

- wafer orientation <100>/<111>
- doping level;

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

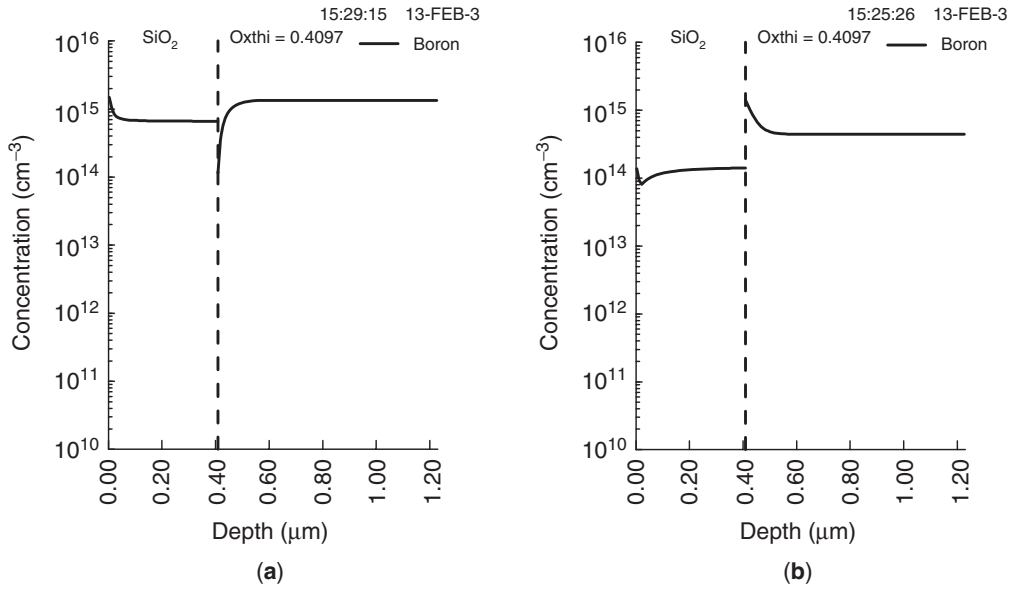


Figure 13.7 Segregation of dopant at silicon–oxide interface during wet oxidation (1000 °C, 60 min): (**a**) boron-doped wafer shows dopant loss at interface and (**b**) phosphorus-doped wafer shows accumulation of dopant at the interface. Substrate resistivity is 10 ohm-cm in both cases

– temperature;
– time;
– oxidizing ambient wet/dry.

For additional model parameters such as oxygen partial pressure (1 atm as default) and high concentration effects, viscous/elastic models can be used instead of default models.

The Deal–Grove model is the default model for wet oxidation, and for thick oxides in general. It is not, however, applicable to thin dry oxides. A power-law model from Nicollian and Reisman can be used for this regime. Oxidation is modelled as

$$x_{ox} = a(t/t_0)^b \tag{13.12}$$

Simulators produce results that are accurate within experimental error for 1D oxidation. Additionally, simulators can account for segregation, the distribution of dopants at the oxide/silicon interface.

13.4.1 Segregation

Dopants that are initially in the silicon are redistributed between silicon and the growing oxide during oxide growth (Figure 13.7), not unlike dopant segregation between solid and melt during crystal growth. Segregation has a major effect on device properties: if the dopant is mostly incorporated in the oxide and depleted in the silicon near the interface, inversion may occur. Segregation proceeds as long as the chemical potentials of the dopants differ in the oxide and silicon. The equilibrium segregation coefficient, m, is defined as the ratio of dopant in silicon to that in oxide.

Dopant atoms have a major impact on oxidation: heavy doping will change oxidation rate significantly. In the case of boron, it is through incorporation of boron into the growing oxide, weakening its bond structure and thus enabling faster diffusion through it.

Metal atoms experience segregation just like the dopants: for example, Al and Ca are segregated preferentially into the oxide (and cause oxide quality problems) whereas Ni and Cu diffuse into bulk (and cause defects that act as lifetime killers).

13.5 LOCAL OXIDATION OF SILICON (LOCOS)

When local oxidation of silicon is needed, silicon nitride mask is used. Nitride will prevent oxygen diffusion, and areas under nitride will not be oxidized. This is known

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

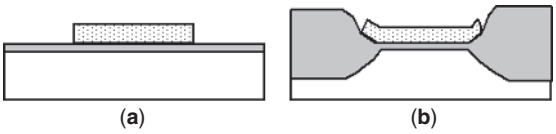


Figure 13.8 LOCOS (**a**) before oxidation: thin pad oxide and patterned nitride and (**b**) after oxidation: no oxidation under nitride but 'bird's beak' at nitride edge

as LOCOS, for *local oxidation of silicon*. LOCOS is pictured in Figure 13.8.

LOCOS process flow

thermal oxidation;
LPCVD nitride deposition;
lithography;
nitride etching;
photoresist strip;
cleaning;
oxidation.

LOCOS variables are pad oxide thickness (10–50 nm), LPCVD nitride thickness (100–200 nm) and oxidation temperature. Pad oxide serves as a stress relief layer, and it diminishes the stress-induced dislocations that a thick nitride exerts in silicon. Nitride acts as a diffusion barrier for oxygen diffusion, and as a mechanical stiffener: the thicker the nitride, the smaller the oxide growth under the mask. This lateral extension is known as *bird's beak*, for obvious reasons. A thinner pad oxide would help minimize bird's beak but at the expense of silicon damage from nitride stress. Recessed LOCOS is used to make the surface more planar after oxidation (Figure 13.9). The etching step involves etching nitride, oxide and silicon, with silicon etched depth approximately half the desired oxide thickness, which then will result in approximately equal surface heights for oxide and silicon.

LOCOS isolation has been used for 30 years for its simplicity. LOCOS has been scaled to much smaller linewidths than anybody thought possible. Numerous modifications have been tried, but most have failed because the added process complexity has not offered enough improvement in isolation.

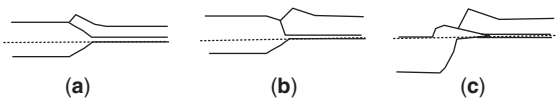


Figure 13.9 Bird's beak in LOCOS (**a**) thin nitride; (**b**) thick nitride and (**c**) recessed LOCOS

13.6 STRESS AND PATTERN EFFECTS IN OXIDATION

Oxide volume is greater than the volume of the silicon it replaces. Oxides are therefore under compressive stresses, and this causes a number of pattern-dependent phenomena that can be either beneficial or disadvantageous. Typical stress values are of the order of 300 MPa. Somewhere between 975 and 1000 °C, the oxide exhibits viscous flow. Oxidation above that temperature will result in reduced stress and wafer bow. Below that temperature, oxide needs to be treated as an elastic material with appropriate elastic constants. Scaling of LOCOS to smaller linewidths meets an inevitable limit at sub-micron dimensions: stresses in the growing oxide prevent full oxidation of narrow gaps. For generations below 0.5 μm linewidths, the isolation method of choice is shallow trench isolation (STI), which will be discussed in Chapter 25.

Thermal oxidation of small silicon wires shows a self-limiting effect due to high stresses and this has been utilized in making nanostructures. This is illustrated in the silicon-on-insulator (SOI) nanowire process (Figure 13.10).

Process flow for silicon nanowires

SOI wafer with 21 nm thick device silicon;
lithography;
silicon etching;
photoresist striping;
oxidation.

Thermal oxidation proceeds for a while, but then a self-limiting effect sets in: a critical stress, which stops oxidation, is ca. 2.6 GPa at 850 °C. After the self-limiting oxide thickness has been grown, no further oxidation takes place. If oxidation is carried out at a higher temperature, say 1000 °C, this stress can be overcome, and the whole structure will be oxidized (Figure 13.10).

Stresses are also responsible for non-uniform oxidation in convex and concave corners as shown in Figure 13.11. Uneven oxide thickness causes problems for reliability because electric field strength is different in corners and planar areas. Etched trenches have concave corners, and therefore both STI and DRAM trench capacitors require fine-tuning of the bottom corners if thermal oxidation is used as the first film in the trench. Etch processes can be tailored to some extent for smoother bottom profiles, but this is a limited option because the top corner needs rounding too. Oxide and nitride can be deposited by conformal CVD, but in very

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

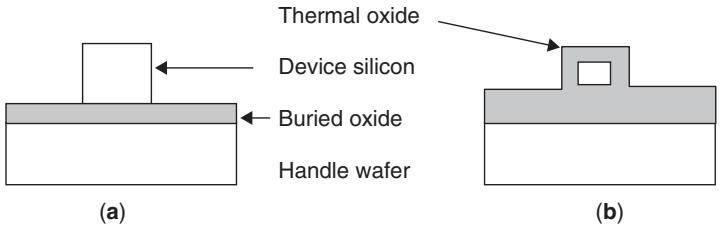


Figure 13.10 Silicon nanowire process on SOI: (**a**) SOI-structure after plasma etching and (**b**) after low-temperature thermal oxidation: unoxidized silicon remains. Redrawn from Heidemayer, H. *et al*. (2000), by permission of AIP

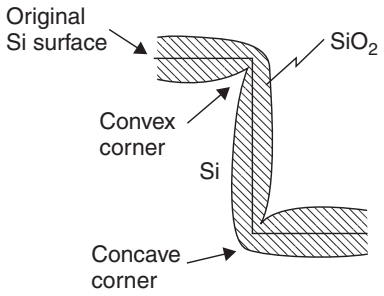


Figure 13.11 Cross section of an oxidized silicon step with oxide thinning at both convex (top) and concave (bottom) corner. Reproduced from Minh, P.N. & T. Ono (1999), by permission of AIP

deep trenches the conformality may not be adequate. Sacrificial thermal oxidation can be used to smooth corners. Second thermal oxidation then provides the actual thin dielectric film, which serves, for example, as a DRAM capacitor dielectric.

Simulation of oxide stresses of KOH-etched V-grooves is pictured in Figure 13.12. This stress-induced oxide thinning has been used to advantage in nanohole fabrication as shown in Figure 13.13. Etching in HF will open the apex only, creating a hole with dimensions in the sub-100 nm range.

13.6.1 Oxidation sharpening

Sharp tips are used as AFM probes and as field emitters in vacuum microelectronic devices, for high resolution in the former application and for low operating voltage in the latter. Such tips can be fabricated by isotropic etching, but the final part of the tip release is difficult: the mask will fall off. Thermal oxidation can help: after initial isotropic (or KOH anisotropic) etching, the final sharpening takes place during oxidation. Mask removal is done by isotropic etching, but this is non-critical, non-patterning etch, Figure 13.14. Thermal oxidation process control is also much tighter than shape control in an etch process. In Chapter 39, a process for AFM cantilever-tip device will be presented.

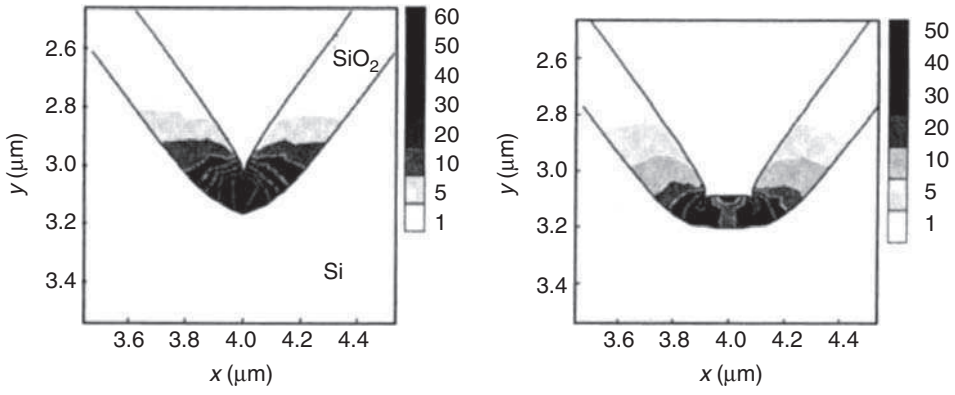


Figure 13.12 Oxide-stress simulation at the apex of etched groove; unit: MPa. Reproduced from Vollkopf, A. *et al*. (2001), by permission of Electrochemical Society Inc

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

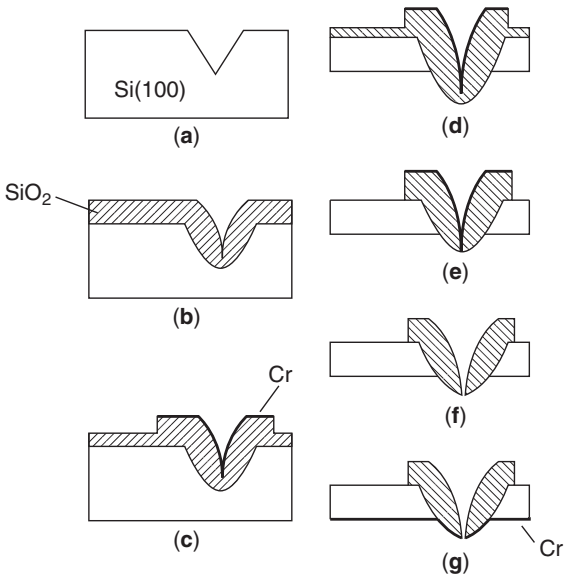


Figure 13.13 Oxide thinning at apex used as a method to fabricate nanoscopic holes: the apex can be etched open while leaving oxide elsewhere because the oxide is thin at the apex. From Minh, P.N. & T. Ono (1999), by permission of AIP

13.7 EXERCISES

1. Holes are etched in 1 μm thick thermal oxide. The wafer is then given 1 h wet oxidation at 1000 °C. All oxide is then etched away. What is the resulting step height in silicon?
2. 250 min wet oxidation results in 1 μm thick oxide. How long will it take to grow 10 μm thick oxide under the same conditions? How long will it take to grow a 0.1 μm thick oxide?

3S. The Deal–Grove oxidation model is not valid for thin oxides. Experimental data for dry oxidation is shown below. Check how your simulator works for thin oxides. Data from Massoud, H.Z. *et al*: *J. Electrochem. Soc.*, **132** (1985), 2685.

Time (min)	850 °C	1000 °C
20	6 nm	26 nm
40	8 nm	42 nm
60	11 nm	56 nm
80	13 nm	68 nm

4S. Phosphorus-doped polysilicon (20–80 ohm/sq) oxidation produces 50 nm thick oxide in 30 min dry oxidation at 1000 °C. At 900 °C, dry oxidation results in 10 nm thick oxide. How do these values compare with single-crystal silicon oxidation?

5S. High-pressure oxidation (HIPOX) increases oxidation rates. Data for dry oxidation at 900 °C is given below. Data from Lie, L.N. *et al*: *J. Electrochem. Soc.*, **129** (1982), 2828.

Pressure (atm)	Time (min)	Thickness (nm)
10	30	40
10	60	65
10	120	100
20	30	55
20	60	100
20	120	180

How does your simulator handle HIPOX oxides?

6S. What is the segregation behaviour of the *n*-type dopants As, P and Sb?

REFERENCES AND RELATED READINGS

Green, M.L. *et al*: Understanding the limits of ultrathin SiO_2 and Si–O–N gate dielectrics for sub-50 nm CMOS, *Microelectron. Eng.*, **48** (1999), 25.

Heidemayer, H. *et al*: Self-limiting and pattern dependent oxidation of silicon dots fabricated on silicon-on-insulator material, *J. Appl. Phys.*, **87** (2000), 4580.

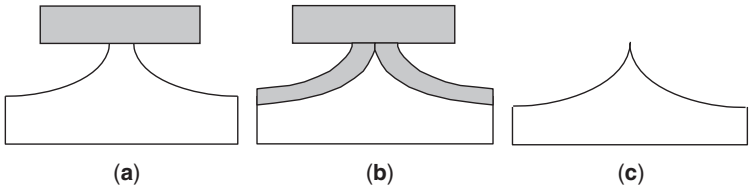


Figure 13.14 Silicon tip fabrication: (**a**) isotropic silicon etching with an oxide mask; (**b**) thermal oxidation and (**c**) silicon tip recovery by HF etching

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Lie, L.N. *et al*: *J. Electrochem. Soc.*, **129** (1982), 2828.
Massoud, H.Z. *et al*: *J. Electrochem. Soc.*, **132** (1985), 2685.
Minh, P.N. & T. Ono: Non-uniform silicon oxidation and application for the fabrication of aperture for near-field scanning optical microscopy, *Appl. Phys. Lett.*, **75** (1999), 4076.
Roy, P.K. *et al*: Synthesis of a new manufacturable high-quality graded gate oxide for sub-0.2 μm technologies, *IEEE TED*, **48** (2001), 2016.
Shimidzu, H.: Behavior of metal-induced oxide charge during thermal oxidation in silicon wafers, *J. Electrochem. Soc.*, **144** (1997), 4335.
Suryanarayana, P. *et al*: Electrical properties of thermal oxides grown over doped polysilicon thin films, *J. Vac. Sci. Technol.*, **B7** (1989), 599.
Vollkopf, A. *et al*: Technology to reduce the aperture size of microfabricated silicon dioxide aperture tips, *J. Electrochem. Soc.*, **148** (2001), G587.

**Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030**

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

14

Diffusion

The power of silicon technology stems from the ability to tailor dopant concentrations over eight orders of magnitude by introducing suitable n- or p-type dopants into the silicon. The upper limit is set by solid solubility of the dopants (ca. 10^{21} atoms/cm^3) (Figure 14.1); the lower limit (ca. 10^{13} atoms/cm^3) by impurities that result from the silicon crystal growth. This enables a wealth of microstructures and devices, witnessed by the multiplicity of diode, transistor, thyristor and other semiconductor device designs.

Dopants can be introduced into silicon by the following five different methods:

- during crystal growth
- by neutron transmutation doping (NTD)
- during epitaxy
- by ion implantation
- by diffusion.

The first two techniques result in doping of the ingot, and epitaxy results in uniformly doped layer all over the wafer. Diffusion and ion implantation are techniques to locally vary the dopant concentration (Figure 14.2), and they are discussed in this chapter and in Chapter 15.

Thermal diffusion is a high-temperature process: diffusion temperatures are in the range 900 to 1200 °C in current silicon technology. The diffusion furnaces are identical to oxidation furnaces, and diffusion is a batch process in which long process times are compensated by a huge load of wafers, 100 or even 200, in a batch. Ion implantation is a room-temperature, high-energy process of accelerating dopant ions and implanting them inside silicon. But dopant activation and damage anneal, which must always accompany ion implantation, are high-temperature processes.

Diffusion is often carried out in two steps: pre-deposition and drive-in. In pre-deposition a known

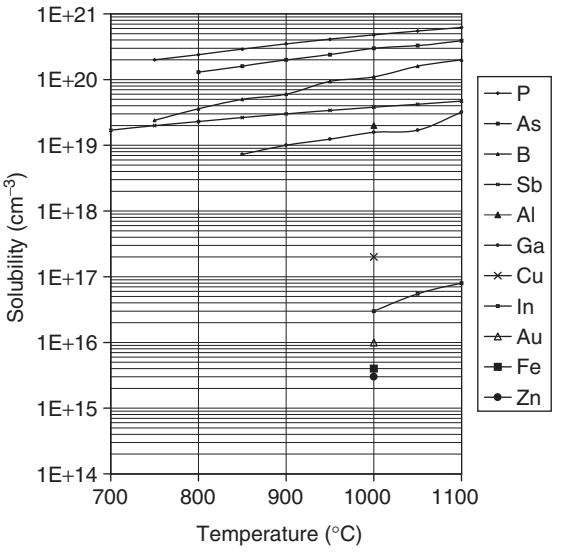


Figure 14.1 Solid solubilities of the most important dopants and impurities in silicon technology. Data from ref. Hull, R. (ed) (1999), by permission of Bell

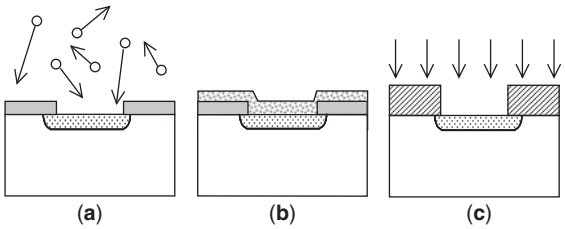


Figure 14.2 Doping processes: (**a**) gas-phase diffusion; (**b**) diffusion from doped solid film and (**c**) ion implantation. Oxide mask shown grey; photoresist mask hatched

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

and limited number of dopants is introduced on the wafer, and during drive-in they will diffuse deeper. Ion implantation and diffusion are strongly interrelated: implantation can be considered as a pre-deposition step for diffusion. Diffusion is, therefore, the general term for doping processes, irrespective of the actual mechanism of dopant introduction. In silicon IC technology dopant diffusion is such a key step that the country of origin of semiconductor devices is defined as the country where diffusions were made.

When local diffusion is done, silicon dioxide is the standard masking material. Even though the dopants do not diffuse through the oxide, they do modify it to the extent that diffusion mask oxides are practically always etched away after diffusion.

Doping can be performed many times over, and silicon doping type may change from p-type to n-type and back again, depending on the process sequence. The device shown in Figure 14.3, an UV-photodiode, is made in a modified npn-bipolar process. UV-photons are absorbed in the top p^+ diffusion layer. We will discuss only the diffusion aspects of the device now.

Process flow for UV-photodiode (lithography, etch and oxidation steps omitted)

p-type substrate wafer
n^+ buried layer diffusion
n epitaxial layer deposition
p^+ substrate contact diffusion
n^+ diffusion to contact buried layer
p^+ base contact enhancement diffusion (under A_{IR})
p base diffusion
n^+ cathode diffusion
p^+ anode diffusion.

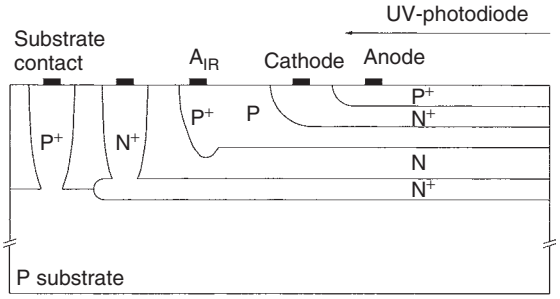


Figure 14.3 UV-photodiode with shallow p+ anode diffusion. The structure is based on npn-bipolar transistor. Reproduced from Zimmermann, H. (1999), by permission of Springer

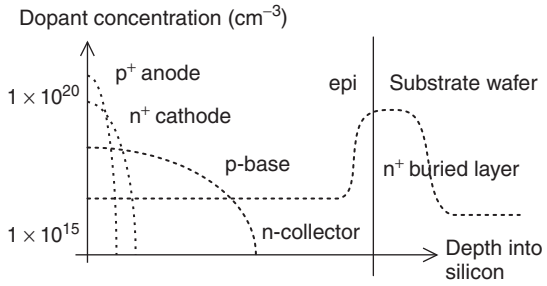


Figure 14.4 UV-photodiode doping profile underneath the anode

The area directly underneath the anode changes its doping type three times: it is originally n-type epilayer, doped by PH_3 gas during epitaxy. Base diffusion changes it to p-type when boron concentration exceeds the phosphorus concentration in the epilayer; the n-cathode diffusion turns it back to the n-type because phosphorus concentration is higher than boron concentration; and finally, the surface anode diffusion with the highest boron concentration of all results in p^+ silicon (Figure 14.4).

14.1 DIFFUSION MECHANISMS

Diffusion is atom movement along concentration gradients. Fairly simple mathematical models can describe concentration profiles in solids, but at the atomistic level diffusion remains to be fully explained. This has consequences for simulators, because mechanisms are not fully known, and therefore, modelling remains inaccurate.

Dopant atoms move with the help of point defects: they jump to vacancies and interstitials. Substitutional dopants are fairly stable without point defects. Vacancies are always present through thermal equilibrium processes: vacancies are thermodynamic defects, and their nature is different from, for example, dislocations and stacking faults, which are 'frozen'. Vacancies as a fraction of all sites can be estimated by

$$f = \exp(-E_a/kT) \tag{14.1}$$

For 1 eV activation energy, it gives ca. 0.01% vacant sites at 1000 °C (1273 K).

Here, we outline some mechanisms for diffusion (Figure 14.5). In interstitial diffusion, atoms jump from one interstitial site to another, which is always available. This is the diffusion mechanism for small atoms, like sodium and lithium. The substitutional/vacancy

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

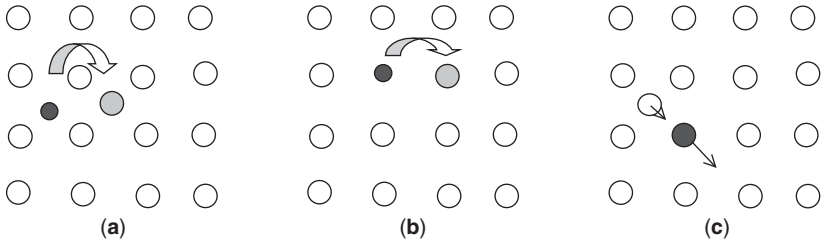


Figure 14.5 Diffusion mechanisms: (**a**) interstitial; (**b**) substitutional/vacancy and (**c**) interstitialcy

diffusion necessitates that empty lattice site is available next to the diffusing atom. At high temperatures substitutional sites are thermally created. Antimony and arsenic demonstrate substitutional mechanisms. The interstitialcy mechanism is related to the substitutional mechanism: the self-interstitial atoms move to the lattice sites, and kick the dopants to the interstitial sites, and from there they move to the lattice sites. Boron and phosphorus are expected to diffuse via interstitialcy mechanism, but there are still some open questions even in diffusion of the best-known dopants.

The substitutional and interstitialcy mechanism with activation energies of ca. 3.5 to 4 eV are the most important for doping in silicon technology. Boron, phosphorus, arsenic as well as antimony, indium and gallium all have activation energies in this range. Therefore, doping by diffusion must take place at a high temperature. Many metallic impurities diffuse with the interstitial mechanism with activation energies round 1 to 1.5 eV, and they are mobile at much lower temperatures than substitutional dopants.

14.2 DOPING PROFILES IN DIFFUSION

Concentration dependent diffusion flux is described by Fick's first law:

$$j = -D(\partial N/\partial x) \tag{14.2}$$

where D is the diffusion coefficient (cm^2/s), N is concentration (in cm^{-3}). The unit of flux is atoms/s*cm^2.

Diffusion coefficients can be presented by

$$D = D_o e^{(-E_a/kT)} \tag{14.3}$$

where

D_o is the frequency factor (related to lattice vibrations, 10^{13} to 10^{14} Hz)

E_a is the activation energy (related to energy barrier that the dopant must overcome)

Table 14.1 D_o and E_a values for boron and phosphorus

	Boron	Phosphorus
D_o (cm^2/s)	0.76	3.85
E_a (eV)	3.46	3.66

k is the Boltzman's constant, $k = 1.38 \times 10^{-23}$ J/K or 8.62×10^{-5} eV/K

T is the temperature in Kelvin.

The boron diffusion coefficient at 950 °C is $4 \times 10^{-15}\,cm^2/s$ and at 1050 °C it is $4.7 \times 10^{-14}\,cm^2/s$ (see Table 14.1). The characteristic diffusion length is given by

$$x \approx \sqrt{4Dt} \tag{14.4}$$

so that at 1050 °C boron diffusion for one hour corresponds to roughly 0.26 μm diffusion depth. This distance is a characteristic length scale only: diffusion profiles are gently sloping and there is no clear cut-off depth.

The sheet resistance of doped layers is given by Equation 14.5a and it is approximated for a box profile by Equation 14.5b.

$$1/R_s = \int_o^{x_j} q\mu(N(x) - N_b)dx \tag{14.5a}$$

$$1/R_s = q\mu x_j N(x) \tag{14.5b}$$

where q is the elementary charge, μ is the mobility, $N(x)$ is the dopant concentration, N_b is the background concentration and x_j is the junction depth. The mobilities of n-type and p-type silicon are ca. $1400\,cm^2/Vs$ and $500\,cm^2/Vs$ respectively, at low concentrations ($<10^{15}/cm^3$) and ca. $50\,cm^2/Vs$ at high concentrations ($>10^{19}/cm^3$), irrespective of dopant. In 1 μm CMOS technology source/drain diffusions are made by $5 \times 10^{15}/cm^2$ ion implant doses, and the depth is ca. 200 nm, which translates to ca. 25 ohm/sq. For more advanced

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

technologies the S/D sheet resistances are rapidly increasing because junction depths are scaled down.

14.2.1 Infinite dopant supply (constant surface concentration of dopant)

The infinite dopant supply corresponds to the gas-phase doping in which a new dopant is constantly being injected into the diffusion tube. A heavily doped thin film (polysilicon or CVD oxide) can act as an approximation to an infinite source when diffusion times and temperatures are moderate. Concentration profile of the dopant in silicon is given by the complementary error function (erfc):

$$N(x, t) = N_o \text{erfc}\,(x/\sqrt{4Dt}) \qquad (14.6)$$

where N_o is the dopant concentration ($1/cm^3$) in the surface layer, x is the depth (cm), t is the time (s) and D is the diffusion coefficient at a given temperature (cm^2/s). Longer doping times will lead to deeper diffusions but the surface concentration is unchanged.

14.2.2 Limited dopant supply (constant dopant amount)

The limited dopant supply case describes the case of pre-deposition: the dopants are definitely in limited supply because no new ones are introduced. This is the case of ion implantation. Longer diffusion times will lead to deeper diffusions but the surface concentration decreases.

The concentration profile is Gaussian:

$$N(x, t) = (Q_o/\sqrt{\pi Dt}) \exp(-(x^2/4Dt)) \qquad (14.7)$$

where Q_o is the total amount of dopant on the surface ($1/cm^2$). The junction depth is given by

$$x_j = \sqrt{4Dt} \times \ln(Q_o/C_{subs}\sqrt{\pi Dt}) \qquad (14.8)$$

This equation cannot be solved in an analytical form for diffusion time. An approximate solution for diffusion time can be obtained by a graphical solution: calculate x_j for a few diffusion times, plot the results and estimate the junction depth from the graph. Simulators are used for more accurate estimates.

14.2.3 Diffusion profile measurement

The diffusion profiles are measured either physically or electrically. The standard physical measurement is secondary ion mass spectrometry (SIMS). The dynamic range of SIMS is six to eight orders of magnitude, that is, dopant concentrations of 10^{14} to $10^{16}/cm^3$ can be detected (silicon atom density is $5 \times 10^{22}/cm^3$). The spreading resistance (SRP) measurement measures resistance with probes at the surface, and then bevelling or anodic oxidation is done in order to have access to the dopants deeper inside the silicon. SRP data needs some heavy calculations before dopant profiles are obtained. Both SIMS and SRP are sample destructive methods.

14.3 SIMULATION OF DIFFUSION

All the high-temperature process steps contribute to diffusion; therefore, diffusion is the omnipresent process to be simulated in the front end of the process. There can easily be tens of steps that contribute to dopant profiles. Segregation effects during oxidation and dopant outdiffusion from free surfaces add to computational and modelling loads.

Simulation of phosphorus diffusion needs to consider at least five species:

- phosphorus (P)
- vacancies (v)
- interstitials (i)
- phosphorus-vacancy pairs (P-v)
- phosphorus-interstitial pairs (P-i).

Vacancies and interstitials are not permanent species like phosphorus atoms, and we must account for annihilation of point defects via the reaction $v + i = nil$. Point defects can also form pairs like v–v. To make the situation even more difficult to analyse, many of the species are charged: diffusion models have to account for equilibrium processes like $P^- + v^o \Leftrightarrow Pv^-$ (charged phosphorus-vacancy pair) or $P^- + i^o \Leftrightarrow Pi^-$. Clustering and precipitation of dopants leads to inactivation. These phenomena are especially important when concentrations are near the solid solubility limit.

A standard simulator requires the following as inputs for diffusion simulation:

- wafer orientation <100>/<111>
- wafer-doping level/resistivity
- dopant type
- concentration of dopant (gas phase/solid phase/implanted)
- temperature
- ambient (oxidizing/inert/reducing).

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

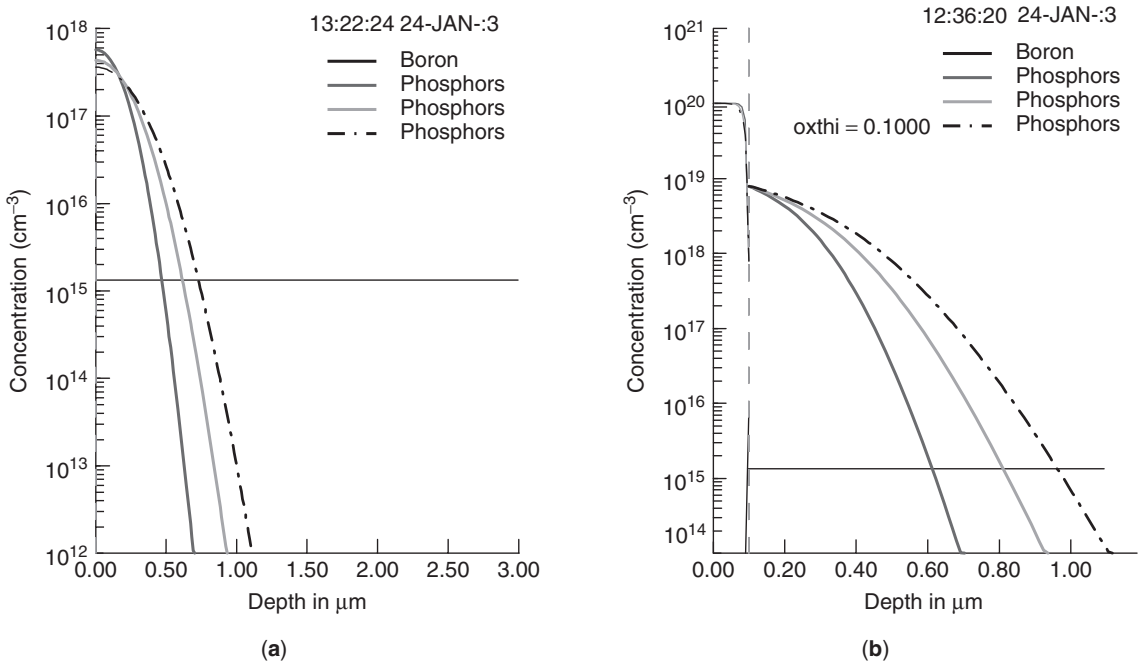


Figure 14.6 Diffusion at 1000 °C, for 100, 200 and 300 minutes in inert atmosphere: (**a**) diffusion from a limited source: implanted dose 10^{13}/cm^2 and (**b**) diffusion from phosphorus doped oxide film (with 10^{20}/cm^3 phosphorus concentration)

Doping profiles shown in Figure 14.6 have been calculated with the simulator ICECREM. The limited dopant supply case leads to lower surface concentrations for longer diffusion times; and the infinite supply case has constant surface concentration. Of course, the latter is just an approximation and it would not be valid for longer diffusion times or higher temperatures.

14.4 DIFFUSION APPLICATIONS

Thermal diffusion is the dominant method for high doping level and/or deep diffusion applications. In IC fabrication, thermal diffusion has largely been replaced by ion implantation because implantation is a more accurate method. But implantation is inherently slow, and therefore many non-critical steps are still done by furnace thermal diffusion: the furnaces are much simpler equipment than implanters. The double-sided nature of thermal diffusion is sometimes advantageous for volume devices.

Gas-phase doping by $POCl_3$ gas for n-type and BBr_3 gas for p-type was used in the early years of semiconductor manufacturing for steps in which a high degree of control was required, for example, bipolar base diffusion. Solid source doping was used when high dopant concentration (near or at solid solubility limit) was required, for example, in bipolar emitters and MOS source/drain. Solid source doping has the drawback that it is often very difficult to remove the dopant source material after diffusion and residues may be left.

Polysilicon deposition is generally done undoped. $POCl_3$ gas-phase doping is often used to dope polysilicon, but there is the alternative method of using solid P_2O_5 wafers: phosphorous oxide wafers and silicon wafers are set in alternating positions in a wafer boat, and at high temperatures the phosphorus will evaporate from P_2O_5 wafers and dope the silicon. Dopants arrive on the wafer from the gas phase, and dopant supply is practically infinite. Polysilicon sheet resistance can be as low as 10 ohm/sq, for 500 nm thick film. Ion-implantation doping will result in one to two orders of higher resistivity.

There are concentration and electric field effects that make actual device diffusions more complex than what the simple Fickian models predict. In emitter-push

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

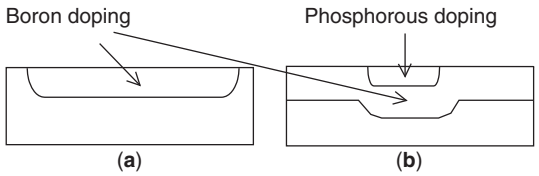


Figure 14.7 Emitter-push effect: (**a**) unimpeded boron diffusion and (**b**) boron diffusion under same conditions when phosphorus is present

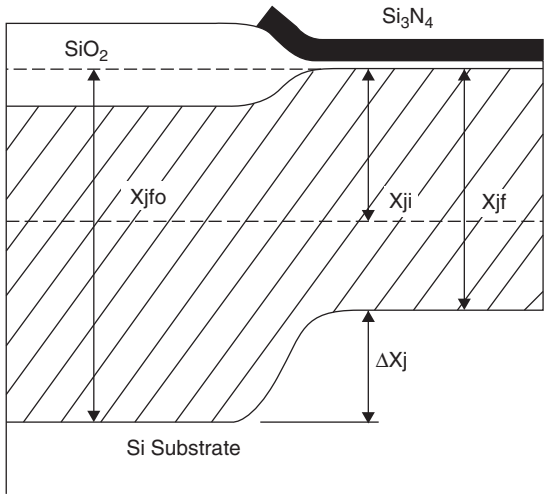


Figure 14.8 Oxidation enhanced diffusion (OED): vacancy injection during oxidation enhances dopant diffusion under oxide. Reproduced from Taniguchi, K. *et al*. (1980), by permission of Electrochemical Society Inc

effect, phosphorus diffusion enhances boron diffusion (see Figure 14.7). Boron diffusion alone would result in a profile predicted by simple theory, but boron diffusion under a phosphorus-doped region is much faster. This is explained by self-interstitial generation in the phosphorus diffusion process, and these interstitials enhance boron diffusion. In oxidation enhanced diffusion (OED) the vacancies generated by volume changes associated with thermal oxidation lead to enhanced diffusion underneath the oxide. This is pictured in Figure 14.8. Simulators can handle emitter-push effect, OED and high dopant concentration effects and other subtleties.

Diffusion is inevitable in all high-temperature steps, but it can be minimized by minimizing the process time. In rapid thermal annealing (RTA; or RTP for rapid thermal processing) wafers are heated rapidly by powerful lamps, and $\sqrt{4Dt}$ is brought down by annealing for very short times at high temperatures: whereas furnace anneal conditions are typically 950 °C, 30 min, corresponding RTA conditions are 1050 °C, 10 s.

14.5 EXERCISES

1. What is the diffusion time required to form a pn-junction at 1 μm depth in 1000 °C, when boron pre-deposition is $10^{14}/cm^2$ and phosphorus-doped wafer ($10^{15}/cm^3$) is used?
2. What is the sheet resistance of diffusion after anneal shown in Figure 2.9?
3. If deep n-type diffusions are needed, which n-type dopant should be used?
4. How far will metallic impurities diffuse during thermal oxidation?

5S. Which is faster, the diffusion of boron or phosphorus?

6S. Boron-doped oxide film (200 nm thick, concentration $10^{21}/cm^3$) is deposited on phosphorus-doped wafer ($10^{15}/cm^3$ phosphorus concentration). What is the junction depth doping after a 300 min, 1100 °C diffusion step?

7S. What is the magnitude of emitter-push effect?

8S. What is the magnitude of OED? Run some simulations to find which process parameters are important.

REFERENCES AND RELATED READINGS

Ghandhi, S.K.: *VLSI Fabrication Principles*, 2nd ed., John Wiley & Sons, 1994.

Taniguchi, K. *et al*: Oxidation enhanced diffusion of boron and phosphorus in (100) silicon, *J. Electrochem. Soc.*, **127** (1980), 2243.

Hull, R. (ed.): *Properties of crystalline silicon*, INSPEC, The Institute of Electrical Engineers (1999).

Zimmermann, H.: *Integrated Silicon Optoelectronics*, Springer, 1999, p. 36.

MRS Bull., **25**(6) (2000), special issue "Defects and diffusion in silicon technology"

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

15

Ion Implantation

Ion implantation is a process in which accelerated ions hit the silicon wafer, penetrate into the silicon, slow down by collisional and stochastic processes and come to rest within femtoseconds at the top micrometre layer. One application, introduction of dopants (As, P, B) into silicon, is by far the most important one, but implantation offers many possibilities. Heavy ions can modify materials by introducing damage and amorphization, which can sometimes be beneficial, even though damage in general is considered to be a drawback of implantation. Implantation of oxygen inside silicon, and subsequent silicon dioxide formation, is used to make SOI wafers.

Ion implantation can be used to produce a great variety of doping profiles inside silicon. Maximum dopant density need not be at the wafer surface; it can be at hundreds of nanometres deep inside the silicon (Figure 15.1). Implantation through the surface layers (e.g., SiO_2) is possible. Neither of these can be done with thermal diffusion. Lateral confinement of implanted dopants is better than in diffusion: sideways spreading under the mask is considerably less, as a rule of thumb, it is one-third of the vertical range, whereas diffusion is an isotropic process in the first approximation.

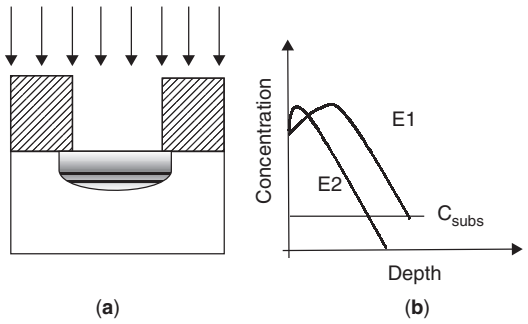


Figure 15.1 (**a**) Implantation with resist mask, with maximum concentration below the surface and (**b**) dopant profile in ion implantation (Energy 1 > Energy 2)

Implantation is a room-temperature process in theory. Photoresist masking is enough, which makes implantation easier than thermal diffusion, but implantation is always connected with a high temperature anneal step because introduction of dopants is not enough; the dopants have to be activated, that is, they have to find the lattice sites. Implantation also damages the silicon crystal, and in order to recover defect-free single-crystalline state, this damage has to be annealed away. Activation of dopants and damage removal can sometimes be one and the same anneal, but as will be discussed in the Chapter 25, this is not always straightforward.

15.1 THE IMPLANT PROCESS

Implanted ions scatter stochastically, travelling a distance R (range). However, we are more interested in the projected range, R_p, the range in the direction of the incident ion beam. Also of interest is the lateral straggle, R_L, or the deviation from the incident direction (Figure 15.2).

Ions are decelerated in the lattice by nuclear and electronic stopping, that is, by collisions with atomic nuclei of atomic number Z and mass M, and by collisions between the electronic cloud, respectively. Under a number of simplifying assumptions (about the nature of material, interaction potentials, energy independence of various variables, etc.,), the Linhard solution to nuclear stopping (S_n) for a projectile (M_1, Z_1) hitting a wafer of (M_2, Z_2) is

$$S_n = 2.8 \times 10^{-15}(Z_1 Z_2 / Z) \times (M_1/(M_1 + M_2))\ \text{unit: eVcm}^2 \quad (15.1)$$

where Z is the reduced atomic number, $Z = (Z_1^{2/3} + Z_2^{2/3})^{1/2}$. The nuclear energy loss is independent of ion

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

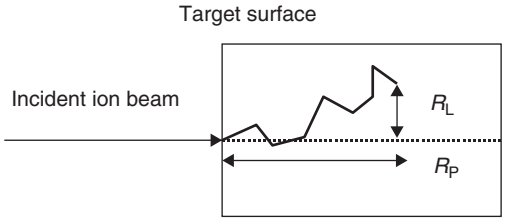


Figure 15.2 Key concepts for implanted ions: R_p projected range, R_L lateral straggle

energy in this approximation (Table 15.1). Electronic stopping is proportional to the square root of energy:

$$S_e = 3.3 \times 10^{-17}(Z_1 + Z_2)(E/M_1)^{1/2} \quad \text{eVcm}^2 \tag{15.2}$$

The total energy loss is calculated as

$$dE/dx = -(S_n + S_e)N \tag{15.3}$$

where N is the silicon atom density, $5 \times 10^{22}\,\text{cm}^{-3}$.

Combined energy loss from nuclear and electronic stopping for 100 keV phosphorus is 724 μm/keV. The range will then be ca. 0.14 μm (100 keV/724 μm/keV). With typical implant energies of 10 to 200 keV ranges are from 10 nm for 10 keV arsenic to 500 nm for 200 keV boron (Figure 15.3(a) and 15.4(a)).

Table 15.1 Energy loss of implanted ions in silicon

Nuclear stopping in silicon (independent of energy) in keV/μm			
Boron	92		
Phosphorus	447		
Arsenic	1160		
Electronic stopping in silicon in keV/μm			
E/keV	Boron	Phosphorus	Arsenic
10	65	88	90
50	145	196	200
100	205	277	283
200	290	391	401

The masking layer thicknesses for ion implantation will thus have to be of the same order of magnitude (Figure 15.3(b)). Photoresists suit ideally, and thermal oxides can be used. But unlike diffusion, oxides need not be grown specifically for implantation masking.

Thin oxides, in the 10 nm range, are grown on silicon before implantation for two reasons: implantation is a high-energy process, and accelerated ions sputter metal atoms from the implanter hardware. The thin oxide prevents these metal atoms from penetrating the silicon.

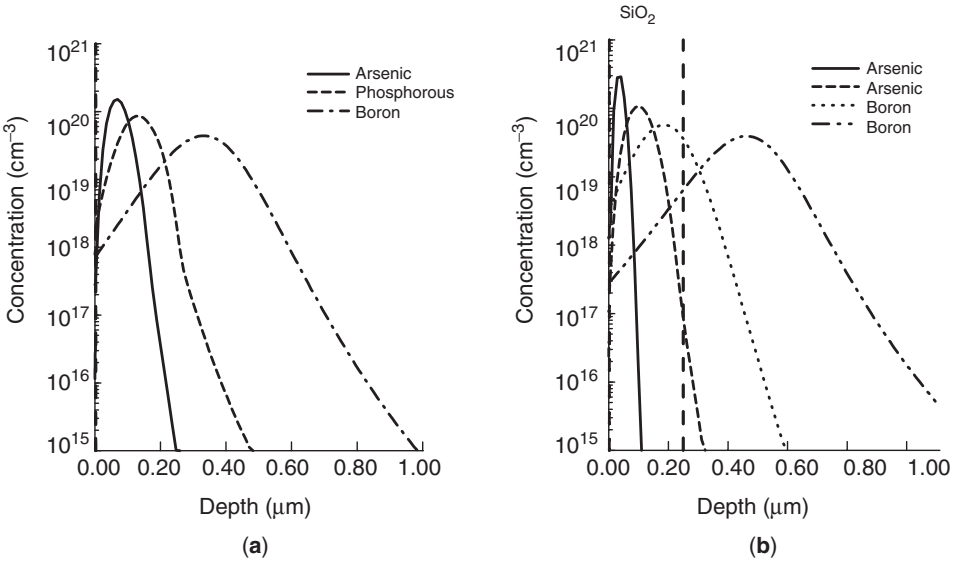


Figure 15.3 (**a**) 100 keV implantation of arsenic, phosphorus and boron: the lighter ions will penetrate deeper and (**b**) implantation through 250 nm thick oxide: most arsenic ions (both 50 keV and 150 keV) will remain in oxide, while boron (both 50 keV and 150 keV) will dope silicon

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

In the post implantation clean, this thin pad oxide and the metals on it can easily be removed by a HF dip. Thin oxides serve also to randomize incoming ions, which might otherwise penetrate deep into the silicon, guided by the crystal planes. This channelling phenomenon will be discussed shortly in connection with implant simulation.

15.2 IMPLANT DAMAGE AND DAMAGE ANNEALING

Nuclear stopping displaces atoms from the silicon lattice: a 100 keV arsenic ion displaces ca. 2000 silicon atoms along its trajectory. Damage creation depends on

- implant species (heavy ions produce more damage);
- energy (more energy, more damage);
- dose (above ca. $10^{14}/cm^2$ extended damage set in);
- dose rate (higher dose rate leads to overlapping collision cascades).

At low doses (below $10^{14}/cm^2$), the predominant damage type is point defects such as vacancies and interstitials, or clusters of point defects. At high doses extended defects are created, and even amorphization can take place. Dislocation loops are created in the crystalline silicon just next to the amorphous/crystalline interface. These are known as *end-of-range* (EOR) defects. If the concentration of dopants is above solid solubility limit, dopants precipitate.

Boron does not cause appreciable amorphization irrespective of dose because it is a light mass ion. High dose phosphorus and arsenic implants can amorphize silicon (Figure 15.4(b)), but if amorphization is needed without doping, germanium can be used. Critical dose for amorphization is ca. $10^{14}/cm^2$.

15.2.1 Measurements for implantation

Implanted wafers can be measured by a four-point probe (4PP) for sheet resistance. It is a natural control measurement for doping. It is, however, a fairly slow feedback loop because the wafer has to be cleaned and annealed before a 4PP measurement. A sheet resistance measurement sees only the electrically active dopants, and annealing is, therefore, not just an auxiliary step for measurement but an essential part of ion implantation doping. What is more, the wafer has to be discarded after a four-point probe measurement because the 4PP makes a metal contact with silicon, which causes contamination.

Alternatively, the dose can be monitored by a modulated photoreflectance (also known as *thermal waves*). A modulated laser beam heats the wafer and the thermal dissipation length is monitored by another

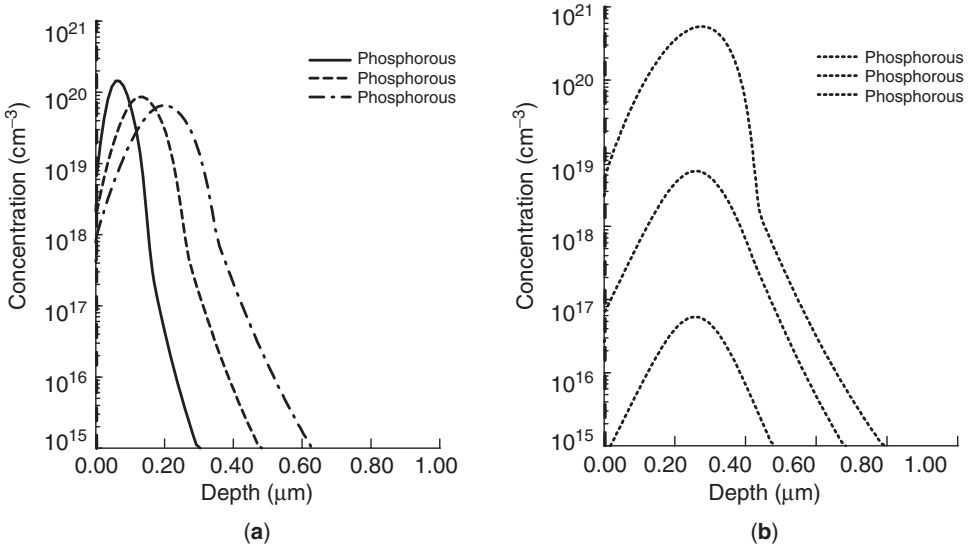


Figure 15.4 (**a**) Phosphorous implantations with different energies: 50 keV, 100 keV and 150 keV (dose constant $10^{15}/cm^2$). (**b**) Phosphorous implantations with different doses: $10^{12}/cm^2$, $10^{14}/cm^2$ and $10^{16}/cm^2$ (energy constant at 200 keV). The shape of dose $10^{16}/cm^2$ is different because it is above amorphization limit, and different stopping parameters are applied for the amorphized region

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

small power laser. The dissipation lengths are correlated to the implant damage, and therefore to the dose. This is a fast, non-contact, non-specific measurement, which needs no wafer preparation, and can be done even on photoresist-patterned wafers.

Point defects created by implantation cannot be seen by physical analysis, but extended defects like dislocations can be seen by TEM. Amorphization can be measured by TEM or by XRD.

15.3 ION IMPLANTATION SIMULATION

Implantation simulation must make a critical first choice in how to treat matter: amorphous matter is easy to model, but silicon really is single crystalline. Many simulators use single-crystal silicon materials parameters, but ignore the actual crystal structure.

The Monte Carlo (MC) simulation offers many advantages over semi-analytical implantation simulations because it can truly take silicon crystal structure into account. Channelling is a phenomenon in which ions are channelled between silicon crystal planes, rather like light in optical fibres. This effect is more pronounced for light ions, and for <100> crystal orientation than for <111>, which has a less open structure (see Figure 4.5). The Monte Carlo simulation can predict not only ranges and straggle, but it also enables physically based damage prediction, including amorphization. The MC simulations are, of course, more computational intensive than the semi-analytic ones. The simulator SRIM (Simulation of Ranges of Ions in Matter) is a widely used MC simulator for implantation and other ion-beam processes.

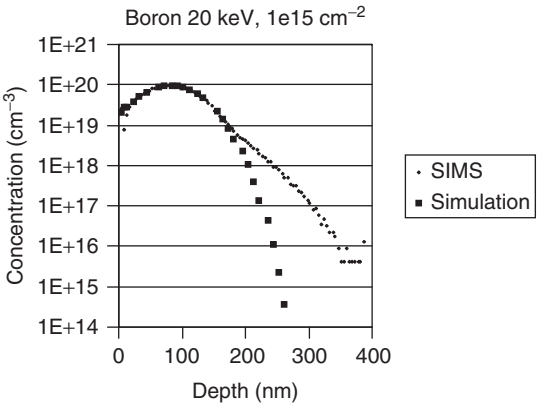


Figure 15.5 Boron implantation into silicon, 20 keV, 1.1015 cm^2. SIMS measured data shown in small markers, ICECREM simulation with large markers. The discrepancy in the tail results partly from ion channelling and partly from model deficiencies. SIMS data courtesy Jari Likonen, by permission of VTT

Input for a prototypical semi-analytical implantation simulation includes:

- wafer type and dopant concentration
- ion specie
- energy
- dose.

The accuracy of the simulation is very good in the peak concentration regime, but worse at the tail of the distribution (Figure 15.5). This is partly due to the ion channelling that is not readily implemented in semi-analytical moment-based simulators. For heavier elements, discrepancies can come from amorphization treatment: a single crystal material parameters may be used initially, but as the dose increases, the simulator adopts amorphous silicon material parameters for further calculations.

15.4 TOOLS FOR ION IMPLANTATION

Ion implantation acceleration voltages used to range from 20 kV to 200 kV, but today low-energy implanters (1 keV minimum) and high-energy implanters (HEI) (max. 2 MeV) exist. Low-energy implants are needed to fabricate shallow source/drain junctions (of the order of 100 nm) in deep submicron CMOS. High-energy implanters implant deep into silicon, one micrometre or even deeper. The ability to fabricate retrograde profiles, that is, to have low concentration at the surface, and high concentration deep down, exactly opposite to thermal diffusion, offers some interesting possibilities, for example, as replacement for buried layers and epitaxy.

Medium current implanters (MCI) are 20 to 200 keV, single-wafer machines, whereas, high-current implanters (HCI) are batch machines with minimum energy of ca. 80 keV. The extraction beam current scales as $V^{3/2}$, which explains why a low voltage HCI is not practical. This scaling means difficulties for low-energy, high-dose implantation that are needed for advanced CMOS source/drain implants.

Implant currents can be anything from 1 μA to 30 mA, and doses range from 10^{11}/cm^2 to 10^{16}/cm^2 in standard use. The beam currents are limited if photoresist is used as a mask: too high currents will damage the resist, and removal of the resist becomes difficult. Cooled wafer stations can be used to minimize the resist damage.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

The scaling down of ion energy involves a number of techniques. One of the oldest techniques is to implant molecular ions instead of ions: BF_2^+ has a mass of 49 versus 11 for that of boron, and its range is ca. a fifth of the boron range in the first approximation. The replacement of B for BF_2^+ is not straightforward, however, because the behaviour of fluorine during annealing and further processing needs to be accounted for. True low energy implanters must accept the fact that a lower beam current is available. In the limit of 1 keV, the sputtering of the surface atoms becomes important: because the low implant energy equals the low penetration depth and every atom layer removed from the surface will affect the final implant profile.

15.4.1 Implanter design and operation

Implantation requires ions, and these are generated in ion sources that are plasma discharges. The dopants have to be vapourized or be in the gaseous state before ionization. The dopant gases in routine use are PH_3, AsH_3 and BF_3, but evaporation of solids in a furnace can also be used, and almost all elements in the periodic table can be implanted. However, efficiency of the solid sources is low and switching between the ions is slow. The ions are extracted from the source by voltage, and enter the selection magnet (Figure 15.6).

Ion selection is based on mass spectrometric separation according to the radius of curvature r in a magnetic field $\boldsymbol{B}$ balanced by the centrifugal force:

$$|\boldsymbol{F}| = |q(\boldsymbol{v} \times \boldsymbol{B})| = m|\boldsymbol{v}|^2/r = qV \qquad (15.4)$$

where m is the mass and q is the charge which can be solved for $B = \sqrt{(2mV/qr^2)}$. By adjusting the magnetic field of the selection magnet, an ion of the desired mass is selected. The magnet selection can be fooled by similar ion masses, termed mass contamination. Doubly charged molybdenum ions Mo^{+2} can pass along with BF_2^+ ions (molybdenum is a common construction material for vacuum equipment). $^{11}BHF^+$ ion behaves like a $^{31}P^+$ ion for the selection magnet. This situation might emerge when PH_3 gas is used after BF_3 gas and some residual gas remains in the ion source. Energy purity refers to the spread of ion energies in the beam, and consequently, their range in silicon.

The acceleration tube must be kept under high vacuum in order to steer the beam to the wafer in a collisionless fashion. After acceleration, either electromagnetic or mechanical scanning spreads the beam over the wafer. Implantation is an inherently slow process because of the scanning nature of the operation. Alternative implantation techniques that work in parallel mode have been devised: plasma immersion ion implantation (PIII) is a process in which the wafer is immersed in plasma, and biased. Very high-dose rates are possible, but the energy purity is sacrificed because the selection magnet has been eliminated from the system. A PIII may have applications in large-area applications like flat-panel displays because of its high throughput.

The wafers will be charged when ions are implanted. The current flows from the beam to the wafer holder, and it passes any oxides on its way. Also, beam nonuniformity between the wafer centre and the edge can cause lateral currents. Charging is compensated by flooding: electron gun generated electrons hit the wafer and neutralize the charges. This approach is prone to overcompensation and problems with electron charging. The plasma discharge, which produces an order of magnitude of higher ion density than the beam, is used in neutralization. Charge neutrality is inherent in the plasma system.

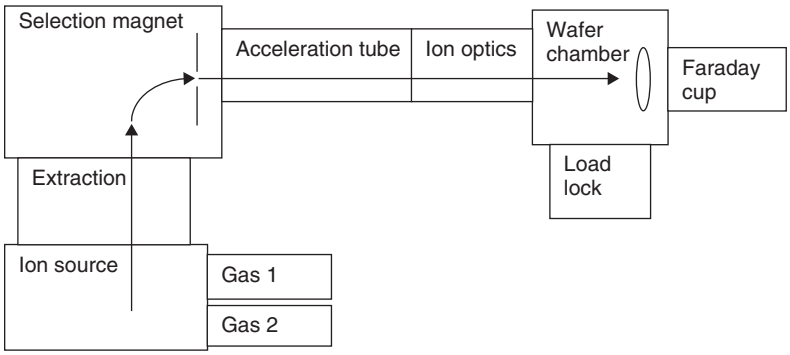


Figure 15.6 The main elements of an implanter: ion generation in the source, extraction of ions, selection by magnet, acceleration, beam shaping and scanning optics and wafer stage. Adapted from Current, M. (1996), by permission of AIP

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Implant dose is monitored during implantation by the Faraday cup current measurement. This is the basis for the high degree of doping control in implantation as compared to diffusion, which has no, whatsoever, *in situ* monitoring method.

15.4.2 Safety aspects

Ion implanters pose a number of safety issues that have to be tackled. The obvious one is the high voltage that is present inside the machines. The second issue is X-rays that are produced as ions decelerate. Lead radiation protection is routinely used around the parts where X-rays are generated. If hydrogen is implanted, as in the Smart-cut process (to be presented in Chapter 17), nuclear reactions are possible at fairly low energies of 150 keV and gamma rays are then generated.

Implant gases AsH_3, PH_3 and BF_3 are extremely toxic. Toxic gas detectors are placed inside the system to sniff for leaks. Operation and maintenance of an implanter can, therefore, be carried out by highly trained staff only. More discussions on safety issues can be found in connection with cleanrooms, in Chapter 35.

15.5 SIMOX: SOI BY ION IMPLANTATION

In SIMOX technology, a SOI structure is realized in two main steps. The first step is oxygen implantation into a silicon wafer and the second step is a high-temperature anneal during which the implanted oxygen atoms form an oxide layer inside the silicon (Table 15.2). This oxide is known as *buried oxide* (BOX). The top silicon layer, known as the *device layer*, becomes insulated from the bottom layer, known as the *handle*.

SIMOX material exhibits inherent defect problems: the device silicon layer is damaged by the implantation process and it cannot be fully recovered during annealing. Its dislocation densities can be a million/cm^2, orders of magnitude more than in bulk silicon. Implantation time poses another limitation: the required doses are two orders of magnitude higher than those in common usage. A low dose SIMOX with $4 \times 10^{17}/cm^2$ implantation helps to minimize both the aforementioned problems. There are further limitations that are inherent to the implant process: with 200 keV maximum energy, the implant depth is fairly shallow and, therefore, the device silicon thickness is rather limited. The thickness of buried is also limited by the implant process.

Table 15.2 SIMOX process

Implant conditions	
Oxygen dose	$2 \times 10^{18}/cm^2$
Oxygen energy	150–200 keV
Wafer temperature	550–650 °C
Anneal conditions	
Temperature	1300–1350 °C
Time	4–6 h
Atmosphere	Ar + 0.5% oxygen

15.6 EXERCISES

1. What will be the implant time for a 200 mm diameter wafer, when arsenic ions are implanted with doses of $10^{15}/cm^2$ and implant current of 100 μA?
2. What is the range of 20 keV $^{11}B^+$ and $^{49}BF_2^+$ ions?
3. How thick a silicon dioxide layer will be formed inside the silicon when the implant dose is $2 \times 10^{18}/cm^2$ in SIMOX?
4. What is the range of 100 keV germanium implantation?

5S. How thick an oxide layer is needed to mask boron implantation? Present your results as a function of boron energy.

6S. Check by simulator the range of 100 keV phosphorus ions and compare it with the simple estimate discussed in the text.

7. At what energy is electronic and nuclear stopping equal for phosphorus?

REFERENCES AND RELATED READINGS

Chanson, E. *et al* Ion beams in silicon processing and characterization, *J. Appl. Phys.*, **81** (1997), 6513–6561.

Cheung, N.: Plasma immersion ion implantation for semiconductor processing, *Mater. Chem. Phy.*, **46** (1996), 132.

Current, M.: Ion implantation for silicon device manufacturing: a vacuum perspective, *J. Vac. Sci. Technol.*, **A14** (1996), 1115.

Izumi, K.: History of SIMOX material, *MRS Bull.*, **23**(12) Special issue on Silicon-on-insulator technology (1998), 20.

LeCoeur, F. *et al*: Ion implantation by plasma immersion: interest, limitations and perspectives, *Surf. Coat. Technol.*, **125** (2000), 71.

White, N.R.: Moore's law: implications for ion implant equipment – an equipment designer's perspective, *Proc. 11th Intl. Conference on Ion Implantation Technology* Austin (1996), p. 355.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

16

CMP: Chemical–Mechanical Polishing

Material removal from a wafer is usually done by etching, but there is the alternative technology of polishing. Polishing is an established technology in silicon-wafer manufacturing where final polishing yields wafers with a root mean square (RMS) roughness of 1 Å, but it emerged in microfabrication only in the late 1980s. In microfabrication, polishing and etching processes can be combined to yield identical final structures via different process sequences, as shown in Figure 16.1: metal lines can be made either in the following sequence:

metal deposition ⇒ metal etching

⇒ oxide deposition ⇒ oxide polishing

or in the sequence

oxide deposition ⇒ oxide etching

⇒ metal deposition ⇒ metal polishing

The latter sequence, known as *damascene*, is used for metals that cannot be plasma-etched, and it is the key technology to copper metallization of ICs.

Polishing in microfabrication is a descendant of glass polishing, which has been an established technology for 400 years. Abrasive particles are dispersed in a suitable liquid to create a slurry, which is fed in between a polishing pad and the piece to be polished. Elevated structures are preferentially removed since the pressure is highest there. In the case of a blanket, wafer-surface irregularities are smoothed out.

Grinding may look similar to CMP, but the two are quite different. In grinding, abrasive particles of 1 to 100 μm in size are mounted in resin, and micrometre-sized chunks of material are removed by crack propagation and brittle fracture. Grinding is fast but also very coarse; the substrate is damaged due to mechanical forces acting on microstructures. This subsurface damage is 5 to 10 μm deep. Grinding is used when hundreds of micrometres need to be removed, as in wafer thinning. CMP removes micrometres only, and the resulting surfaces are very smooth and defect free. In CMP, abrasive particles of 10 to 300 nm are dispersed in a slurry. The mechanism is different from grinding: CMP works in the atomic regime. Atomic bonds are weakened or broken, and removal is based on the interaction between the slurry and the mechanical effect of the abrasive particles. Surface roughness after CMP is in the nanometer range, while grinding results in hundreds of nanometres.

16.1 CMP PROCESS AND TOOL

The CMP tool consists of a solid, extremely flat platen, on which the polishing pad is glued. The wafer chuck, which holds the wafer upside down, is situated on a spindle. A slurry introduction mechanism feeds the slurry on the pad. Both the platen and the spindle are rotated, and the linear velocity (used in Preston's equation) is the sum of two velocities (Figure 16.2).

There are four major elements in a CMP process:

- topography
- materials
- polishing pad
- slurry.

Down force is an average force, but local pressure is needed to understand removal mechanisms. It depends on the contact area, which in turn depends on both the structures on the wafer and on the pad structure. Pads are rough, with say 50 μm roughness, and contact is made by asperities, and the contact area is only a fraction of the wafer area (Figure 16.3).

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

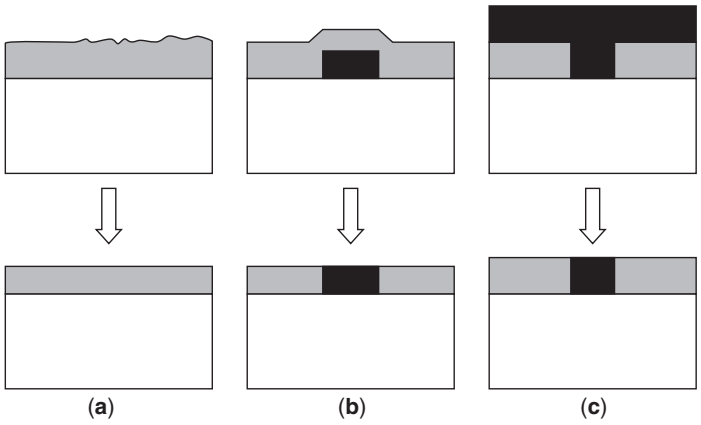


Figure 16.1 Applications of polishing: (**a**) smoothing; (**b**) planarization and (**c**) damascene

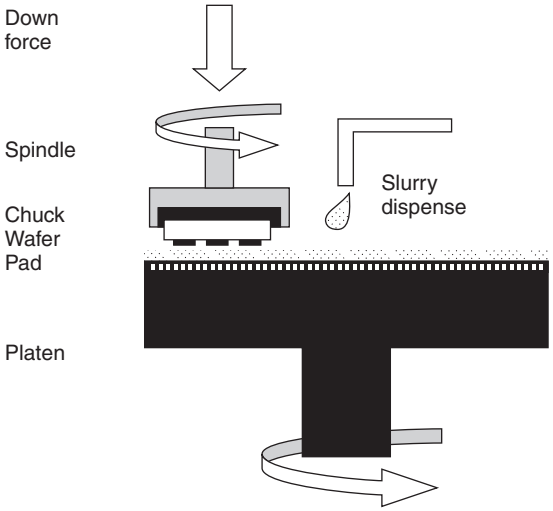


Figure 16.2 Schematic structure of a rotary CMP equipment

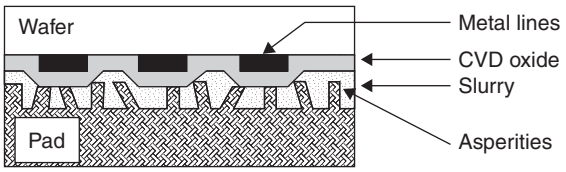


Figure 16.3 Close-up of CMP set-up: wafer, upside down, is pressed against the pad with slurry in between. Pad asperities make contact with the wafer

Structure height obviously affects CMP, but pattern density is also important because it determines effective contact area: denser patterns are polished at a lower rate due to lower pressure. Polishing of a single material is easier than polishing stacks of materials, or structures with different materials present simultaneously. The mechanical properties of the wafer itself must also be considered: if it is bowed, the pressure will be different at the centre and the edges, leading to non-uniform polishing. Pressure can be applied through the chuck to the wafer backside: this will equalize centre–edge differences and compensate for wafer bow.

The pad should be rigid so that it uniformly polishes the wafer. However, such a rigid pad will have to be aligned and kept in alignment with the wafer surface at all times. Therefore, real pads are often stacks of soft and hard materials that conform to wafer topography to some extent. Pads are porous polymeric materials (with 30–50 μm pore size) that are consumed in the process and must be reconditioned regularly. Polyurethane is commonly used for pads. Pads are very much proprietary, and people usually refer to pads by their trade names, rather than by chemical or other unambiguous properties.

Slurries incorporate both mechanical elements via abrasive particle size and hardness, and chemical effects via reactivity and pH of the fluid. Typical slurry materials are silica (SiO_2) and alumina (Al_2O_3), with some experiments being carried out on cerium oxide (CeO_2). Abrasive particle-size distribution is related to smoothness: monodisperse slurry leads to smoother surfaces. Copper can be polished in ammonia-based slurry with 2% NH_4OH and abrasive particles of Al_2O_3 at 2.5%wt concentration. Slurries are a cause of concern for post-CMP: particles must be cleaned away after polishing. Like pads, slurries are often proprietary, and the information given is often restricted to pH value, base liquid (for instance, NH_4OH-based) and abrasive particle size. Slurries can be buffered against

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

consumption in the process (cf. etching in buffered HF). At the end of CMP, a soft polishing step is often done: no slurry is used, just water. This step does not remove solid material but is effective in washing away abrasive particles and corrosive chemicals.

CMP tool input variables include the following:

– platen rotation	10–100 rpm
– velocity	10–100 cm/s
– applied pressure (load)	10–50 kPa
– slurry supply rate	50–500 ml/min

Pad type, compressibility, hardness and elastic modulus, conditioning, pore size and ageing can be considered variables too. Because there is a chemical component in CMP, temperature will have an effect on polishing results.

CMP process factors resemble those encountered in etching:

- polish rate
- selectivity
- overpolish time
- pattern density effects
- uniformity across wafer
- wafer-to-wafer repeatability.

Plasma etching and CMP resemble each other also in the sense that both depend on interaction between chemical and physical processes: in etching, ion bombardment removes reaction products from surface; in CMP, mechanical abrasion removes surface layers that have been modified chemically, for instance, by oxidative slurries.

Polish rate can be limited by transport of reactants, or by surface processes, just like etching. This can be found out by varying the input variables: if the rate is unaffected by change in a variable, it cannot be the rate-controlling factor. Another similarity is pattern dependency: small pattern density leads to higher rates. Pattern size effect is, however, opposite: in CMP, small patterns are polished faster, but, in etching, small patterns will be etched slower than large ones. This will be discussed in Chapter 20.

16.2 MECHANICS OF CMP

There are three modes in polishing, depending on the degree of contact between the pad and the wafer. In the direct contact (boundary lubrication) mode, the pad makes contact with the wafer, resulting in high and constant friction because there is no lubrication from the slurry. Polish rate is very high. In the rolling contact mode (mixed lubrication mode), slurry particles occasionally roll on the wafer surface. In the non-contact mode (hydrodynamic lubrication mode), slurry particles are accelerated hydrodynamically and they impart energy to the wafer surface, weakening the surface so that chemical attack can occur. Hydrodynamic lubrication takes place at high velocities at which the load is borne by the fluid, and the system is well lubricated. Friction force between the pad and the wafer is very different in these modes and it is classified in a Stribeck diagram (Figure 16.4).

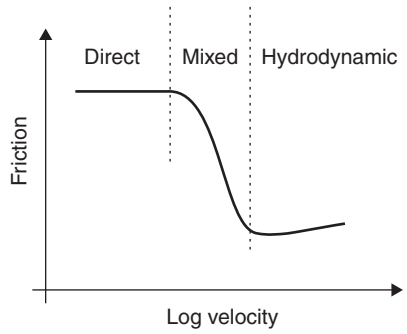


Figure 16.4 Stribeck diagram of CMP: three different lubrication modes

The penetration of the abrasive particles into the substrate is very small indeed: this is the reason for smooth surfaces with no visible grooves or scratches. Penetration depth is given by

$$R_s = (3/4)d(P/2kE)^{2/3} \qquad (16.1)$$

where d is the abrasive particle diameter (e.g., 100 nm), k is the filling factor of abrasive particles (for instance, 50%), P is the local pressure (not down force, which is 10–50 kPa) and E is Young's modulus of the surface being polished. Penetration depths are of the order of nanometres, which is similar to surface roughness after polishing, as would be expected. Increasing pressure will lead to deeper penetration but also to higher removal rate. Sometimes, the abrasive particles agglomerate into huge chunks, and this leads to much larger penetration depths and will result in microscratches that are tens of nanometres deep.

16.2.1 Preston model

Polish rates have been measured experimentally by Preston (in 1927) to obey the following equation:

$$R = \Delta H/\Delta t = K_p P(\Delta s/\Delta t) \qquad (16.2)$$

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

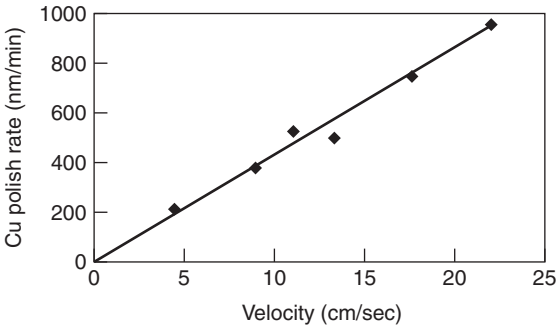


Figure 16.5 Copper polish rate as a function of velocity (15 kPa pressure). Reproduced from Steigerwald, J.M., S.P. Murarka & R.J. Gutman (1997), by permission of John Wiley & Sons

ΔH = change in the height of the surface
P = pad pressure
K_p = Preston coefficient
$(\Delta s/\Delta t)$ = linear velocity of the pad relative to the wafer.

Experimental results show a fairly good fit for Preston's equation, especially in the low-pressure/low-velocity regime, that is, in the direct contact mode (Figure 16.5).

The Preston coefficient is related to the elastic properties of the material, and it can be approximated by

$$K_p = 1/(2E) \tag{16.3}$$

where E is Young's modulus.

With Young's moduli in the range of 100 GPa for many inorganic and metallic solids, K_ps are of the order of 10^{-11} Pa^{-1}. Applied pressures are of the order of 10 kPa, and velocities, of the order of 0.10 m/s, which leads to polish rates of the order of 10 nm/s or 600 nm/min, which is the correct order of magnitude. This estimate is, however, not accurate enough to be of predictive use. It explains, however, many basic features of polishing; for instance, the fact that hard materials are polished at a lower rate than soft materials.

Local polishing pressure is load-divided by contact area. For a flat wafer, pressure is low because the load is evenly distributed over the whole geometrical area, but on a structured wafer, the effective contact area is only a fraction of wafer area, and the local pressure is much higher. Polishing rate is thus not constant: when the contact area is small, local pressure is high, and polishing rate is high. As polishing continues, steps are reduced and contact area increases, leading to rate decrease.

16.3 CHEMISTRY OF CMP

In chemical–mechanical polishing, there are two components: in addition to the mechanical pressure, chemical modifications and etching take place. For instance, a tungsten surface is turned into tungsten oxide according to the following equation:

$$W + 6Fe(CN)_6^{3-} + 3H_2O \longrightarrow WO_3 + 6Fe(CN)_6^{4-} + 6H^+$$

Tungsten oxide has two important roles: it is a protective layer, and, in the valleys, it protects the tungsten from further chemical attack. However, it is a mechanically weaker and more brittle material than tungsten, and, in the high points, it can be removed by mechanical abrasion. The same mechanism is at work in copper polishing: Cu_2O is removed by mechanical action while copper is not. For hard materials like tungsten and tantalum, the mechanical effects are usually important, whereas for soft materials like aluminium and polymers, the chemical effects often dominate.

When WO_3 is removed by polishing, the underlying metal is etched according to

$$W + 6Fe(CN)_6^{3-} + 4H_2O \longrightarrow WO_4^{2-}(aq) + 6Fe(CN)_6^{4-} + 8H^+$$

Possible corresponding reactions in copper polishing are

$$Cu \Leftrightarrow Cu^{2+} + 2e^-$$
$$2Cu^{2+} + H_2O + 2e^- \Leftrightarrow Cu_2O + 2H^+$$

Copper polishing is carried out with slurries based on $Fe(NO_3)_3$ and H_2O_2. Hydrogen peroxide oxidizes copper, which enhances removal rate. Typical rates are 100 to 1000 nm/min, selectivity to oxide ranges from 40:1 to 200:1 and residual step height, 100 to 300 nm. Copper polishing uniformities can be 10 to 15%, which is among the worst uniformities of any microfabrication process.

Aluminium polishing can be done in acidic solutions, for instance, phosphoric acid (pH ca. 3–4) with alumina abrasive. Aluminium CMP proceeds by aluminium oxidation and mechanical removal of the oxide, not unlike copper and tungsten polishing. Selectivity to oxide can be 100:1.

Oxide polishing slurries are ammonia or KOH-based, for instance, 1 to 2% NH_4OH in DI-water, with up to 30% silica abrasives of 50 to 100 nm. Oxide polishing slurries are mildly alkaline, with pH values of ca. 11. The oxide polishing mechanism depends on surface

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

modification of the oxide: leaching of oxide by the slurry softens the top layer, and the mechanical abrasion rate goes up.

CMP slurries etch without mechanical polishing, just like fluorine etches silicon without plasma; but in both etching and CMP, it is the interaction between different processes that leads to the desired total process: slurry etch rates of 10 nm/min are typical, but CMP removal rates of 500 nm/min are standard.

16.4 APPLICATIONS OF CMP

Conformal deposition processes replicate the underlying topography dutifully. Such processes are useful in gap filling: small spaces between lines are completely filled without any voids. However, this argument does not hold for larger linewidths: step height is unchanged after conformal deposition, as shown in Figure 16.6(a).

Some deposited CVD films flow, or have flow-like profiles, resulting in profiles like the one shown in Figure 16.6(b). Spin-on dielectrics flow over the topography, but the planarization length (Figure 16.7) defined as

$$R = h / \tan\theta \tag{16.4}$$

is in the range of micrometres or tens of micrometres in the maximum, as shown in Figure 16.6(c). CMP is the closest you can get to global planarity.

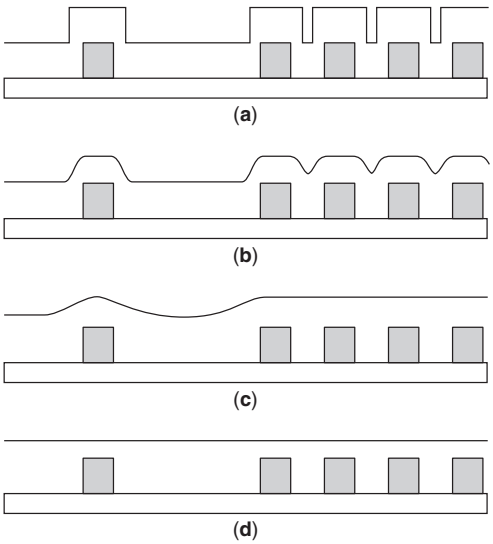


Figure 16.6 Planarity: (**a**) conformal deposition, no planarization; (**b**) surface smoothing during deposition; (**c**) local planarization by spin-film and (**d**) global planarization by CMP

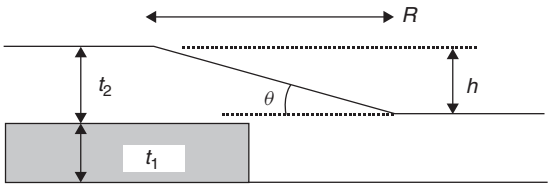


Figure 16.7 Planarization relaxation distance R

Polishing rate and planarization rate are two different concepts. Polishing rate is applicable to one material. Planarization rate is the rate of decrease in step height: the high peaks are polished, which decreases step height, but some material is removed from the valleys too, which decreases the planarization rate. Towards the end of the process, the planarization rate drops to zero, even though the overall polishing rate is still finite.

Selectivity in CMP bears close resemblance to etching: we need to know the polish rates of the top and bottom films in order to calculate, for instance, substrate loss during overpolishing. Identically to etching, it is sometimes beneficial to have the same 1:1 selectivity between films, but, most often, it is desirable to remove one film relatively rapidly, and to have high selectivity against the bottom film, which can then be processed in a separate step.

Oxide polishing is the oldest and most widely practiced CMP process. Its main application is planarization in multi-level metallization in advanced ICs, where it provides a planar surface that makes subsequent lithography and deposition steps easy. One problem with oxide polishing is the lack of endpoint: there is no clear end for polishing. This is called *blind polishing*. The opposite is *stopped polishing*, in which, for instance, a nitride layer acts as a polish stop (cf. etch-stop layer) but selectivities are not necessarily very high.

Tungsten polishing is another CMP process that was adopted rapidly. Contact holes and via holes are filled by CVD tungsten, which is then removed from planar areas, leaving just the contact plug filled with metal (Figure 16.1(c)). The same structure can, of course, be obtained by tungsten etchback, and the first implementations of tungsten plug process did use etchback. CMP has proven to be better with respect to plug loss: at etching end point, the etchable area decreases dramatically and the etchant will attack the tungsten in the plug, leading to severe plug recess. CMP is much better in this respect, but, naturally, process optimization with either technology can bring about improvements.

CMP is used whenever global planarity is required. In addition to multi-level metallization for ICs, other applications have sprung up. In superconducting quantum

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

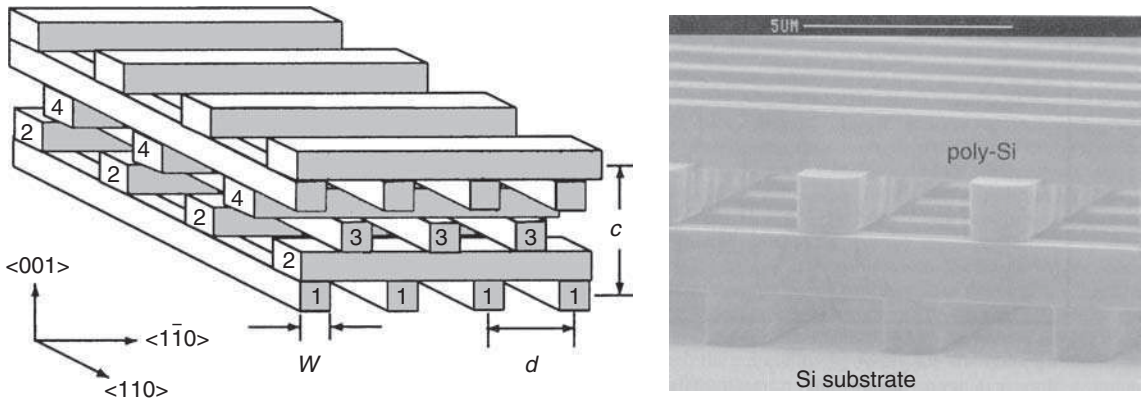


Figure 16.8 Infrared wavelength selective photonic lattice has been made with the help of CMP: oxide deposition, oxide trench etching, polysilicon LPCVD trench filling and polysilicon CMP have been repeated five times to create the lattice. As the last step, all oxide has been etched away in HF. Reproduced from Lin, S.Y. *et al*. (1998), by permission of Nature

interference devices (SQUIDs), CMP planarization of PECVD oxide is performed before metallization to eliminate step coverage problems and conductor cross-section variation to ensure high and constant current density, up to 10^7 A/m^2.

Photonic crystals (photonic band gap materials) are artificial lattices in which electromagnetic wave propagation is selectively restricted due to forbidden energy levels. There are many ways to fabricate photonic lattices (recall Figure 11.3), and CMP is just one approach. Grooves are etched in oxide, and filling material is deposited by CVD; polysilicon and tungsten are typical materials. CVD film is then chemical–mechanical polished and the process is continued until the desired number of layers has been made. Oxide is finally etched away to create the air gaps (Figure 16.8).

16.5 CMP CONTROL MEASUREMENTS

Top view microscopy, either optical or SEM, can be used for cross-checking CMP. Stains from slurry residues, scratches, layer peeling and other coarse problems can be identified. Scanning probe methods, mechanical stylus and AFM, are widely used to study micrometer-scale phenomena (Figure 16.9). Sub-micron resolution is needed because many CMP effects are strongly feature size dependent. Many optical, electrochemical, mechanical, thermal and acoustic methods are being developed to monitor CMP in real time.

16.6 NON-IDEALITIES IN CMP

CMP is an interplay between many process factors. Pressure, velocity, slurry composition and so on can be varied for optimization, but device design cannot usually be changed (even though sometimes dummy patterns are made, in order to make CMP and etching processes easier). Polish stop layers add process complexity too, but improved process control can balance the cost. Polish selectivities are similar to etch selectivities: they range from 1:1 to 200:1; for example, copper to oxide selectivities are 40:1 to 200:1, and copper to tantalum selectivities are so high that measurements are difficult. Oxide to nitride selectivities can be 50:1, and this is useful in shallow trench isolation, which will be discussed in Chapter 25.

Because of finite selectivity, some underlying layer loss is unavoidable. This is termed *erosion* and is pictured in Figure 16.10. Another non-ideality is the *dishing*. It is caused by two factors: the pad conforms to some extent to the structures on the wafer and softer material is polished faster than the surrounding hard material. Recess etching is a chemical effect. Recess in CMP can be as low as few tens of nanometres and, in this respect, CMP is superior to etchback.

Copper dishing is strongly feature size dependent, but rather insensitive to pattern density. Oxide erosion, on the other hand, is strongly pattern density dependent, but feature size independent.

On the practical side, slurry cost is a major problem. Slurries are consumables with very low utilization:

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

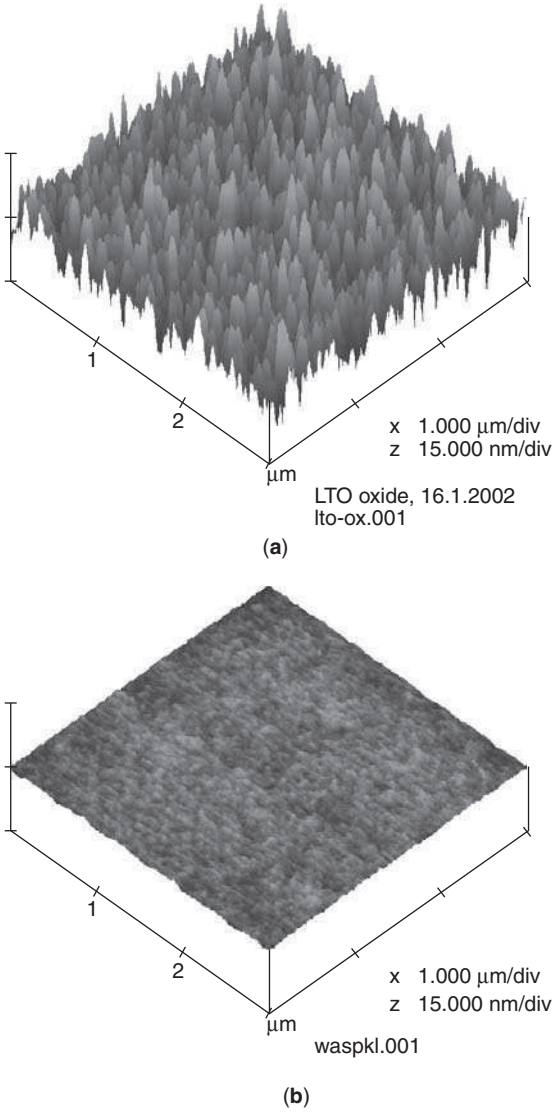


Figure 16.9 Surface roughness of CVD oxide by AFM: (**a**) as deposited film peak-to-valley height is 26 nm, with RMS roughness of 3.3 nm and (**b**) after CMP peak-to-valley is 2 nm and RMS roughness is 0.2 nm. Figure courtesy Kimmo Henttinen, by permission of VTT

in some processes, it is estimated that only 2% of slurry actually participates in the process, the rest is swept away by platen rotation. Various solutions to this problem are being investigated: structured pads with grooves and channels of various shapes retain the slurry better, and also result in more uniform slurry distribution, leading to better uniformity. Another solution is to use fixed abrasive: the abrasive particles

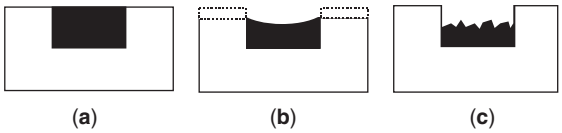


Figure 16.10 (**a**) Ideal CMP result; (**b**) erosion and dishing and (**c**) plug recess (chemical attack)

are attached to the pad, and the slurry is replaced by particle-free chemicals.

Temperature is not constant during CMP: friction easily leads to 10 °C temperature rise, which is detrimental to reproducibility and uniformity. Rates of chemical reactions go up as expected, and this temperature rise can easily double the removal rate. Pad hardness decreases as temperature goes up, which leads to more asperities in contact with the wafer and reduced local contact pressure. This effect, is, however, not significant compared to chemical rate increase.

16.6.1 Post-CMP cleaning

The introduction of CMP was obviously resisted by many people because the very idea of bringing zillions of particles, intentionally, on the wafer was against all accepted cleanroom and manufacturing policies. Post-CMP cleaning was, and remains, a topic of paramount importance. Brush cleaning and other physical cleaning techniques are good for rather large particles, but as always, the smaller particles pose problems. RCA-1 cleaning is efficient in particle removal, but its use is limited on metallized wafers. In addition to the particle problem, there is metal contamination: potassium hydroxide is a common slurry liquid, and copper residues may be embedded in PSG, which is a soft material. HF etching can remove a thin top layer of PSG, and reduce the amount of copper. In order to minimize particle and chemical contamination from spreading, the CMP section is usually separated from the rest of the fab, and DI-water is drained immediately after use, even though used DI-water is normally recycled.

16.7 EXERCISES

1a. What is the Preston's coefficient for copper on theoretical grounds?
1b. What is the experimental value of Preston's coefficient? Use data from Figure 16.5.
2. How do the polish rates of tungsten, silicon dioxide and polymers compare with each other?
3. How do polish-rate and planarization-rate measurements differ from each other?

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

4. If a 20 nm thick titanium layer is used as a polish stop underneath 500 nm thick tungsten, and film thickness non-uniformities are ±5% and CMP non-uniformity is ±10%, what must polish selectivity be?
5. Work out a step-by-step fabrication process for the photonic crystal shown in Figure 16.8.

REFERENCES AND RELATED READINGS

Evans, D.R.: Slurry admittance and its effect on polishing, *Mater. Res. Soc. Symp. Proc.*, **767** (2003), F5.1.1.

Hernandez, J. *et al*: Chemical mechanical polishing of Al and SiO_2 thin films: the role of consumables, *J. Electrochem. Soc.*, **146** (1999), 4647.

Jindal, A. *et al*: Chemical mechanical polishing of dielectric films using mixed abrasive slurries, *J. Electrochem. Soc.*, **150** (2003), G314.

Kiviranta, M. *et al*: Dc and un SQUIDs for read-out of ac-biased transition-edge sensors, *IEEE Trans. Appl. Supercond.*, **13** (2003), 614.

Lin, S.Y. *et al*: A three-dimensional photonic crystal operating at infrared wavelengths, *Nature*, **394** (1998), 251.

Steigerwald, J.M., S.P. Murarka & R.J. Gutman: *Chemical Mechanical Planarization of Microelectronic Materials*, John Wiley & Sons, 1997.

Stine, B.E. *et al*: Rapid characterization and modeling of pattern-dependent variation in chemical-mechanical polishing, *IEEE TSM*, **11** (1998), 129.

Wrschka, P. *et al*: Chemical mechanical planarization of copper damascene structures, *J. Electrochem. Soc.*, **147** (2000), 706.

Yasseen, A.A. *et al*: Chemical-mechanical polishing for polysilicon surface micromachining, *J. Electrochem. Soc.*, **144** (1997), 236.

Zhang, F. *et al*: Particle adhesion and removal in chemical mechanical polishing and post-CMP cleaning, *J. Electrochem. Soc.*, **146** (1999), 2665.

**Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030**

17

Bonding and Layer Transfer

Wafer bonding has emerged in many different applications in microfabrication: two wafers can be bonded together to create a more versatile starting wafer; bonding creates cavities and seals channels and enables highly 3D structures. In layer transfer, structures are processed on one wafer, then detached and bonded to another wafer. This enables completely different technologies and materials to be merged. Devices can be processed on silicon for convenience, and transferred to, for example, glass or quartz for transparency and insulation, or to a plastic substrate for flexibility. MEMS parts or III-V semiconductor optical devices can be transferred on silicon IC wafers that contain drive or readout electronics. The transferred layers are often very thin, of the order of micrometres, and their handling is very delicate. Therefore, they are usually bonded to another wafer even before detachment from the original wafer.

Two wafers can be joined by a number of methods, but two main classes can be distinguished:

- direct bonding
- indirect bonding with deposited layers ('glue').

Direct bonding involves bare or oxidized silicon and glass wafers. It results in strong chemical bonds across the bonding interface, so strong that breakage happens inside the wafers, and not at bond interface. The bonded wafers can be processed further as if it were one wafer. Indirect bonding uses a great variety of materials as 'glues': metals, glass and polymers (Table 17.1). Bonding methods differ mostly in their temperature range and permanency. Direct bonding is usually hermetic and permanent. Bonding with intermediate layers is done at low temperatures, $<400\,^{\circ}C$, and it may or may not form a hermetic seal. 'Glue' limits the process temperatures and ambients. Some of these methods applicable to both wafer bonding and chip attachment, like adhesive bonding.

The driving force for bonding can be temperature, pressure, electric field or a combination of these.

Table 17.1 Bonding techniques

• Fusion bonding (FB)	Si/Si, SiO_2/Si, glass/glass
• Anodic bonding (AB)	Si/glass, glass/Si/glass
• Thermo-compression bonding (TCB)	Si/glass frit; metal/metal
• Adhesive bonding	Si/polymer/Si

Fusion bonding temperature range is up to $1200\,^{\circ}C$ for silicon and quartz, and ca. $600\,^{\circ}C$ for glasses. Anodic bonding and thermo-compression bonding are performed typically in the range of 300 to $500\,^{\circ}C$, and adhesive bonding, below $200\,^{\circ}C$.

Similar and dissimilar wafers can be bonded. Bonding silicon to oxidized silicon, resulting in silicon-on-insulator, SOI, structure, and bonding silicon to glass, also resulting in permanent bond, are two typical applications. Whereas epitaxial deposition is possible only on top of a crystalline substrate, we can, in principle, bond single crystalline material on any substrate. However, because bonding involves elevated temperatures, differences in thermal expansion have to be accounted for.

At least theoretically, a wafer of any material can be bonded at room temperature to another wafer of any material via van der Waals intermolecular forces. This bonding requires that the bonding surfaces are sufficiently smooth, flat, clean and terminated by a bonding species on the surface. A strong bond can then develop across the bonding interface upon annealing. There is constant progress towards lower and lower bonding temperatures, that is, for lower temperatures without sacrificing bond strength.

Bonding can be done at almost any phase of the process:

- at the wafer manufacturer, as a way to make more advanced wafers;

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

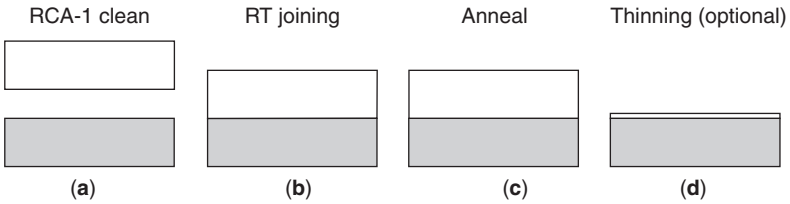


Figure 17.1 Prototypical steps in wafer bonding (**a**) surface preparation; (**b**) room temperature joining; (**c**) annealing for bond strengthening and (**d**) top wafer thinning (optional)

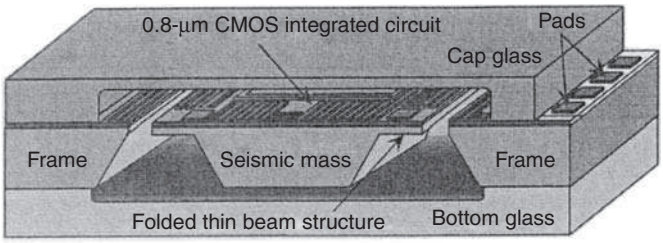


Figure 17.2 Accelerometer by glass–silicon–glass bonding. Reproduced from Takao, H. *et al*. (2001), by permission of IEEE

- in device processing as a process step like any other;
- at the end of the process for cavity formation and encapsulation (zero-level packaging).

If the bonding is done by the wafer manufacturer, the user sees the bonded wafer as any other wafer, except that its special properties will be utilized in the process. Silicon-on-insulator technology is an example of bonded wafer application (bonding is only one way to make SOI). In bonded SOI, the top wafer is thinned down to 10 to 50 μm. It is known as the *device wafer*, and the bottom wafer, of standard thickness, is known as the *handle wafer*. Bonding is not limited to two-wafer joining. More and more wafers can be bonded, yield allowing. Of course, the price will go up.

The basic requirements for good wafer bonding are (1) the materials being bonded form a chemical bond across their interface, (2) high stresses are avoided and (3) no interface bubbles develop. Thermal expansion coefficients of the two materials have to be matched and various glasses have been tailored to match silicon coefficients of thermal expansion CTE. To achieve these requirements, the following processing steps are usually involved in wafer bonding (Figure 17.1).

Prototypical steps in bonding:

– surface cleaning
  particle removal
  hydrophilic surface finish treatment
– room temperature joining
  initiation of bonding at centre or wafer flat
– anneal for bond energy improvement
– top wafer thinning (optional).

In microturbine fabrication (Figure 1.10), five structured wafers are bonded one at a time to form a final device. In blanket wafer bonding, alignment is trivial but in structured wafer bonding it is critical, and it will be discussed in Chapter 28. No wafer thinning is required for turbine application: blade thickness is equal to wafer thickness, 380 μm.

In the final encapsulation, bonding serves many functions: it protects free-standing mechanical parts in the dicing process and it forms cavities for pressure sensors and resonators (Figure 17.2). With all the sensitive, delicate micromechanical parts covered by a capping wafer, dicing, encapsulation and other packaging operations can be generic, whereas packaging of unprotected chips with beams and air gaps would have to be developed for each and every design separately.

17.1 SILICON FUSION BONDING

Silicon-to-silicon bonding can yield abrupt pn-junctions when p-type and n-type wafers are bonded without oxide. This is utilized in power semiconductor fabrication. The alternatives are epitaxial deposition of 100 μm

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

thick p-type layers, or 100 μm deep diffused junctions. While 100 μm deep aluminium diffusions can be made, diffusion times are very long and junctions are not very abrupt.

Fusion bonding, like all bonding processes, begins with a cleaning step. RCA-1 cleaning with ammonia–peroxide mixture takes care of two requirements at the same time: it is effective in particle removal and it leaves the surface in a hydrophilic condition with silanol groups (Si–OH). RCA-1 cleaned surfaces are extremely smooth, <0.5 nm, which is essential for good bonding. Wafers cleaned with HF-last process result in Si-H terminated surfaces, which are rougher and prone to attract particles. Deposited films are usually not smooth enough for bonding, but CMP polishing can be done to achieve surface roughness below 1 nm required for successful bonding (see Figure 16.9).

Surface energy is the energy required to break a bond and to create two new surfaces. It can be estimated from bond strengths and bond densities:

$$\gamma = (1/2) E_{\text{bond}} d_{\text{bond}} \qquad (17.1)$$

The factor 1/2 comes from the fact that when a bond is broken, two surfaces are created. Two wafers in close contact are bonded by hydrogen bonds, as shown in Figure 17.3. We can get an estimate for surface energies from silicon atom surface density, ca. $10^{15}\,\text{cm}^{-2}$, and hydrogen bond energies, 25 to 40 kJ/mol, which translate to ca. 200 to 350 mJ/m^2. Measured values for room temperature–bonded silicon wafers are between 50 to 80 mJ/cm^2. This indicates that less than 100% of the area is in contact with hydrogen bonds. This is understandable because the wafer surfaces are neither perfectly flat nor smooth but have local roughness and waviness, and hydrogen bonds have short range. Even if RMS surface roughness is 0.2 nm, peak-to-valley heights are typically 10 times more, ~2 nm. The saturation value of surface energy after mild thermal treatment or extended time has been measured to be ca. 250 mJ/m^2.

The reaction that takes place during storage or anneal is siloxane bond (Si–O–Si) formation (Figure 17.4).

$$\text{Si–OH} + \text{HO–Si} \longrightarrow \text{Si–O–Si} + \text{H}_2\text{O} \qquad (17.2)$$

Siloxane bonds are much stronger than silanol hydrogen bonds, and measured surface energies are ca. 1300 mJ/m^2. This surface energy is almost constant from 150 to 800 °C (Figure 17.6).

However, surface energies calculated from Si–O bond energies (4.5 eV/bond or 430 kJ/mol) translate to ca. 3000 mJ/m^2. This discrepancy is due to the fact that the surfaces are not fully bonded but have some

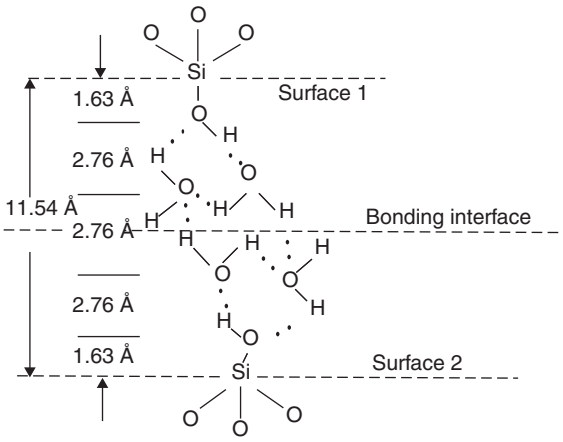


Figure 17.3 Bonding of hydrophilic silicon surfaces. Source: Tong, Q.Y. & U. Gösele, Semiconductor Bonding, © Wiley, 1999. This material is used by permission of John Wiley & Sons, Inc

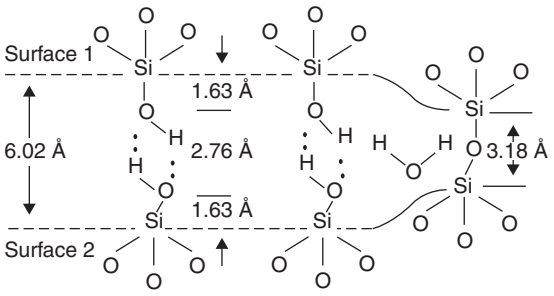


Figure 17.4 Water removal and siloxane bond formation at 110 to 150 °C. Source: Tong, Q.Y. & U. Gösele, Semiconductor Bonding, © Wiley, 1999. This material is used by permission of John Wiley & Sons, Inc

areas that bond via silanol bonds only (as shown in Figure 17.4), but somewhere above 800 °C, the oxide becomes viscous and flows, which increases contact area and leads to higher surface energy, as shown in Figure 17.5. Fusion bonded interface is seen in the TEM micrograph, Figure 2.2. Surface energies of 3000 mJ/m^2 are not encountered in experiments, however, because wafer breakage will take place inside silicon because Si–Si bonds are weaker than the Si–O bonds.

The water released during the formation of Si–O–Si bonds will oxidize silicon further ($\text{Si} + 2\text{H}_2\text{O} \rightarrow \text{SiO}_2 + 2\text{H}_2$; wet oxidation). The thinner the oxide on the wafers, the more important is the effect of this oxide; if wafers with thick oxides are bonded, water diffusion will be

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

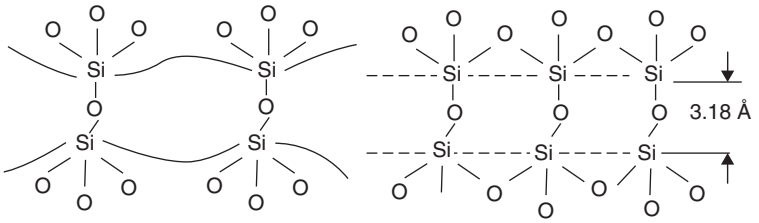


Figure 17.5 Viscous flow of oxide (800 °C for native oxide, 1000 °C for grown oxides). Source: Tong, Q.Y. & U. Gösele, Semiconductor Bonding, © Wiley, 1999. This material is used by permission of John Wiley & Sons, Inc

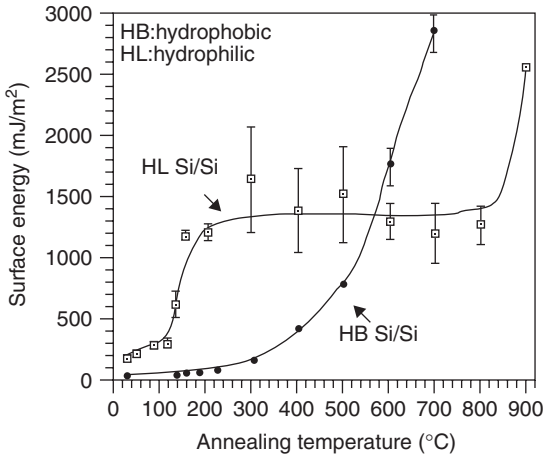


Figure 17.6 Surface energies for hydrophilic (HL) and hydrophobic (HB) bonding. Source: Tong, Q.Y. & U. Gösele, Semiconductor Bonding, © Wiley, 1999. This material is used by permission of John Wiley & Sons, Inc

slow and the additional oxidation, minuscule. A combination of thin (or native) oxide wafer and a thick oxide wafer is a compromise: oxidation will proceed according to the aforementioned equation, strengthening the bond, and hydrogen can dissolve in the oxide, preventing build-up of interfacial stresses.

In the case of hydrophobic (–Si–H terminated) surfaces, roughness is of the order of 5 Å and their bonding properties are much worse. Hydrogen bonds between HF-units are small and bonding is weak. Hydrogen will evolve as a product of hydrophobic bonding:

$$\equiv \text{Si–H} + \text{H–Si} \equiv \longrightarrow \equiv \text{Si–Si} \equiv + \text{H}_2 \qquad (17.3)$$

Hydrogen will diffuse along the bonding interface, and not dissolve into the bulk below 500 °C. Bond energies of hydrophobic bonding are much lower than those of hydrophilic bonding at low temperatures (as shown in Figure 17.6), but they can be improved by annealing. Hydrophilic bonding, however, is the main approach.

Surface preparation by wet cleaning solution is the traditional method but alternatives have been explored, and plasma activation, especially, seems to offer excellent bond strengths at very low temperatures, even below 200 °C.

17.2 ANODIC BONDING

Anodic bonding of silicon to glass (also known as *field-assisted thermal bonding*, FATB), is the oldest bonding technique in microfabrication. It has many features that make is easy: glass is a soft material that will conform at 400 to 500 °C bonding temperatures, sealing structures and irregularities of up to 50 nm hermetically. Native oxides, and thin grown or deposited oxides, do not prevent bonding. Anodic bonding can be visually checked through the glass side: bonded surfaces look black and non-bonding areas are seen as lighter.

Not all glasses are amenable to anodic bonding. Thermal mismatch between silicon and glass needs to be considered at two temperatures: bonding temperature and room temperature/operating temperature of the device. Glasses have higher coefficients of thermal expansion than silicon, but the match at two temperatures is approximately met with glasses like Schott 8339 and 8329 and Corning 7070 and 7740 (Pyrex). CTE of 7740 is almost constant $3.3 \times 10^{-6}/°\text{C}$ from room temperature to 450 °C, and that of silicon increases from 2.5 to $4 \times 10^{-6}/°\text{C}$.

When glass is heated to ca. 400 °C, sodium oxide (Na_2O) decomposes into sodium and oxygen ions. The bonding process uses −300 V to −1000 V applied to the glass wafer. Sodium ions (Na^+) move towards the glass top surface and oxygen ions (O^{2-}) towards the silicon wafer (Figure 17.7). This will create a depletion layer and electrostatic force pulls the glass and the silicon wafer together. The resulting electrostatic forces are very strong: if the thickness of the depletion region is 1 μm, field is E = 500 MV/m (500 V/1 μm); and electrostatic force is proportional to E^2.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

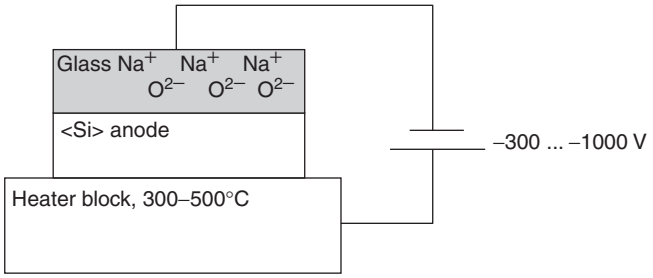


Figure 17.7 Anodic bonding: mobile ions in glass move in the electric field, and a depletion region is established, leading to a large electrostatic force which pulls the wafers together

Oxygen ions react at the glass/silicon interface according to

$$Si + 2O^{2-} \longrightarrow SiO_2 + 4e^- \qquad (17.4)$$

and sodium ions are neutralized at the cathode. If higher temperatures are used, sodium atoms will diffuse faster, and the depletion width is greater, leading to stronger bonds.

Bonding initiation is by applying pressure at the wafer centre, but, if bonding is done in vacuum, it is possible to bond without an initiation point. Current increases rapidly at the initiation of bonding because contact area increases and then decreases exponentially as oxygen ions react at the interface to form SiO_2, and the oxide becomes thicker. When the current has dropped to 10% of its peak value, bonding is termed finished. Typical bonding times are 10 to 30 min. This is fairly long for a single-wafer operation, and special wafer holders have been designed so that wafer loading and unloading can be done while another wafer is being bonded.

A sizable area of silicon is needed for good bonding. At least a 200 μm 'collar' around a cavity or recess is necessary for hermetic sealing, but there are no standardized design rules for wafer bonding.

Anodic bonding of multilayer structures is also possible: glass/silicon/glass systems can be made in a single bonding step. Heating uniformity is important, and double side heating is usually employed. Contacting the middle wafer electrically can be difficult.

17.2.1 Anodic bonding with intermediate deposited layers

Bonding of two silicon wafers or two glass wafers by anodic bonding is not possible as such, but deposited films in between enable bonding. Sputtered Pyrex glass on silicon is a standard approach. Silicon nitride and silicon carbide can be used for silicon wafers, and deposited silicon for glass wafers. Doped spin-on glass has also been experimented with. It is important for anodic bonding that a depletion layer be formed at the interface, and this requires that the intermediate layer acts as an ion barrier.

17.3 OTHER BONDING TECHNIQUES

17.3.1 Thermo-compression bonding (TCB)

Thermo-compression bonding (TCB) applies pressure and heat simultaneously on the samples. This is the standard bonding technique for attaching gold leads to ICs. Gold is suitable because it is noble metal: there are no gold oxides on the surface prevent TCB, and the low yield point of gold is also advantageous. Typical pressures and temperatures for wafer level TCB with metals are in the range 1 to 10 MPa at 300 to 400 °C. Bonding times are then minutes or tens of minutes. Nitrogen atmosphere prevents metal oxidation during bonding.

Wafer-level TCB is made possible by deposition of thin films, with film thicknesses corresponding to the eutectic composition, for example, 80%wt Au, 20% Sn or Si 3%wt, 97% Au. Static pressure may be applied during annealing in hydrogen. Interdiffusion can take place at temperatures below the eutectic temperature.

Glass–frit bonding is another example of TCB. Certain glasses melt under pressure at 500 °C and form hermetic bonds. Glass-frit bonding is similar to anodic bonding, except that pressure is mechanical and not electrostatic. Glass-frit bonding is utilized in many bulk micromechanical applications such as pressure sensors.

17.3.2 Polymer adhesive bonding

Adhesive bonding with a polymeric intermediate layer offers many advantages for bonding as follows:

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

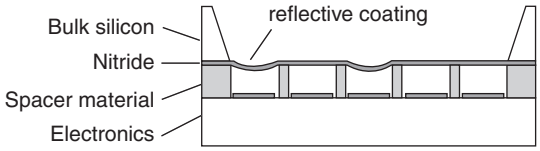


Figure 17.8 Aluminium mirror on nitride membrane is addressed pixelwise by electronics in the bottom wafer. Photoresist serves the roles of both spacer and adhesive. From Sakarya, S. *et al*. (2002), by permission of Elsevier

- temperatures around 100 °C
- tolerant to (some) particle contamination
- structured wafers can be bonded easily
- low cost, simple process.

Because polymers are soft materials they conform to particles, and there will be less problems with voids, compared to stiffer materials like silicon. The main problem with adhesive bonding is limited long-term stability and limited thermal range, with ca. 400 °C maximum. Because of low temperatures and benign processes, CMOS wafers can be used as substrates. A mirror array with individually addressable pixel elements steered by electronics in the bottom wafer is shown in Figure 17.8.

Prototypical steps in adhesive bonding are

- surface cleaning and adhesion promoter application
- spin coating of polymer
- initial curing (solvent bake)
- join the wafers (vacuum may be used)
- final curing of the polymer: pressure and/or heat.

The final curing temperature has to be above the glass transition temperature of the polymer, otherwise no bonding will take place. For CYTOP-fluoropolymer bonding at 160 °C for 30 minutes results in 4 MPa bond strength; bonding below 108 °C glass transition temperature results in no bonding.

Chip bonding can be done similarly: capping chips with polymeric ring structures can be bonded to a substrate in a flip-chip–like way, creating a cavity, which can enclose, for example, a micromechanical resonator that needs to be operated in a protected atmosphere.

17.4 BONDING MECHANICS

Bonding requires flatness and smoothness. Flatness specification is a global/large area concept measured over chip or wafer area, whereas smoothness is a local concept, measured with an atomic force microscope AFM at a $5 \times 5\,\mu\text{m}$ site. Because of non-idealities, the two wafers will not touch fully (Figure 17.9). It is possible to estimate the dimensions of cavities that can be closed in the bonding process. The same equations also govern the closure of micromachined cavities.

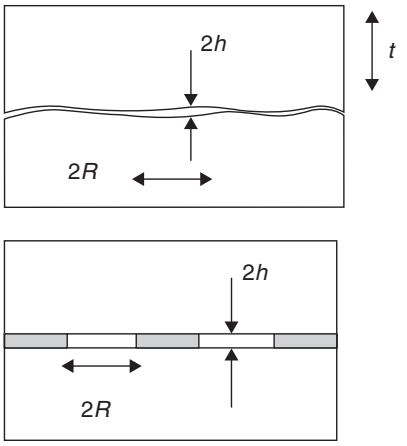


Figure 17.9 Geometry for analysing closing of cavities for the case $2h \ll 2R$. t is wafer thickness

Gap closing is a function of wafer thickness (t), wafer mechanical strength determined by Young's modulus (E), Poisson ratio (ν) and surface energy (γ) (ca. 100 mJ/m^2 for room temperature bonding). Cavities of radius R (in the plane of the wafer) will be closed if the distance between the wafers, h, is

$$h < R^2/(2Et^3/3\gamma(1-\nu^2))^{1/2} \quad \text{for cavities } R > 2t,\ R \gg h \qquad (17.5)$$

$$h < 3.5(R\gamma(1-\nu^2)/E)^{1/2} \quad \text{for cavities } R < 2t,\ R \gg h \qquad (17.6)$$

Particles between wafers cause non-bonding areas (voids) because wafers cannot conform abruptly to particles. The radius of the non-bonding area (see Figure 17.10(a)) is given by

$$R = (2Et^3/3\gamma(1-\nu^2))^{1/4} \times \sqrt{h} \qquad (17.7)$$

Below a critical size h_{crit}, the wafers can conform to particles, and the void size is practically identical to the particle size. This critical size is given by

$$h_{\text{crit}} = 5(t\gamma(1-\nu^2)/E)^{1/2} \qquad (17.8)$$

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

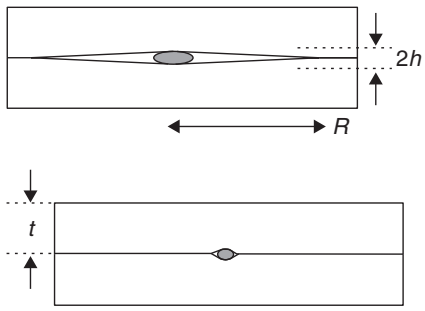


Figure 17.10 Particle-caused void in bonding (**a**) a large particle leads to non-bonded area much larger than the particle itself and (**b**) wafers conform to small particles below critical size

17.4.1 Bond quality measurements

Cleanliness is paramount in wafer bonding: particles at the bond interface will prevent bonding locally. Voids can be detected either destructively or non-destructively. Debonding the wafers and visual or microscopy examination reveal bond interface quality. Bond strength can also be checked by pull tests: successful bonding will result in breakage within either material, but not at the bond interface.

Anodic bonding can be observed through the glass side easily, but if the wafers are not transparent, infrared optical measurement through the wafer is possible. For silicon, this translates to 1.1 μm wavelength and above. The height of voids can be inferred from interferometric rings, with $\lambda/4$ as the minimum detectable height, or ca. 0.28 μm for silicon.

Acoustic microscopy can be used to check voids of the finished wafer stack non-destructively. The wafer to be measured is immersed in water and high-frequency ultrasound is aimed at it. Higher frequency would offer better resolution but energy losses in water increase with frequency, and anyway, acoustic microscopes cannot see the particles but can see only the voids caused by particles.

17.5 BONDING OF STRUCTURED WAFERS

Bond tightness can be measured by gas leakage. When patterned and etched wafers have been fusion bonded, etched depths of 6 nm can be sealed gas-tight, but 9 nm grooves will result in leakage. Higher anneal temperature will seal slightly better. Anodic bonding is much more flexible: even 50 nm grooves can be sealed in a gas-tight manner. Glass will elastically deform to seal the grooves. Higher bonding voltage and temperature will result in better sealing.

We have seen that silicon fusion bonding reaction products are hydrogen in the case of hydrophobic bonding and water in hydrophilic bonding. If there are cavities on the wafers, these gases will be trapped in the cavities. When the temperature is increased, hydrogen and water behave differently: hydrogen dissolves into silicon but water oxidizes silicon. Other gases found in cavities are probably desorption products from wafer surfaces, and not trapped during bonding in gaseous form. In anodic bonding, oxygen diffuses towards the interface (Equation 17.4), and oxygen gas accumulates in the cavity. The desorbed species can also be found in the cavity. Titanium is known to be an oxygen getter, and titanium is sometimes sputtered/evaporated in the cavities to maintain pressure.

Bonding pressure needs some attention when anodic bonding is done on wafers with cavities. At millitorr pressures, a glow discharge can be initiated in the cavity. Therefore, either a good vacuum or atmospheric pressure is desirable. Bonding chamber pressure can usually be varied from atmospheric down to high vacuum, and the chamber can be filled with a chosen gas with selected pressure. This is important for resonating microstructures because damping will depend on gas pressure.

Pressure inside microcavities can be measured from diaphragm bending. Thin diaphragms will bend, and it is possible to relate this bending to pressure. Alternatively, the chips can be placed in a vacuum chamber, and the flat diaphragm condition is equated to gas pressure inside the cavity. The ideal gas law is a good approximation for gas pressures inside cavities.

Oxidizable metal films like aluminium can be sealed between glass and silicon if the films are thin enough (<300 nm). Metals like gold or chromium will prevent bond formation because either they do not oxidize (Au) or their oxides are conductive (CrO). Signal lines out of a bonded structure can be made by diffused lines in the silicon wafer. Resistivity will be high, but the surface is perfectly planar. This method is also suitable for fusion-bonded wafers.

The alternative method for cavity formation is deposition. This will be discussed in Chapter 23. Deposition avoids the main drawback of bonding, which is the fact that an extra wafer is needed in the process.

17.5.1 Bonding by deposition

Bonding of structured wafers can be done by metal deposition: wafers are brought to contact so that an

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

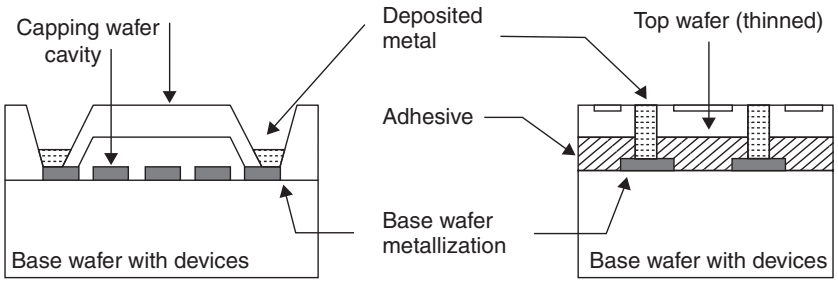


Figure 17.11 (**a**) Microriveting: joining by electrodeposition. Redrawn after Shivkumar, B. & C.-J. Kim (1997), by permission of IEEE and (**b**) adhesive joining with W-CVD via plugs making electrical connection between the wafers. Redrawn after Ramm, P. *et al*. (1997), by permission of Elsevier

opening in the top wafer matches a metal pad on the bottom wafer (Figure 17.11). The wafers are joined by adhesive bonding before W-CVD. Metal deposition then creates contact between the two wafers. Multi-wafer ICs have been made by W-CVD filling of vias that connect the wafers. In microriveting, wafers are bonded by selective electrodeposition. Compared to most other bonding methods, microriveting offers the lowest temperature. Liquid tightness before metal deposition remains to be clarified.

17.6 BONDING FOR SOI WAFER FAB

Bonding is a straightforward way to make SOI structures. Bonded SOI technique uses bonding of two wafers (one or both oxidized) followed by thinning. One of the bonded silicon wafers has to be thinned down to the desired thickness.

Wafer bonding allows independent optimization of the top device layer and the supporting substrate. The substrate (*handle wafer*), is chosen for mechanical support, thermal compatibility, micromachining, doping level or some other property. Device layer can have material, crystal orientation, doping level or thickness tailored to the particular device design, irrespective of handle wafer properties. Oxide thicknesses range from 0.3 to 4 μm, with the upper limit coming from the practical thermal oxide thickness. Bonding of wafers with deposited oxides has been actively studied, but the films are generally not smooth enough for good bonding. If CMP is used to polish the surface, the process cost increases rapidly.

There are two possibilities for the pair to be bonded: a silicon wafer and an oxidized wafer, or two oxidized wafers. The latter results in reduced bond strength, just 70 to 80% of the former, but the resulting structure is symmetric with respect to interfaces. In MEMS applications where the oxide between silicon wafers is etched away during processing, symmetry or asymmetry of the bonding interface is important because etch fronts can travel fast along the bonding interface. In SOI wafer specifications, it is stated which wafer has thermal oxide on it.

Thinning of the device wafer involves grinding, polishing and etching. Thinning down to 10 μm thickness is reasonably easy, and thinning down to 5 μm can also be done. For layers thinner than this, special techniques are required: either real time–thickness monitoring during final polishing or etch-stop layers. Epitaxial layers with different etching properties have to be grown on the device wafer before bonding. Grinding removes the bulk of silicon, and selective etching removes the remaining material until the etch-stop layer is met. High boron doping ($\geq 10^{20}\,\mathrm{cm}^{-3}$) can be used as the etch stop but because of its high dislocation density, a second epitaxial layer is grown on it. The highly doped etch-stop layer can then be removed by, for example, 1–3–8 etchant (a mixture of HF, HNO_3 and CH_3COOH in the volume ratio of 1:3:8), which does not etch a lightly doped material. Etch-stop layers enable fabrication of 100 nm thick device silicon layers with ±5 to 10 nm variation.

17.7 LAYER TRANSFER

Layer transfer is practised along two different lines: in cutting methods, thin layers are separated from substrates and transferred onto other substrates; in sacrificial wafer methods, the processed wafer is bonded to a carrier wafer and the original wafer is dissolved.

Hydrogen bubble–induced layer splitting is based on hydrogen implantation (Figure 17.12). Gas bubbles

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

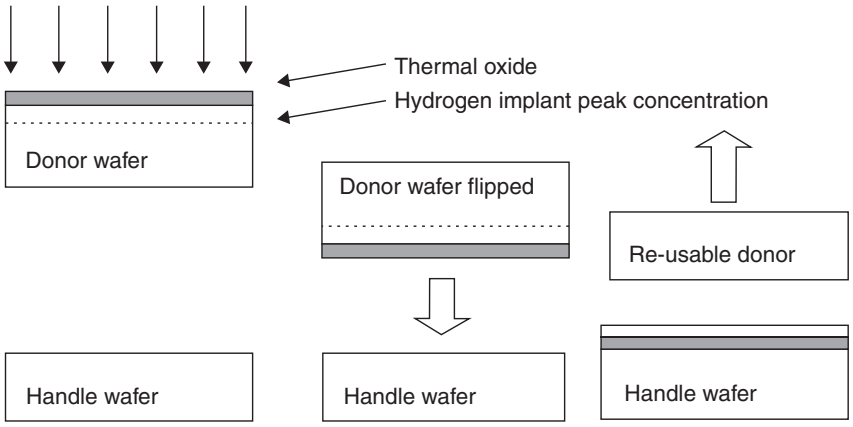


Figure 17.12 Hydrogen implantation layer transfer (**a**) H^+ implantation into an oxidized donor wafer; (**b**) donor wafer is bonded to a handle wafer and (**c**) cleavage along ion implanted maximum concentration depth results in an SOI wafer

form at the depth of maximum hydrogen concentration. These bubbles lead to mechanical weakening of the silicon material, and microcracks lead to cleavage of the implanted layer when suitable thermal treatment or mechanical pressure is applied.

Hydrogen implantation method is patented, and called Smart-cut™, and wafers manufactured with the method are marketed as Unibond™.

Smart-cut™ process flow

thermal oxidation of donor wafer;
H^+ implantation into donor wafer;
hydrophilic bonding at room temperature;
anneal at 400 to 600 °C to split the wafers;
high-temperature anneal at 1100 °C, 2h strengthen the chemical bonds;
final polishing.

The hydrogen dose required for bubble formation is 3.5×10^{16} to 10^{17} cm^{-2}, much less than the oxygen dose in SIMOX. The thickness of the splitting layer is related to the H^+ energy, which can accurately and easily be controlled. Low-temperature annealing is used to split the wafers, and the donor wafer can be reused. CMP is necessary to eliminate the microroughness of the SOI layer, even though the layer thickness just after splitting is homogeneous to a few nanometres.

An alternative way of detachment is mechanical force. Water jets or pressurized gas can be used. Bonding energy at the bonding interface is much higher than that in the H-implanted region, which is embrittled. Thus, even at room temperature, the H-implanted layer can be peeled off from the donor wafer.

17.8 EXERCISES

1 (a). What is the non-bonded area caused by a 0.3 μm particle on 150 mm wafers?
(b). If 150 mm wafers are specified to have 50 particles of 0.3 μm size, what fraction of the wafer area will be non-bonded?
2. What is the critical particle radius for 100 mm silicon wafers?
3. What is resolution of a 160 MHz acoustic measurement of voids?
4. What dimension of microfluidic channels shown in Figure 17.9 will remain open in fusion bonding?
5. Which measurements can reveal the role of sodium ion depletion in anodic bonding?
6. What is the maximum device silicon thickness in (a) SIMOX and (b) Smart-cut if 200 keV implanter is used?
7. Calculate the gas pressure inside an anodically bonded cavity when bonding has been done at 400 °C.

REFERENCES AND RELATED READINGS

Berthold, A. *et al*: Glass-to-glass anodic bonding with standard IC technology thin films as intermediate layers, *Sensors Actuators*, **82** (2000), 224.

Cheng, Y.T., L. Lin & K. Najafi: Localized silicon fusion and eutectic bonding for MEMS fabrication and packaging, *J. MEMS*, **9** (2000), 3–8.

Gui, C. *et al*: Present and future role of chemical mechanical polishing in wafer bonding, *J. Electrochem. Soc.*, **145** (1998), 2198.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Han, A. *et al*: A low temperature biochemically compatible bonding technique using fluoropolymers for biochemical microfluidic systems, *Proc. IEEE MEMS* (2000), p. 414.

Henttinen, K. *et al*: Mechanically induced Si layer transfer in hydrogen-implanted Si wafers, *Appl. Phys. Lett.*, **76** (2000), 2370.

Huff, M.A. *et al*: Design of sealed cavity microstructures formed by silicon wafer bonding, *J. MEMS*, **2** (1993), p. 74

Jourdain, A. *et al*: Investigation of the hermeticity of BCB-sealed cavities for housing (RF-)MEMS devices, *Proc. IEEE MEMS* (2002), p. 677.

Lee, B. *et al*: A study on wafer level vacuum packaging for MEMS devices, *J. Micromech. Microeng.*, **13** (2003), 663.

Mack, S. *et al*: Analysis of bonding-related gas enclosure in micromachined cavities sealed by silicon wafer bonding, *J. Electrochem. Soc.*, **144** (1997), 1106.

Niklaus, F. *et al*: Low-temperature full wafer adhesive bonding, *J. Micromech. Microeng.*, **11** (2001), 100–107.

Ramm, P. *et al*: Three dimensional metallization for vertically integrated circuits, *Microelectron. Eng.*, **37/38** (1997), 39.

Sakakuchi, K. *et al*: Current progress in epitaxial layer transfer (ELTRAN®), *IEICE Trans. Electron.*, **E80-C** (1997), 378.

Sakarya, S. *et al*: Technology of reflective membranes for spatial light modulators, *Sensors Actuators*, **A97–98** (2002), 468.

Shivkumar, B. & C.-J. Kim: Microrivets for MEMS packaging, *J. MEMS*, **6** (1997), 217–225.

Singh, A. *et al*: Batch transfer of microstructures using flip-chip solder bonding, *J. MEMS*, **8** (1999), 27.

Takao, H. *et al*: A CMOS integrated three-axis accelerometer fabricated with commercial CMOS technology and bulk micromachining, *IEEE TED*, **48** (2001), 1961.

Tong, Q.-Y. & U. Gösele: *Semiconductor Wafer Bonding*, John Wiley & Sons, 1999.

Tsau, C.T., S.M. Spearing & M.A. Schmidt: Fabrication of wafer-level thermocompression bonds, *J. MEMS*, **11** (2002), 641–647.

Varma, C.M.: Hydrogen-implant induced exfoliation of silicon and other crystal, *Appl. Phys. Lett.*, **71** (1997), 3519.

**Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030**

18

Moulding and Stamping

Moulding and stamping are age-old techniques that have recently been given new twists by microtechnologies. The printing industry depends on stamping the inked typeface against paper for transferring the ink. The very same process has now been adopted in microfabrication, with sophisticated tools and materials for micrometre and even nanometre dimensions. Moulding of metals, plastics and ceramics can be extended to novel applications by microfabrication techniques.

Thomas Alva Edison used sputtered gold seed layer, wax mask and gold electroplating to fabricate phonograph masters. The technology entered production in 1901 and it could replicate 125 μm pitch (200 grooves/inch), 25 μm thick structures. Electroplating is still a major method for mould-master fabrication. In microfluidic applications, dimensions are not much smaller than in Edison's time; in fact, traditional machine tools could, in principle, be used to fabricate the masters, but most often the surface finish is too rough and the pattern complexity makes machining throughput low but it is useful for quick turnaround time prototyping.

Moulding and stamping have different material flows: in moulding, material is being transported into the mould (Figure 18.1(a)). The traditional method is casting and is still in use in microfabrication: thick polymethyl methacrylate (PMMA) resists and polydimethyl siloxane (PDMS) elastomers are cast. But our usage includes various transport and deposition processes: injection of thermoplastics, electroplating of metals, CVD of polysilicon or diamond or sol-gel of PZT. In stamping, there is no transport of material: the polymeric material, which is on the wafer to begin with, is modified locally by the stamp (Figure 18.1(b)).

Moulding can be further divided into methods that use reusable or disposable moulds (Figure 18.2). In stamping, we can distinguish two cases: 2D-surface processes and 3D-volume processes, which have rather different requirements for stamp masters.

Terminology in the field of micromoulding and stamping is not established because the field is new and rapidly expanding. Sometimes the field is known as *soft lithography*, but this really applies to surface stamping only. Microcontact printing (μCP) is a surface stamping method that relies on alkanethiol inks on gold surfaces. Hot embossing is the name used for volume stamping of MEMS structures, and is sometimes referred to as *hot embossing lithography* (HEL). The same technique is called *nanoimprint lithography* (NIL) in communities that aim at ultimate resolution. The name step-and-stamp is used when NIL is performed analogously to step-and-repeat lithography, that is, one chip is exposed at a time followed by a mechanical movement to fill the wafer with patterns.

18.1 MOULDING

Materials of all classes can be used as moulds: resist mould for electroplated nickel, electroplated nickel mould for PDMS, PDMS mould for ceramics, or single-crystal silicon for polysilicon, diamond and PZT. Of course, thermal and other limitations apply, but clearly the choices are many. There is a plethora of variants of these techniques, and this chapter discusses just the basic issues involved in the replication technologies.

Injection moulding is applied for micrometre dimensions in mass manufacturing: molten plastic is injected into a mould insert to fabricate compact discs (CDs). However, from a general microfabrication point of view, CD is an easy application because the aspect ratios are ca. 0.2 only, the pattern density is quite uniform and the pattern sizes are not dissimilar. Circular symmetry with injection from the centre is beneficial for stress minimization.

Moulding can be continued to further generations: instead of using the moulded piece itself, it can be used

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

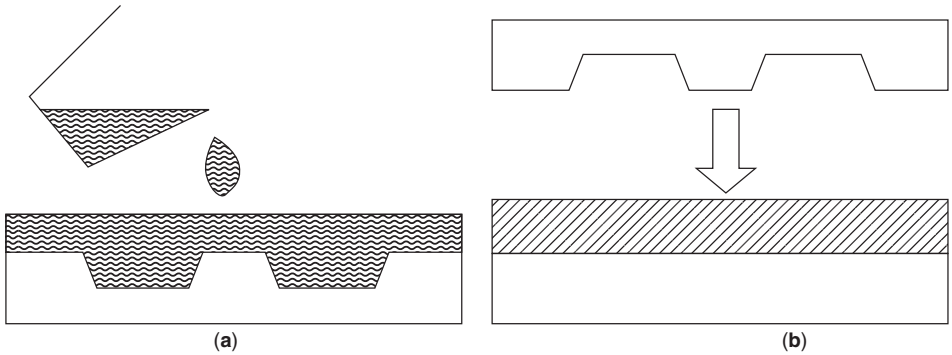


Figure 18.1 (**a**) moulding: material flow into mould master and (**b**) stamping: the stamp modifies material already on the wafer

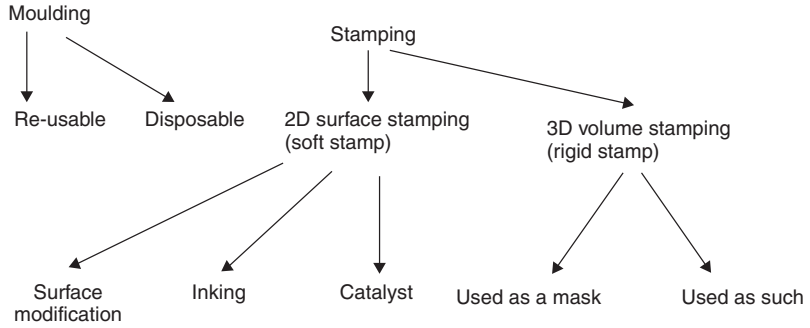


Figure 18.2 Classification of replication technologies

as a new mould. This process can be continued at least till the fourth generation in certain applications, before the quality of moulded pieces becomes unacceptable. However, each generation results in a reverse polarity structure of its parent, so it is necessary to decide beforehand which generation is going to be used.

18.1.1 Disposable moulds

Photoresist is the standard disposable mould, and electroplating into a resist structure is its typical exemplification. Thick resists (e.g., PMMA, SU-8) are used in LIGA technique (LIGA is short for German Lithographie, Galvanoformung, Abformung; for lithography, plating, moulding). In X-ray-LIGA millimetre high structures can be made, while UV-LIGA can be used for 500 μm structures. X-ray LIGA enables higher aspect ratios, and sidewalls that are vertical and smooth, both properties of importance for mould masters.

Hard-to-etch materials can be made into patterns by a few methods: for instance, ion milling, which is a brute-force method. Ion milling has an inherent problem with mask erosion: all materials are sputtered to some extent and selectivity is hard to obtain. Selective deposition depends critically on chemical surface processes that are hard to control. Moulding is rather a universal process because so many different ways of transporting the material are available. The reverse of the final pattern is fabricated in silicon and filled with the desired material and then the silicon is removed. The diamond structures shown in Figure 18.3 are made by etching a silicon mould and then filling it with CVD diamond, followed by silicon wafer dissolution.

The etch selectivity between silicon and the moulded material limits the use of this method: the usual silicon etchants, hot concentrated KOH or HF:HNO_3 mixtures, are very aggressive solutions. Alternatively, silicon can be removed by SF_6 plasma etching or by XeF_2 dry etching. No plasma is needed in XeF_2 etching as it will dissociate into free fluorine in vacuum and etch silicon spontaneously. A number of devices have been made with silicon moulds: AFM tips of Si_3N_4, PZT-ultrasonic transducers and parylene needles.

Backing or bulking is often needed in connection with mould removal: some mechanical support layer is needed to make the structure rigid enough. A typical

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

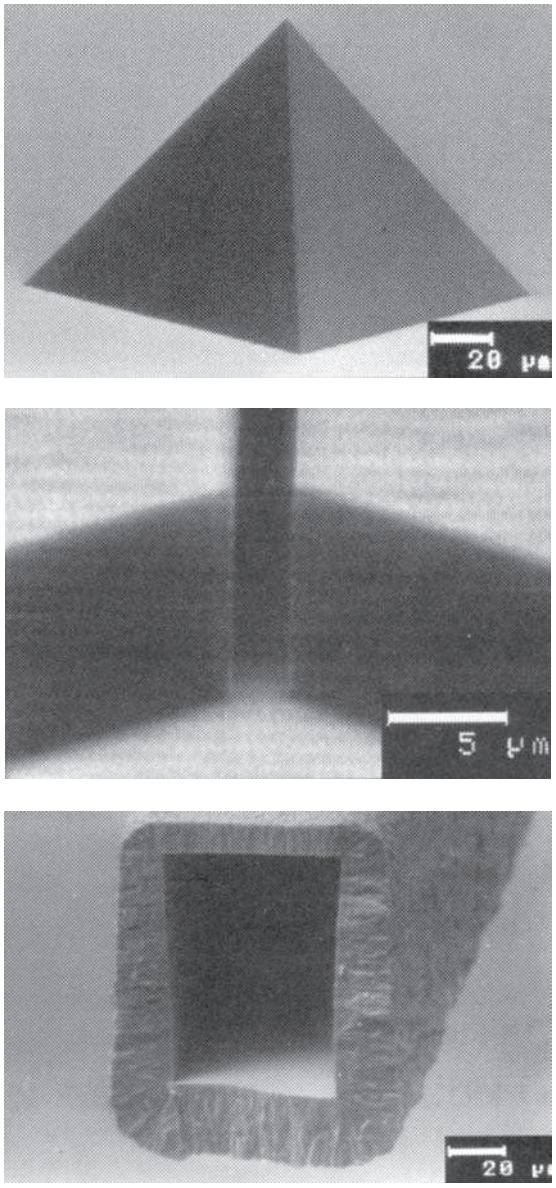


Figure 18.3 Diamond microstructures made with silicon wafer disposable moulds. Reproduced from Björkman, H. *et al*. (1999), by permission of Elsevier

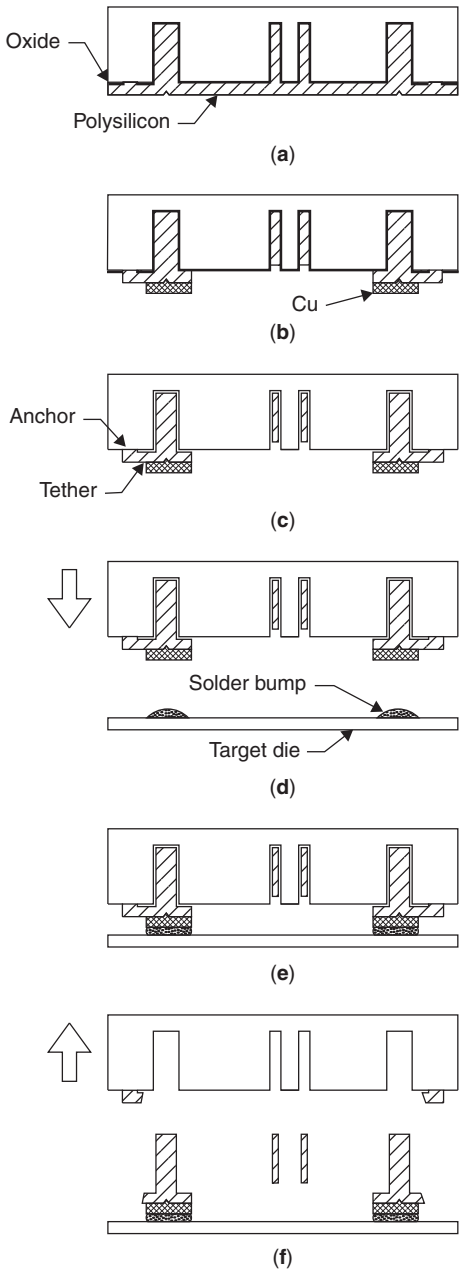
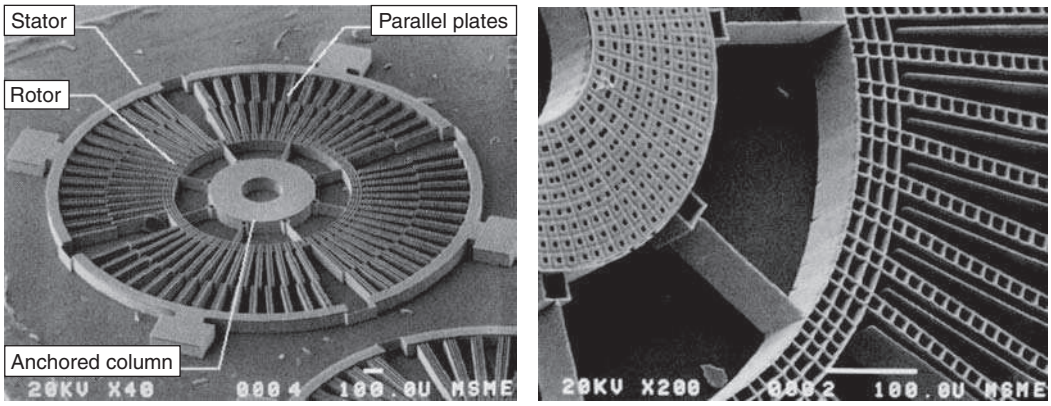


Figure 18.4 Polysilicon moulding in HexSil process: (**a**) Deep reactive ion etching (DRIE) of trenches; CVD release oxide, LPCVD polysilicon structural layer deposition; (**b**) poly patterning and metallization; (**c**) oxide pre-release etch; (**d**) alignment to carrier wafer bumps; (**e**) attachment to carrier solder bumps and (**f**) final release etch. Reproduced from Horsley, D.A. *et al*. (1998), by permission of IEEE

approach would be to deposit a thin metal layer on top of a device material and then use electroplating to deposit a thick ($>100\,\mu$m) backing layer.

Heavy boron doping forms the basis of dissolved wafer process. The p^{++}-doped regions form the structural

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030



(**a**) (**b**)

Figure 18.5 HexSil moulded and released polysilicon pieces attached to a carrier wafer. Reproduced from Horsley, D.A. *et al*. (1998), by permission of IEEE

parts, and the rest of the wafer is etched away. In a sense, the wafer itself is a sacrificial mould. The process begins by standard etching and doping steps, and ends up with KOH/TMAH etching. Owing to mechanical fragility of thin p^{++} structures, bonding to glass or to another wafer is often done before dissolution.

When the mould will is completely removed, freedom of shape is unlimited. If the material to be moulded can fill retrograde features, these pose no problem in release. With reusable moulds, retrograde shapes are not allowed because the mould has to be released.

18.1.2 Reusable moulds

Silicon wafers with etched structures, electroplated metals and SU-8 epoxy structures are typical materials for reusable moulds. The release process must damage neither the mould nor the moulded piece. This can be helped by a couple of methods: the mould can be coated with a material that eliminates reactions between the materials, or an anti-stiction surface coating can be applied. Diamond would be a good choice for a mould for both the above-mentioned reasons. Several Teflon-like fluoropolymer coatings, such as deposition from CHF_3 or C_4F_8 gases in a plasma and vacuum desiccator treatment with tridecafluoro-1,1,2,2-tetrahydrooctyl-1trichlorosilane, have also been utilized. Another way to go is to deposit a sacrificial layer on the mould master and release the structures by etching. The mould can be reused after another sacrificial layer deposition. The HexSil process (Figure 18.4 and 18.5) makes use of a CVD oxide–release layer and a LPCVD polysilicon as the structural material.

Poly dimethylsiloxane is a favourite material for many microdevice applications because it is chemically inert, transparent down to 250 nm and flexible. PDMS is used in microchannels and microreactors, and it is widely used as the master for 2D-surface stamping.

Because PDMS is a polymeric material, its processing does not necessitate elevated temperatures, and a variety of materials can be used as moulds. PDMS pre-polymer is poured over the mould, and cured, for example, at 80 °C for 10 h. PDMS will demould easily because of its inertness. However, because of its coefficient of thermal expansion of ca. 300 ppm/ °C, PDMS is not suitable for applications that require accurate pattern positioning.

18.2 2D SURFACE STAMPING

Surface stamps are soft, elastic materials, like polymer PDMS. These stamps conform to surfaces, but detach easily and retain their shape even after intimate contact. Both elastic constant and surface energy are important considerations for soft stamps. Stiffer materials offer higher resolution but worse contact. Hybrid stamps with a stiff mechanical backing and a soft stamping surface have been devised in order to have the best of both worlds.

The contact area plays an important role: light field structures, with a small contact area, are non-problematic because separation force is small. Structures with aspect ratios not too far from unity and structures with fairly uniform pattern densities, such as periodic structures, are less prolematic than if the aspect ratios of structures to be stamped differ from unity or from each other considerably, when stamping becomes

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

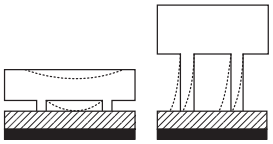


Figure 18.6 (**a**) sagging of low AR structures and (**b**) lateral collapse of high AR structures

problematic. Structures with ca. 1:1 aspect ratios and uniform pattern densities, such as periodic structures, are less problematic than structures with either very low or very high aspect ratios, or a mix of different aspect ratios or pattern densities (Figure 18.6).

18.2.1 Microcontact printing (μCP)

Microcontact printing is a microlithographic version of ink-and-stamp patterning: a polymeric stamp is wetted by 'ink', for example, alkanethiol $CH_3(CH_2)_{15}SH$ or octadecyltrichlorosilane (OTS), and the wet stamp is pressed against a gold surface (Figure 18.7). A reaction between thiol and gold leaves a self-assembled monolayer (SAM) pattern on the wafer. A stamp is most often made of PDMS.

SAMs are usually only 2 to 3 nm thick, and their usefulness as plating, etch or lift-off masks, needs to be improved; even though 20 to 30 nm etched depths have been demonstrated, this is clearly not enough for the majority of applications. Techniques similar to top surface imaging (TSI) (see Figure 10.7) allow wider use of this technique.

18.2.2 Stamping non-planar objects

PDMS is flexible, and this opens up special applications: patterns can be contact-printed on curved surfaces. Gratings on optical fibers have been realized. Similarly, a round object can be rolled over a PDMS stamp and a spiral structure created. Microcoils have been made in this way. Alternatively, the PDMS piece can be curved and used as a mould. Polyurethane moulded into a curved PDMS results in a curved, rigid piece of polyurethane.

18.3 3D-VOLUME STAMPING

Volume stamps are rigid. Silicon wafers make excellent stamp masters: they combine thermal and mechanical stability with the possibility of fabricating elaborate shapes with good surface finish. Electroplated metals are also widely used stamp materials.

Polymers are stamped at temperatures 5 to 100 °C above their glass transition temperatures, which translates to 50 to 200 °C. Both the stamp surface and the sidewalls make intimate contact with the polymer. The 3D nature of the rigid stamp is of paramount importance: not only the surface smoothness but also the sidewall angles are important for stamp release. The surface roughness should be less than 100 nm for successful release. Sacrificial layers for release are not used, because interactions with the polymer might result in unwanted reactions at elevated temperatures.

3D stamp masters are true 3D objects: all their features are replicated, whereas with 2D masters the third dimension does not print. This has crucially important implications for releasing: 3D masters must not have retrograde sloping walls, whereas, the detailed sidewall structure of 2D masters is not an issue. Depending on application, stamped polymeric patterns can be used as final devices or as photoresist-like masks for further processing steps, usually etching or deposition.

18.3.1 Hot embossing

Hot embossing involves pressing a master against a polymer at a temperature slightly above the polymer

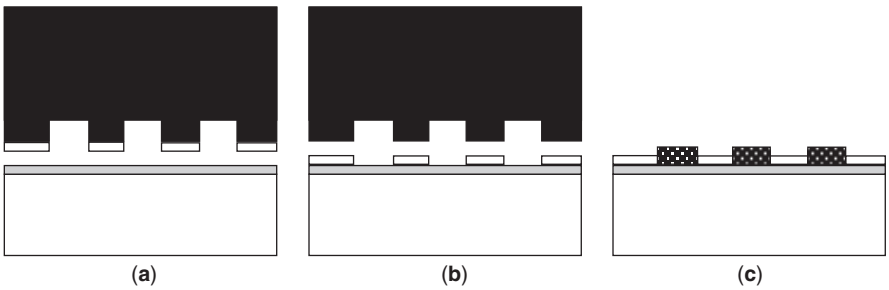


Figure 18.7 Microcontact printing on a gold-coated surface: (**a**) alkanethiol-inked PDMS master; (**b**) alkanethiol attached to gold surface; PDMS stamp lifted and (**c**) metal plating on gold

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

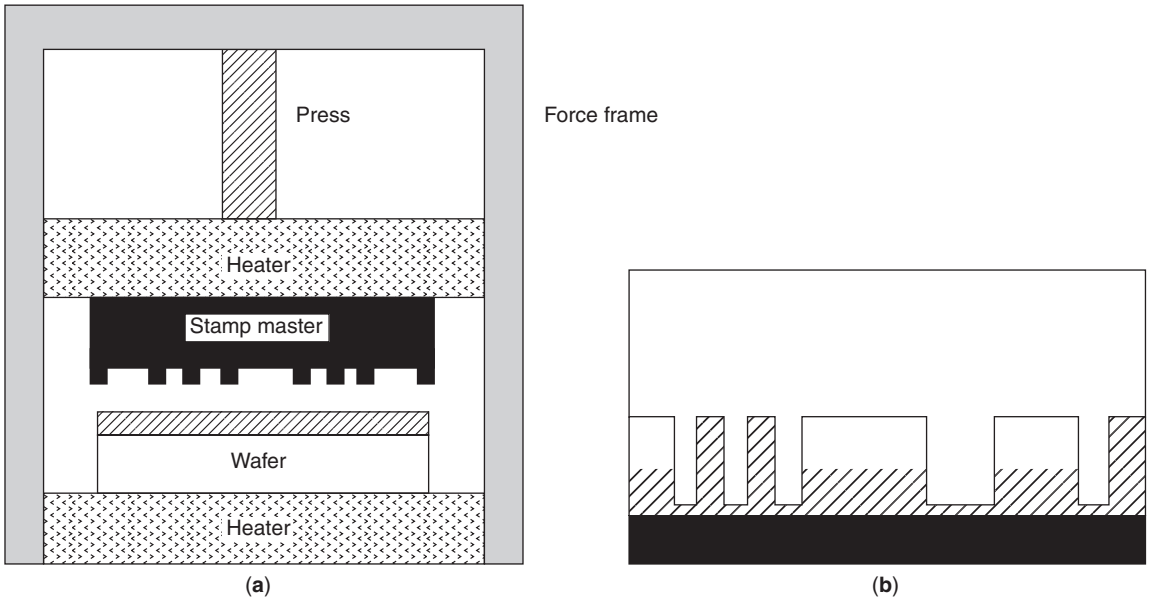


Figure 18.8 (**a**) Schematic hot embossing equipment and (**b**) unequal stamp cavity filling of variable aspect ratio structure

glass transition temperature. The equipment for hot embossing is shown in Figure 18.8. The process has three major issues: filling of structures by polymer (Figure 18.8(b)), reproduction fidelity and master separation and de-embossing.

Both the wafer and the master stages are heated above the polymer glass transition temperature T_g. Widely used polymers such as PMMA have a T_g of 106 °C and polycarbonate (PC) has a T_g of 150 °C. The master is then pressed against the polymer. The embossing force is of the order of 20 to 30 kN and the hold time is of the order of one minute. De-embossing takes place after cooling below the glass transition temperature.

Polymeric materials have coefficients of thermal expansion (CTE) of the order of 20 to 100 ppm, whereas silicon has a CTE of 2.6 ppm and nickel, a typical electroplated master material, 13 ppm. Thermal cycling is mandatory for hot embossing but it should be minimized to around T_g to avoid thermal mismatch cracking.

The thickness of hot embossed structures can be varied enormously, from 150 nm to 150 μm. There is no resolution limit, and embossing can replicate structures down to 10 nm size; making the master becomes the limiting factor. The aspect ratios of embossed structures can be as high as 20:1, and up to 50:1 when special release coatings have been applied.

Hot embossing is suitable for simple structures, preferably involving only one patterning step. Various microfluidic and biomedical microdevices fall under this category, especially if they need to be cheap enough to be disposable.

18.3.2 Imprint lithography

Imprint lithography (also known as *nanoimprint lithography*) involves physical pressing of the master against a polymer-coated wafer, followed by a master release. It is a hot embossing process that is used to make lithography-like structures, which necessitates removal of the polymer from the bottom of the structure (Figure 18.9). The thickness contrast is the ratio of the original polymer thickness to the residual thickness at feature bottom. This value ranges from 2:1 to 6:1.

Imprint lithography is a very simple process for making submicron structures: if mask making can be subcontracted, the printing equipment costs a fraction of a 1X optical system.

If a single-layer pattern is needed, imprint lithography is very cost effective. Magnetic storage devices have been suggested as an application. If alignment between successive layers is needed, the complexity of the equipment increases considerably.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

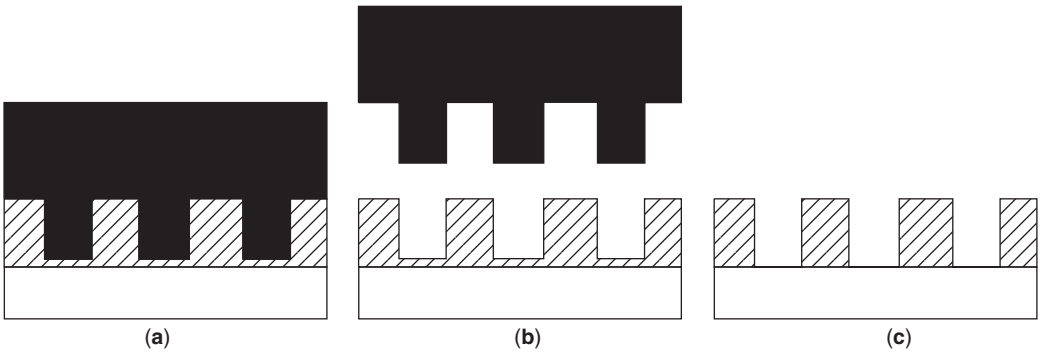


Figure 18.9 Imprint lithography: (**a**) embossing; (**b**) mould release (de-embossing) and (**c**) bottom clearing by RIE

18.4 COMPARISON WITH LITHOGRAPHY

In optical lithography, the mask can be in contact with the resist, but most often contact printing is avoided and proximity printing is used instead. When optical contact lithography was the mainstay of lithography, mask makers had a big business in making replicates of masks (work masks) from the master mask. The movie business uses a similar approach: the original film is never projected, just copies of it (or rather, slave masters are made from the original, and theatre copies are made from the slave masters). Printing industries have been using contact printing for centuries, so the basic problem is not the contact itself. The release process has to be designed into the materials of the master and the film to be imprinted.

Replication masters need to be made with the final dimensions, just like 1X optical or X-ray lithography masks. Replication masters resemble X-ray lithography masks in the sense that they are 3D objects, whereas optical masks are basically planar 2D objects. Therefore, the fabrication of 3D masters is more difficult than photomask fabrication.

18.5 EXERCISES

1. If a PDMS stamp master with a CTE of 300 ppm/°C is made by moulding over a 100 mm silicon wafer, what is the positional accuracy that can be achieved?
2. Design fabrication processes and layouts for the silicon moulds that have been used to make the diamond microstructures shown in Figure 18.3.
3. If 20 μm thick nickel pillars are needed as masters, and master fabrication is by photolithography, what is the smallest feature size that can be fabricated?
4. What are the dimensional limitations of the HexSil process?
5. How can you make hemispherical microlenses by moulding/stamping methods?

REFERENCES

Becker, H. & C. Gärtner: Polymer microfabrication methods for microfluidic analytical applications, *Electrophoresis*, **21** (2000), 12–26.

Bernard, B. *et al*: Printing meets lithography: soft approaches to high resolution patterning, *IBM J. Res. Dev.*, **45** (2001), 697.

Biebuyck, H.A. *et al*: Lithography beyond light: microcontact printing with monolayer resists, *IBM J. Res. Dev.*, **41** (1997), 159.

Björkman, H. *et al*: Diamond replicas from microstructured silicon masters, *Sensors Actuators*, **73** (1999), 24.

Chou, S.Y. *et al*: Sub-10 nm imprint lithography and applications, *J. Vac. Sci. Technol.*, **B15** (1997), 2897.

Horsley, D.A. *et al*: Design and fabrication of an angular microactuator for magnetic disk drives, *J. MEMS*, **7** (1998), 141.

Waits, R.K.: Edison's vacuum coating patents, *J. Vac. Sci. Technol.*, **A19** (2001), 1666.

Wang, D. *et al*: Nanometer scale patterning and pattern transfer on amorphous Si, crystalline Si and SiO_2 surfaces using self-assembled monolayers, *Appl. Phys. Lett.*, **70** (1997), 1593.

Wang, S.N. *et al*: Novel processing of high aspect ratio structures of high density PZT, *Proc. IEEE MEMS* (1998), p. 223.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Part IV

Structures

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

19

Self-aligned Structures

Lithography is most often discussed as a resolution question: how small a structure can be printed on the wafer? Alignment is equally important: how closely can the structures on the different mask levels be aligned with each other? Device-packing density is clearly dependent on both.

Self-alignment is a process by which two structures are aligned to each other non-lithographically. The existing structures act as masks for subsequent steps. Unlike photoresist, these structures are fixed and are integral parts of the device. Self-alignment offers inherently accurate alignment between two structures because alignment is not determined by the optomechanical lithography tool but by the structures and materials themselves.

In this chapter, the examples are related to CMOS but self-alignment is not limited to CMOS: it can be applied widely in microdevice fabrication. More examples will be presented in chapters on sacrificial structures (Figure 22.11), bipolar technology (Figure 26.3), processing on non-silicon substrates (Figure 29.3) and Moore's law (Figure 38.2).

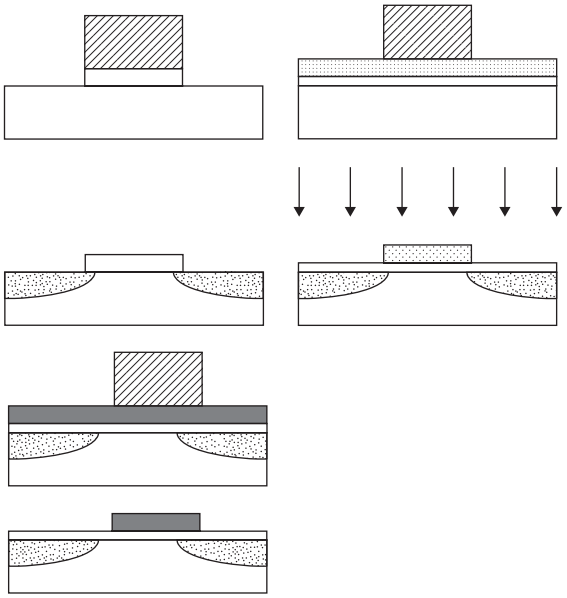


Figure 19.1 Non-self-aligned Al-gate versus self-aligned polysilicon gate MOS. Leftside is Al-gate, right side polygate

19.1 MOS GATE MODULE

Aluminium gate MOS is an example of a non-self-aligned transistor. Its gate module fabrication flow shown below is highly simplified (Figure 19.1). After aluminium gate, the self-aligned polysilicon gate process will be presented.

Al-gate MOS process flow

thermal oxidation of silicon; thick oxide for diffusion masking;
lithography #1: photoresist pattern formed on oxide;
oxide etching in BHF;
photoresist stripping;
boron diffusion at 1000 °C;
thick diffusion mask oxide is etched away in HF;
wafer cleaning
gate oxidation;
aluminium sputtering;
lithography #2: aluminium gate pattern;
aluminium etching;
photoresist stripping.

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Polygate MOS process flow

The first major self-aligned structure to be implemented was the polysilicon gate, which rapidly replaced the non-self-aligned aluminium gate.

Process flow for polygate

gate oxidation
polysilicon LPCVD
polysilicon doping with phosphorus
lithography #1: polysilicon gate pattern
etching of polysilicon
stripping of the photoresist
boron ion implantation
wafer cleaning
implant anneal.

The polysilicon gate blocks ion implantation and source and drain areas are doped (the polysilicon will be implanted too, but it has been so heavily doped by phosphorus in the preceding step that its resistivity or doping type will not change). The boron-doped areas are automatically aligned to the gate. Aluminium (melting point 653 °C) cannot be used in a self-aligned process because it does not tolerate the post-implant anneal.

19.2 SELF-ALIGNED TWIN WELL

In a twin-well CMOS, both n-type and p-type wells are used. With this approach, both NMOS and PMOS transistors can be optimized independently. Wells can be made sequentially with two lithographic steps, or with one lithographic step in a self-aligned sequence (Figure 19.2).

Process flow for a self-aligned twin well

thermal oxidation of the pad oxide (40 nm)
LPCVD nitride (150 nm)
lithography
nitride etching (selective against oxide)
phosphorus ion implantation
(no penetration of 190 nm thick nitride/oxide stack)
photoresist strip
cleaning
thermal oxidation (500 nm)
boron implantation
(no penetration of 500 nm thick oxide)
oxide etch.

However, when the thick oxide is removed, the n-well and the p-well will not be in the same focus plane, but the n-well will be somewhat lower. A standard twin well with two lithography steps does not have this problem.

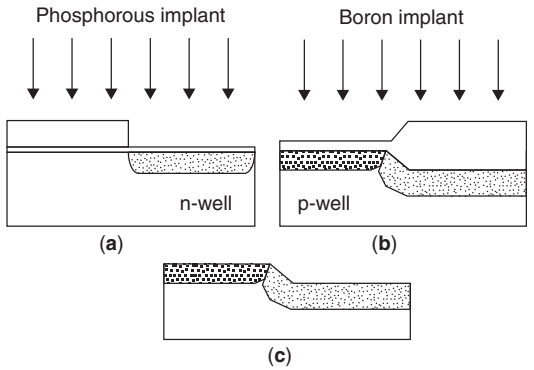


Figure 19.2 Self-aligned twin well: (**a**) phosphorus implant blocked by nitride; (**b**) boron implant blocked by thick thermal oxide and (**c**) after all oxide is etched away

19.3 SPACERS AND SELF-ALIGNED SILICIDE (SALICIDE)

The self-aligned polygate has further evolved into the self-aligned-silicide (salicide) structure: not only the source/drain implantations are self-aligned to the gate, but also the source, drain and gate are metallized in a self-aligned fashion (Figure 19.3). The key innovation is the sidewall spacer: spacers separate the metallized areas, and this separation can be considerably smaller than the minimum lithographic dimension. Cobalt silicide formation is described below.

Process flow for self-aligned cobalt silicide gate

polysilicon gate etching
photoresist strip
wafer cleaning
dry oxidation (10 nm)
CVD oxide deposition
spacer etching (in CHF_3 plasma)
HF-dip

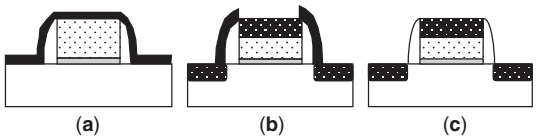


Figure 19.3 Self-aligned metallization: (**a**) metal deposition; (**b**) annealing forms silicide on polysilicon gate and single-crystal silicon source/drain areas and (**c**) unreacted metal is selectively etched away. Silicide (black with dots), metallic titanium (black), polysilicon (dotted)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

cobalt deposition
annealing in argon to form CoSi at 550 °C
cobalt etching
annealing in argon to form $CoSi_2$ at 650 °C.

The silicide reaction takes place where the metal and the silicon are in contact, but no reaction takes place on the oxide. However, there is the possibility of bridging: some silicon (from either the source/drain area or the polysilicon gate) diffuses over the spacer, and the silicide reaction will then take place there as well. This is highly undesirable, because S/D/G would then be electrically contacted. Annealing in two steps avoids this: the first, low-temperature-annealing step, forms monosilicide CoSi, which enables selective etching of the unreacted cobalt. The second annealing is done to lower the resistivity of the silicide, and in the case of cobalt, $CoSi_2$ has the lowest resistivity (for nickel, NiSi is the desired final state, and $NiSi_2$ formation has to be avoided).

The silicide thickness is determined by the metal thickness, and a compromise between two factors must be made: thick silicide would have lower sheet resistance, but it is not compatible with shallow junctions and leads to increased leakage currents. In theory, 1 nm of metallic titanium will result in 2.2 nm of silicide, all of it below the original surface. Cobalt silicide, $CoSi_2$, will consume even more silicon: the silicide thickness is ca. 3.5 times the cobalt thickness. Cobalt silicide formation can be measured by RBS, as shown in Figure 19.4. In as-deposited sample, a signal at 1550 keV is obtained from the top surface of the cobalt, and a signal at 1100 keV is obtained from the silicon at the Si/Co interface. In an annealed sample, the cobalt leading edge is unchanged at 1550 keV because it comes from the cobalt atoms at the surface, just like in an as-deposited sample, but the trailing edge is at 1420 keV because some cobalt atoms have diffused into the silicon during reaction. Similarly, some silicon atoms have diffused to the surface, and the silicon leading edge signal is at 1150 keV. Note that the area under the cobalt signal is unchanged, because no cobalt atoms are lost in the silicidation process.

The surface needs to be cleaned before metal deposition. An HF-dip removes the native oxide, but it will, however, also etch the CVD oxide spacer, and therefore its duration must be carefully optimized. The nitride spacer width would remain intact because a LPCVD nitride has very high selectivity against dilute HF. It is also possible to remove the native oxide in the sputtering system by RF sputter etching. However, argon ion bombardment is prone to produce damage, for example, gate oxide charging and charge-induced

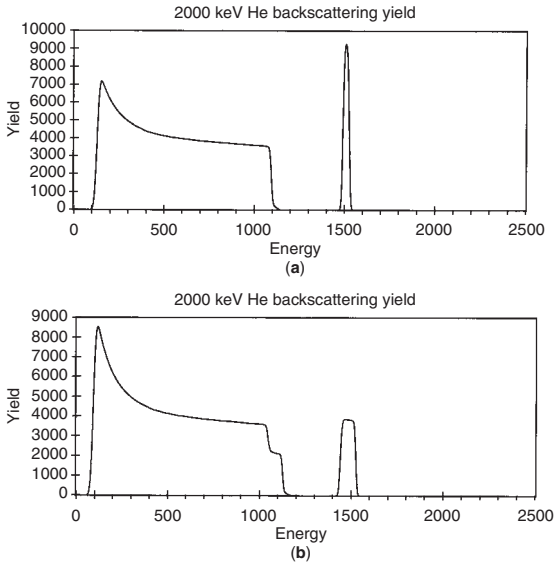


Figure 19.4 RBS spectra of cobalt silicide formation: (**a**) ca. 30 nm cobalt on silicon and (**b**) ca. 100 nm $CoSi_2$ on silicon. Figure courtesy Jaakko Saarilahti, VTT

breakdown, and it is a delicate process. Titanium can reduce oxides, and thin oxide does not prevent the silicidation reaction, but cobalt and nickel do not reduce oxides, and a clean surface is of paramount importance. Titanium salicide presents other novel features, which are discussed below.

Titanium salicide process flow

spacer etching
HF-dip
titanium deposition
annealing in nitrogen to form $TiSi_2$ and TiN at 750 °C
titanium and TiN etching
annealing to reduce $TiSi_2$ resistivity.

Titanium is annealed in nitrogen. The surface of titanium will react with nitrogen to form TiN, and this TiN film will suppress lateral growth of the salicide over the spacers. A simple one-step anneal in argon, which would produce a predictable thickness of titanium silicide, is not possible because of excessive lateral growth over the spacers. Furnace annealing is not practical because residual oxygen in furnace incorporates into titanium and prevents silicidation reaction. Rapid thermal annealing (RTA) equipment is better suited to applications where gas phase impurities must be tightly controlled. Control measurement for the first anneal is the silicide sheet resistance. First annealing has to be optimized so that

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

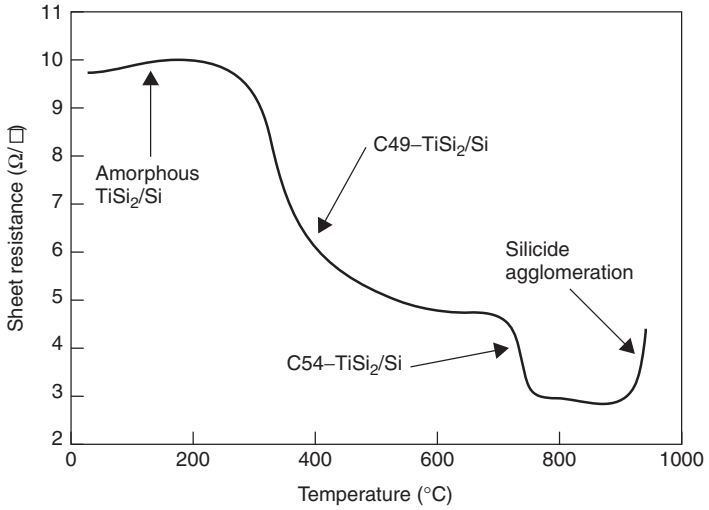


Figure 19.5 $TiSi_2$ phase transitions C-49 to C-54 to agglomeration. Reproduced from Mann, R.W. *et al*. (1995), by permission of IBM

silicon/titanium reaction ($TiSi_2$ formation) at the interface is faster than the gas phase nitridation of titanium into TiN. This, together with lateral overgrowth minimization, leads to first anneal temperatures of ca. 700 to 750 °C.

In the case of nitrogen anneal, we have to remove not only the unreacted metallic titanium but also TiN, so we need to know the selectivity for both Ti:$TiSi_2$ and TiN:$TiSi_2$ pairs. The thickness of titanium cannot be calculated simply from titanium, silicon and $TiSi_2$ densities because dome titanium is consumed by the TiN formation reaction. $TiSi_2$ thickness is also reduced by the fact that selective etches are not infinitely selective: some $TiSi_2$ is lost during titanium etching (see Table 5.8 for selective etches). If titanium thickness is scaled down and the rest of the process is unchanged, $TiSi_2$ thickness will decrease more than predicted by a simple metal-to-silicide relation because the surface nitride thickness is independent of titanium thickness.

The first anneal results in C49 phase $TiSi_2$, which has fairly high resistivity. The second anneal transforms silicide into C54 phase, which has resistivity of ca. 15 μohm-cm. This anneal is limited from above by $TiSi_2$ thermal stability and from below by the need to effectuate the phase transformation: 850 °C, 30 s is usually used. At higher temperatures the silicide tends to ball up, that is, it minimizes its surface energy by agglomerating into ball-shaped crystals and film continuity is then lost (Figure 19.5). Contact resistance and junction leakage current measurements characterize completed silicide processes.

The silicidation reaction is not necessarily identical on polysilicon gate and single-crystal silicon S/D areas. Dopants may also behave differently: for example, heavy boron doping might lead to TiB_2 formation.

19.4 SELF-ALIGNED JUNCTIONS

In the process sequence, where junctions are formed before the silicide, there is always the possibility that the silicide will reach the junction and destroy the device. Silicides can be doped much like polycrystalline silicon. If the salicide gate process is performed in the following order, the junction will be vertically self-aligned to the silicide (Figure 19.6).

Process flow for self-aligned junctions

implantation (low energy, low dose)
spacer formation
silicide formation
ion implantation (high dose)
dopant outdiffusion from silicide during annealing.

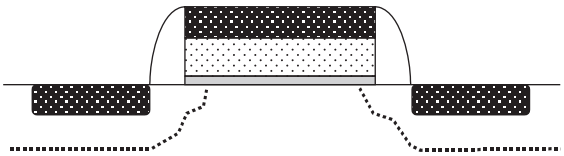


Figure 19.6 Junction diffusion from self-aligned silicide

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

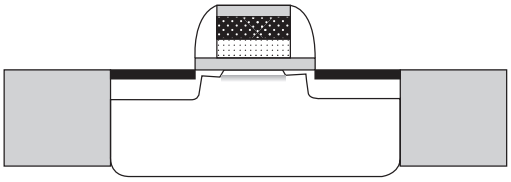
19.5 EXERCISES

1a. How thick a titanium silicide layer will be formed from a 100 nm thick titanium layer under argon annealing?

1b. Where is the surface of $TiSi_2$ relative to original silicon surface?

2. What was the original titanium thickness in Figure 19.5?

3. Analyse the fabrication steps of the dual-silicide structure shown below. Oxide is grey; silicides are black and dotted black. A thick deposited and etched silicide on gate; and a thin, self-aligned silicide on source/drain areas.



4. Estimate the final $TiSi_2$ film thickness for a two-step nitrogen annealing process given that the initial titanium thickness is 50 nm.

REFERENCES AND RELATED READINGS

Gambino, J.P. & E.G. Colgan: Silicides and ohmic contacts, *Mater. Chem. Phy.*, **52** (1998), 99–146.

Hou, T.-H. *et al*: Improvement of junction leakage of nickel silicided junction by a Ti-capping layer, *IEEE EDL*, **20** (1999), 572.

Kittl, J.A. *et al*: Salicides and alternative technologies for future ICs: Part I, *Solid State Technol.*, (1999), 81; Part II August 1999, p. 55.

Lasky, J.B. *et al*: Comparison of transformation to low-resistivity phase and agglomeration of $TiSi_2$ and $CoSi_2$, *IEEE TED*, **38** (1991), 262.

Mann, R.W. *et al*: Silicides and local interconnections for high-performance VLSI applications, *IBM J. Res. Dev.*, **39** (1995), 403.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

20

Plasma-etched Structures

Plasma etching is a technology that enables narrow linewidths and high aspect ratios. It has completely replaced wet etching for feature patterning in modern ICs and it is mandatory in polysilicon surface micromechanics. It has also been applied to structures and applications that are not at all possible with wet etching. For instance, plasma etching without resist mask is essential for planarization and spacer formation.

20.1 MULTI-STEP ETCHING

Etching a single layer structure can be accomplished in a single step, but multi-step etching can be used for improved process control. In polysilicon gate etching, a three-step process is typical:

Step 1: *Native oxide breakthrough*:
- low oxide selectivity;
- a few nanometres of native oxide are quickly removed in CF_4/Ar;
- some polysilicon is etched too.

Step 2: *Bulk etching*:
- optimized for high rate and vertical profile: HCl/HBr.

Step 3: *End point and overetch*:
- the last 50 nm of poly etched in HCl/HBr;
- high selectivity to oxide.

Note that the underlying oxide loss is a sum of four different factors:

1. polysilicon film (non)uniformity;
2. polysilicon etch process (non)uniformity;
3. poly:oxide selectivity;
4. overetch time.

Aluminium etching incorporates similar native oxide, bulk, end point and overetch steps. Etching of silicon oxide selectively against silicon is a heavily polymerizing process and selectivity depends on this polymerization. A three-step oxide etch process consists of a bulk etching step, an end point step which is highly selective (and polymerizing), followed by a third, low-power step that removes polymeric residues: a few extra nanometres of silicon are lost in the low-power etch step but wafer cleaning that follows will be much easier (Figure 20.1).

A combination of anisotropic and isotropic etching steps can be used to make free-standing structures with vertical walls (Figure 20.2). One version is known as SCREAM (for Single CRystal Etching And Metallization) and it consists of the following steps:

- anisotropic plasma-etching for the trench (oxide hard mask);
- spacer oxide deposition by CVD;

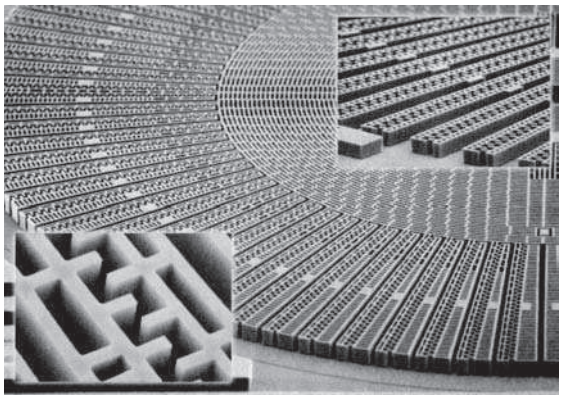


Figure 20.1 RIE of silicon for hard disk drive read/write head positioning actuator. Reproduced from Murari, B. (2003), by permission of IEEE

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

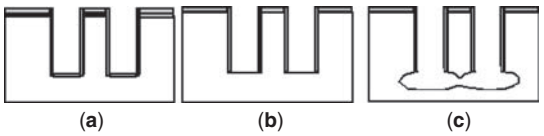


Figure 20.2 (**a**) DRIE of silicon with oxide/nitride mask; followed by oxide deposition to protect the sidewalls; (**b**) anisotropic etching of bottom oxide and (**c**) isotropic undercut etching

- anisotropic spacer etching (oxide removed at bottom and on top of mask oxide);
- isotropic undercutting etching;
- metallization (undercut regions will automatically prevent metal shorts).

Release etch of underlying silicon is clearly not selective relative to the silicon bridge, which will inevitably lead to loss of some material. Furthermore, this loss is coupled with bridge width.

20.2 MULTI-LAYER ETCHING

Thin-film functionalities are often enhanced by stacked layers of different materials. This is bad news for etch engineers, because there is no guarantee that the materials behave similarly at all in etching.

It seldom happens that both (or all) layers can be etched with the same process parameters and it may well be that completely different etch chemistries must be used. In two-step double layer etching, an end point signal must be obtained so that etching can be stopped, or else etch chemistry must provide high selectivity. High selectivity, however, is not always beneficial: if TiN on top of aluminium is etched in fluorine plasma, etching will definitely stop once the underlying aluminium is met, but the aluminium surface will turn to AlF_3, which is a very stable material, and initiation of the aluminium etch step is endangered. Etching of the bottom layer has all the usual requirements about rate, selectivity and profile, and the extra requirement of not etching the top layer. Of course, the acceptable profile in either of the layers calls for engineering judgement (Figure 20.3).

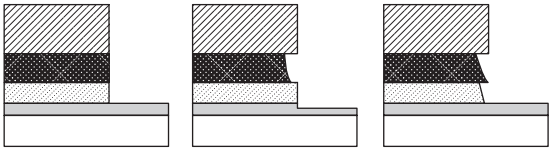


Figure 20.3 Double layer plasma etching: ideal and non-ideal profiles. Photoresist still in place

20.2.1 WSi_2/polysilicon (polycide) etching

Step 1: *WSi_2 etching*: $Cl_2/He/O_2$ for WSi_2;
Step 2: *Poly etching*: Cl_2/HBr for poly;
Step 3: *Poly end point step*: $HBr/He/O_2$ for etching last 20 nm of poly;
Step 4: *Overetch step*: $HBr/He/O_2$ optimized for high oxide selectivity.

Problems with films stacks that require different etch chemistries (chlorine versus fluorine) has led to multi-chamber etch reactors, with each chamber reserved for one material and/or specific etch chemistry. This will be discussed in Chapter 34.

20.2.2 Etching with a hard mask

In deep sub-micron processes, resist thickness has to be scaled down for maximum lithographic resolution, but these thin resists are not always suitable as etch (or implant) masks. Many wet- and dry-etching processes utilize hard masks because resists are simply not tolerant enough under harsh etch conditions. 'Harsh' can mean aggressive chlorine plasmas, very long etch times or hot acids and bases.

Polysilicon gate etching can be done with an oxide hard mask. Because poly etching is highly selective against gate oxide, it is also highly selective against oxide hard mask, therefore a very thin oxide hard mask is enough, and very thin photoresist can be used to etch this hard mask. Elimination of carbon (i.e., elimination of photoresist) from the reaction brings about a major selectivity improvement: selectivity between poly and oxide can be as high as 300:1 compared with 30:1 with resist mask, keeping all plasma parameters, RF power, pressure and gas flows constant. In the presence of carbon, CO is formed because it is energetically favourable, and the source of oxygen for CO formation is the gate oxide, therefore the low selectivity. In the absence of carbon, no CO is formed.

Hard masks offer some interesting options to scale features narrower. A thin photoresist is used to pattern a thin hard mask. Before resist stripping, the hard mask is made narrower by isotropic etching. The hard mask sidewall will be vertical, however, because the isotropic etch sees only the sidewall of the hard mask. The photoresist is stripped only after the hard mask narrowing etch, and the actual film etching then takes place with the narrowed hard mask.

In SF_6-based deep RIE processes, in which etching depths go down to 500 μm (through the wafer), either thick photoresists or CVD-oxides are used as masks.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

DRIE processes that use Cl_2 chemistry use metals such as chromium or nickel as etch masks. Etching of thick oxide structures ($>10\,\mu m$) (for optical waveguides or capillary electrophoresis channels) uses thick polysilicon, amorphous silicon or metal masks.

However, the use of metal masks poses a problem in plasma etching. Even though the mask is stable, it is always etched somewhat under ion bombardment. Re-deposition of these non-volatile sputter-etched species on the surfaces leads to non-etchable areas. This is called *micromasking*. In the case of perfect anisotropy, micromasking leads to formation of high aspect ratio pillars.

20.3 RESIST EFFECTS ON ETCHING

20.3.1 Resist selectivity

Usually, a vertical walled resist is desirable and necessary for the best dimensional control in plasma etching. Most often the resist is, however, slightly sloped, for example, 86° or 88° (positive slope), or even negative (retrograde). If the resist bake temperature is too high (above the glass transition temperature T_g), the resist will flow, and the shape is determined by surface forces. In the 'ideal' case, a hemispherical resist drop will be formed (and in some applications resist lenses are very useful).

Resist selectivity can affect the etched profile. Slight deviation from the vertical does not usually show if selectivity between film and resist is reasonable, say 3:1. But if the resist profile is sloppy, and resist selectivity is 1:1, then etching will transfer the resist profile into the underlying film. A hemispherical initial shape in resist results in hemispherical microlenses in the film material (Figure 20.4).

20.3.2 CD gain

Etching usually results in a slight narrowing of the lines compared to the resist line. The opposite case of line widening, also know as *CD gain*, is also possible (Figure 20.5). CD gain is typical of plasma-etching processes when there is heavy ion bombardment,

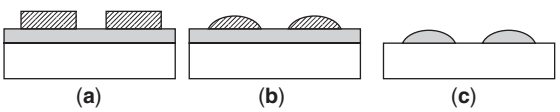


Figure 20.4 Microlens fabrication: (**a**) initial resist profile; (**b**) after resist flow at $T > T_g$ and (**c**) after etching by a 1:1 selectivity etch process

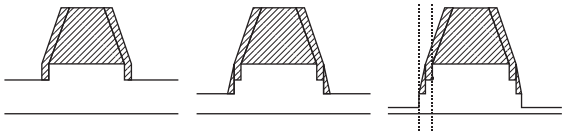


Figure 20.5 CD gain (linewidth increase): resist erosion products and platinum redeposit on resist sidewalls. This debris acts as additional mask, leading to wider lines

which leads to physical sputter etching and severe resist erosion, like in chlorine plasma-etching of platinum. Sputtered (non-volatile) etch products and eroded resist redeposit on the sidewalls of the already etched structures, making them apparently wider. This debris acts as additional masking when etching continues.

20.4 NON-MASKED ETCHING

Plasma etching replaced wet etching because of less undercut and better CD control. But this argument applies to patterning etching only; there are plenty of applications in which etching is done without photoresist or hard mask pattern. Spacer formation is one. It relies on etching anisotropy. Spacers are sometimes regarded as residues (bridging neighbouring metal lines) but sometimes regarded as useful elements, depending on the following process steps.

Spacers are formed when a conformal film is anisotropically etched. If the underlying structures are lines or dots, spacers result in apparently wider structures; but if the original structures are holes or trenches, spacers will make them smaller. Inside spacers (Figure 20.6) make features smaller by 2X film thickness. Inside spacers can be used to study structures smaller than the lithographic capability; for example, in studying scaling of contact resistance, contact holes can be made smaller than the optical lithography limit, without resorting to electron beam lithography.

In etchback process, a thin film is etched immediately after deposition with no patterning step in-between. CVD tungsten fills contact plugs (Figure 20.7), and it is needed in plugs only. Etchback removes tungsten from planar areas. Initially, etchable area is 100% of

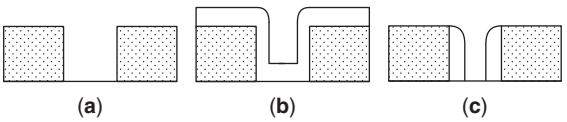


Figure 20.6 Inside spacer (**a**) initial structure; (**b**) after conformal deposition and (**c**) after anisotropic etching

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

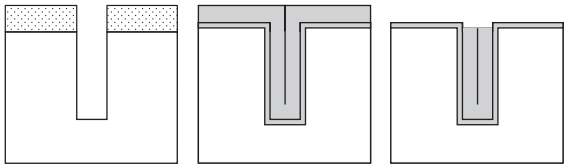


Figure 20.7 Trench/plug fill (**a**) trench etching; (**b**) thin liner plus thick conformal (CVD) deposition and (**c**) etching will result in planar surface (with some plug recess)

the wafer area, but at etching end point the situation changes dramatically: the plugs may represent only a few percent of the wafer area, and the etch rate will go up as all the etch gases attack the tungsten in the plugs.

20.4.1 Etchback planarization

Etchback planarization (Figure 20.8) depends on two factors: smoothing of the surface by spin-coated film, and transfer of this smoothed surface into the underlying layer by etching. When etch selectivity between the spin-coated layer and the underlying layer is 1:1, a true replication of the topography will take place.

Both polymeric and inorganic spin-films are used for planarization. Smoothing is similar for both materials, but etching is very different: glass-like materials (for example SOG) are fairly close to CVD oxides as far as etching is concerned, and 1:1 selectivity can be achieved. With polymers, selectivity tailoring is much more difficult.

Some inorganic spin-films can be left as permanent parts of the device and this is a great simplification in processing, but an additional CVD oxide deposition is still needed: more oxide needs to be deposited in order to obtain the correct thickness of dielectric. If spin-films are left as structural parts, there is the problem of outgassing: during subsequent vacuum deposition steps, spin-films outgas and these outgassing products may interfere with vacuum deposition of metal. *Via poisoning* is the name for poor electrical quality of vias due to outgassing.

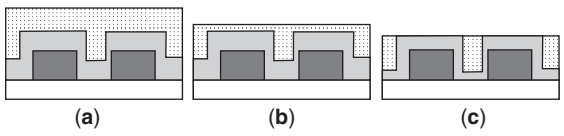


Figure 20.8 Etchback planarization (**a**) planarizing film deposition; (**b**) etchback mid-way and (**c**) at the end of the etch back process planarizing film remains in the gaps

The planarization wavelength of spin-film is a few micrometres or tens of micrometres in the lateral direction. They are thus methods for local planarization only. Etchback with dummy patterns can provide global planarization, at the expense of more complex design and processing.

20.5 PATTERN SIZE AND PATTERN DENSITY EFFECTS

20.5.1 Loading effects

Loading effect or area-dependent reaction rate is a common phenomenon in chemical reactions. For a process optimized for a certain etchable area, the flow may not be high enough to supply reactants to keep the etch rate identical when area is increased by, for example, changing designs: this is a major problem for ASIC manufacturers who face hundreds of different designs.

Loading effect is very general and it operates in all etching processes. It manifests itself when reactions are under mass-transport/diffusion-limited regime. Surface reaction–controlled reactions do not exhibit loading effects.

Loading effects operate at various scales:

- in batch reactors, the etchable area changes because the number of wafers changes;
- in single-wafer reactors, different chip designs have different etchable areas;
- local patterns on the chip are different in every design.

Microloading manifests itself as an etch-depth difference between isolated and array features: there is more material to be etched in arrays, therefore, the rate is lower (Figure 20.9(a)). Microloading can also manifest itself as profile microloading: the lines at the edges of arrays will have a different slope from those in the middle. Microloading results in different etched depths for identical linewidths, dependent on neighbouring structures. Other pattern dependencies discussed below are deceptively similar, yet different.

20.5.2 RIE-lag and aspect-ratio dependent etching (ARDE)

Plasma etching of 1:1 aspect ratio structures is fairly straightforward but at an aspect ratio somewhere around

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

2:1, a phenomenon known as *RIE-lag* manifests itself: smaller features etch slower than larger features. Gas conductance in deep narrow holes is low and the reactants simply cannot reach the bottom effectively (similarly, reaction product removal is hindered). RIE-lag is not related to RIE-reactors; it is present in all plasma-etching systems irrespective of actual reactor design.

RIE-lag can be seen from a single SEM cross-sectional micrograph: one etch time but many different linewidths are compared (Figure 20.9(b) and (c)). Aspect ratio–dependent etching (ARDE) is a dynamic effect: aspect ratio increases as etching proceeds, for every linewidth. At a high aspect ratio, etching slows down because reactant-transport into (and reaction product transport out of) high aspect ratio structures is hindered. The basic reason for RIE-lag and ARDE is thus the same. In order to see ARDE, many wafers have to be etched, with different etch times.

DRIE is fairly straightforward for structures with aspect ratios of 10:1 while 20:1 is more demanding. And even though 40:1 has been demonstrated in the lab, it is not to be considered a standard fabrication step. For 380 μm wafers, these numbers translate to ca. 40 μm, 20 μm and 10 μm trench widths in through-wafer structures, and holes have even more severe dependency on aspect ratios than long trenches. In bonded SOI wafers, device layer thicknesses range from 5 μm upwards. Feature size is then limited by lithography and undercutting of pulsed (Bosch) process rather than by aspect ratio effects.

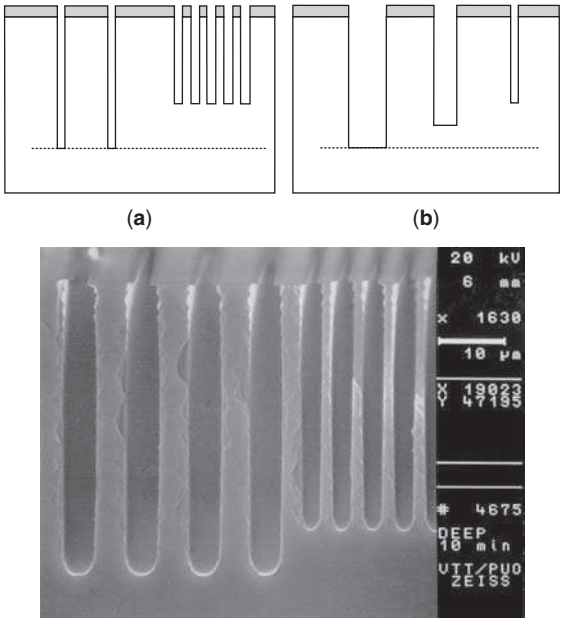


Figure 20.9 (**a**) Microloading effect: etch rate is lower for lines in dense arrays compared with isolated lines of the same width; (**b**) RIE-lag schematic: narrow patterns etch at slower rate than wider patterns and (**c**) RIE-lag SEM micrograph (sidewall undulation is typical of Bosch process with pulsed etching)

20.6 ETCH RESIDUES AND DAMAGE

Many etching reactions rely on polymer deposition for anisotropy. It is usual that, for example, CF_2* radicals that are formed in the discharge polymerize on the sidewalls of the etched features and protect the sidewalls from etching. Removal of these polymers can be extremely difficult. Often, etch products are incorporated into a sidewall polymer film. Sidewall polymer films often require multi-step removal, for example, plasma stripping in oxygen followed by a NH_4OH:H_2O_2 wet clean (RCA-1).

Etchability is intimately related to vapour pressure of the etch products. $AlCl_3$ has a fairly low vapour pressure and aluminium is thus difficult to etch. Aluminium has poor electromigration resistance and copper is often added to aluminium films to improve electromigration resistance. But copper chlorides are even less volatile than $AlCl_3$, and often leave residue. Ion bombardment can sputter them away, but at the expense of decreased resist and oxide selectivity. A balance has to be found between electromigration resistance and copper residues: 2%wt Cu in Al is often chosen as a compromise.

Charge can accumulate on isolated conductors, and the oxide beneath these conductors can be damaged by this charge accumulation. Not only plasma etching but all plasma processes, PECVD and sputtering contribute to this damage.

20.7 EXERCISES

1. Molybdenum etching in Cl_2/O_2 plasmas results in oxychlorides such as $MoOCl_4$. The etch rate is 300 nm/min, molybdenum film thickness is 300 nm and film non-uniformity and etch process non-uniformity across the wafer are both 5%. The selectivity of Mo:oxide is 20:1. Calculate oxide loss as a function of overetch time.
2. Determine the DRIE single-crystal silicon etch rate from the following trench etching data.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Etch time (min)	Etched depth (μm)		
	80 μm wide	40 μm wide	12 μm wide
20	109	104	85
40	205	193	156
60	292	278	215

3. Redo exercise 11.8 with resist effects included. Draw cross-sectional figures of the shown structure under the following etch conditions, for two etch times: right at etch end point; and after 50% overetch.

A etch	A:B	A:S
Process	Selectivity	Selectivity
Anisotropic	1:1	∞
Anisotropic	5:1	5:1
Isotropic	1:1	∞
Isotropic	5:1	5:1

4. What is the difference in making inside versus outside spacers by anisotropic etching?
5. How much etch non-uniformity can native oxide cause in polysilicon RIE?
6. What must SF_6 gas flow be in a DRIE reactor if the silicon etch rate is 10 μm/min, wafer size is 150 mm and etchable area is 20%?

REFERENCES AND RELATED READINGS

Armacost, M. *et al*: Plasma-etching processes for ULSI semiconductor circuits, *IBM J. Res. Dev.*, **43** (1999), 39.

Chen, K.-S. *et al*: Effect of process parameters on the surface morphology and mechanical performance of silicon structures after deep reactive ion etching (DRIE), *J. MEMS*, **11** (2002), 264.

Franssila, S. *et al*: Etching through silicon wafer in inductively coupled plasma, *Microsyst. Technol.*, **6** (2000), 141.

Gottscho, R.A. *et al*: Microscopic uniformity in plasma etching, *J. Vac. Sci. Technol.*, **B10** (1992), 2133–2147.

Kiihamäki, J. & S. Franssila: Pattern shape effects and artefacts in deep silicon etching, *J. Vac. Sci. Technol.*, **A17** (1999), 2280.

MacDonald, N.C.: SCREAM MicroElectroMechanical Systems, *Microelectron. Eng.*, **32** (1996), 49.

Murari, B.: Lateral thinking: the challenge of microsystems, *Transducers '03* (2003), p. 1.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

21

Wet-etched Silicon Structures

Microsystems technology relies on anisotropic wet etching of silicon for many major applications. Bulk micromechanics depends on silicon crystal plane–dependent etching, and many surface micromechanical and SOI devices make use of silicon wet etching for auxiliary structures, even though main device features are defined by plasma etching. Because <100> silicon is the workhorse of microsystems, the discussion concentrates on it. Both <110> and <111> etching will be reviewed briefly.

21.1 BASIC STRUCTURES ON <100> SILICON

Etched grooves, trenches and wells exemplify the basic features of crystal plane–dependent etching. They can be used as sample wells and flow channels in microfluidics, or as optical fibre-alignment fixtures. Other basic structures are diaphragms (membranes), beams and cantilevers. Mechanical devices such as pressure sensors, resonators and AFM cantilevers rely on these basic elements. Through-wafer structures include nozzles and orifices, for example, for ink jets or micropipettes.

Anisotropic etching relies on aligning the structures with wafer crystal planes (Figure 21.1). The primary flat, which is along the [110] direction, is used as a reference. Rectangular structures with concave corners are easily made, with four (111) sidewalls and the (100) plane as the bottom. If the slow etching (111) planes meet, etching will be self-limiting. This process results in inverted pyramids, which were already seen in Figure 1.6(a).

Self-limiting depth is the depth at which the slow etching (111) planes meet. The angle between (100) and (111) planes is 54.7° and the self-limiting depth is given by $\tan 54.7 = d/(W_m/2)$, which gives $d = W_m/\sqrt{2}$ for a mask opening of W_m.

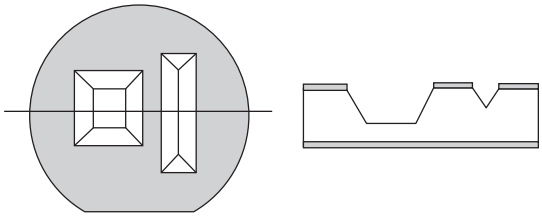


Figure 21.1 Orientation of structures relative to wafer crystal planes is paramount for anisotropic wet etching: (**a**) top view of rectangular shapes on <100> wafer and (**b**) cross-sectional view shown along cut linewidth (oxide mask shown in grey)

21.2 ETCHANTS

A number of alkaline etchants have been tried for crystal plane–dependent etching but KOH has emerged as the main etchant. 1 μm/min is a typical etch rate, which translates to 6 to 7 h for through-wafer etching of 380 μm wafers. KOH poses a contamination hazard for CMOS work, and therefore CMOS-compatible etchants are desirable. Tetramethyl ammonium hydroxide, $(CH_3)_4NOH$, usually known as TMAH, is such a compound. In fact, both NaOH and TMAH are used as photoresist developers, in diluted concentrations and at room temperature, so the contamination danger can be handled with proper working procedures. Organic amines have also been used for anisotropic etching, most notably ethylene diamine ($(NH_2)(CH_2)_2NH_2$) mixture with pyrocathecol and water, known as *EDP* or *EPW*. Hydrazine (N_4H_2) has also been tried. Both amines pose occupational safety and health hazards, and they are not widely used. Ammonia has been shown to etch silicon reasonably well, but the stability of ammonia etch baths during extended etching needs special attention.

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

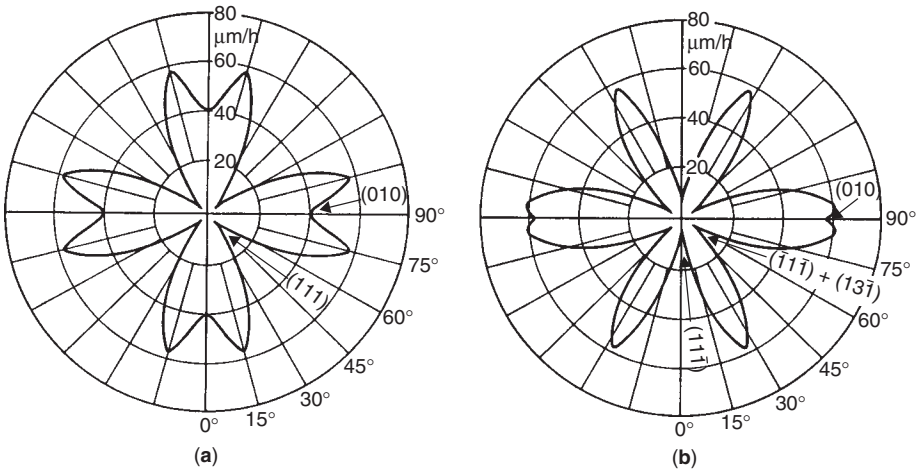


Figure 21.2 Etch rates in different crystal directions in 50% KOH at 78 °C: (**a**) <100> Si: fast, but not maximum etching in (010) direction and (**b**) <110> Si: (010) near maximum etch rate. Reproduced from Seidel, H. *et al*. (1990), by permission of Electrochemical Society Inc

Even though all the alkaline etchants share the same basic features of etching (100) planes fast and (111) planes slowly, the actual selectivity between the crystal planes needs careful attention. KOH has selectivities between (100) and (111) of the order of 200:1, whereas TMAH only exhibits 30:1. These selectivities are dependent on etchant concentration and temperature. But when other crystal planes are considered, even more differences pop up: when planes such as (110) and high-index planes such as (311) are studied, the differences multiply. Figure 21.2 shows etch rates for <100> and <110> silicon in KOH. Identifying minima and maxima etch rate planes is essential for prediction of etched shapes.

Early investigations on etch selectivities were sometimes misleading because wafer miscut will confound etch rate measurement. Discrepancies of a factor of 2, compared with present values, are not unusual.

Isopropanol (IPA) addition into KOH will change the relative etch rates of crystal planes, and depending on exact conditions, either of the (100) or (110) planes will be the maximum etch rate planes.

Because etch times are rather long, evaporation and decomposition of etchant must be prevented. Dissolution of excess silicon in TMAH before etching eliminates changes due to silicon dissolution during etching. Pyrocathecol is employed in EDP for similar reasons: decomposition of ethylene diamine releases small amounts of pyrocathecol, which changes etchant composition, but if pyrocathecol is added in large amounts to begin with, the decomposition has a negligible effect.

21.3 ETCH MASKS AND PROTECTIVE COATINGS

Silicon dioxide and silicon nitride are the common masking materials for anisotropic wet etching. KOH etches oxides fast, while TMAH and EDP, hardly at all. Nitride is more resistant than oxide in both solutions. Mask etch rates depend on temperature and concentration just like silicon etch rates, but some general guidelines can be given. An oxide thickness of 2 μm is needed for through-wafer etching in KOH, whereas 200 nm is enough in TMAH or EDP. Thermal oxide etch rate is slower than that of CVD oxides. Silicon nitride is a better masking material than silicon dioxide, and LPCVD nitride is hardly etched at all, while PECVD nitride etch rates are strongly deposition condition dependent, as is usual with CVD films.

LPCVD nitride is usually under very high stress; gigapascal-range tensile stresses are not atypical. This leads to defects in the underlying silicon, and defects will change etch rates; (100) to (111) crystal plane selectivity can change by a factor of 3. For this reason, pad oxides are employed: as discussed in connection with LOCOS oxidation (Chapter 13), a thin, 10 to 50 nm thermal oxide is grown first, and LPCVD nitride is deposited on this pad oxide in order to eliminate stresses to the substrate.

As a practical issue, it should be noted that thermal oxide and LPCVD nitride are furnace processes and film is grown/deposited on both sides of the wafer so that the backside of the wafer is protected. This is important when deep etching is done. PECVD deposition is usually on the front side of the wafer only.

Page 227

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

All silicon etchants etch aluminium, which means that either aluminum deposition has to be done after silicon etching, or aluminium has to be protected during silicon etching. In some cases aluminum can be replaced by another metal, such as gold. Some relief can be achieved by saturating TMAH solution with silicon, but typically only very short alkaline etchings are done after metallization.

21.4 ETCH RATE AND ETCH STOP

KOH rate can be made very high: the boiling point of 50% KOH is ca. 150 °C, which translates to ca. 10 μm/min etch rate for (100) planes. But in addition to rate, other factors must be considered: surface roughness increases in alkaline etching beyond bonding quality, so the surfaces to be bonded must be protected by oxide or nitride mask during KOH etching. There have been experiments with ammonia etching with arsenic oxide: etch rates of 1.5 μm/min at 70 °C have been demonstrated, with high selectivity against oxide and aluminum masks and very smooth surfaces, 2.4 nm RMS roughness, whereas typical KOH-etched surfaces exhibit 5 to 10 nm RMS roughness. Arsenic and antimony additions to KOH have shown similar results of improved surface smoothness and increased rate. Standard etch processes are compared in Table 21.1. Practical etch rates are in the range 0.5 to 1 μm/min.

Table 21.1 Alkaline anisotropic etchants: some main features of etchants

Etchant	KOH	TMAH	EDP
Rate (at 80 °C) μm/min	1	0.5	1 (at 115 °C)
Typical concentration	40%	25%	80%
Selectivity (100):(111)	200:1	30:1	35:1
Selectivity Si:SiO_2	200:1	2000:1	10 000:1
Selectivity Si:Si_3N_4	2000:1	2000:1	10 000:1
Etch stop factor (10^{20} cm^{-3})	25	10	50

Etch stop is an idealization; infinite selectivities are not met with in the real world. High selectivity is termed *etch stop* when selectivity is so high that etch timing becomes non-critical. Etch stop can happen through various mechanisms.

Etch rate of boron-doped silicon decreases rapidly when the doping level exceeds 10^{19} cm^{-3} (Figure 21.3). The exact mechanism is unknown but high stresses in heavily doped silicon may play a part. Boron etch stop is frequently used in bulk micromechanics, as a way to fabricate simple mechanical structures. The silicon microbridge shown in Figure 2.1(b) was done by p^{++} etch

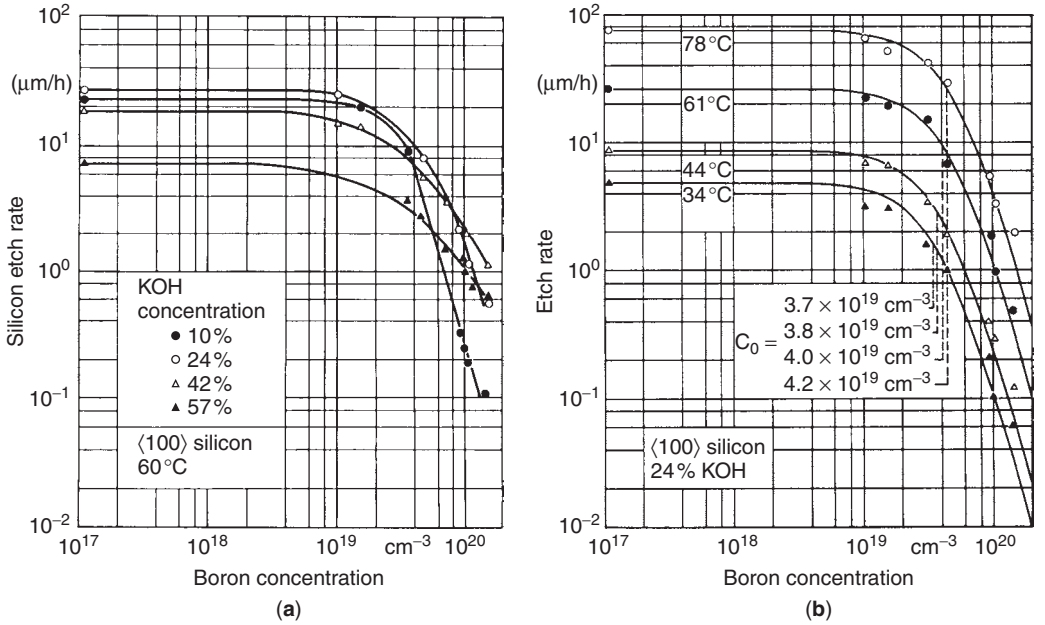


Figure 21.3 p^{++} etch stop: (**a**) with KOH concentration as a parameter and (**b**) with etch temperature for 24% KOH as a parameter. Reproduced from Seidel, H. *et al*. (1990), by permission of Electrochemical Society Inc

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

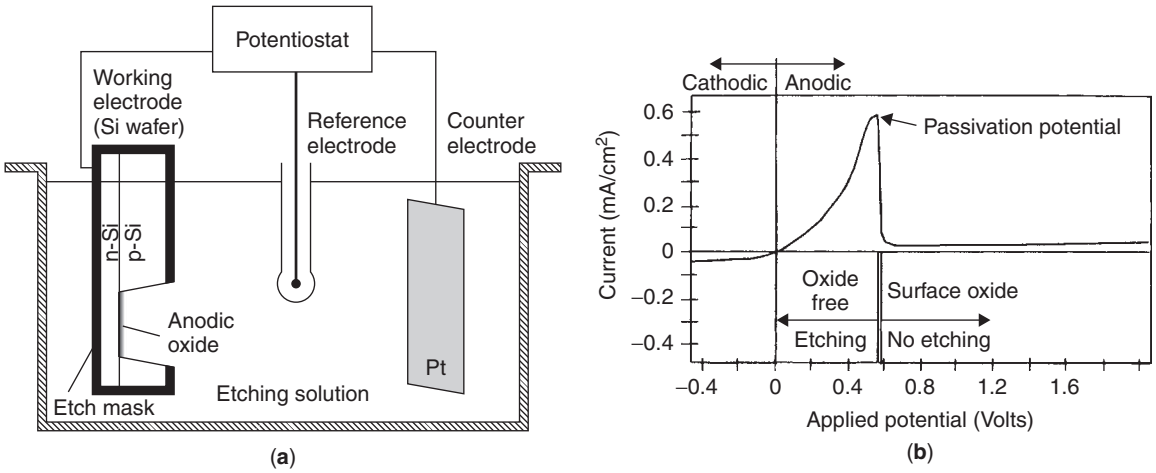


Figure 21.4 (**a**) Electrochemical cell for silicon electrochemical etching in KOH: p-type silicon etched; n-silicon passivated by anodic oxide. Reproduced from Wong, S.S. *et al*. (1992), by permission of Electrochemical Society Inc and (**b**) passivation potential and anodic oxidation regime. From Collins, S.C. (1997), by permission of IEEE

stop. It is, however, not possible to fabricate electrical devices on such a highly doped material. For instance, piezoresistors cannot be made by doping because the p^{++} etch stop doping level is higher than the piezoresistor doping level. The stresses in p^{++} doped structures make them mechanically inferior to lightly doped material. Furthermore, slips are introduced in silicon because of high stresses, and this makes bonding of highly doped wafers difficult.

21.4.1 Electrochemical etch stop

When a silicon wafer is an anode in an alkaline-etching solution biased positively above passivation potential, the surface will be oxidized, which stops silicon dissolution. The n-type layer of a pn-structure can similarly be protected. Positive potential, above passivation potential, is applied to the n-type layer (Figure 21.4). Etching of p-type silicon continues until the diode is destroyed, and n-type silicon is then passivated.

21.5 DIAPHRAGM FABRICATION

There are two basic diaphragm (membrane) structures: either the diaphragm is made of a deposited film or it is made of single-crystal silicon. In the first case, etching is quite simple: all the silicon is removed and the thin film remains. There are two main considerations for the membrane material: it has to be (slightly) tensile-stressed because a compressively stressed film would buckle and a too highly tensile-stressed film would crack. The film has also to be resistant to alkaline etchants. Silicon nitride fulfils both requirements, and it is almost universally used. It is also electrically (and thermally) insulating so that resistors can be readily deposited on it, and it is optically transparent.

Silicon diaphragm fabrication, pictured in Figure 21.5(b), relies on timed etching, but this is a very unsatisfactory approach if thin membranes are needed. Depending on the device requirement on the membrane, 40 μm is the thinnest that can reasonably be made by timed etching in a manufacturing environment.

p^{++} etch stop has two variants: either the p^{++} layer is made by diffusion (or implantation) or it is an epitaxial layer. Because the doping levels required for etch stop are very high, diffusion p^{++} is limited to very thin membranes. If pn-junction etch stop is utilized, we have again the same alternatives: diffusion doping and epitaxy. Additionally, the n-layer has to be electrically contacted, and this contact has to be protected from the alkaline silicon etchant. Holders of various designs have been invented, with the drawback that part of the wafer front side is used for sealing the holder, leading to silicon

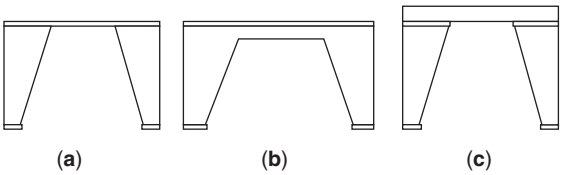


Figure 21.5 Nitride, bulk silicon and SOI diaphragms

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

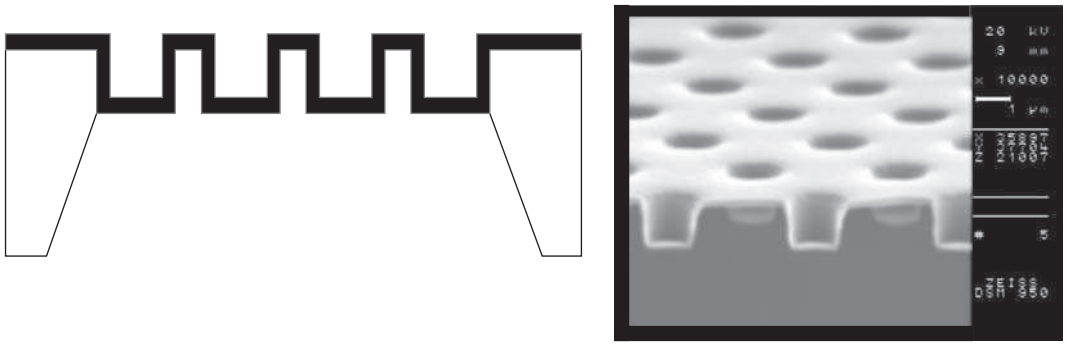


Figure 21.6 Corrugated diaphragm: grooves etched in silicon, filled with membrane material, released by backside etching. Diaphragms can be made of silicon nitride or parylene, for example. SEM micrograph courtesy Kestas Grigoras, Helsinki University of Technology

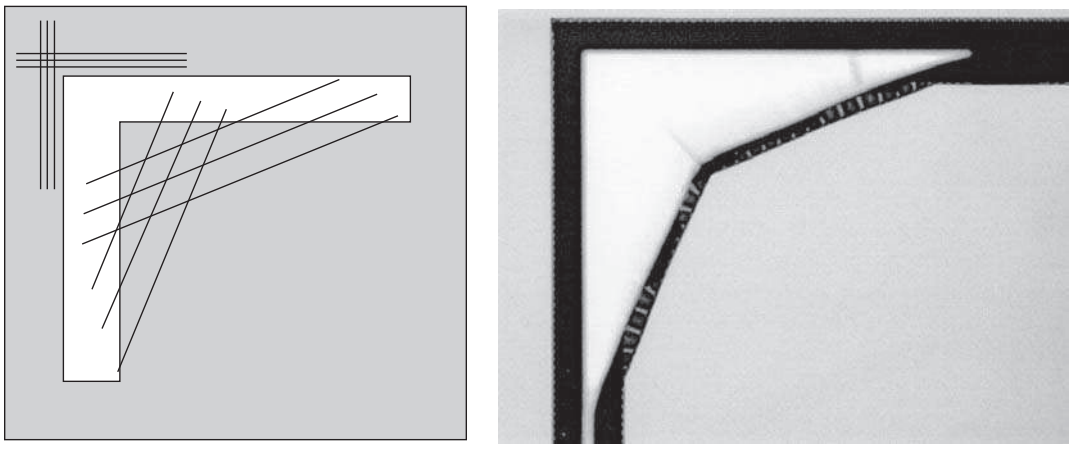
real estate loss of sometimes up to 20% fewer chips than in free etching.

SOI wafers offer an elegant but somewhat expensive way of making membrane structures (Figure 21.5(c)). The buried oxide of SOI acts as an etch-stop layer, leaving the SOI device layer untouched by the etch process. Bonded SOI device layer thicknesses are usually specified at ca. 10%, so that a 10 μm membrane with $\pm 1\,\mu$m thickness variation results.

Corrugated membranes (Figure 21.6) (and U-shaped beams) are stiffer than planar ones, and these can be made by one extra lithography step: patterning of the grooves. Membrane etching is identical to planar membrane etching but step coverage and film quality on the sidewalls may introduce some problems.

21.6 COMPLEX SHAPES BY <100> ETCHING

The etch rate of (100) planes is high relative to that of (111) planes. When simple concave shapes are etched, the fast etching planes will disappear and the slow etching (111) planes will dominate in the final structure. The fastest etching planes, usually (110) and some high-index planes such as (311), are not present in the simple rectangular wells, channels and nozzles, which have only concave 90° inside corners. Convex corners reveal these high etch rate planes, and rapid corner rounding takes place, as shown in Figure 21.7. The etched shape is initially determined by the fast etching planes, but the structures will finally be limited by the slow etching (111) planes.



(**a**) (**b**)

Figure 21.7 Convex corner (270°) reveals fast-etching high-index planes leading to rapid corner undercut; concave corner (90°) will be etched slowly because (111) planes are exposed. Optical microscope image after etching. Photo courtesy Seppo Marttila, Helsinki University of Technology

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

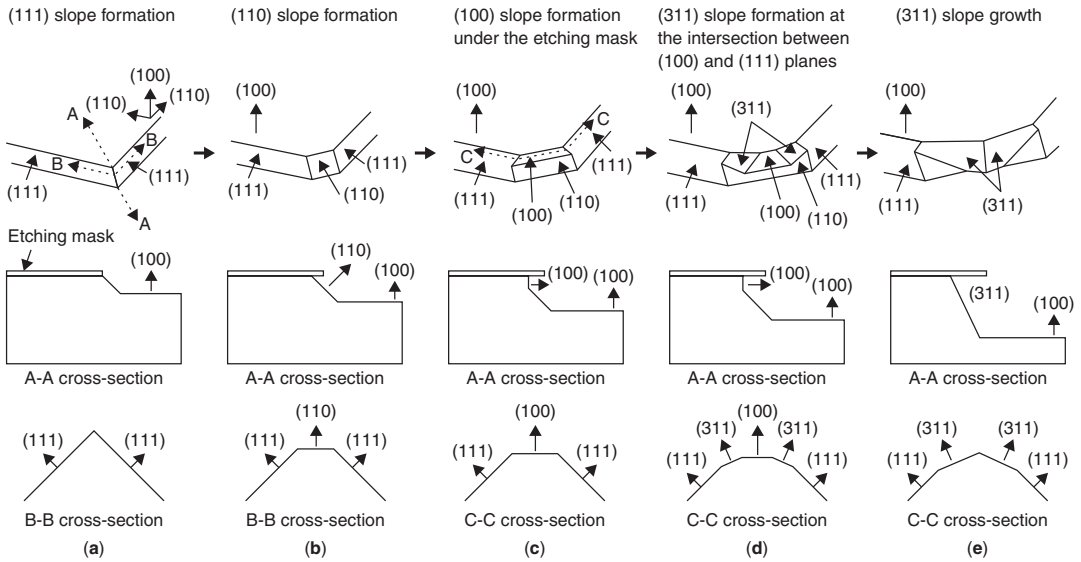


Figure 21.8 Convex corner undercutting time evolution. Reproduced from Shikida (2001), by permission of Springer

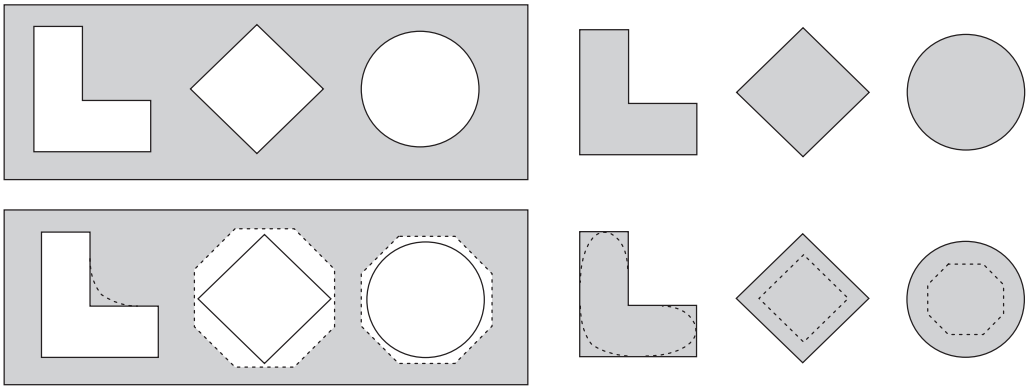


Figure 21.9 The effect of mask polarity on shape: top row; initial mask opening; bottom row and etched shape (oxide mask shown grey)

Time evolution of various structures, with convex and concave corners, are shown in Figures 21.8, 21.9 and 21.10.

If the structures are aligned along the [100] direction (45° relative to wafer flat) instead of the usual flat direction [110], new possibilities arise. For instance, 45° walls suitable for fibre coupling mirrors and 90° sidewall mesas can be made. These structures depend on relative etch rates of (100) and (110) planes according to Conditions 21.1 and 21.2:

$$\text{rate}\{100\}/\text{rate}\{110\} < 1/\sqrt{2} \qquad 90^\circ\text{walls} \quad (21.1)$$

$$\text{rate}\{100\}/\text{rate}\{110\} > \sqrt{2} \qquad 45^\circ\text{walls} \quad (21.2)$$

Condition 21.1 leads to vertical walls that are (100) planes, and Condition 21.2 leads to 45° walls that are (110) walls. This is shown in Figures 21.11 and 21.12. KOH etchant, 25 to 50%, fulfils Condition 21.1, and KOH–IPA solution is an example of Condition 21.2. When the rate condition is close to limit values, as is the case with <25% TMAH, inadequate stirring or some other disturbance can lead to unexpected changes in final shapes.

If double-sided lithography and etching is done (to be discussed in more detail in Chapter 28), more elaborate shapes appear, for example, vertical sidewalls and inward slanted (111) planes. This is illustrated in Figure 21.13.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

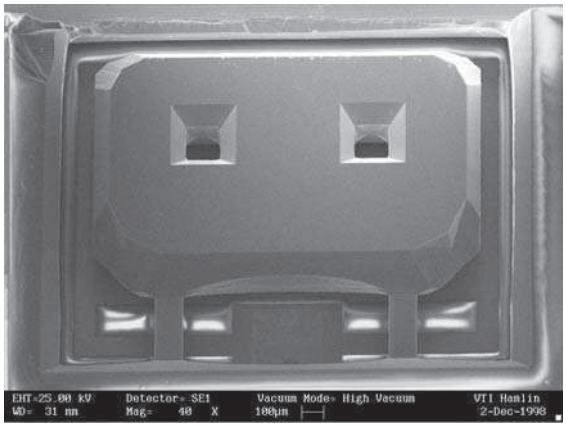


Figure 21.10 Bulk silicon micromachined accelerometer: a 380 µm thick wafer has been etched through: concave holes show familiar <111> limited sidewalls, but at convex corners fast etching planes have been revealed. Photo courtesy Risto Mutikainen, VTI Technologies

Simulation of anisotropic wet etching has been around for years but until recently it has not had a major impact. New simulation tools such as MICROCAD can take into account most of the crystal plane effects and double side etching as well. MICROCAD is a geometric simulator based on experimentally determined etch rates of crystal planes. The alternative is the atomistic approach: bond directions, bond breakage and bond energies are analysed. Atomistic simulators can explain surface roughness, which is beyond the capabilities of geometric simulators.

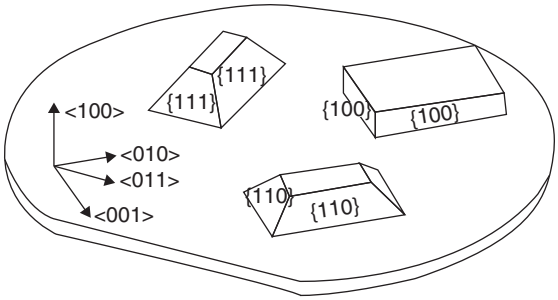
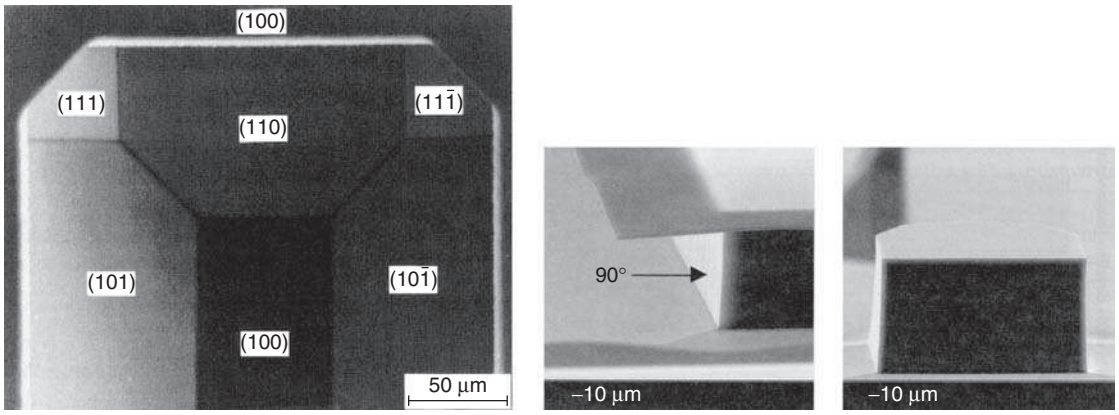


Figure 21.11 Orientation of structures on (100) wafer. Alignment to wafer flat leads to 54.7° angles and {111} sidewalls. Alignment 45° relative to flat leads to {110} walls and {100} vertical walls result when rates of {110} relative to {100} fulfil Conditions 21.1 and 21.2. Reproduced from Powell, O. & H. Harrison (2001), by permission of IOP

21.7 FRONT SIDE BULK MICROMACHINING

Cantilevers and bridges can be made by front side micromachining by undercutting. Either convex corners are designed into release etch openings (Figure 21.14), or else the structures are aligned not to main axes of silicon, but for example 45° off, so that fast etching planes appear. This method was used to make the silicon bridge in Figure 2.1(b).

All structures made on the bridges, membranes or cantilevers have to be processed before the silicon release etch because topology and topography do not allow lithography after release. Piezoresistors, thermopiles and AFM tips are typical devices on



(**a**) (**b**)

Figure 21.12 (**a**) 45° slanted sidewalls in <100> wafer by 45° degree off-orientation. Reproduced from Strandman, C. *et al*. (1995), by permission of IEEE and (**b**) 90° angles in <100> wafer, before and after etch-mask removal. Note the severe undercut that is unavoidable to make vertical walls in <100>. From Vazsonyi, E. *et al*. (2003), by permission of IOP

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

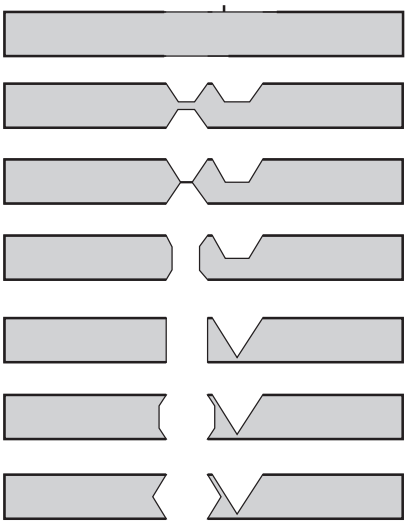


Figure 21.13 Etching through <100> silicon from two sides simultaneously. Reproduced from Nijdam, A.J. *et al.* (1999), by permission of IOP

cantilevers. Structures already made, resistors, junctions, tips, have to be covered during silicon etching, but because etch times are short compared to backside through-wafer etching, CVD oxide films of standard thickness ($<1\,\mu$m) can be used as protective coatings.

21.8 CORNER COMPENSATION

We noted in Section 21.6 that convex corners are dominated by (311) planes (Figure 21.8). In many designs, it would be very useful to have sharp corners. This is possible with a little extra effort in mask design by adding compensation structures, shown in Figure 21.15.

The fast etching planes start to erode at convex corners. But the final convex corner is protected by this sacrificial structure so that after the compensation structure has been etched away, a rectangular corner remains.

Timing is the difficult part: if etching is stopped too early, a peak remains on the corner. Overetching leads to a structure with an undercut corner, similar to the non-compensated case but with less undercut. Even though this method looks perfect in two dimensions, it leaves some small <311> surfaces in three dimensions, as seen in Figure 21.8. Another shortcoming of this method is that it takes a lot of space to form these compensation structures.

21.9 <110> ETCHING

Silicon of <110> orientation offers an interesting possibility to anisotropically wet etch perfectly vertical walls when the mask is aligned so that slow-etching (111) planes form the sidewalls (Figure 21.16). However, just as in the case of <100> silicon etching, the relative rates of different crystal planes can be changed by etchant concentration and temperature. It is possible to find conditions in which square bottom profile can be achieved, for instance, KOH (23% wt)-H_2O-isopropanol (10–15% wt) at 85 °C or 30% KOH at 70 °C.

Under other etch conditions (for instance with 40% KOH at 70 °C), a self-limiting shape, U-groove, is met (Figure 21.17). U-grooves are self-limiting just like V-grooves on (100) wafers, when planes that etch slower than (110) appear. Etching will proceed until the six

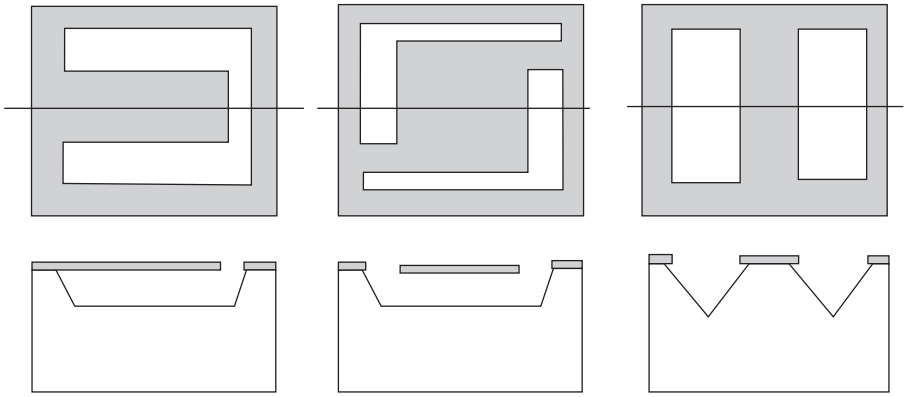
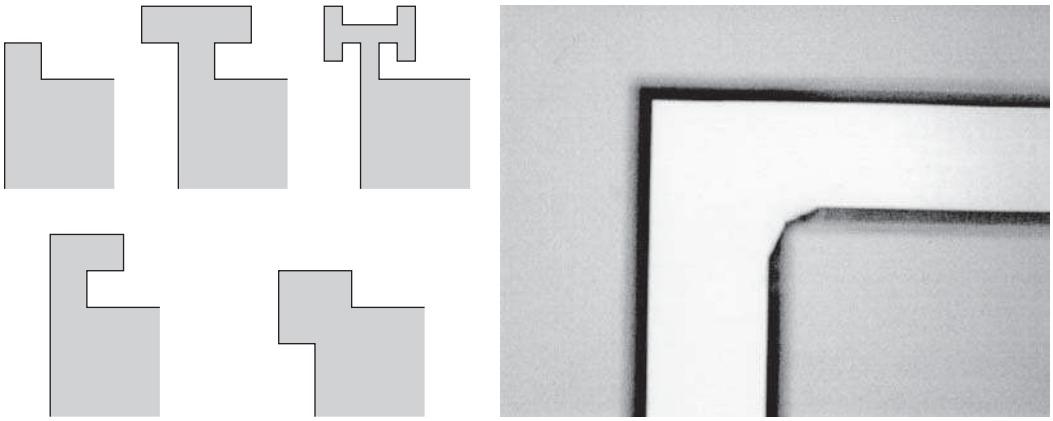


Figure 21.14 Cantilever and bridge structures by front-side etching. Underetching from convex corners is used, with structures aligned to the [110] main axes on a wafer. Simple rectangular holes along [110] axis result in V-grooves only

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030



(**a**) (**b**)

Figure 21.15 (**a**) Different designs for corner compensation. Figure courtesy Ville Voipio, Helsinki University of Technology and (**b**) optical microscope image of a compensated corner after etching. Photo courtesy Seppo Marttila, Helsinki University of Technology

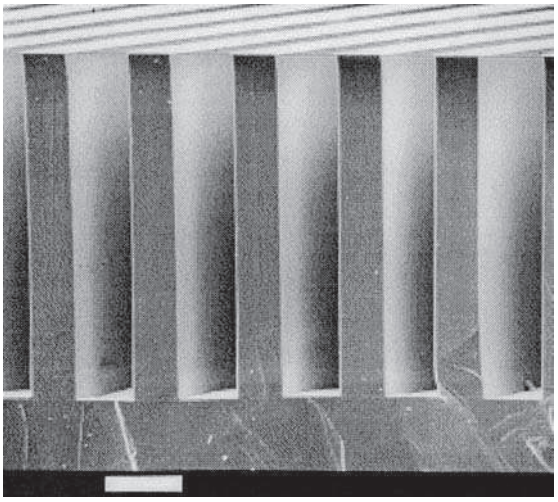


Figure 21.16 Rectangular groove bottoms in KOH–IPA etching of <110> silicon. Reproduced from Dwivedi, V.K. *et al*. (2000), by permission of Elsevier

slow etching (111) planes meet. U-grooves' self-limiting depth D is given by Equation 21.3 for initial mask opening sizes a and b (Figure 21.18)

$$D = (a + b\sqrt{2})/2\sqrt{6} \qquad (21.3)$$

A major limitation of vertical walled structures on (110) silicon is that only diamond shaped structures (with 70.5 and 109.5 degree angles) will have all four walls vertical. Rectangular shapes will turn into hexagons, but diamond oriented along crystal axes will retain their shape in the etching process (Figure 21.18).

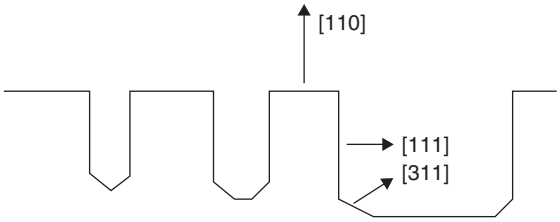


Figure 21.17 Etching of <110> silicon: slow etching (111) planes form vertical sidewalls. Depending on etchant concentration, composition and temperature, slow etching planes start limiting the groove (compare with Figure 21.1)

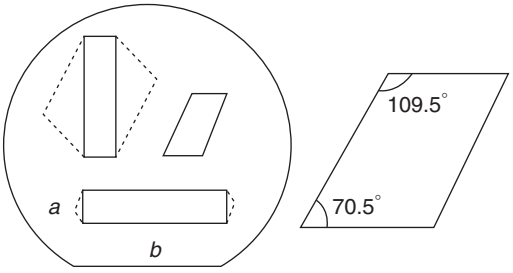
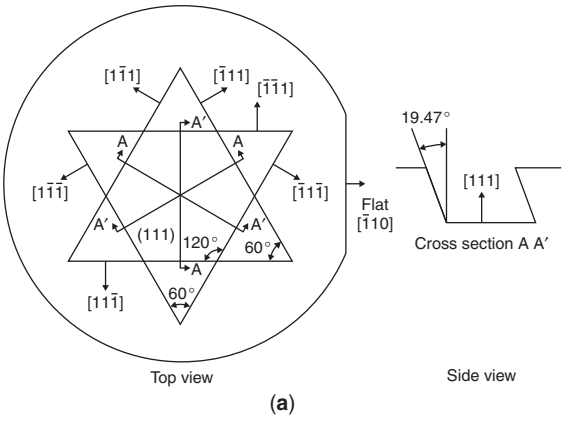


Figure 21.18 <110> etched shapes: solid lines indicate mask openings; dashed lines final etched shapes. Diamond oriented along major crystal axes retain their shape

21.10 <111> SILICON ETCHING

<111> silicon wafers cannot be etched in KOH because (111) planes are the slow etching planes. If, however, initial trenches are opened by plasma etching, other

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030



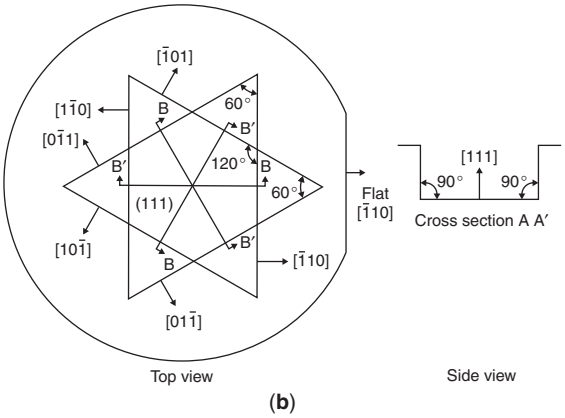


Figure 21.19 <111> silicon crystal planes. Note the hexagonal symmetry. Not all walls are bound by slow etching (111) planes. Reproduced from Park, S. *et al*. (1999), by permission of Institute of Pure and Applied Physics

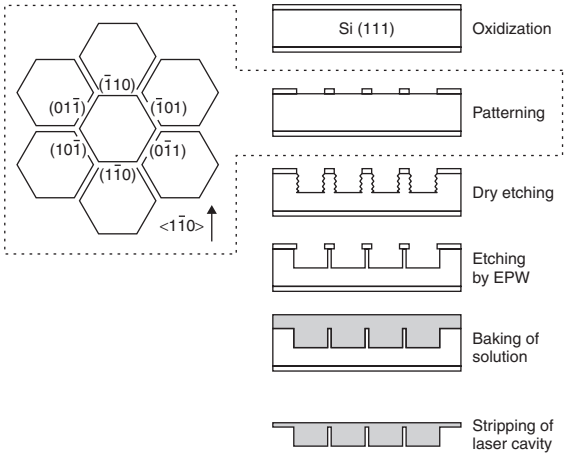


Figure 21.20 Hexagonal symmetry of <111> silicon is utilized in making vertical sidewall structures of (110) planes which are local etch rate minima planes in EPW. Reproduced from Sasaki, M. *et al*. (2000), by permission of Institute of Pure and Applied Physics

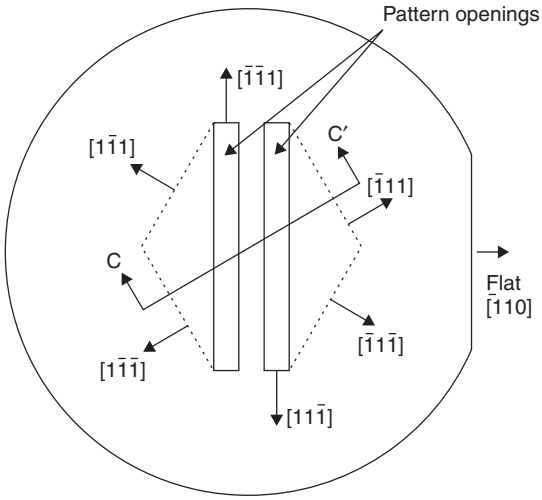


Figure 21.21 Etching of <111> silicon bridge: two rectangular pattern openings are undercut, and etching will proceed until slow etching (111) planes are met. Undercutting to the left and right of the bridge is large compared to bridge width. Reproduced from Park, S. *et al*. (1999), by permission of Institute of Pure and Applied Physics

crystal planes will be exposed. The depth of the structure is determined by the initial plasma etch step because the bottoms are (111) planes just like the wafer surface and they do not etch further in KOH.

The sixfold symmetry that was seen in the vertex view of the silicon crystal (Figure 4.5) is evident in <111> wafers (Figure 21.19). Triangular and hexagonal patterns will retain their shapes if oriented properly (Figure 21.20). The sidewalls will be either 70.5° or 90°. Rectangular structures will end up as hexagons when (111) planes meet (Figure 21.21).

Sidewalls of (111) are very smooth compared to plasma-etched sidewalls, and in some applications, wet etching is used as a self-limiting, self-aligned smoothing method after DRIE. Figure 21.20 shows a honeycomb-shaped trench pattern that acts as a master for polymer optical-device casting.

Free-standing thin-film structures can be made by etching an initial release hole, and then continuing with

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

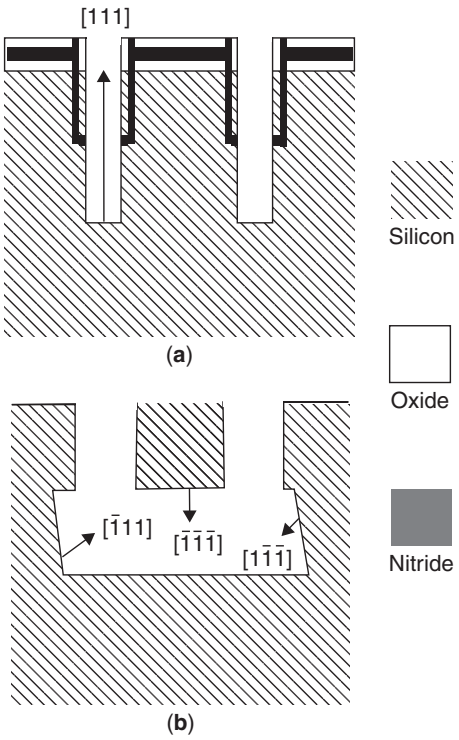


Figure 21.22 Silicon bridges in (111) silicon: First RIE defines silicon-bridge thickness. A spacer is formed before the second RIE step, which defines the release gap. The spacer protects the bridge during undercutting etch in KOH. Reproduced from Park, S. *et al*. (1999), by permission of Institute of Pure and Applied Physics

anisotropic wet etching. Complete undercutting leads to free-standing structures not unlike those made on (100) silicon. However, lateral undercutting in some directions is fairly large, as shown in Figure 21.21.

If free-standing silicon bridges and beams need to be made, an approach similar to that shown in Figure 20.2 can be used: sidewall oxide protection results in silicon bridges without heavy p^{++} doping. Bridge thickness is determined by the first RIE step and release gap thickness by the second RIE step, as shown in Figure 21.22. The depths of the RIE steps are not very accurate but since the bridge roof and ceiling are slow etching (111) planes, surface quality is excellent.

21.11 COMPARISON OF <100>, <110> AND <111> ETCHING

If an initial trench has been etched in the wafer by anisotropic plasma etching (i.e., vertical sidewalls),

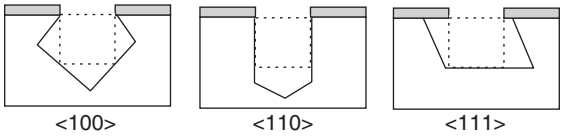


Figure 21.23 Initial plasma etched groove shown by dotted lines; wet etched final shape by solid lines. Other shapes are possible depending on structure orientation relative to wafer flat

anisotropic wet etching will proceed until slow etching (111) planes are met. On a (100) wafer, this will result in a rhombohedric structure with 54.7° angles. On a (110) wafer, the flat bottom will be further etched, and depending on relative etch rates in the etchant in question, either the flat bottom remains or the U-groove sets in. On (111) wafers, either vertical or slanted walls will result, depending on pattern orientation (Figure 21.23).

21.12 EXERCISES

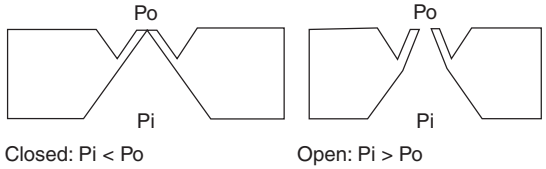
1. Silicon <100> wet etch rate in 25% KOH at 90 °C has been measured to be 2.5 μm/min, and the activation energy was determined to be 0.61 eV (59 kJ/mol). If 340 μm deep structures need to be etched and the etch bath temperature is controlled to ±1 °C, what uncertainty does this introduce in the etch time?
2. Rate vs. temperature data for <110>; silicon etching in 30% KOH is given below. What is the activation energy?

30	40	50	60	70	80	90	100 °C
4.7	9.8	19.4	37	68	121	209	350 μm/h

3. Micromechanical pressure sensor chips have 40 μm thick diaphragms that are 1 × 1 mm in area. How many such chips can be made on
   (a) 380 μm thick 3 inch wafers?
   (b) 525 μm thick 100 mm wafers?
   (c) 675 μm thick 150 mm wafers?
4. <110> wafer-etch selectivity between (110) and (111) planes is measured from SEM cross sections: etched depth and mask undercut are recorded. How does finite mask etch rate affect the result?
5. What is the angle between the (111) and (311) planes shown in Figure 21.17?
6. Design 'corner compensation' structures for etching a circular hole in a <100> wafer.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

7. Design the process and mask for fabrication of silicon bridges on (110) wafers.
8. Design a process to fabricate the duckbill valve shown below.



REFERENCES AND RELATED READINGS

Asaumi, K. *et al*: Anisotropic etching process simulation system MICROCAD analyzing complete 3D etching profiles of single crystal silicon, *Proc. IEEE MEMS* '97 (1997), p. 412.

Collins, S.C.: Etch stop techniques for micromachining, *J. Electrochem. Soc.*, **144** (1997), 2242.

Dwivedi, V.K. *et al*: Fabrication of very smooth walls and bottoms of silicon microchannels for heat dissipation of semiconductor devices, *Microelectron. J.*, **31** (2000), 405.

Elwenspoek, M. & H. Jansen: *Silicon Micromachining*, Cambridge University Press, 1998.

Gosalvez, M.A. *et al*: Anisotropic wet chemical etching of crystalline silicon: atomistic Monte-Carlo simulations and experiments, *Appl. Surf. Sci.*, **178** (2001), 7.

Hannemann, B. & J. Fruhauf: New and extended possibilities of orientation dependent etching in microtechnics, *Proc. IEEE MEMS '98* (1998), p. 234.

Hoffmann, M. & E. Voges: Bulk silicon micromachining for MEMS in optical communication systems, *J. Micromech. Microeng.*, **12** (2002), 349.

Laurell, T. *et al*: Silicon microstructures for high-speed and high-sensitivity protein identifications, *J. Chromatogr., B*, **752** (2001), 217.

Mihalcea, C. *et al*: Improved anisotropic deep etching in KOH-solutions to fabricate highly specular surfaces, *Microelectron. Eng.*, **57–58** (2001a), 781.

Mihalcea, C. *et al*: Ultra-fast anisotropic silicon etching with resulting mirror surfaces in ammonia, *Transducers '01* (2001b), p. 608

Nijdam, A.J. *et al*: Velocity sources as an explanation for experimentally observed variations in Si{111} etch rates, *J. Micromech. Microeng.*, **9** (1999), 135.

Oosterbroek, R.E. *et al*: Etching methodologies in <111>-oriented silicon wafers, *J. MEMS*, **9** (2000), 390.

Park, S. *et al*: Mesa-supported, single-crystal microstructures fabricated by the surface/bulk micromachining process, *Jpn. J. Appl. Phys.*, **38** (1999), 4244.

Powell, O. & H. Harrison: Anisotropic etching of {100} and {110} planes in (100) silicon, *J. Micromech. Microeng.*, **11** (2001), 217.

Sasaki, M. *et al*: Anisotropically etched Si mold for solid polymer dye microcavity laser, *Jpn. J. Appl. Phys.*, **39** (2000), 7145.

Seidel, H. *et al*: Anisotropic etching of crystalline silicon in alkaline solutions I, *J. Electrochem. Soc.*, **137** (1990), 3612.

Seidel, H. *et al*: Anisotropic etching of crystalline silicon in alkaline solutions II, *J. Electrochem. Soc.*, **137** (1990), 3626.

Shikida, M. *et al*: Differences in anisotropic etching properties of KOH and TMAH solutions, *Sensors Actuators*, **80** (2000), 179.

Shikida, M. *et al*: A new explanation of mask undercut in anisotropic silicon etching: saddle point in etching rate diagram, *Transducers '01* (2001), p. 648.

Strandman, C. *et al*: Fabrication of 45° degree mirrors together with well-defined V-grooves using wet anisotropic etching of silicon, *J. MEMS*, **4** (1995), 214.

Tanaka, H. *et al*: Fast wet anisotropic etching of Si{100} and Si{110} with smooth surface in ultra-high temperature KOH solutions, *Transducers '03*, (2003), p. 1675.

van Veenendaal, E. *et al*: Simulation of anisotropic wet chemical etching using a physical model, *Sensors Actuators*, **84** (2000), 324.

Vazsonyi, E. *et al*: Anisotropic etching of silicon in a two-component alkaline solution, *J. Micromech. Microeng.*, **13** (2003), 165.

Wong, S.S. *et al*: An etch stop utilizing selective etching of n-type silicon by pulsed potential anodization, *J. MEMS*, **1** (1992), 187.

Proceedings of the IEEE, (1998), Special issue on integrated sensors, microactuators and microsystems.

**Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030**

22

Sacrificial and Released Structures

In many cases, films and structures are used intermittently, only to be disposed of in the next process step. Photoresists are an obvious example. Cleaning by oxidation is another: a surface that has been damaged (for example, by plasma etching) is oxidized, and the oxide film is immediately etched away in HF to reclaim the perfect silicon surface. However, sacrificial layers enable more complex structural shapes than standard two-dimensional patterning. Hollow structures and free-standing structures can be made by deposition of structural and sacrificial layers and by selective removal of the sacrificial layers. Nanofilter (Figure 22.1(a)) pass size is determined by thickness of thermal oxide on polysilicon: HF etching removes this polyoxide, opening up channels with dimensions determined by the oxide thickness, not by lithography. In vacuum microelectronic "triode", (Figure 22.1(b)) the anode metal is deposited on PSG layer, which is later removed to create a cavity around the silicon emitter tip.

When SOI wafers are used, buried oxide can act as an etch-stop layer for either the device layer or handle-wafer etching, or both, and it can also be used as a sacrificial layer for releasing structures. The photonic crystal structure (Figure 11.3) is fabricated this way.

In this chapter we will, however, concentrate on deposited films as sacrificial and structural layers. Deposited polycrystalline films cannot match the mechanical properties of single crystals (for example, the SOI device layer), but they offer a much wider range of possibilities because multiple structural and sacrificial layers can be deposited. These processes are single-sided: release etching takes place on the front of the wafer. No double-sided processing is involved, which is a great simplification. Standard single-side polished wafers can be used.

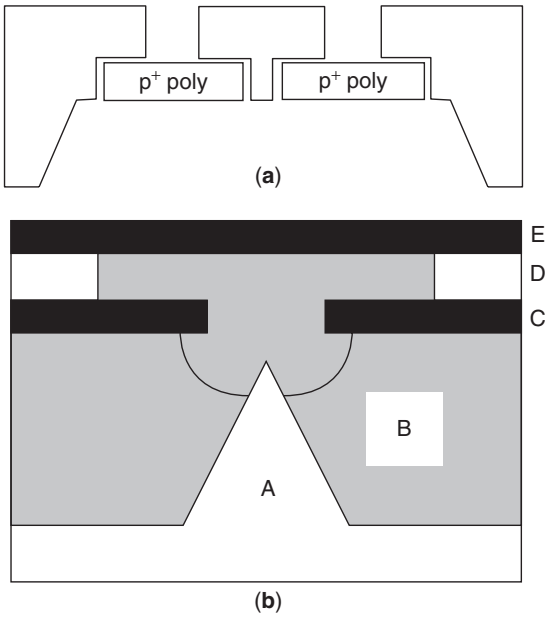


Figure 22.1 (**a**) Nanofluidic filter made by etching the polyoxide away. Inlets are lithographically defined but filter action depends on the polyoxide thickness, which can be much smaller than the lithographic minimum dimension. Redrawn after Chu, W.-H. *et al*. (1999), by permission of IEEE. (**b**) Microvacuum triode on silicon (cross sectional view): anisotropically etched emitter tip (A), PSG insulators (B,D) and polysilicon grid (C) and anode (E). Final etching of the PSG creates the microcavity around the tip. Redrawn after Orvis, W.J. *et al*. (1989), by permission of IEEE

22.1 STRUCTURAL AND SACRIFICIAL LAYERS

The structural layer needs to be of sufficient mechanical strength and proper stress state when released. Depending on film mechanical properties, anything from

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Table 22.1 Materials for released structures

Structural film	Sacrificial film(s)	Technology/application
Polysilicon	CVD oxide, PSG	Surface micromechanics
Silicon nitride	CVD oxide	Thermal isolation
Electroplated nickel	Cu, resist	LIGA
Al	Resist, PECVD oxide	Post-CMOS processing
Au	Cu, resist	Air bridges in RF circuits
Parylene	Resist	Microfluidics
SU-8	Cu, Al	Microfluidics
Cu	Resist	Post-CMOS processing

10 μm span lengths (for electroplated gold) to centimetres (for silicon nitride) are possible for released lateral structures.

Free-standing beams and plates will bend depending on their stress state, as shown in Figure 7.14. A series of beams with different lengths can act as a stress monitor. Compressively stressed beams (both ends clamped) will buckle after the critical compressive stress is exceeded. Strains of 0.001 in annealed polysilicon films translate to ca. 120 μm critical length for buckling, and 3×10^{-4} strain to ca. 220 μm buckling lengths. Tensile stresses are preferred for free-standing structures. For vertical structures, low stresses and stress gradients are similarly important in preventing a collapse.

The sacrificial layer has to fulfil two major requirements: it has to tolerate the deposition conditions of the structural layer and be removable selectively with respect to the structural layer. Table 22.1 lists some commonly used pairs of structural and sacrificial layers. Silicon surface micromechanics utilizes LPCVD silicon as a structural layer and CVD oxides, usually PSG, as sacrificial layers. LPCVD nitride can be used as an additional structural or insulating layer. LIGA is usually practised with nickel, copper and resist as the main materials.

If silicon dioxide is used as a sacrificial material, the removal etch has to be HF-based. This limits the metals that can be used for device metallization; or else metals need protective layers, which have to be removed after sacrificial etching. However, sacrificial etching is preferably the very last process step because the released structures may bend, resonate, stick, break or otherwise be damaged in further processing steps.

22.2 SINGLE STRUCTURAL LAYER

Free-standing released microstructures can be used as resonators, force sensors, switches, relays, movable mirrors and as inductor coils with minimized substrate capacitance, among others.

In its simplest form, a free-standing cantilever can be made in a single-mask process. The process flow is simple: deposition of the sacrificial layer, deposition of the structural layer, patterning of the structural layer and release etch. This is shown in Figure 22.2(a).

The one-mask process depends on timed etching: too much overetching would eliminate the anchor altogether and detach the cantilever from the substrate. Cantilever and anchor dimensions are closely related: the etch undercut must be long enough to release the cantilever but short enough for the anchor to remain.

In the two-mask process (Figure 22.2(b)), the structural layer is attached to the substrate and the etch timing becomes irrelevant because the structure acts as its own anchor. Extended overetching does not destroy the structure, but poor etch selectivity between the layers may change the dimensions of the structural layer.

The photoresist can act as a sacrificial layer for electroplated structures (Figure 22.3). Etch selectivity between the resist and the metal is practically infinite but large structures are difficult to release because of long etching times involved.

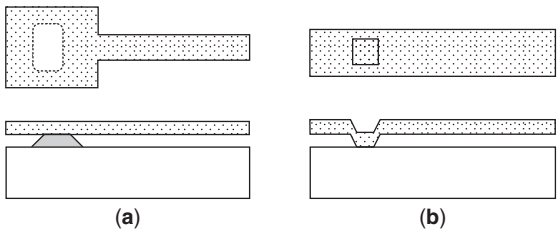


Figure 22.2 Cantilever fabrication; top views and side views of (**a**) a single, photomask cantilever process, with oxide serving both as an anchor and as a sacrificial material and (**b**) two-mask, cantilever process with the structural layer anchored directly to the substrate

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

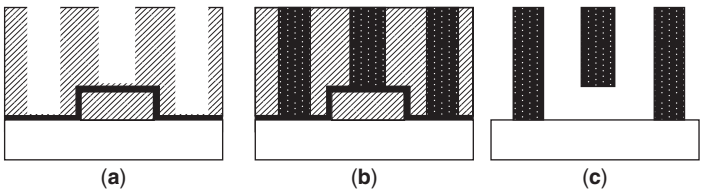
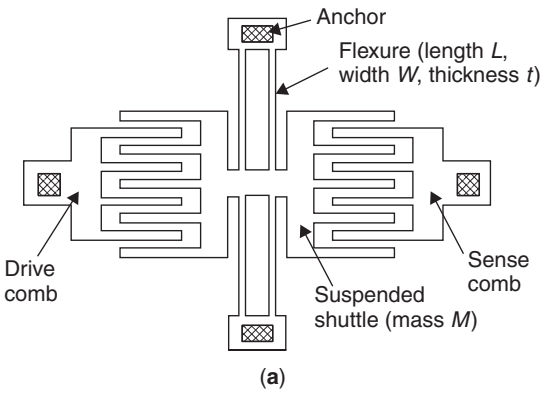


Figure 22.3 Electroplated free-standing structure: (**a**) first resist patterning and seed metal deposition; followed by a second, thick resist patterning; (**b**) electroplating and (**c**) development of the second resist, seed metal etching and removal of the first resist



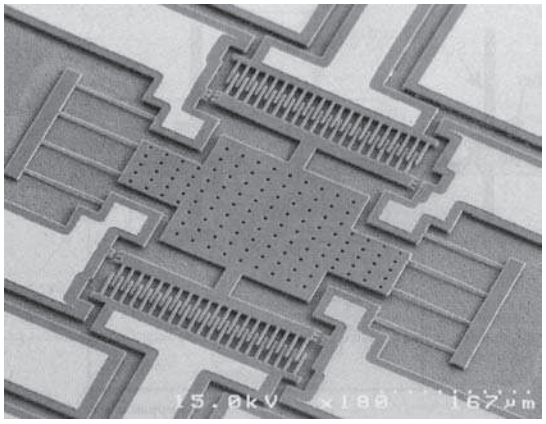


Figure 22.4 (**a**) Comb drive with suspended shuttle mass. From Bustillo, J. *et al*. (1998), by permission of IEEE. (**b**) SiC comb drive on silicon wafer. Plate release has been aided by using perforations in the plate. Reproduced from Roy, S. *et al*. (2002), by permission of IEEE

A comb drive with interdigitated fixed and movable (released but anchored) electrodes is a versatile sensor and actuator (Figure 22.4). Comb drives can be made in a single structural layer process, though multiple structural layers are often used, which will be discussed shortly.

22.3 STICTION

The release etch process looks like a simple isotropic etch but it has many difficulties not associated with isotropic patterning etching. Etch time control is difficult because etch front propagation under the structural layer cannot usually be observed. The etch process is diffusion limited in nature and it slows down in long and narrow release gaps.

A serious limitation for a wet release process comes from stiction (from 'sticking + friction'): during drying, the capillary force strength exceeds the spring force of the released structures and the free-standing cantilever/bridge/diaphragm makes contact with the substrate and adheres to it.

Stiction prevention has the following three alternative approaches:

1. *Dry release*: If silicon is used as sacrificial material, isotropic SF_6 plasma and XeF_2 gas are suitable. If oxide is used, anhydrous HF vapour can be used, but its etch rate is lower than that of aqueous HF. If photoresist is used as the sacrificial material, then oxygen plasma can be used for removal.
2. *Surface engineering*: Stiction depends on surface smoothness (on microscale), flatness (on macroscale) and surface chemistry (just like wafer bonding). Corrugated or otherwise patterned surfaces can prevent stiction. This approach requires extra process steps that need to be integrated into the process flow (Figure 22.5). Alternatively, the surfaces can be coated with hydrophobic coatings, for example, self-assembled monolayers (SAMs) or plasma-deposited fluoropolymers.
3. *Phase engineering*: Sublimation and supercritical drying sidestep normal liquid drying. In sublimation,

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

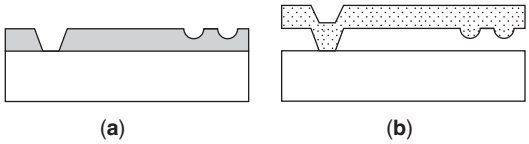


Figure 22.5 Three-mask process for cantilever with dimples: (**a**) first mask step for anchor area etching; second mask step for dimple etching and (**b**) structural-layer deposition, lithography and etching

rinsing water is replaced by tert-butanol, and then frozen. Heating is performed under reduced pressure in a regime where solid tert-butanol turns to vapor directly (sublimation). This route is shown in Figure 22.6 as FD, for freeze drying. In supercritical drying liquid, CO_2 replaces the rinsing solvent (methanol). After heating into supercritical region under pressure, a pressure drop vaporizes CO_2. This is shown as route SD, for supercritical drying. Normal drying is indicated as ND.

Avoiding stiction during the fabrication process is one thing; avoiding it during device operation is another. RF switches operate by making a contact between two surfaces. Both metal-to-dielectric contacts (as shown in Figure 22.7) and metal-to-metal contacts are used.

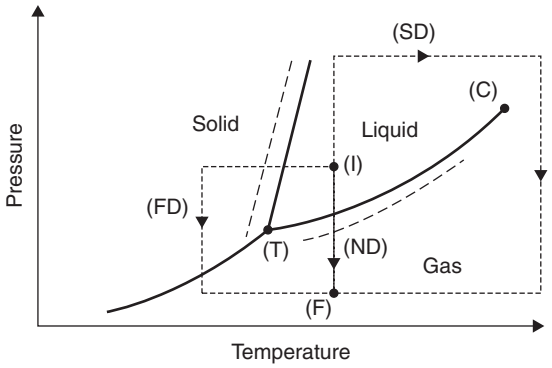


Figure 22.6 Thermodynamics of drying: I = initial stage; F = final stage; ND = normal drying; FD = freeze drying; SD = supercritical drying. Reproduced from Bellet, D. & Canham, L. (1998), by permission of Wiley-VCH

Some switches even conduct current while metals are in contact, which may lead to welding together of the two metals.

22.4 TWO STRUCTURAL–LAYER PROCESSES

A comb-drive actuator can generate sizable forces when the number of interdigitated fingers is made

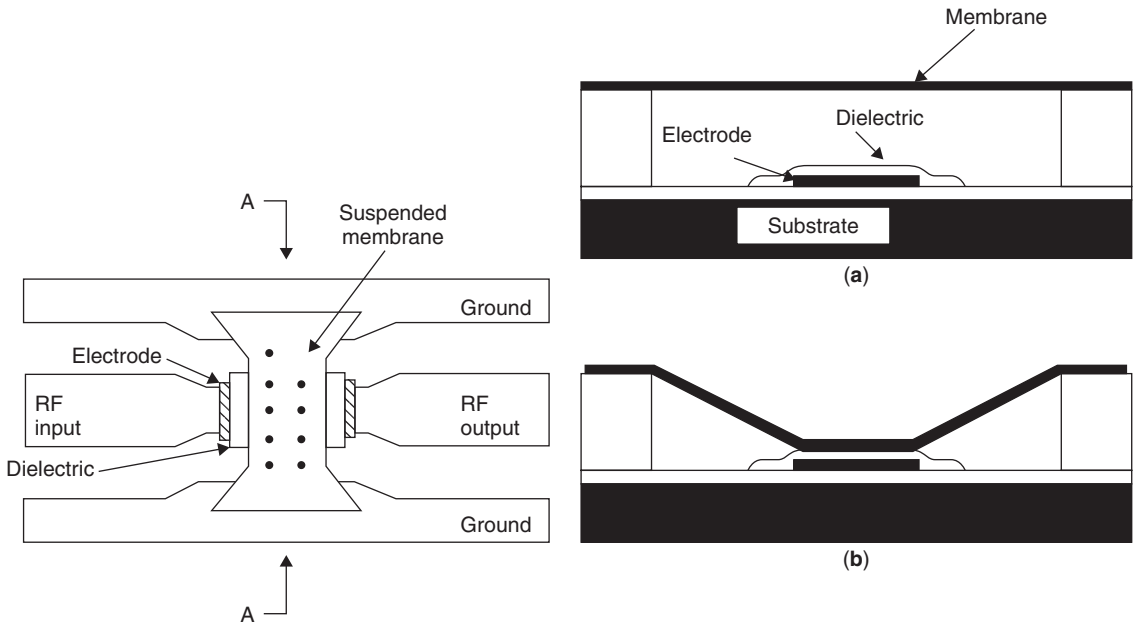


Figure 22.7 RF switch: (**a**) top view and (**b**) cross-sectional view along AA in off-state (up) and on-state (down). Reproduced from Yao, Z.J. *et al*. (1999), by permission of IEEE

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

large. Alternatively, capacitance change between the finger plates can be used for sensing, for example, in accelerometers and gyroscopes.

It is possible to make such a comb drive in a one mask, single structural-layer process if the fixed comb dimensions were designed to be much larger than those of the movable comb; in fact, the whole fixed comb should be considered as an anchor. However, such a process has too many design limitations for it to be useful. A two-layer, four-mask process described in Figure 22.8 and outlined below offers a robust fabrication process for comb drives.

Comb-drive process flow

(a) oxide + nitride insulation
(b) lithography #1: contact to substrate
(c) poly1 deposition (300 nm thick, heavily n^+ doped)
(d) lithography #2: poly1 patterning
(e) deposition of sacrificial PSG, 2 μm thick
(f) lithography #3: anchors for poly2
(g) deposition of poly2, 2 μm thick
(h) second PSG deposition, anneal and etch
(i) lithography #4: patterning of poly2
(j) etching of PSG for release of poly2.

The second polysilicon is doped by PSG from top, eliminating dopant gradient effects. In addition to doping, the annealing step also has the role of poly2 stress optimization. Both the fixed and the movable comb are defined in the same photolithography step, and thus their spacing is free of alignment errors.

Two structural–layer processes offer similar device and fabrication benefits in metal micromechanics. Electroplated metals can serve both as structural layers and as sacrificial layers, for example, copper can be selectively removed under nickel or gold, enabling elaborate 3D structures to be made, Figure 22.9.

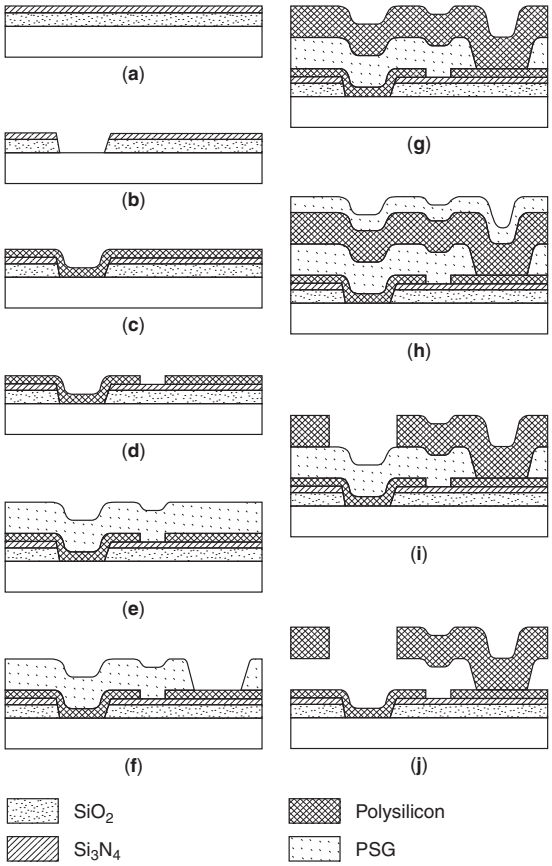


Figure 22.8 Fabrication of a comb-drive structure in a two structural–layer process. Reproduced from Tang, W.C. *et al*. (1989), by permission of IEEE

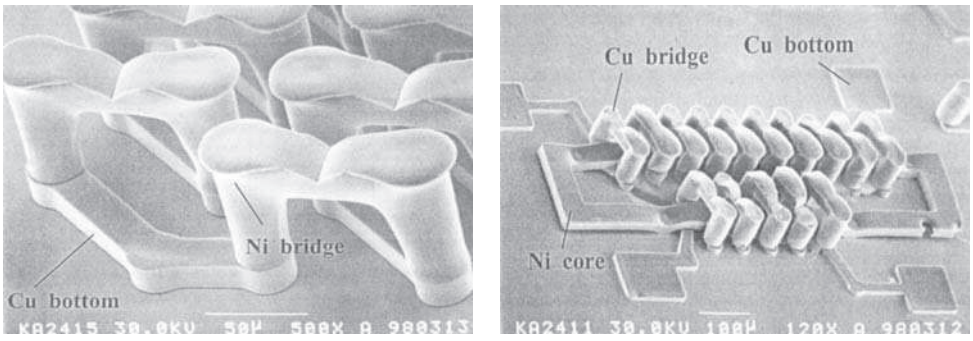


Figure 22.9 (**a**) 3D inductor coil with copper bottom and nickel bridge structural layers and (**b**) 3D transformer with Cu-bottom and copper bridge with Ni-core by three structural layers. Reproduced from Yoon, J.-B. *et al*. (1998), by permission of Institute of Pure and Applied Physics

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

22.5 ROTATING STRUCTURES

Two structural layers enable rotating structures to be made. The centre-pin process utilizes two structural and two sacrificial layers (Figure 22.10). In contrast to the previous comb-drive example, poly1 becomes the movable element, and poly2 serves as the fixed element that bounds the rotating element made of poly1. The first sacrificial layer defines the gap between substrate and poly1, and the second sacrificial layer defines interpoly gap.

The concept of self-alignment is useful in released structures as well. The centre-pin and the rotor can be self-aligned. It depends on the relative thickness of the structural and sacrificial layers. Poly2 pin can be made to limit the movements of poly1 rotor in the lateral direction. In the opposite case, the rotor can wobble because the centre-pin is too high (Figure 22.11).

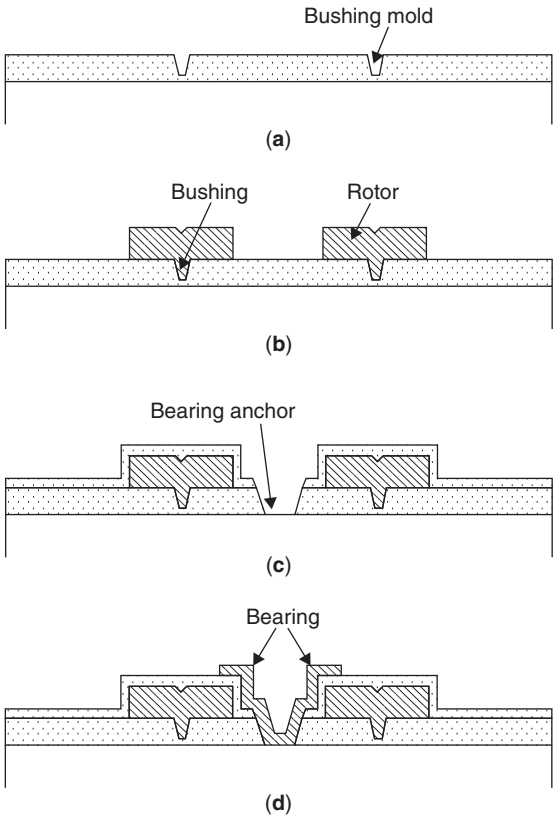


Figure 22.10 Cross-sectional schematics demonstrating the centre-pin bearing process: (**a**) after patterning of the bushing mould in the first sacrificial layer; (**b**) after deposition and patterning of poly1; (**c**) after deposition of the second sacrificial layer and anchor region definition and (**d**) deposition and patterning of poly2, followed by oxide etching. Reproduced from Mehregany, M. & Dewa, A.S.: http://mems.cwru.edu/shortcourse/ by permission of Case Western Reserve University

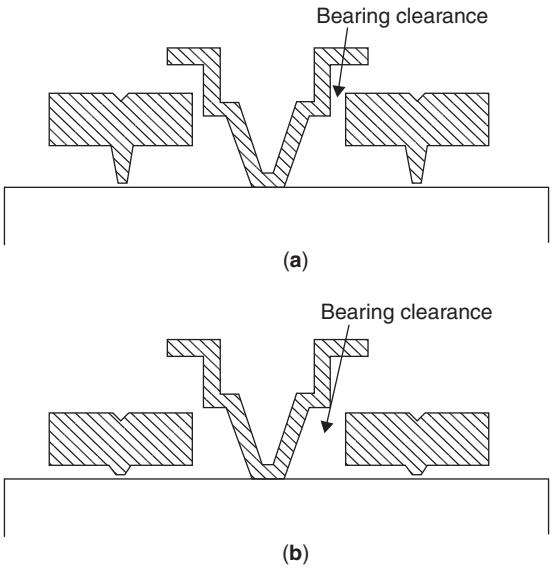


Figure 22.11 Cross-sectional schematics demonstrating two types of centre-pin bearings that may result after release: (**a**) self-aligned and (**b**) non-self-aligned. Reproduced from Mehregany, M. & Dewa, A.S.: http://mems.cwru.edu/shortcourse/ by permission of Case Western Reserve University

22.6 HINGED STRUCTURES

Structures that pop up from the plane of the wafer can be made by various methods. Mechanical hinges can be made in a two structural-layer process or with polymeric hinges in a one-layer process. In the polymeric-hinge process, a polyimide hinge is patterned on top of the structural layers (Figure 22.12). The movable plate dimensions have to be smaller than those of the anchor, which can be helped by making perforations for release etching. Upon release, the movable poly plate can be actuated by, for example, thermal expansion of the imide.

Alternative hinge technology is based on two polysilicon layers: poly1 forms the moving element and poly2 forms a staple that lets the poly1 structure rotate upwards from the plane of the wafer but confines it otherwise (Figure 22.13).

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

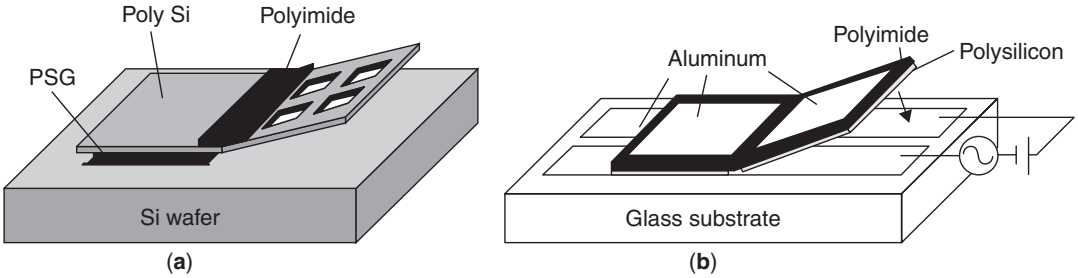


Figure 22.12 (**a**) A polyimide hinge joins static and moving polysilicon plates and (**b**) polyimide hinged, electrostatically actuated mirror. Reproduced from Suzuki, K. *et al*. (1994), by permission of IEEE

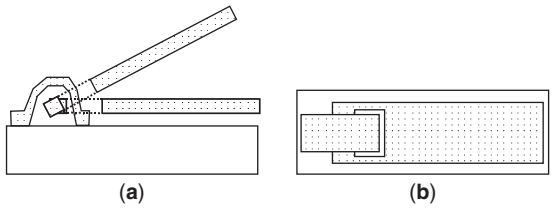


Figure 22.13 Two-poly staple hinge: (**a**) side view and (**b**) top view. Adapted from Pister, K. *et al*. (1992), by permission of Elsevier

22.7 SACRIFICIAL STRUCTURES USING POROUS SILICON

The electrochemical etch rate of n-type silicon (10–20 ohm-cm) in an HF electrolyte is very low compared to p-type silicon or low-resistivity n-type silicon (ca. 0.01 ohm-cm) (Figure 22.14). Doping (by diffusion or epitaxy) can, therefore, be used to create porous silicon patterns. Alternatively, protective etch masks can be used, as in any other etching process. Photoresist, silicon nitride, amorphous silicon and silicon carbide are candidates; silicon dioxide cannot be used because of the HF electrolyte, and photoresists are limited to cases with diluted HF.

The material of the structural layer can be, for instance silicon nitride, but epitaxial silicon can also be used. Porous silicon is single-crystalline silicon and it is possible to grow epitaxial film on it.

Porous silicon is a mechanically weak material, and it can be destroyed by the capillary forces during drying (cf. stiction where capillary forces pull free-standing structures together upon drying). Porous silicon can be destroyed by gas bubbles as well: KOH etching releases hydrogen (Equation 11.1), and if gas evolution is rapid, the bubbles can burst porous structures. For this reason dilute KOH, 0.1 to 1%, is used rather than 20 to 50%, which is typical of silicon anisotropic etching.

In a modification of the above scheme, a free-standing structure can be made of bulk single-crystal silicon. The n-type silicon is intact in electrochemical etching and the p-type silicon underneath is fully transformed into porous silicon (Figure 22.15).

22.8 EXERCISES

1. What etch selectivity is needed to release a 1 μm thick silicon nitride plate of 50 μm width by sacrificial-oxide etching (49% HF, rate 2 μm/min) if plate thickness variation due to etching has to

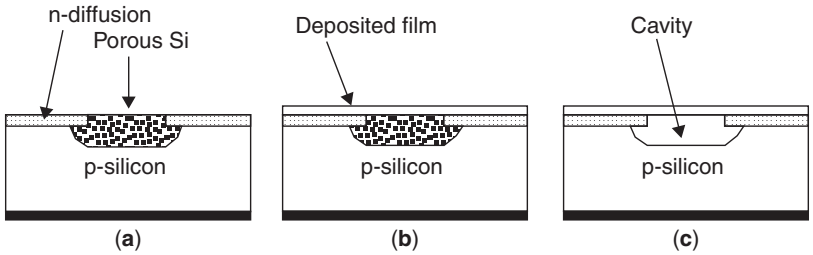


Figure 22.14 Fabrication of a free-standing bridge on a p-type substrate: (**a**) n-diffusion of selected areas, followed by electrochemical etching; (**b**) bridge material deposition and (**c**) removal of porous silicon in dilute KOH resulting in a bridge over a cavity. Reproduced from Hedrich, F., Billat, S. & Lang, W. (2000), by permission of Elsevier

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

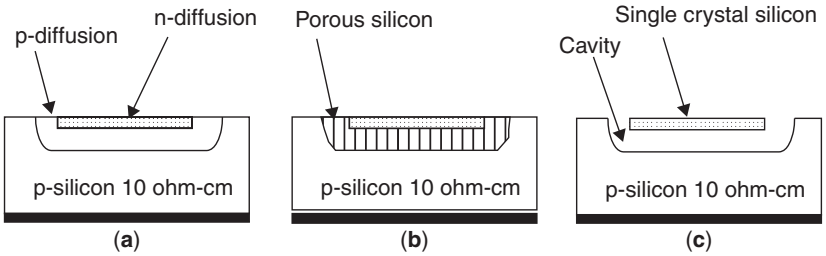
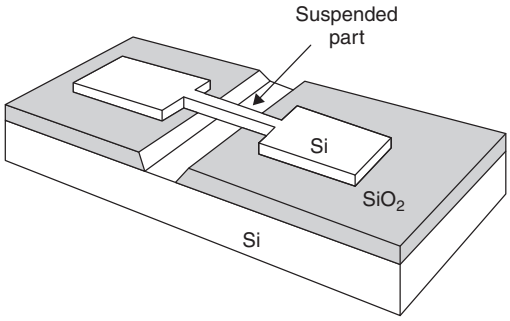


Figure 22.15 (**a**) A shallow n-diffusion and a deeper p-diffusion; (**b**) lateral porous silicon formation in the heavily boron-doped region and (**c**) dilute KOH sacrificial etching releases a single-crystalline n-silicon bridge. Redrawn after Lee, C.-S., Lee, J.-D. & Han, C.-H. (2000), by permission of Elsevier

be smaller than nitride deposition non-uniformity of 3%?

2. Design a fabrication process for the suspended silicon bridge shown below. Consider two cases: a bridge made of LPCVD polysilicon and a SOI device silicon layer bridge.



From Bruschi, P. *et al*. (2001), by permission of Elsevier.

3. Comb-drive fabrication tolerance: resonant frequency of a surface micromachined resonator with straight flexures (see Figure 22.4(a)) is given by

$$f_0 = (1/2\pi)\{(4EtW^3/ML^3) + (24\sigma_r Wt/5ML)\}^{1/2}$$

where E is Young's modulus, σ_r is residual stress in polysilicon, M is shuttle mass, t is poly thickness, L is flexure length and W is flexure width. What is the effect of fabrication tolerance on resonance frequency? Consider poly thickness and lithography/etching variation for some realistic dimensions.

4. Design proper thicknesses and etched depths to make the self-aligned rotor shown in Figure 22.11.
5. How many photolithography steps are needed to make the polysilicon-hinged mirror structure shown in Figure 22.13?
6. Design a fabrication process for the polymer hinged mirror shown in Figure 22.12(a).
7. Design a fabrication process for the fluidic filter shown in Figure 22.1. Also draw the photomasks that show how the filter is anchored to the substrate.
8. What are the lithography steps and sacrificial layers needed to make a 3D coil with a Ni core (transformer) shown in Figure 22.9(b)?

REFERENCES AND RELATED READINGS

Bellet, D. & Canham, L.: Controlled drying, *Adv. Mater.*, **10** (1998), 487.

Bruschi, P. *et al*: Micromachined silicon suspended wires with submicrometric dimensions, *Microelectron. Eng.*, **57–58** (2001), 959.

Bustillo, J. *et al*: Surface micromachining for microelectromechanical systems, *IEEE Proc.*, **86** (1998), 1559.

Chu, W.-H. *et al*: Silicon membrane nanofilters from sacrificial oxide removal, *J. MEMS*, **8** (1999), 34.

Hedrich, F., Billat, S. & Lang, W.: Structuring of membrane sensors using sacrificial porous silicon, *Sensors Actuators*, **84** (2000), 315.

Lammel, G. & Renaud, Ph.: Free-standing mobile 3D porous silicon microstructures, *Sensors Actuators*, **85** (2000), 356.

Lee, C.-S., Lee, J.-D. & Han, C.-H.: A new wide-dimensional freestanding microstructure fabrication technology using laterally formed porous silicon as a sacrificial layer, *Sensors Actuators*, **84** (2000), 181.

Löchel, B. *et al*: Ultraviolet depth lithography and galvanoforming for micromachining, *J. Electrochem. Soc.*, **143** (1996), 237.

Mehregany, M. & Dewa, A.S.: http://mems.cwru.edu/shortcourse/, Case Western Reserve University.

Orvis, W.J. *et al*: Modeling and fabricating microcavity integrated vacuum tubes, *IEEE TED*, **36** (1989), 2651.

Pister, K. *et al*: Microfabricated hinges, *Sensors Actuators*, **A33** (1992), 249.

Roy, S. *et al*: Fabrication and characterization of polycrystalline SiC resonators, *IEEE TED*, **49** (2002), 2323.

Suzuki, K. *et al*: Insect-model based microrobot with elastic hinges, *J. MEMS*, **3** (1994), 5.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Syms, R.R.A. *et al*: Improving yield, accuracy and complexity in surface tension self-assembled MOEMS, *Sensors Actuators*, **A88** (2001), 273.

Tang, W.C. *et al*: Laterally driven polysilicon resonant microstructures, *Proc. IEEE MEMS* (1989), p. 53.

Wang, S.N. *et al*: Novel processing of high aspect ratio 1–3 structures in high density PZT, *Proc. IEEE MEMS* (1998), p. 223.

Yao, Z.J. *et al*: Micromachined low-loss microwave switches, *J. MEMS*, **8** (1999), 129.

Yoon, J.-B. *et al*: Monolithic fabrication of electroplated solenoid inductors using three-dimensional photolithography of a thick photoresist, *Jpn. J. Appl. Phys.*, **37** (1998), 7081.

Proc. IEEE, **86** (1998), special issue on integrated sensors, microactuators & microsystems (MEMS).

**Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030**

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

23

Structures by Deposition

The standard approach in microfabrication is to deposit film all over the wafer and then remove unwanted parts by etching or polishing. In this chapter, various techniques for direct and localized structure formation by deposition are presented. They are for the most part, niche applications, and not mainstream.

Processes come in two forms: directional and diffuse (Figure 23.1). The former includes processes in which beams of atoms, photons, electrons or ions impinge on the wafer (such as lithography, evaporation and implantation); the latter includes immersion processes in which wafers are surrounded by vapours, gases or liquids (such as wet etching, oxidation or CVD). In order to prevent immersion processes acting on the whole wafer, selected areas can be protected by masking layers. These layers are deposited and patterned on the wafer. This also applies to directional processes: masking layers will stop ions, absorb photons and prevent atoms from reaching the substrate. However, directional processes can also be blanked above the wafer by absorbers, collimators or stencil masks.

Localized processing comes in two major variants: focused beam processing and microstructure-assisted processing (Figure 23.2). In both cases energy is supplied locally and reactions take place only where the beam or the microstructure provides energy. This energy can be, for example, photonic energy from a laser beam or thermal energy from a resistor.

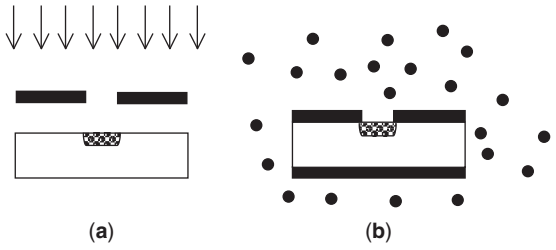


Figure 23.1 (**a**) Directional process blanked by a stencil above the wafer and (**b**) diffuse process blanked by a masking layer on the wafer

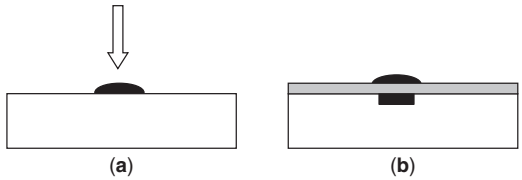


Figure 23.2 Localized processing: (**a**) focused beam supplies energy and (**b**) microstructure provides energy

23.1 PLATED STRUCTURES

Electroplating is a prototypical process in which deposition leads to the final structure in one step (but, of course, more complex structures can be made if several steps are made in sequence) (Figure 23.3). An electrically conducting layer is needed to initiate plating. This seed layer (also known as the plating base or field metal) can be very thin, tens of nanometres, and is usually deposited by sputtering.

The seed layer needs to be removed after plating because otherwise it would electrically short-circuit all the metallized structures. Often, the deposited metal itself can act as an etch mask for seed-layer removal because the seed layer is always very thin compared to the plated metal; in many cases, seed-layer thickness is less than plating thickness variation. Thickness uniformity of plated metals is ca. 5 to 10%, so that 50 nm seed-layer thickness is less than thickness fluctuation of 1 μm-thick plated metal.

Electroplating is a prototypical process where deposition leads to the final structure in one step (Figure 23.4), but of course more complex structures can be made if several steps are made in sequence. If X-ray lithography

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

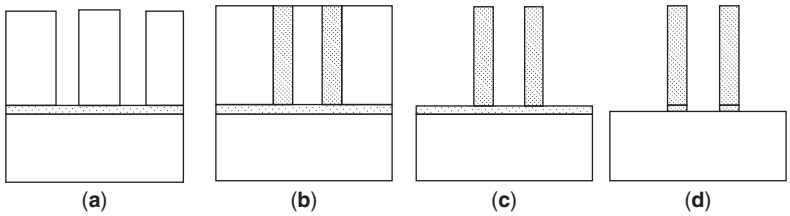


Figure 23.3 Resist masked plating: (**a**) seed layer deposition and photolithography; (**b**) plating to fill resist patterns; (**c**) resist stripping and (**d**) seed layer removal

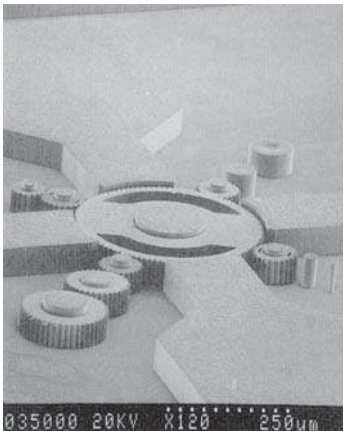


Figure 23.4 Nickel gear structures on silicon made by electroplating. Reproduced from Guckel, H. (1998), by permission of IEEE

has been used to pattern the resist with 100:1 aspect ratios structures, for example, 500 μm thick, 5 μm wide filling by plating is not a problem. Thermal CVD processes (LPCVD nitride, TEOS oxide or LPCVD poly) can fill similar aspect ratios, but at elevated temperatures and not at room temperature with photoresists.

Usually, filling is allowed to proceed till the resist top surface level but not above (Figure 23.5(a)). It is, however, possible to overplate, and to form mushroom-shaped structures (Figure 23.5(b)). After resist stripping, such a mushroom can be annealed (reflown) to form ball-like bumps. Bumps of Sn-Pb and In are used for flip-chip packaging. Alternatively, plating can be continued until the metal fronts touch (Figure 23.5(c)). Removal of the resist underneath results in free-standing metal bridges. Such bridges have uses as transformer coils or air bridges in RF-circuits.

Plating of the active wire structure without masking results in sloped-walled structures and free-form 3D shapes, depending on currents and voltages in the wires, but dimensional control is difficult.

23.2 LIFT-OFF METALLIZATION

Lift-off is metallization with sacrificial resist: after lithography, metal deposition is done on the resist pattern, followed by resist dissolution in solvent and lift-off, with all the metal that is not in contact with the substrate being removed (Figure 23.6). There is always some deposition on the sidewalls too, but if films are thin, they are discontinuous and resist dissolution can take place.

Lift-off is very general: all metals, their alloys and multi-metal stacks can be patterned with the same basic process; there is no need for etch-process development when metallization is changed. Lift-off is especially suited for hard-to-etch metals, such as gold and platinum.

The deposition process has, however, many photoresist-imposed limitations: it must take place under ca. 120 °C temperature because of resist thermal stability.

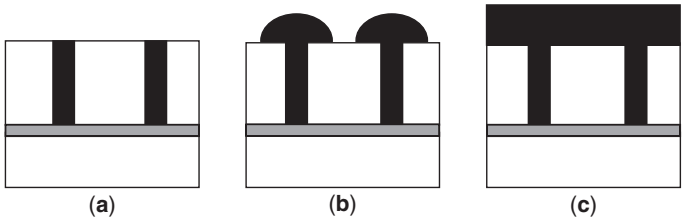


Figure 23.5 Aspect ratio preserving (**a**) plating; (**b**) overplating and (**c**) backplating

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

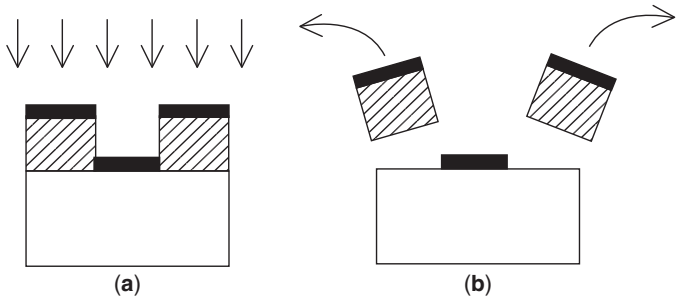


Figure 23.6 Lift-off process (**a**) metal deposition on resist pattern and (**b**) resist dissolution and metal lift-off

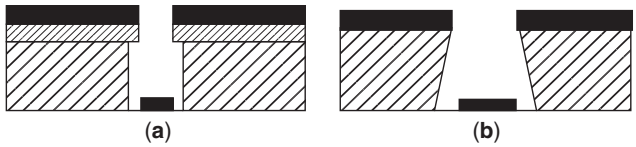


Figure 23.7 Profile tailoring for lift-off: (**a**) bi-layer resist and (**b**) retrograde resist profile

The deposition should have poor step coverage, which is a very special requirement. Evaporation, which is a line-of-sight method, is best suited for lift-off metallization. Poor step coverage, however, forbids lift-off metallization for samples with complex topography because the metal would be discontinuous over other steps as well.

Resist profile can be tailored to minimize sidewall deposition (Figure 23.7). Two-layer resists with an overhang profile or retrograde profiles (typical of negative resists) are useful. Two-layer structures can be true bi-layer resists, or the top layer of a single layer resist can be hardened so that its development rate is slower. The hardening can be a chemical benzene soak or some other surface treatment.

Lift-off is not limited to resist masking: bi-layer masks of two thin films can be used. This has been used for unetchable films or for materials with harsh deposition conditions, for example, diamond. Stresses in the deposited films must be low enough so that the overhang layer is not deformed.

23.3 SPECIAL DEPOSITION APPLICATIONS

Directionality of evaporation, its line-of-sight deposition geometry, is favourable for lift-off and if this is combined with a tilted sample, very small structures can be deposited on sidewalls (Figure 23.8). Some of the smallest ever MOSFETs have been demonstrated by oblique angle evaporation.

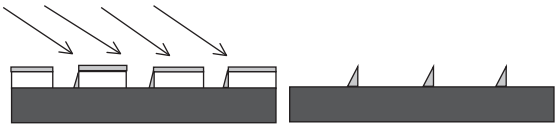


Figure 23.8 Oblique angle evaporation; followed by etching away the support structure

23.3.1 Shadow masks

Sometimes films are so sensitive that their deposition has to be the very last process step, for example, (bio)chemical sensor films. Application of the photoresist on these films is not possible and acetone dissolution, as in lift-off, cannot be used.

Shadow masks (also known as stencil masks) are mechanical aperture plates. Shadow-mask patterning is basically lift-off with a mechanical mask instead of a resist mask. The shadow mask is aligned to and attached to the substrate, and this stack is then positioned in the deposition system (Figure 23.9).

If the shadow mask and wafer can be aligned to each other in a bond-aligner, micrometre alignment accuracy is possible; but often shadow masks are only used for non-critical applications where manual $\pm 10\,\mu$m alignment is enough. Minimum linewidths that are possible with shadow masks are in the $10\,\mu$m range, with silicon-wafer masks fabricated by standard lithography and anisotropic etching processes. One special limitation of shadow masks is the impossibility of doughnut-shaped structures.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

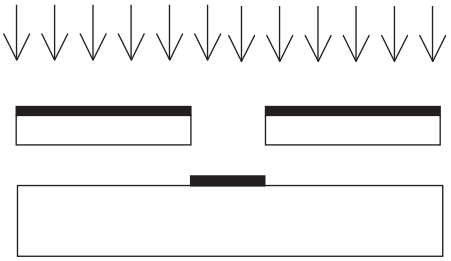


Figure 23.9 Deposition with a shadow mask

23.3.2 Sidewall lithography (edge-defined structures)

Sidewall spacers remain on the sidewall after anisotropic etching of a conformal film. Extended overetch can remove them but an alternative approach calls for removal of the original structure after spacer formation, leaving the spacers intact. This is shown in Figure 23.10. Stand-alone spacers can be used as very narrow etch masks or as a high surface area cylinder over which CVD films can be deposited. This is used in 'hollow crown' DRAM capacitors as a way to increase capacitor area.

Spacer width is determined by conformal deposition thickness. Deposition thickness is easily controlled, even in the sub-100 nm range, and extremely narrow lines have been made by the sidewall spacer technique. Stability of sidewall pillars is determined by stresses in the film and pillar length-height-width ratio. Aspect ratios of 5:1 can be made fairly easily. Small holes and apertures can be made by sidewall spacer removal, as shown for nanofilter of Figure 22.1.

23.4 LOCALIZED DEPOSITION

Most thin film deposition methods are blanket depositions, that is, film deposits everywhere on the wafer. A handful of techniques provide selective area deposition. Chemical differences in microstructures form the basis for selectively depositing material on just one of the surfaces. Selective deposition has many attractive features, simplicity of process integration being the foremost.

23.4.1 Selective deposition

Both CVD and electrochemical processes can be used for selective deposition, with electroless copper and CVD tungsten being the most studied ones.

Silicon surface reduction process allows selective CVD tungsten in contact holes

$$2WF_6\ (g) + 3Si\ (s) \longrightarrow 2W\ (s) + 3SiF_4\ (g) \quad (23.1)$$

This reaction is selective because SiO_2 does not reduce WF_6. However, ca. 20 nm of silicon is consumed,

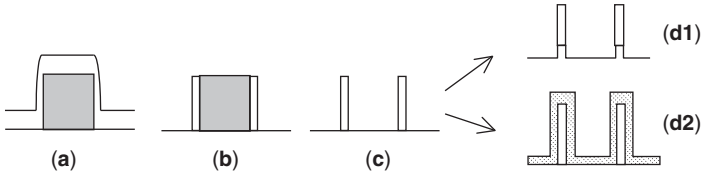


Figure 23.10 Cross-sectional view of sidewall spacer structures (**a**) after conformal film deposition; (**b**) after spacer etching; (**c**) after removal of the original structure; (**d1**) spacers used as an etch mask and (**d2**) spacers used as a deposition template

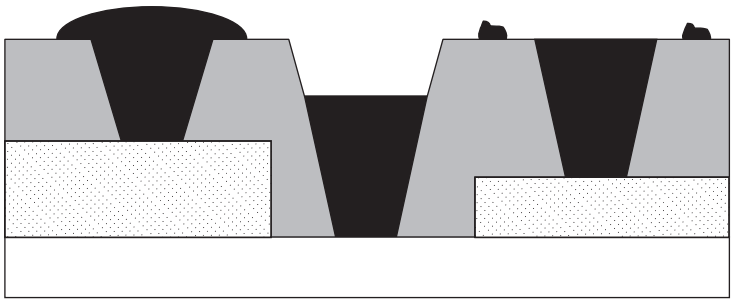


Figure 23.11 Problems with selective deposition: unequal hole depths and loss of selectivity

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

and the reaction is self-limiting: WF_6 cannot diffuse through the growing tungsten layer. Tungsten deposition is continued by silane reduction of tungsten hexafluoride on tungsten according to

$$WF_6\ (g) + 2SiH_4\ (g) \longrightarrow W\ (s) + 3H_2\ (g) + 2SiHF_3\ (g) \qquad (23.2)$$

This reaction, however, is transport limited and difficult to control. Additionally, it faces problems when contact holes of different depths have to be filled: some are underfilled, some are overfilled (Figure 23.11).

Plug fill can be achieved by continuing deposition in hydrogen reduction mode:

$$WF_6\ (g) + 3H_2\ (g) \longrightarrow W\ (s) + 6HF\ (g) \qquad (23.3)$$

There is always the problem of selectivity loss. It is usually connected with residues from preceding process steps, for instance, incomplete resist removal. Selective deposition processes are rare in volume manufacturing even though they sometimes offer enormous simplifications in process integration.

23.4.2 Localized deposition by external excitation

Localized deposition depends on some sort of local excitation, thermal, ion beam or photon flux, and is used to induce growth just at a localized spot. There are three regimes for heating: in adiabatic regime, thermal energy is limited to a few micrometres on wafer surface because there is no time for heat diffusion; in thermal flux regime, the bulk of the wafer heats up but wafer backside is still at ambient temperature; in isothermal regime, the wafer is in thermal equilibrium.

Focused beams can be used either directly or indirectly. In photomask writing, they draw the pattern in resist film, which then serves as a mask for chrome etching, however, now we are interested in beam interaction with the wafer (and the surrounding gaseous atmosphere) to form the pattern directly.

Focused ion beam (FIB) can be used to etch features on a wafer, for example, to remove erroneous chrome spots from the photomask or to deposit films in the presence of suitable source gases. Repair of missing features on the photomask can be done by depositing tungsten according to

$$W(CO)_6\ (g) \longrightarrow W\ (s) + 6CO\ (g) \qquad (23.4)$$

There are two mechanisms in laser-CVD: photolytic and photothermal. In photolytic deposition, laser light interacts with gaseous species, which then deposit on the wafer. In a photothermal process, the laser heats the surface and elevated local temperature drives chemical reactions, but often both elements are present simultaneously. The chemical reactions are the same as those in traditional CVD deposition; for example, silane source gas for (poly)silicon deposition.

It is possible to fabricate 3D structures by changing the focal point of the focused beam in space. Electron microscopes and FIB systems have been used in many 3D-deposition applications. Structures such as out-of-plane nanoneedles and microcoils with ca. 10 μm-wire diameter and 50 μm coil diameter have been made by electron beam-induced CVD of carbon. In stereomicrolithography, a laser beam solidifies polymer at the focal spot. After a single layer has been drawn, focus shifts up and the next level of polymer is solidified. Elaborate 3D shapes can be drawn, but like all direct writing techniques, stereomicrolithography has low throughput.

23.4.3 Microstructure-assisted local processing

Electrical and thermal modification by microstructure-assisted processing is also possible in the field after the device processing has been completed, whereas beam processes are done in wafer fab at wafer level or chip level.

Heat dissipation in microstructures is not very amenable to macroworld intuition because surface-to-volume ratios in microstructures are very different from macroscopic objects. A silicon wire sandwiched between glass wafers and heated up to 1400 °C will lead to a 40 °C temperature rise 15 μm away.

Microfuses are one-time programmable elements that can be used to store chip identity data or calibration curves, to trim resistors or to cut off malfunctional circuit blocks and to connect redundant spare blocks. Both normally-on and normally-off fuses exist. A normally-on fuse has a thin metallic/conductive part that can be broken. The mechanism for breakage differs: chemical reaction can turn the metal film into an insulator, a phase change can alter its resistivity or electromigration can create a void in the wire. Antifuses can be made, for example, of high-resistivity undoped amorphous silicon that will crystallize and become conductive when a programming pulse is driven through it. Gigaohm versus 100 ohm off- and on-resistances (10^7 on–off ratio) are possible.

Local (chip-scale) sealing of cavities has been demonstrated with a microfabricated polysilicon resistor on the wafer supplying energy for CVD of the sealing

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

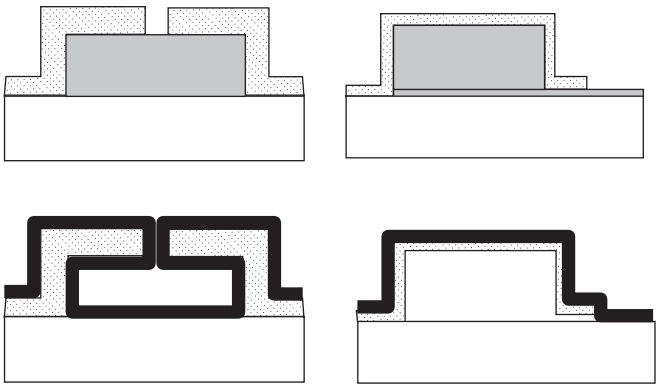


Figure 23.12 Cavity formation by etching of sacrificial oxide (gray) and (**a**) deposition sealing of a lithographically defined, plasma-etched, vertical access hole and (**b**) sealing of a horizontal-access hole defined by film deposition: very little deposition takes place inside the cavity when the access channel is long and narrow

material. Generally, however, sealing is done at a wafer level.

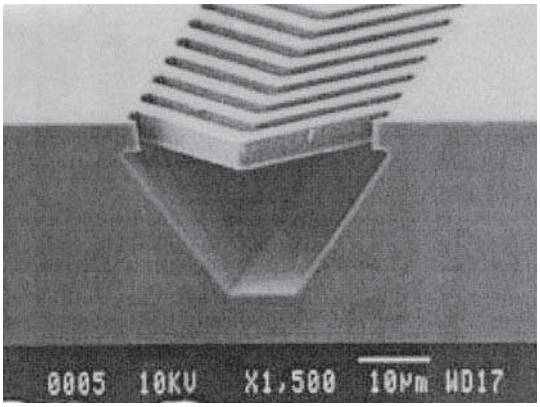
23.5 SEALING OF CAVITIES

Cavities are closed structures with a controlled atmosphere inside. Absolute pressure sensor is a simple example: the cavity holds the reference pressure. In resonating structures, such as accelerometres and gyroscopes, squeeze-film damping requires cavity pressure to be reduced from atmospheric pressure. This can be done in a bonding process or in a deposition process.

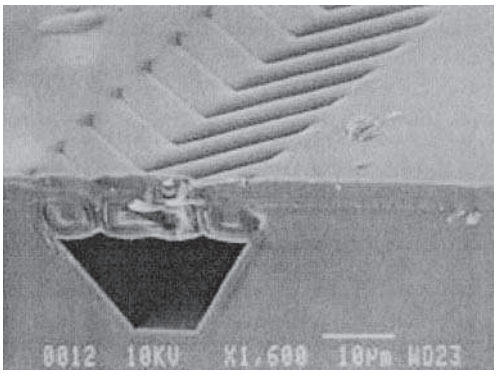
CVD processes with conformal deposition are well suited for cavity sealing, but conformality also means that a film will deposit on the inner walls of the cavity. CVD processes with high surface mobility of adatoms and long mean free paths are best candidates for sealing. Schematic CVD sealing is shown in Figure 23.12 and SEM micrographs are shown in Figure 23.13.

In order to reduce the influence of the sealing film on the structural films, the sealing film should be as thin as possible. This is often best achieved with horizontal-access holes rather than with plasma-etched vertical holes. Horizontal-access hole minimum dimension is determined by film thickness, which can be made small easily compared to lithographically determined plasma-etched access holes.

If ultimate vacuum is needed inside the cavity, evaporation is the method of choice. Contrary to CVD sealing, no (potentially) harmful gases will be incorporated into the cavity. Owing to the directional nature of evaporation, horizontal-access holes have to be used.



(**a**)



(**b**)

Figure 23.13 Cavity sealing by CVD: plasma-etched, chevron-shaped access holes are closed by LPCVD nitride deposition. Reproduced from Chen, J. & K.D. Wise (1997), by permission of IEEE

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

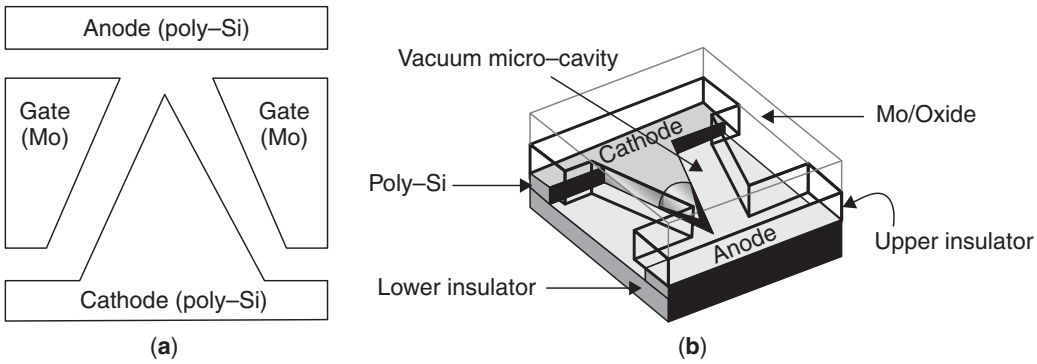


Figure 23.14 Lateral microtip emitter. Reproduced from Lim, M.-S. *et al*. (2001), by permission of IEEE

Measurement of cavity pressure is no easy task because of leaks and gettering. In fact, resonant microstructures in the cavity are used as vacuum gauges; because frequency is very sensitive to pressure, it can be used for vacuum measurement. This, of course, depends critically on the stability of the resonator: any drifts in mechanical quality factor, surface charging or film deposition on the resonator will change resonant frequency.

Fabrication of a lateral field emitter calls for a six-layer stack of nitride/oxide/n+ poly/oxide/nitride/oxide (Figure 23.14). The top oxide layer acts as a hard mask for stack RIE etching. RIE removes the layers all the way to the bottom nitride (lower insulator). The approximate shape of the cathode is determined by lithography and bottom polysilicon etching, but oxidation of polysilicon will shorten and sharpen the cathode tip and determine its final distance from the anode poly.

The initial structure can be made with $2\,\mu$m lithography, and poly oxidation (of $1\,\mu$m/side) sharpens the tip. HF etching removes polyoxide, and creates the vacuum microcavity. Cathode–anode separations in the sub-100 nm range can be made. Vacuum microtip emitter has been sealed with evaporated metal (molybdenum in this case). The pressure inside the microcavity is the same as the base pressure in the evaporation chamber (e.g. 10^{-6} torr).

23.6 EXERCISES

1. (a) How does shadow-mask thickness affect dimensional control?
   (b) What effect does the contact versus proximity-mode operation have, on shadow-mask resolution?
2. When test capacitors are made, it is usual to deposit the top electrode through a shadow mask because of speed and simplicity. If the capacitors are used to measure the dielectric constant ε, how much will ε values be affected if shadow-mask dimensional control is $100\,\mu\text{m} \pm 5\,\mu\text{m}$?
3. If DRAM capacitor is made on a planar surface with $0.35\,\mu$m lithography, its area is ca. $0.35^2\,\mu\text{m}^2$. Calculate the capacitance increase that is offered by the hollow crown structure shown in Figure 23.10(d2).
4. Create a process flow for the horizontal-access hole structure shown in Figure 23.12.
5. It has recently been proposed to use shadow masks in ion implantation. Explore the issues that need to be addressed for such an approach.

REFERENCES AND RELATED READINGS

Bischofberger, R. *et al*: Low-cost HARMS process, *Sensors Actuators*, **A61** (1997), 392.

Chen, J. & K.D. Wise: A high-resolution silicon monolithic nozzle array for inkjet printing, *IEEE TED*, **44** (1997), 1401.

Cheng, Y.T. *et al*: Localized silicon fusion and eutectic bonding for MEMS fabrication and packaging, *J. MEMS*, **9** (2000), 3.

Cheng, Y.T. *et al*: Vacuum packaging technology using localized aluminum/silicon-to-glass bonding, *J. MEMS*, **11** (2002), 556.

Guckel, H.: High aspect ratio micromachining via deep X-ray lithography, *Proc. IEEE*, **86** (1998), 1586.

Hartstein, A. *et al*: A metal-oxide-semiconductor field-effect transistor with a 20 nm channel length, *J. Appl. Phys.*, **68** (1990), 2493.

Hing, S. *et al*: Multiple ink nanolithography: toward a multiple-pen nanoplotter, *Science*, **286** (1999), 523.

Hunter, W.R. *et al*: A new edge-defined approach for submicrometer MOSFET fabrication, *IEEE EDL*, **2** (1981), 4.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

LaDuca, A.J.: Amorphous silicon based anti-fuse, *Proc. IEEE Bipolar Circuits and Technology Meeting* (1993), p. 20.

Liang, C. & Y.-C. Tai: Sealing of micromachined cavities using chemical vapor deposition methods: characterization and optimization, *J. MEMS*, **8** (1999), 135–145.

Lim, M.-S. *et al*: In-situ vacuum-sealed lateral FEAs with low turn-on voltage and high transconductance, *IEEE TED*, **48**, (2001), 161.

Proceedings of the IEEE, **90** (2002), special issue on lasers in microelectronics manufacturing.

**Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030**

Part V

Integration

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

24

Process Integration

Process integration is the task of putting together individual process steps to create functional devices. This necessitates interfacing device design and processing, knowledge of process capability and device operation, understanding materials interactions and being prepared for equipment limitations – all aspects of microfabrication.

Process integration is about questions such as the following:

Wafer selection:

- Should n-type or p-type wafers be used?
- Can epitaxial or SOI wafers contribute to device performance?
- Are mechanical wafer specifications important, or electrical, or both?

Materials compatibility:

- Are the interfaces stable at process temperatures?
- Will the thermal expansion coefficient mismatches create stresses?
- Do the metals withstand the wet cleaning solutions?

Process-device interactions:

- How do thermal treatments add to diffusion profiles?
- Is etch profile critical?
- How does lithographic linewidth variation affect device performance?

Equipment and process capability:

- How much of the underlayer is lost during overetching?
- What is the step coverage of sputtered films in contact holes?
- Can thick stacks of bonded wafers be inserted into tools?

Design rules:

- What is the minimum width allowed for lines?
- How closely can you place structures?
- How much area should be allowed for misalignment tolerances?

Mask considerations:

- Which photomasks are critical, which are noncritical?
- Does etch undercutting need to be compensated on the mask?
- How much area should be reserved for test chips and how much for device chips?

Order of process steps:

- Does the stress relief anneal affect structures already fabricated?
- Can any steps be done after thin membrane formation?
- Should front-side processing be completed before backside processing?

Reliability:

- Do current densities in wiring need to be limited?
- How do stresses build up when more layers are deposited?
- What is the breakdown voltage of thin oxides?

24.1 PROCESS INTEGRATION ASPECTS OF A SOLAR-CELL PROCESS

The simple solar-cell process described in Figure 24.1 features some important interactions between process steps that arise when complete processes are put together.

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

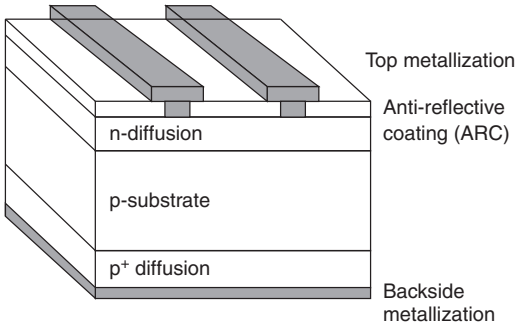


Figure 24.1 Solar cell cross-section

Process flow for solar cell: (cleaning steps omitted)

wafer selection
thermal oxidation
photoresist spinning on front
backside oxide etching
photoresist stripping
p^+ backside diffusion
oxide etching
n-diffusion
(optional thermal oxidation + backside oxide etching)
metal sputtering on the backside
anti reflective coating of PECVD nitride
contact-hole lithography
contact-hole etching
photoresist stripping
metal deposition on the front-side
lithography for front metal
metal etching.

All processes begin with substrate selection. P-type silicon is chosen, and the pn-junction is made by n-diffusion. However, it is advantageous to make a backside contact enhancement p^+ diffusion before the pn-junction. The heavy p+ diffusion on the back is unaffected by the light diffusion on the front-side because the difference in doping is three orders of magnitude. If n-diffusion was done first (with the backside protected by oxide), another oxidation would be needed to protect the lightly n-doped front-side during the heavy p^+ backside diffusion.

Oxidation and diffusion steps are high-temperature steps, and they must be finished before any silicon-to-metal contacts are made. After the first metal deposition (backside metallization), the process temperatures must be limited to ca. 450 °C. This rules out many deposition processes for the antireflective coating (ARC), for example, thermal oxide, TEOS CVD oxide or LPCVD nitride.

Backside metallization is done before the front-side ARC and metal. This is because the front-side is more important for device operation, and we would not like to clamp the wafer in a sputtering system face down after front-side processing is completed. It is possible to add a thermal-oxidation step after n-diffusion, or to perform the diffusion in oxygen, which will result in oxide growth. This oxide passivates the front surface and protects it during backside metal sputtering. However, the oxide has to be removed from the backside before sputtering, while leaving it on the front, which adds a few steps. Backside oxide could, of course, be removed by plasma etching, which only etches one side of the wafer. Solar cells are, however, devices driven by extreme cost-reduction objectives, and plasma etching is expensive compared to wet etching.

PECVD nitride ARC is deposited at 300 °C. We now have to open holes in this nitride to make contact with silicon. If the top metal was of the same size as the contact holes, perfect alignment and zero undercut etching would be needed for the metal to cover the hole completely. Because such processes do not exist, the top metal is designed to be somewhat wider than the contact hole to make sure that minor misalignment or linewidth loss in etching will not result in structures in which some silicon (in n-diffusion) would be exposed to the ambient air. If this was the case, cell performance would rapidly deteriorate as humidity and other environmental agents would get in contact with the pn-diode. Nitride ARC (with index of refraction ≈ 2) serves not only as an optical matching layer between the air ($n = 1$) and the silicon ($n \approx 4$) but it also protects from scratches, moisture and mobile ions.

24.2 WAFER SELECTION

Wafer selection and process design go hand-in-hand. In many cases, either n- or p-type silicon can be used, but then the doping steps need to be designed accordingly. If epitaxial wafers are used, then process design offers greater freedom because some bulk effects can be ignored, but it also introduces some limitations and incurs extra wafer costs. SOI wafers usually require full process rethinking in order to realize their full potential in reducing the number of process steps or enhancing device performance.

For MOS and bulk micromechanics, <100> material is used. For MOS, the motivation is silicon/oxide interface quality: less trapped charge and interface defects are generated in the oxidation of <100> silicon than of <111> silicon. For MEMS, anisotropic etching of <100> silicon is standard technology. In bipolar

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

technology <111> is used. When both MOS and bipolars are on the same chip (BiCMOS), <100> wafers are used because oxide for the MOS-part is more critical than <111> special features of the bipolar part. If there are no special requirements for silicon electrical or mechanical properties, <100> silicon is usually used because of its wide availability and low cost.

Crystal orientation need not be exactly along the major axis. Intentional off-axis cut (miscut) is beneficial for silicon epitaxy. <111> surface is atomically flat but the miscut introduces terraces that are favourable nucleation points in epitaxy (see Figure 6.4). A large miscut of 4° changes the apparent lattice constant of the silicon and offers possibilities to grow epitaxial oxides Y_2O_3 or $SrTiO_3$ on silicon. However, for anisotropic wet etching, wafers need to be cut as closely to the main crystal axis as possible. Whereas the standard cut is ±1°, MEMS wafers have a ±0.2° specification.

Wafer thickness increases with the diameter to improve mechanical strength. Mechanical strength is important especially during the high-temperature steps of oxidation, diffusion and epitaxy, especially at and above 1100 °C because thermally-generated stresses must not destroy the wafers. The occurrence of slip dislocations upon uneven cooling is a major concern. Thick wafers are also generally easier to handle.

In many applications thin wafers are needed. Solar cells would be cheaper if they used less silicon; wet etched bulk MEMS devices with 54.7° angle require less area in thorough-wafer etching, and in power transistors, resistive losses are minimized by using thin wafers. Wafer thicknesses down to 200 μm are quite readily available but they require special attention during processing. Wafers can also be thinned down to final thickness after all the device processing is done. This improves flexibility of the silicon dice and helps in packaging in applications such as smart cards.

24.2.1 Wafer specifications

24.2.1.1 Electrical specifications

Czochralski wafers are available over a wide range of dopant density, or alternatively stated, over a wide range of resistivities. Typical CZ-resistivities are listed in Table 24.1. If high-resistivity silicon (in kilo-ohm-cm range) is needed, CZ-wafers are not available and float zone (FZ) must be used.

24.2.1.2 Mechanical and surface specifications

Wafers come in standard sizes and thicknesses: for example, 100 mm and 525 μm, or 200 mm and 725 μm.

Table 24.1 CZ-silicon resistivity ranges (more extreme values can be obtained but then only part of the ingot will be within specifications)

Boron	0.002 – 4000 ohm-cm
Phosphorus	0.001 – 1000 ohm-cm
Antimony	0.008 – 0.1 ohm-cm
Arsenic	0.002 – 0.01 ohm-cm

In IC fabrication or many thin-film devices, wafer thickness is not an issue, but in bulk MEMS applications through-wafer etching is standard, and it depends critically on wafer thickness control.

Thickness refers to wafer centre point thickness only, and other numbers are needed to account for thickness variation and geometric distortions. Total thickness variation, TTV, is defined as the difference between the maximum and minimum values of thickness encountered in the wafer (Figure 24.2). Total indicator reading (TIR) concerns a front-side referenced measurement. TIR is defined as the sum of the maximum positive and negative deviation from a reference plane. If this reference plane is chosen to coincide with the focal plane of the mask aligner, focal plane deviation, FPD, is defined as the largest deviation, positive or negative, from this plane (Figure 24.3).

Bow and warp relate to shape deformations of free, unclamped wafers. Wafers can be concave, convex or undulating. Bow may be eliminated by clamping, that is, forcing the wafer flat on a chuck. Warp is the difference between the maximum and minimum distances of the median surface. Warp is a bulk property, in contrast to flatness, which is a surface property. Warp and bow can

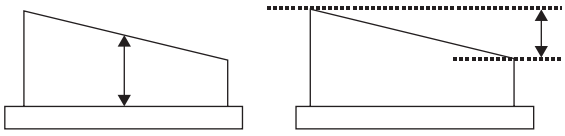


Figure 24.2 Thickness and total thickness variation (TTV). Wafer flattened to chuck; that is, backside reference

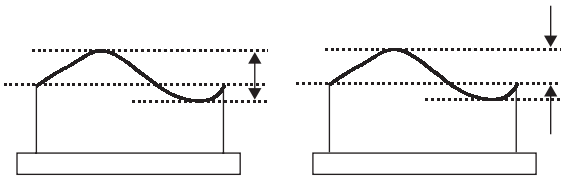


Figure 24.3 Total indicator reading (TIR) and focal plane deviation (FPD)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

develop during high temperature process steps or result from ingot sawing and lapping operations. The presence of excessive thickness variation and warp, will affect the lithographic performance via depth-of-focus problems.

Wafer surface topography can be divided into a few distinct scales: roughness is in the micron scale, flatness is in the chip scale and bow and warp are in the wafer scale. Smoothness and flatness are essential parameters for fusion bonding: wafers with 0.1 nm roughness are preferred for fusion bonding. Anodic bonding is more forgiving to surface roughness, and wafers with 0.5 nm roughness are fine for anodic bonding.

Flatness is measured over an area that is relevant to the lithography process and chip size. It directly impacts linewidth variation through lithographic depth-of-focus. Lithographic processes utilizing 1X full wafer-imaging systems are sensitive to global flatness, whereas step-and-repeat imaging systems are sensitive to local site flatness, over an exposure area, for instance, 20×20 mm.

24.2.2 Wafer behaviour in thermal treatments

Gettering is the trapping of impurities either intrinsically inside the wafer or extrinsically by a wafer backside layer. Gettering collects impurities in known and designed regions, where they do not interfere with device operation. In solar-cell fabrication, the costs are reduced by cheaper fabrication processes and looser cleanliness specifications, and cleanliness is not comparable to that in the IC industry. Gettering is incorporated in a few critical steps to reduce metal contamination. The IC industry uses gettering as extra insurance, in addition to high overall cleanliness.

Intrinsic gettering (IG) is closely related to bulk microdefects (BMD) and the thermal cycles that the wafer will experience during processing. Oxygen precipitates act as precipitation sites for other impurities, creating an impurity gradient that drives impurities towards designed precipitation sites. Wafer oxygen concentration is, thus, critical for internal gettering. IG is determined, by and large, when wafer processing begins. Oxygen precipitation has other effects too: it can cause stacking faults and dislocation loops, which lead to changes in <100>:<111> selectivity in KOH etching.

Extrinsic gettering on the wafer backside can be achieved by a number of techniques: both damage layer (laser or sand blasting damage), thin films (polysilicon) and phosphorous doping (diffusion or ion implantation) are possible. The number of gettering sites increases in these steps, or metal diffusion is modified, as in the case of phosphorus. Extrinsic gettering can be added to a process flow before critical oxidation steps.

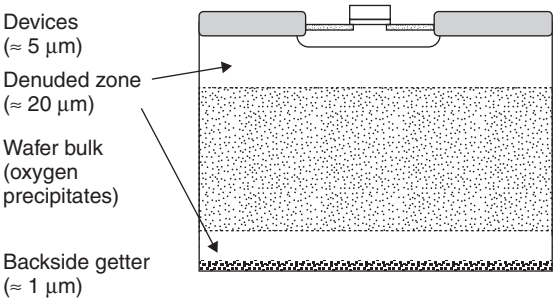


Figure 24.4 Wafer cross-section with denuded zone (not to scale)

In order to improve surface layer properties, oxygen is depleted in the surface layers by the creation of the so-called *denuded zone* (DZ) (Figure 24.4). Denuded zone, which has low oxygen concentration and minimized oxygen induced defects, is formed in three steps:

1. Outdiffusion step (1100–1200 °C; 1–4 h) in which oxygen diffuses out of the surface region, leaving <5 ppma oxygen.
2. Nucleation step at 600 °C, SiO_x is formed homogeneously throughout the wafer volume.
3. SiO_x precipitates growth and gettering (950–1200 °C, 4–16 h).

The denuded zone depth depends strongly on device requirements and it can range from 10 to 40 μm.

A DZ is not suitable for volume devices because of the vertical non-uniformity it introduces. If both ICs and MEMS devices are made on the same wafer, it is beneficial to have small, uniform oxygen precipitates as a compromise that satisfies to some extent the demands of both internal gettering and anisotropic etching.

24.2.3 Epitaxial wafers

Epitaxial wafers offer extreme purity: carbon and oxygen, which are always present in CZ-wafers, are practically absent in epitaxial layers. There are no COPs in epitaxial layers, meaning higher crystalline perfection of epi material. Epitaxial layers are not defect free, however, and stacking faults are the largest yield limiters in epitaxy. While CZ-wafers have cylindrical symmetry because of the rotation during crystal pulling, epitaxial deposition is uniform. Epitaxial doping uniformity is typically <4% and thickness uniformity around 1%.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Table 24.2 Epitaxial wafer applications

Technology	Subst	Epi	ρ (ohm-cm)	Thick (μm)	Motivation
CMOS	p^+	p	5–10	5–20	Latch-up prevention
Power-MOS	n^+	n	5–10	10–20	On-state conductivity
Analog bipolar	p^+	p	1–20	10–100	Speed performance
MEMS	p	n	1–10	7–150	Electrochem. etch stop
MEMS	p	p^{++}/p	0.005/1–10	3/3–30	Etch stop/device layer

Epitaxial deposition is reproducible, both for resistivity and thickness.

Minimum thickness by CVD homoepitaxy is around 0.5 μm, and the maximum thickness is determined by the economics of epitaxial growth, not by physics and chemistry. Epitaxial wafers have applications in almost all areas of microfabrication (Table 24.2), but epiwafer costs limit their use to expensive applications only.

24.2.4 SOI wafers

Several technologies have been developed for SOI-wafer fabrication. Each has its characteristic SOI device-layer thickness as well as typical buried oxide (BOX) thickness (Table 24.3). Epitaxial deposition on the SOI-device layer can be done to get almost any desired thickness, but this is an expensive approach because it combines epitaxy and SOI, both of which are expensive.

SOI technology offers improvements in many ways, and one of them is the reduction of the number of process steps because more processing has been done to the wafer to begin with. Compared to bulk materials, the most obvious advantage of all the SOI devices is dielectric isolation. Integrated circuits fabricated in SOI material consist of single-device islands dielectrically isolated from each other (lateral isolation) and from the underlying substrate (vertical isolation). Similarly, each and every piezoresistor fabricated on SOI is isolated from other resistors. This means that leakage currents through the bulk are eliminated. SOI MOS transistors and SOI piezoresistors can operate at ca. 300 °C, as opposed to bulk devices, which fail above ca. 125 °C due to increased leakage currents.

Table 24.3 SOI-wafer applications

Device technology	<Si> device layer	Buried oxide	SOI technology
CMOS	10–200 nm	200–400 nm	Smart-cut, SIMOX
Bipolar	1–10 μm	0.1–1.0 μm	Various
MEMS	5–50 μm	0.5–4 μm	Bonded SOI
Power IC	1–100 μm	1–4 μm	Bonded SOI

SOI-wafer cost is ca. 10 times the cost of bulk wafers. This cost disadvantage has to be compensated by other factors like smaller chip size, higher performance, easier processing (less process steps) or special features like radiation hardness for space and military applications. SOI-wafer availability is also an issue: SOI-wafer manufacturers use very different technologies, and wafers from different manufacturers are not substitutes for each other like bulk wafers are (in the first approximation).

24.2.5 Non-silicon substrates

Using non-silicon wafers can have various reasons. Quartz and fused silica are dielectric and fully compatible with silicon processing, but they are more expensive and fragile than silicon. The main reason against use of glass wafers is contamination danger from sodium in the glass. However, the alternatives are not ideal either: high-resistivity silicon is still somewhat conductive, and capacitive losses will occur. Processing on non-silicon substrates will be discussed in Chapter 29.

24.3 PATTERNS

The lithography tool must be specified early on in process design, because with the tool, exposure wavelength, mask size, wafer size and chip size become fixed. Wavelength sets limits on photoresist selection, mask plate material and resolution. In 1X exposure tools, the mask size is somewhat larger than the wafer size, for example, 5″ for 100 mm wafers and 7″ for 150 mm wafers. With 1X aligner the chip size is limited by wafer size and edge exclusion. With step-and-repeat lithography tools the chip size is limited by exposure field size, which is ca. 20 × 20 mm. Optimization is needed to fit many small chips in the field or alternatively, stitching is needed to make larger chips.

Photoresist polarity, negative or positive, needs to be selected before mask making. It is possible to

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

design the patterns in one polarity and to invert polarity computationally in the mask making process, but once the physical mask plates have been drawn, the mask and resist are tied together. Exposure wavelength also limits mask plate materials: at 436 nm (g-line), soda-lime glass is acceptable, but at 365 nm (i-line) and below, quartz becomes the material of choice.

It is possible to mix lithographic techniques: this approach is known as mix-and-match. Not all lithography steps are equal: some are more critical than others. Critical levels determine device functionality in a critical way, for example, CMOS gate mask determines gate length, which affects transistor speed and leakage. CMOS contact holes are critical because they have to be aligned very closely to the active area and the gate. A single linewidth-critical level may be written by an e-beam, while the rest are exposed by optical lithography. This approach saves money by eliminating a new optical tool with better resolution, and enables devices and chips to be made for R&D purposes or small volume production. In the production of 0.35 μm technology, the critical levels can be exposed by 4X, 248 nm deep UV stepper and the non-critical levels by 5X, 365 nm i-line stepper, or in 0.50 μm technology, the critical levels are exposed by 365 nm 5X stepper and the non-critical levels on a 1X tool. This approach is investment related: some additional work from mix-and-match (e.g., in alignment scheme) is traded for major savings in equipment purchase prices.

The design data format that is generally used in photomask fabrication is GDSII. Similar standards for plastic masks made by photoplotters for printed circuit boards are Gerber and HPGL. If designs are made in other formats, conversion is required. This may introduce pattern errors and should be carefully checked.

In CMOS, the complementarity of NMOS and PMOS can be utilized to reduce mask design work: once an n-well mask is finished, its complement can be made and used as a p-well mask because all areas on the wafer that are not n-well are p-well or isolation areas. Such a mask is termed an *automatically generated mask*.

Imperfections in the patterning process can be partly compensated in the mask making process. Proximity effects, or effects of neighbouring structures, can be eliminated or reduced by optical proximity correction (OPC) techniques. OPC calculation determines the exposure dose on the basis of pattern size, shape and spacing of neighbouring structures, and compensates for non-idealities by fine-tuning pattern shapes. OPC calculations are massive and the implementation requires extra writing time in mask making.

Undercutting in wet etching can be compensated by biasing the photomask. The patterns on the mask are made wider by the amount of etch undercutting for light-field structures, and narrower for dark-field structures. This procedure is process dependent, in the sense that it yields good results for one film thickness. Mask biasing can be done in a global fashion: all structures on an aluminium level can be biased wider by, for example, twice the designed aluminium film thickness. For a 3 μm nominal linewidth, this translates to 5 μm wide patterns (assuming 1 μm aluminium thickness), and thus 1 μm etch undercutting per side. If the resolution of the lithography tool is 6 μm (capable of printing 3 μm lines with 3 μm spaces), mask biasing cannot be done because 1 μm spaces would need to be resolved. Mask biasing wastes silicon real estate, and the resolving power of the lithography tool is not fully utilized for increasing device-packing density.

On a 1X mask there are usually three elements: device chips, test structures and alignment marks (Figure 1.13). The area usage between these elements depends on process and device maturity. In early phase development, the mask includes mostly test structures and a few devices; in volume manufacturing, device chips take up practically all the area, with test structures embedded in the scribe lines between the chips. Test structures include both device-specific and process-specific measurements. The latter are identical in all runs using the same process, and they are used for collecting information on process performance, stability, drifts and variation for statistical process control (SPC).

The speed and flexibility of direct write lithographies have some niches to themselves, in R&D and in the manufacturing of extremely specialized devices, in which only a handful of chips are needed. Optical lithography is not completely out of that market either: it is possible to write, on a single mask plate, as many different chip designs as the area allows. If wafer stepper exposure area is 20×20 mm, it is possible to fit six designs of ca. 0.6 to 0.7 cm^2 on one reticle. This multi project chip (MPC)/multi project wafer (MPW) approach is often used in R&D when only 10 to 20 chips are needed for functionality checking or system-design experiments. Of course, all chips on the mask will see exactly the same fabrication process. This is usually not a limitation for CMOS ICs, but MEMS processes are usually very idiosyncratic and cannot easily be shared by different designs.

24.4 DESIGN RULES

Design rules are statements about allowed structures with regard to linewidths and spacings, overlap and

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

layer-to-layer positioning. These are often referred to as layout rules, as opposed to electrical design rules that include information about sheet resistances, current density limitations, contact resistances and so on. Layer-thickness design rules are needed in a capacitor design: oxide thickness determines capacitance density, both when the oxide is used as a capacitor dielectric as such, and when it is used as a sacrificial layer in the fabrication of an air-gap capacitor. Device models (for transistors, resistors, capacitors) are additional higher-level abstractions of the process for circuit designers. Design rules and models are always process specific. They are also company specific: 0.13 μm CMOS processes from different suppliers have different sets of rules and models.

24.4.1 Layout rules

Layout design rules are formal geometric rules that relieve the designer from the details of the fabrication process (Figure 24.5). The process engineer has distilled the physical capabilities and limitations of the fabrication process into design rules with the aim of making the process more robust. Sometimes breaking the rules leads to zero yield and sometimes subtler effects are encountered. Design rules are often divided into compulsory and advisory rules, the latter being hints of known good practices.

Minimum size and spacing are basic layout rules. Three elements contribute to them

- lithographic process capability;
- structure widening in subsequent process steps;
- device interactions.

Lithographic capability involves the optical tool, photomask quality, resist properties and resist thickness. If the lines are not accurate on the mask, then the design width cannot be obtained on the wafer. Breaking the minimum line and space rules will lead to catastrophic failures.

Very often, minimum space is different from minimum linewidth. For one thing, lithographic resolution (pitch) is not usually divided equally between line and space: it is typical that, for example, a 0.5 μm linewidth process has a 0.5 μm minimum line and a 0.7 μm minimum space. Sometimes processes are specified by half-pitch: the previous process would then be classified as a 0.6 μm process.

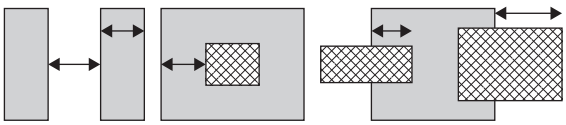


Figure 24.5 Layout design rules: spacing, linewidth, enclose, cut-in and cut-out

The final structure width is determined by process step properties. Diffusion is an isotropic process and a 3 μm diffusion depth leads to ca. 3 μm lateral spreading. Similarly, isotropic etch undercutting necessitates similar design concerns: equal spacing of 10 μm wide, 5 μm deep grooves would result in touching of the neighbouring grooves.

Device interactions come in many guises and they are device and process specific. Transistors need to be isolated from each other, and this isolation takes up space. Inductive devices must be placed far away from each other because of magnetic field coupling over distance. It is also important to understand and to limit structures that can be placed between two coils as these can couple into the magnetic field.

Different mask levels may have different linewidth rules: for example, one mask level contains critical structures, and narrow lines are allowed, but other levels may have only non-critical structures: pads for wire bonding are, for example, 50 × 50 μm or 100 × 100 μm and design rules are then more relaxed, with, for instance, a 5 μm minimum overlap rule while a 0.3 μm overlap rule might be used for critical levels.

24.4.2 RCL elements

As an example of design rules, let us consider three devices, resistors, capacitors and inductors (RCL). Analog components are more demanding than digital ones, with absolute values of resistance; for instance, in digital MOS transistors a 10% linewidth variation will not affect the on/off action, but it changes the resistance of a resistor by 10%. A gate oxide thickness change of 10% will not ruin a MOS transistor even though its threshold voltage and leakage current will differ from the design values, but for an analog capacitor, the variation is there to stay. In many cases, absolute values of resistance or capacitance are not used, but instead the ratios of two resistances or capacitances are. Deposition process non-uniformity is usually taken as ±5% across the wafer but it is very good locally.

Inductors exemplify linewidth and spacing rules (Figure 24.6 and Table 24.4): linewidth determines resistance and spacing is important for inductance. Narrow spaces would be advantageous for real estate savings, but lithographic resolution sets limits there. Narrow lines will lead to increased resistive losses and are thus counterproductive.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

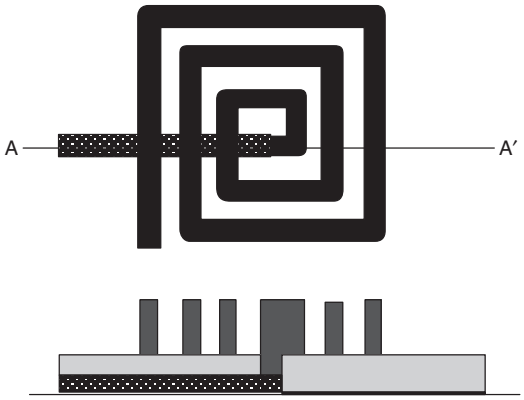


Figure 24.6 Inductor coil (black): top view and cross-sectional view along cut line AA′. Lower metal (dotted) makes contact with the coil metal at the centre

Table 24.4 Design rules for inductor

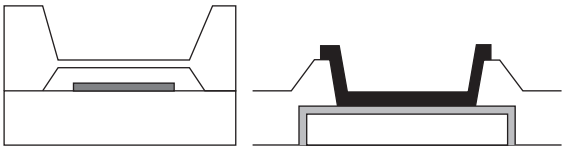
Minimum linewidth	5 μm
Minimum space	3 μm
Distance from unrelated inductor	50 μm
45° corners recommended	
90° corners allowed	

Resistance is determined by linewidth, linelength, thickness and resistivity (the latter two are usually taken together via sheet resistance $R_s \equiv \rho/t$). High resistance values call for thin resistors, long lines, narrow lines or high-resistivity material. Resistor linewidths are seldom the minimum linewidths that are available in the process, but are rather large in order to improve the absolute value control. Long, straight resistors complicate circuit topology and meandering resistors are usually employed. However, meandering structures need some special rules of their own because corners do not contribute to resistance equally with the linear parts. Thinning down the resistor is not without problems because of process control and reproducibility, not to mention the fact that thin-film resistivity is thickness dependent, which leads to a new characterization of the material.

Design rules for resistors must, therefore, include linewidth and spacing rules and sheet resistance rules, with appropriate rules for meander corners (Table 24.5). For thin-film resistors that are made by etching, the spacing rule is determined by the etch process and it can be made very small. Diffused resistors always require allowance for lateral spreading. Unlike inductors, two resistors can be placed with minimum space between them because resistors do not interact over distance like inductors.

Table 24.5 Design rules for a polysilicon thin-film resistor

Resistor lines	3 μm
Space	3 μm
High-resistivity poly	5000 ohm/sq
Low-resistivity poly	500 ohm/sq
Only 90° corners allowed in meandering resistors	



(**a**) (**b**)

Figure 24.7 Capacitor area determined by the bottom electrode in a micromechanical air-gap capacitor (**a**) and by top electrode in a metal-to-polysilicon capacitor with polyoxide as the capacitor dielectric (**b**)

Capacitance per unit area is the basic electrical rule for a capacitor ($C/A = \varepsilon/d$). Capacitor rules are very much two-layer rules: both the bottom and top electrodes need attention. It is important to specify which electrode determines the capacitor area. Two cases are shown in Figure 24.7.

24.4.3 Layer-to-layer placement rules

Placement of the top electrode over the bottom electrode must be limited by the design rules: Figure 24.8 shows ideal and misaligned capacitors.

The misaligned top electrode is undesirable not only because it introduces uncertainty in capacitor area but also because the film quality on the sidewall is different from planar areas. The breakdown voltage of the dielectric is, for one thing, different on the sidewalls, along with many other electrical reliability measures. The design rules must demand the capacitor top plate to be smaller by a margin that ensures planar capacitors, as shown in Figure 24.7.

A similar argument is the basis for edge location rules on two different layers in general. It is not advisable to



Figure 24.8 Cross-sectional views of a capacitor: top and bottom electrodes perfectly aligned (**a**) and misaligned (**b**)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

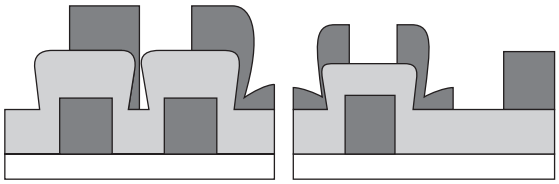


Figure 24.9 Coincident structures on two different levels will lead to serious topography evolution due to misalignment. The spacing rule of unrelated structures must also account for interlayer thicknesses to avoid crevasses

place two structures exactly on top of each other because misalignment (and lithographic and etch uncertainties) will always introduce some uncertainty into the edge position (Figure 24.9).

24.4.4 Overlap rules

When structures on two different layers need to coincide, overlap rules must be invoked. Overlap rules make sure that the layers that need to touch will do so irrespective of process variation. Alignment of structures on different levels depends on the following three factors:

- lithography tool alignment performance;
- pattern placement accuracy;
- alignment sequence.

Tool alignment performance is usually taken as 1/3 of minimum linewidth for 1X tools and 1/5 for steppers. If a 1X tool with 3 μm minimum capability is used to print 3 μm wide contact holes, 1 μm alignment tolerance needs to be designed in. If the underlying resistor is of the same width as the contact hole, this misalignment will lead to a severe crevasse formation: when the contact hole is etched into CVD oxide, misaligned contact exposes the underlying oxide, which will also be etched (Figure 24.10). The subsequent metal sputtering and/or CVD process will have difficulties in filling the crevasse.

In order to make sure that the contact hole will touch the resistor, the resistor contacting area is made larger to accommodate any misalignment. This is termed *collar* or *border* or *dogbone*. This wastes area but it is necessary for process robustness.

The second contribution to alignment accuracy between levels comes from pattern placement on the mask: the masks for two different layers are two separate physical objects and the exact position of the structures on the mask plate is subject to its

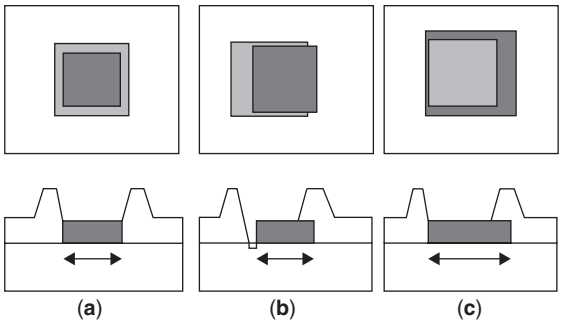


Figure 24.10 Top view mask images and cross-sectional view of contact-hole alignment are: (**a**) perfect alignment of contact hole (grey) to the underlying structure (black); (**b**) misaligned contact without misalignment allowance and (**c**) misalignment with collar in the underlying structure

own statistical variation. If image placement error on the mask is 1/10 of the minimum linewidth, its contribution is $\sqrt{(x_1^2 + x_2^2)} \approx \sqrt{2}\ x$, if mask errors are identical on both plates. This translates to ca. 14%, usually less than the contribution from misalignment.

Alignment sequence is the third factor. In Figure 24.11, contact holes are aligned to the resistor, and the metal is also aligned to the resistor: the whole idea of the structure is to make the metal-to-resistor contact. If the metal was aligned to the contact hole, we would have to account for two tool misalignment tolerances: one for contact hole-to-resistor alignment and another for contact hole-to-metal alignment. Assuming Gaussian distribution, this leads to an alignment tolerance of $\delta\sqrt{n}$, where n is the number of alignments involved.

If the first process step is diffusion or implantation, there will be nothing visible (or something barely visible) on the wafer, and the second lithography

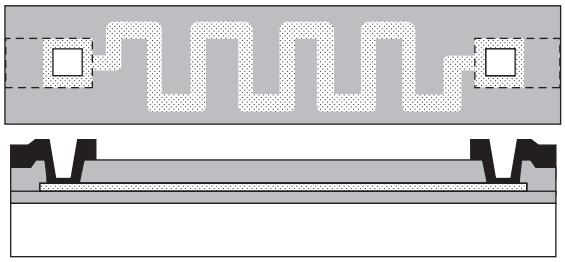


Figure 24.11 Thin-film resistor: top view and cross-sectional view. Both contact hole and metal are aligned to resistor. Resistor (dotted) has collars to ensure contact hole overlap; similarly, metal collars ensure overlap of contact

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

step – the first alignment – cannot be done. Therefore, it is common practice to etch special alignment marks into silicon at the very beginning of the process. This is called *zero level*, and it adds a little complexity to the process, but on the other hand it makes alignment more robust. Planarization later in the process may smear alignment marks, and it might be that in some process steps the alignment marks must be protected in order to maintain them.

When isotropic wet etching is used in the resistor process, etch undercutting of the resistor and contact holes work in opposite directions: the resistor is a light-field structure that is narrowed by etch undercutting, whereas contact holes are dark-field structures that become wider. These processes add up and the overlap rule has to accommodate that. In a similar fashion, contact hole and metal etching work in opposite directions. In general, overlap rules for plasma-etched processes are much tighter than those of wet-etched processes. Plasma etching increases device-packing density not only by its ability to make narrower lines but also through smaller overlap requirements.

In multilevel metallization or in multilayer surface micromechanical processes, it would often be advantageous to place many holes (contact holes or release etch holes) on top of each other to save area and to simplify design work. This is called *stacking* (Figure 24.12). However, it rapidly leads to serious step coverage problems in the deposition steps that follow. A simple solution is to make the upper-level contact larger. This alleviates some problems related to misalignment and to sputtering step coverage because a larger contact hole has a lower aspect ratio. Most often design rules forbid stacked contact holes. Area is then lost because the holes must be placed side by side. In Chapter 27, we will see how replacement of sputtered aluminium by CVD tungsten can overcome this problem at the expense of increased process complexity.

When a circuit with a few devices is made (e.g., in a student lab) the effects of misalignment might be shrouded by process noise and other variations, but in manufacturing with millions of devices on a chip, statistical variation will always produce some misaligned structures. Some of these are fatal, but some are hidden. Misalignment can cause unintentional etching and gaps that are deeper and/or wider than expected, which can leave a void when gap filling fails, with potential reliability problems during device lifetime in the field.

Automatic checking of design rules is a standard procedure for advanced chips. Design rule checking (DRC) includes both individual level checks (dimensional rules) as well as layer-to-layer checks (overlap rules, positioning rules).

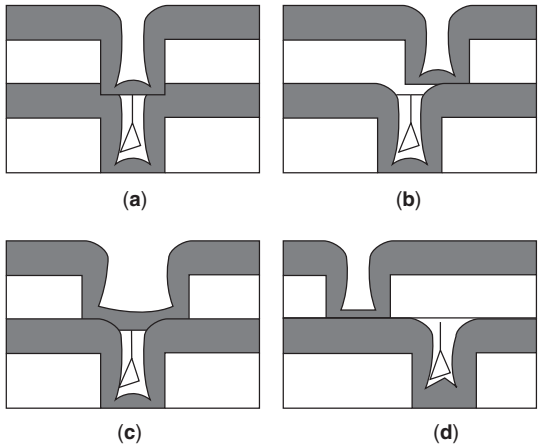


Figure 24.12 (**a**) Stacked contacts – perfect alignment; (**b**) stacked contacts – misalignment; (**c**) stacked contacts – wider upper contact and (**d**) non-stacked contacts

Table 24.6 Electrical design rules for a 1 μm analog–digital CMOS process

Layers	R_s (ohm/sq)	Contacts	Contact res (ohm)
Gate poly	100 ± 20	Metal 1 to diffusion	15*
Resistor poly	200 ± 20	Metal 1 to poly	10*
Resistor poly, hi res	1000 ± 100	Metal 2 to metal 1	0.2*
Metal 1	0.1		
Metal 2	0.03		

**Note*: Contact resistances are for 1.2 μm × 1.2 μm contact size.

24.4.5 Electrical design rules

Electrical design rules for a 1 μm analog CMOS process are given in Table 24.6. Circuit designers can use these values when assessing wiring resistances and timing delays, and to evaluate current densities.

24.4.6 RCL chip

For a simple device, the order of process steps is sometimes obvious, but for more complex devices there are many possible variations in the order of steps. An integrated passive chip (RCL chip) with four different devices is shown in Figure 24.13. Molybdenum is

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

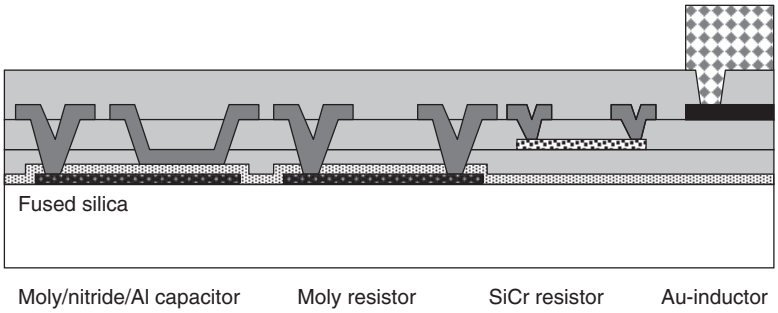


Figure 24.13 RCL chip on a fused silica substrate: four metallic layers (Mo, Al, SiCr, Au) and four insulator layers are used (a LPCVD nitride and three CVD oxides). Adapted from VTT Microelectronics annual review 2000

used for low-resistivity resistors (Mo $\rho \approx 10\,\mu$ohm-cm), SiCr for high-resistivity resistors ($\rho \approx 2000\,\mu$ohm-cm), moly-nitride-aluminium for capacitors and gold coils for inductors. The chips are processed on fused silica substrates. LPCVD nitride is used for capacitor dielectric, and three layers of CVD oxide insulate the devices from each other.

RCL-chip process flow: (cleaning steps omitted)

wafer selection
molybdenum deposition
photomask #1: molybdenum resistor and capacitor bottom plate
molybdenum etching (strip resist)
nitride deposition (LPCVD)
CVD oxide-1 deposition
deposition of SiCr high-resistivity resistor
photomask #2: SiCr resistor pattern
SiCr etching (strip resist)
CVD oxide-2 deposition
photomask #3: contact holes to molybdenum
plasma etching of CVD-ox-2/CVD-ox-1/nitride (strip resist)
photomask #4: contact holes to SiCr resistor and to capacitor top
wet etching of CVD-ox-2/CVD-ox-1 (strip resist)
aluminium deposition
photomask #5: aluminium pattern
aluminium etching (strip resist)
CVD oxide-3 deposition
photomask #6: contact holes to aluminium
etching of CVD-ox-3 (strip resist)
photomask #7: Inductor coil pattern
gold electroplating (strip resist).

24.5 CONTAMINATION BUDGET

Wafer cleaning can be viewed as an important stabilization tool: surfaces will be in a known state after wafer cleaning. Cleaning steps are the most numerous of all process steps: most other major steps are both preceded and followed by cleaning steps.

Cleaning processes need to be tailored for the particular process steps that follow: processes have different tolerances for different kinds of contamination. Thermal oxidation will clear organic residues, but it is very sensitive to metal contamination because metals diffuse rapidly at elevated temperatures and some metals are incorporated into the growing oxide. Epitaxy requires crystal information and it is extremely sensitive to native oxides or other surface layers. Wafer bonding is a major challenge for particle cleaning.

The processes generate contamination themselves: ion implantation and sputtering, where energetic ion bombardment is present, and produce metallic contamination by sputtering metals from shield plates; deposition processes generate films and particles form when unwanted films on reactor walls flake; lithography is done with organic films and lithography chemicals (HMDS, photoresists) are major sources of organic contamination, as is plasma etching where carbon from etch gases and etched resist are abundant.

Contamination is partly a materials selection problem: some materials are allowed and some are forbidden. This can be either device related or tool related: in the RCL example in Figure 24.13, a separate LPCVD nitride tube must be used for nitride-on-molybdenum deposition and another LPCVD tube is reserved for non-metal processes. Copper causes a serious minority carrier lifetime degradation in silicon, but its superior electrical properties warrant its use in high-performance applications. Copper, therefore, puts very high demands on barrier properties.

Cleaning strategies are also process integration issues. Iron contamination increases oxide defect density and results in lower oxide breakdown voltage. Use of p-type

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

wafers differs from n-doped wafers because some iron is held immobile by Fe-B pairs. Contamination is strongly oxide-thickness dependent, and the pre-oxidation cleaning strategy must be designed accordingly. Use of ultra-high purity chemicals in a 20 nm gate oxide process is financial waste but an absolute must in a sub-10 nm oxide process.

Photoresist developers are hydroxides, and NaOH-based developers were once the mainstay, also in MOS-fabs, but organic developers such as TMAH do not pose alkali contamination risks. MEMS fabrication with KOH etching tends to be strictly separated from all MOS activities. If MEMS fabrication is done in a MOS fab/lab, TMAH etchant is used to eliminate alkali ion contamination risk. However, TMAH and KOH etching processes are similar only in their gross features, and all details of rates, selectivities and etch stop properties need to be redone, as discussed in Chapter 21.

Wet cleaning baths must also be dedicated to certain processes only. Pre-gate cleaning is very critical, and only wafers that are very clean to begin with can be processed in pre-gate cleaning baths. Gate oxide usually has an oxidation tube of its own; not shared even with other front-end oxidation processes. Wet etching baths may additionally be divided by no-resist/resist division. For example, of two HF-baths one is used for sacrificial oxide removal and the other for pattern etching.

24.6 THERMAL PROCESSES

24.6.1 Film modification

Metal films have limitations both because of presence of metal/silicon interfaces, and because the top surface can oxidize. Sputtering, evaporation and electrochemical deposition are basically room temperature processes, and even mild thermal treatments, at and below 400 °C can modify film properties dramatically. Electroless copper can have resistivity of 4 μohm-cm as-deposited, but 400 °C anneal in N_2/H_2 can bring it down to 2 μohm-cm. This results from grain growth and void annihilation. Grain growth is proportional to square root of anneal time, indicative of a diffusion limited process (cf. thermal oxidation).

CVD films (and PECVD films in particular) and spin coated films are often porous and unstable. PECVD films may contain up to 30 at. % hydrogen, which will diffuse during subsequent processing. Inert anneal at 900 °C will densify (PE)CVD oxide film into more thermal oxide –like state. Thickness reduction of 10% is not unusual. This densification is seen as etch rate and polish rate reduction. There is room for high temperature annealed (PE)CVD oxides because thermal oxide thicknesses are limited by the diffusion-controlled parabolic growth law, whereas (PE)CVD film thickness increases linearly with deposition time. PECVD deposition of 2 μm thick film plus annealing can be completed in ca. two hours, whereas thermal oxidation would require two days. Thick oxides (>1 μm) are needed as mask oxides in MEMS and in optical devices as waveguides.

Deposited films may need stoichiometry tailoring, and for oxide films, oxygen anneal can result in more stoichiometric films. Sputter and MOCVD deposited Ta_2O_5 films are often annealed at 700 °C in oxygen. This causes crystallization and oxygen deficiency is compensated. Dielectric constant of amorphous Ta_2O_5 is ca. 25, whereas crystalline Ta_2O_5 has ε of ca. 35.

Annealing will crystallize amorphous LPCVD silicon into polycrystalline silicon at ca. 600 °C. This polycrystalline film is not identical to the film which has been deposited at 600 °C and which is polycrystalline to begin with: its grain size and grain size distribution are different, its surface morphology and stress state are different. When those films are doped, they will end up with different resistivities, because dopant diffusion in a polycrystalline film is dependent on grain size and grain size distribution. Diffusion in polycrystalline films is mainly along the grain boundaries, with a minor contribution from bulk diffusion inside grains. Diffusion of dopants in polysilicon is, therefore, much faster than diffusion in single-crystalline silicon.

24.6.2 Surface modification

Silicon nitride is the standard masking material for localized thermal oxidation of silicon (LOCOS). The surface of nitride will react with oxygen, even though oxygen cannot diffuse through the nitride. This modified surface layer is termed *oxynitride*. Its thickness is limited to a few nanometres. Somewhat similar, extremely etch-resistant material can be deposited by PECVD, using a process that has features of both oxide and nitride deposition.

Nitridation in molecular nitrogen can sometimes take place, even though N_2 is usually regarded as an inert gas and often employed in place of argon. When wafers are loaded into oxidation furnace, nitrogen is used as a curtain gas and some nitridation of silicon surface is possible because the temperatures are fairly high.

Intentional nitridation is usually done with ammonia. Oxide can be nitrided in NH_3. Oxynitride film

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

has a higher dielectric constant and better electrical quality than pure oxides. Films such as this are known as NO, ONO and RONO, or nitrided oxide, oxidized nitrided oxide and reoxidized nitrided oxide, respectively. These films are standard CMOS gate dielectrics in deep sub-micron technologies where oxide thickness is below 10 nm.

The unintentional surface modification most commonly encountered is oxidation: some residual oxygen or moisture in a furnace atmosphere will lead to oxidation. Copper annealing in a moist atmosphere will result in copper oxide, and 5 ppm water vapour is enough to disturb titanium silicide formation. Oxidation is sometimes done to protect the surface: for example, aluminium oxide is chemically much more stable than aluminium, and it is preferable to oxidize the aluminium surface. Room temperature plasma oxidation (i.e., RIE etching step with oxygen) will do the job.

24.7 THERMAL BUDGET

The thermal budget concept is a central to front-end process integration. Diffusion of dopants takes place in all high-temperature steps: in addition to diffusion itself, it manifests itself during epitaxy, oxidation, densification anneal and implant damage annealing. The final doping profile is the sum of diffusion in all these steps. Effective Dt, which is a measure of diffusion distance, is calculated as

$$(Dt)_{\text{eff}} = \Sigma D_n t_n \qquad (24.1)$$

where Ds are diffusivities under appropriate conditions and ts are times for the high-temperature steps.

In an aluminum gate CMOS process (Figure 19.1), source/drain diffusions are done before gate oxidation, and dopants will, thus, diffuse further during gate oxide growth. In a self-aligned polygate process, gate oxide growth is done before S/D formation, and therefore shallower junctions are possible because there are fewer high-temperature steps after source/drain formation.

A thermal budget sets limits on possible process steps. PSG and BPSG film flow was once a standard technique to make the topography smoother in CMOS processes above 1 μm generations. Of course, it was only applicable after polysilicon, not after metal deposition. However, the required annealing (ca. 950–1000 °C, dependent on boron and phosphorous content) causes dopant diffusion, and as junction depths were scaled down with linewidth, glass flow became non-usable in sub-micron technologies.

Dopant segregation must be taken into account when designing a fabrication process. Segregation of dopants between silicon and oxide can seriously deplete the interface of dopants, but this segregation is dependent on annealing/oxidation atmosphere: wet oxidation, dry oxidation, inert anneal in nitrogen or reducing anneal in hydrogen rich ambient can behave differently.

Ion implantation annealing has two different elements: activation of dopants and damage removal. Activation energies for these processes are different, and depending on the temperature, damage removal can either be accomplished in a few seconds or it can take hours. Transient enhanced diffusion has major implication for diffusion profiles, as will be discussed in connection of shallow junctions in Chapter 25.

24.8 METALLIZATION

All electrical devices need at least one level of metallization in order to connect to the outside world and so do most mechanical, thermal, fluidic and bio-devices, because electrical sensing and actuation are widely used.

Metal to semiconductor contacts come in two basic varieties: ohmic (resistive) or diode-like (Schottky) (Figure 24.14). Even the ohmic contacts have some diode character because metal and semiconductor work functions are never exactly equal. If the semiconductor doping level is low ($<10^{19}/\text{cm}^3$), charge carriers will have to overcome the barrier (which is proportional to metal workfunction−semiconductor electron affinity difference $\varphi_{\text{metal}} - \chi_{\text{semiconductor}}$) by thermionic emission. In a heavily doped semiconductor, the situation is different: charge carriers can tunnel through the barrier because the barrier is thin. Barrier thickness is related to depletion width in the semiconductor (which is proportional to $1/N_{\text{D}}$).

Aluminium is the most widely-used ohmic contact between metal and silicon. The silicon doping level needs to be in excess of $10^{19}/\text{cm}^3$ for good ohmic contact. Aluminium, which is a p-type dopant for silicon, can also be used to make an ohmic contact with a lightly doped p-type silicon: during contact anneal (in forming

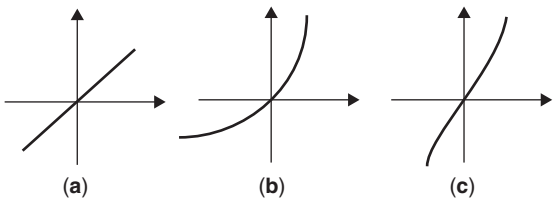


Figure 24.14 Metal-semiconductor contact I-V-curves (**a**) ohmic; (**b**) diode-like (Schottky) and (**c**) real metal-semiconductor contact

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

gas at 450 °C), aluminium will dope the top surface of the silicon and good contact is made. Schottky contacts to silicon are usually made with PtSi.

Contact resistance R_c is given by

$$R_c = \rho_c / WL \tag{24.2}$$

where ρ_c is the contact resistivity, and W and L are the contact dimensions.

Contact resistivity depends on barrier height (0.55 eV half bandgap of silicon) and silicon doping concentration ($2 \times 10^{20}/cm^3$ maximum dopant solubility), which cannot be changed. Therefore, metal-to-silicon contact resistivities cannot be much less than 10^{-7} ohm-cm^2. This translates to ca. 0.1 ohm for $1 \times 1\,\mu m$ contacts. Metal-to-silicide and metal-to-metal contact resistivities are in the 10^{-8} ohm-cm^2 range, and this is one added benefit of silicides in sub-micron technologies.

24.9 RELIABILITY

Final passivation provides protection against the environment. There are mechanical elements of passivation such as scratch resistance, chemical aspects such as moisture resistance and gettering and physical effects such as prevention of sodium diffusion.

The standard passivation materials are PSG and PECVD nitride, either alone or as a two-layer stack. Phosphorous doping of a CVD oxide film is beneficial for sodium ion gettering, but too much phosphorus makes the oxide hygroscopic, so there is a delicate balance. Usually, phosphorus content is ca. 5% wt. The nitride provides mechanical strength and chemical resistance, but this chemical stability translates to plasma etching for bonding pad opening, whereas oxide passivation can be etched in HF-based solutions (not, however, without difficulty because HF-water solutions attack aluminum: see Table 11.3 for etch selectivities).

Reliability has both built-in and operational features. Oxide thickness non-uniformity results in a permanent non-uniformity that may pose, for example, breakdown voltage variation. During the MOS transistor operation high-energy electrons, scattered from the channel into the gate oxide, cause oxide charge there, leading to wear-out. This degradation depends on the operating voltage. Similarly, step coverage is frozen in but its effects on reliability depend on the current density.

24.9.1 Oxide defects and electrical quality

Even though the interface between silicon and thermally-grown silicon dioxide can be reproducibly fabricated, it is far from ideal. The interface-trapped charges are caused by broken bonds (from structural defects, oxidation induced defects and contamination). Because they are at the interface, the potential in silicon will charge or discharge them. An interface-trapped charge can be reduced by forming gas anneal. There is always some positive fixed charge in the vicinity of the interface, and it is related to silicon ionization during the oxidation process. There are also trapped charges, which can be positive or negative, caused by energetic electrons from ionizing radiation, and there can be mobile charges from contamination, most notably Na^+ ions.

The electric field that oxide can sustain is usually reported by the breakdown voltage: 10 MV/cm is considered to be the intrinsic breakdown field. This is also termed *C-mode failure*. B-mode failures happen at 2 to 8 MV/cm and A-mode below 2 MV/cm. An example of oxide breakdown statistics is shown in Figure 24.15.

A-mode failures are gross defects: pinholes and voids (Figure 24.16). COPs in silicon lead to oxidation of microscopic pits, which will lead to oxide integrity loss. B-mode failures are more benign and more subtle, like oxide thinning, trapped charges or metal contamination induced defects. C-mode failures are intrinsic to the oxide structure, but can be affected by nanoscopic defects such as increased surface and interface roughness. A-mode failures are seen as yield loss in fabrication and B-mode failures as reliability problems in accelerated testing or in the field.

Metals are responsible for many of the defects described above. If the surface is contaminated, silicates like $MgSiO_4$ or silicides CuSi and NiSi can be formed, rather than silicon dioxide. Their formation consumes silicon and, therefore, the oxide will be locally thinner.

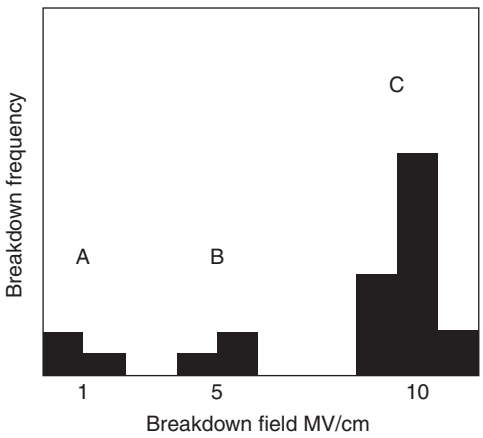


Figure 24.15 Oxide breakdown distribution: A-mode at low field; B-mode at medium field and C-mode at high field

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

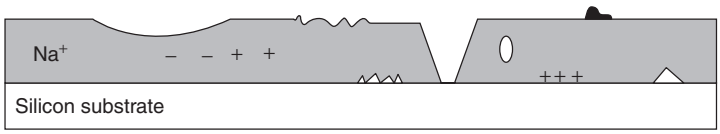


Figure 24.16 Oxide defects (left to right): Na^+ mobile charge, thinning, fixed charge, surface and interface microroughness, pinhole, void, interface charge, particle, stacking fault. Adapted from Schröder, D.K. (1998), by permission of John Wiley & Sons

Unreactive metals dissolve in the growing oxide, which leads to decreased intrinsic breakdown strength. Sodium (Na) contamination leads to increased oxidation rate; whereas iron (Fe) and aluminium (Al) lead either to increase or decrease depending on the level of contamination and time. Metals can also catalyse the reaction SiO_2 (s) + Si (s) $\rightarrow$ 2 SiO (g) (which takes place under low oxygen partial pressure, e.g., during ramp-up in a furnace), leading to oxide evaporation and pinhole-like defects.

Oxide dielectric strength is tested by a number of different experimental set-ups:

- Ramped voltage: the voltage between MOS gate and substrate is linearly increased (0.1 or 1 V/s) until the oxide breaks down. Breakdown voltage V_{BD} is defined as the voltage where a sudden voltage drop occurs.
- Time-to-breakdown under constant current (TTBD; t_{BD}): constant, preset current is fed into the insulator, and the voltage is recorded as a function of time. TTBD is the time when a sudden voltage drop occurs.
- Charge-to-breakdown (Q_{BD}): in constant current test $Q_{BD} = J_{injected} \times t_{BD}$. Good oxides exhibit values of $10\,C/cm^2$, but this is dependent on the injected current.

24.9.2 Electromigration

Electromigration (recall page 58) depends on a large number of factors: macroscopic factors include geometry of the lines, and their width, shape and area. Microscopic factors include grain size, texture, and alloy solutes and their precipitation at the grain boundaries and interfaces. Solutes like copper in aluminium (e.g., in Al-2 wt% Cu) increase resistance to electromigration because copper atoms block diffusion at grain boundaries (Figure 24.17). What is more, grain size and linewidth are not independent: when grain size and linewidth become equal (typically when thickness-to-width ratio is about unity), the number of grain boundaries is strongly reduced, leading to the so-called bamboo structure with one grain extending across the line. In polycrystalline material, grain boundary diffusion is important and the elimination of grain boundaries will affect electromigration.

Mean time to failure (MTF) due to electromigration is given by

$$MTF = AJ^{-n} \exp^{(Ea/kT)} \qquad (24.3)$$

where A is a constant dependent on wire geometry and metal microstructure, J is the current density and E_a the activation energy. The factor n is not known accurately, but $n = 1.7$ is a usable value for aluminium.

For aluminium thin films E_a is of the order of 0.5 to 0.8 eV, whereas for bulk aluminium it is 1.4 to 1.5 eV. As a general trend, the higher the activation energy, the better the electromigration resistance. It can be roughly estimated on the basis of metal melting point T_m: the higher the melting point, the higher the electromigration resistance. To put it in another way: high melting point equals high bond energy. At room temperature, which is $T_m/3$ for aluminium, aluminium atoms have a reasonable probability for diffusion. For tungsten, room temperature corresponds to $T_m/10$, and electromigration is less by orders of magnitude. Copper falls between the two. For short lines and/or for low current densities, electromigration is not an issue.

24.9.3 Stress migration

Electromigration is studied by accelerated tests under higher-than-normal current densities at elevated temperatures. However, voids appear in metal lines at elevated temperatures even when no current runs through them. This is known as *stress-induced voiding* or *stress migration*. The driving force is the gradient in the strain field: some atoms find it energetically favourable to move to voids.

The source of stress is thermal expansion mismatch between metal and the encapsulating (PE)CVD dielectric. Strain (elongation) is proportional to CTE and temperature difference, which translates, for aluminum, to 1% linear elongation or ca. 3% volume

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

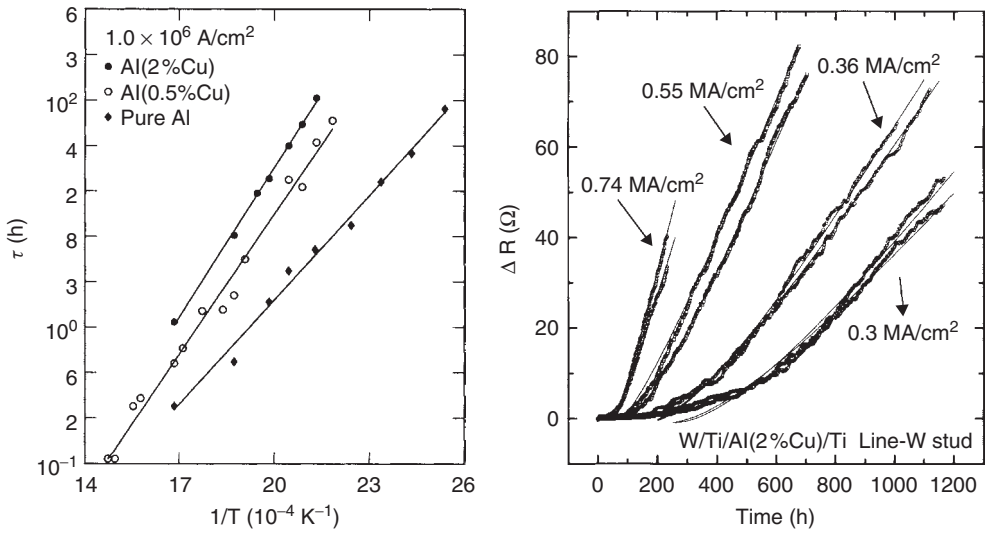


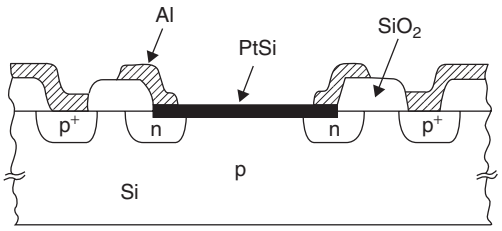
Figure 24.17 (**a**) mean time to failure of 2.5 μm wide Al, Al (0.5 wt% Cu) and Al(2 wt% Cu) lines at different temperatures with 1 MA/cm^2 current density. Reproduced from Hu, C.-K. *et al*. (1993), by permission of AIP. (**b**) incubation time before resistance increase sets in at 255 °C. From Hu, C.-K. (1995a), by permission of Elsevier

change when 300 °C PECVD is done. This elongation corresponds to stresses over 1 GPa (the order of magnitude can be estimated by Equation 4.1). Aluminium lines expand during PECVD, and they are fixed at their elongated state because of mechanical stiffness of deposited oxide/nitride layers. This high tensile stress can be relaxed by cracks, and once a crack is formed, it tends to grow.

Compressive stresses in aluminium can be relaxed via hillock formation. Hillocks are small protrusions. Their size can be up to micrometres, which is equal to insulator thickness between two levels of metallization. If some mechanically stiffer film prevents relaxation in the vertical direction, then hillocks can grow laterally, and again, a micrometre is a very typical size for metal line spacing. In both cases, hillocks can short-circuit the two metal lines. Low-temperature processing helps in reducing hillocks (and stress and electromigration). Alloying aluminium with copper is also helpful in minimizing hillock formation because it blocks grain boundary diffusion.

24.10 EXERCISES

1. How many lithography steps are needed to fabricate the solar cell shown in Figure 1.6?
2. Draw the photomasks (e.g., on transparency film) required to fabricate the RCL chip of Figure 24.13. Include design features such as spacing rules and dogbones.
3. Create a fabrication process for the platinum silicide Schottky diode shown below. Platinum silicide is formed by metal/silicon reaction, not by etching. From Chen, C.K. *et al*: Ultraviolet, visible and infrared response of PtSi Schottky-barrier detectors operated in the front-illuminated mode, *IEEE TED*, **38** (1991) 1094, fig. 2.



4. How do diffused resistor design rules differ from the thin-film resistor case?
5. Integrated passive chip (Figure 24.13):
   (a) What is the nitride thickness if areal capacitance density is 4 nF/mm^2, and nitride $\varepsilon_r = 7$?
   (b) Why is the first contact etching by plasma and the second by wet etching?
   (c) SiCr thin-film resistor resistivity is 2000 μohm-cm. Design a 5 kohm resistor.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

6. Which methods can you use for the following measurement tasks:
   - oxide pinhole density;
   - thickness of nominally 30 nm thick titanium;
   - photoresist thickness uniformity;
   - sputtered aluminium step coverage;
   - implanted arsenic dose;
   - particle removal efficiency in NH_4OH/H_2O_2 wet cleaning;
   - Ta_2O_5 film deposition;
   - ion implantation of boron into a phosphorous doped wafer;
   - silicon dioxide thinning in etching;
   - mask oxide undercutting in KOH etching of <100> silicon;
   - copper electroplating;
   - photoresist sidewall angle.
7. DRAM trench capacitors are cylindrical holes with high aspect ratios. What is the aspect ratio in a 0.15 μm linewidth process if the capacitor oxide thickness is 5 nm and capacitance is 40 fF?
8. Capacitor nitride deposition uniformity across the wafer is ±1%, and across the batch it is ±2%. The top electrode area is defined by etching the CVD oxide (thickness and etch non-uniformity ±5%) against the capacitor nitride. If the oxide thickness is 200 nm and nitride thickness is 10 nm, plot the capacitance variation as a function of the oxide:nitride etch selectivity.
9. Redo Exercise 9.3, this time for 5X step-and-repeat lithography and quartz masks.
10. If the TiW/Al (50 nm/400 nm) line experiences a void in aluminium, how much will the line resistance increase?
11. If Al (2% wt. Cu) lines have MTF of 400 hours at 255 °C, what is their expected lifetime under standard operating conditions?
12. A micromechanical air gap parallel plate capacitor (Figure 24.7(a)) has 1 mm^2 area and 1 μm air gap. What is the capacitance? If femtofarad capacitance change can be measured, what is the corresponding displacement of the movable capacitor plate?

REFERENCES AND RELATED READINGS

Chen, C.K. *et al*: Ultraviolet, visible, and infrared response of PtSi Schottky-barrier detectors operated in the front-illuminated mode, *IEEE TED*, **38** (1991) 1094, fig. 2.

Fair, R.B., Conventional and rapid thermal processes, in C.Y. Cheng & S.M. Sze (eds.): *ULSI Technology*, McGraw-Hill, 1996.

Gardner, D.S. & Flinn, P.A.: Mechanical stress as a function of temperature in aluminum films, *IEEE TED*, **35** (1988), 2160.

Hu, C.-K. *et al*: Electromigration of Al(Cu) two-level structures: effect of Cu kinetics of damage formation, *J. Appl. Phys.*, **74** (1993), 969.

Hu, C.-K.: Electromigration failure mechanism in bamboo-grained Al(Cu) interconnections, *Thin Solid Films*, **260** (1995a), 124

Hu, C.-K. *et al*: Electromigration and stress-induced voiding in fine Al- and Al-alloy thin-film lines, *IBM J. Res. Dev.*, **39** (1995b), 465.

Istratov, A.A. *et al*: Advanced gettering techniques in ULSI technology, *MRS Bull.*, **25**(6) (2000), 33.

Leslie, T. *et al*: Photolithography overview of 64 Mbit production, *Microelectron. Eng.*, **25** (1994), 67.

Muller, T. *et al*: Assessment of silicon wafer material for the fabrication of integrated circuit sensors, *J. Electrochem. Soc*, **147** (2000), 1604–1611.

Schröder, D.K.: *Semiconductor Material and Device Characterization*, 2nd ed., John Wiley & Sons, 1998.

Yue, J.T., Reliability, in C.Y. Cheng & S.M. Sze (eds.): *ULSI Technology*, McGraw-Hill, 1996.

**Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030**

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

25

CMOS Transistor Fabrication

CMOS remains the most voluminous microfabricated device by a wide margin. Many of the process steps of microfabrication were developed originally for CMOS fabrication, and later adapted to other microdevices. In the last 30 years, linewidth scaling has been driven almost exclusively by CMOS. Ion implantation was a technique for high-resolution nuclear spectroscopy in the 1960s, but today CMOS doping is its main application. Thin oxides, down to 2 nm today, are really nanostructures in volume production, and major CMOS wafer fabs produce these oxides by square metres a day.

CMOS linewidths were in the 5 μm range in the mid 1970s. This may sound like old-fashioned technology, but it was the time when CMOS got its present-day appearance and diverged dramatically from older generation aluminium gate processes. The 5 μm process exhibits most of the essential process steps that characterize CMOS: it is an oxide-isolated, ion-implanted, plasma-etched, self-aligned gate process (Table 25.1). Advanced CMOS features and processes will be discussed later in this chapter after the basic polygate process has been presented.

The main modules of CMOS fabrication are shown in Figure 25.1. Front end is about diffusions and doping profiles. It is high-temperature processing. The gate module involves gate oxidation and gate poly deposition, lithography and etching, plus the source/drain diffusions. Contact defines the division between the front end and the back end: after the metal–silicon interface has been formed, process temperatures become limited to ca. 450 °C. The number of metallization levels has increased steadily: 5 μm CMOS had one level, 2 μm CMOS two levels, 0.8 μm CMOS three levels and with 0.13 μm generation has seven levels of metal.

Table 25.1 Al versus polygate CMOS

	Al-gate	Polygate
Linewidths	>5 μm	<10 μm
Doping	Thermal diffusion	Ion implantation
Isolation	pn-junction	Oxide (LOCOS)
Gate material	Aluminium	Doped polysilicon
Gate process	Non-self-aligned	Self-aligned
Gate etching	Wet/isotropic	Plasma/anisotropic

25.1 5 μm POLYSILICON GATE CMOS PROCESS

Process integration begins with wafer selection. n-type silicon, 4 ohm-cm (phosphorus concentration ca. 1.5×10^{15} cm^{-3}) is chosen as the starting material. This will mean that NMOS transistors will be made in p-well, and PMOS transistors in the substrate directly. The choice of p-type starting material would lead to a reversed configuration.

In Figure 25.2, the top view of the photomask is shown, together with a cross-sectional view of the device at a specified stage of the process.

Wafers are cleaned, and a pad oxide of 40 nm is grown in dry oxygen and followed by LPCVD nitride deposition (100 nm). These films will be used in making the LOCOS isolation structure. The first lithography step defines transistor-active areas. Nitride will cover transistor-active areas, and it will be etched away from areas that will become isolation oxide. Nitride etching in CF_4 plasma stops on pad oxide. By stopping the etch at the oxide, the silicon surface is not damaged and cleaning of the wafer will be easy. It is possible to etch through the nitride/oxide stack and into silicon to create an isolation structure known as *recessed LOCOS*. Recessed LOCOS has the advantage that the surface will be approximately planar when the silicon-etched depth is ca. 50% of LOCOS oxide thickness.

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

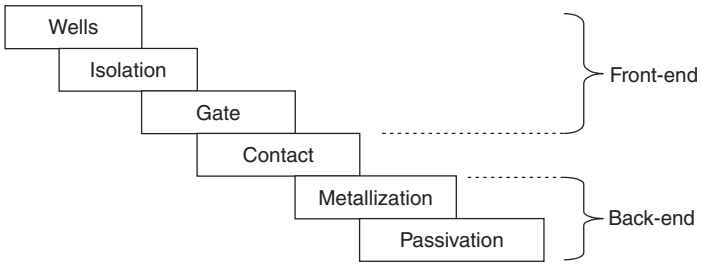


Figure 25.1 Main modules of a CMOS process

The second lithography step defines p-well areas (Figure 25.2(b)). Boron is implanted with a dose of $2 \times 10^{13}\,\mathrm{cm}^{-2}$ and energy of 40 keV. There are three distinct areas on the wafer; the resist-covered areas will not be implanted, and no boron will penetrate through the resist. Boron will traverse thin pad oxide areas and dope silicon. Some boron will penetrate through the nitride/oxide stack, but the dose reaching silicon will be small and short range.

After photoresist strip, arsenic is implanted with energy of 50 keV and dose $10^{12}\,\mathrm{cm}^{-2}$ (Figure 25.2(c)). This low energy, coupled with the heavy mass of arsenic, leads to shallow implanted depth under the pad oxide areas and no penetration of nitride/oxide stack. Arsenic will thus be confined to areas that will be under thick field oxide in the final device. This field oxide implant improves isolation between neighbouring transistors. Drive-in diffusion is performed next: a short oxidation step (50 min at 950 °C, dry oxidation) is followed by a 500 min, 1150 °C diffusion in nitrogen. Diffused layer sheet resistance is monitored by four point–probe measurements.

Note that arsenic and boron implants overlap. The overlap could be eliminated by an extra lithography step, but there is no need for that: the p-well area remains p-type because the boron ion implantation dose is twenty times more than the arsenic dose.

LOCOS oxidation then follows: 360 min at 1050 °C, wet oxidation. This will result in ca. 1.2 μm thick oxide (Figure 25.2(d)). p-well is diffused to a depth of ca. 4 μm. After oxidation, the nitride/oxide stack is removed in three steps: nitride surface is oxidized during LOCOS wet oxidation, and HF is used to remove this oxynitride; phosphoric acid (H_3PO_4) etches nitride; and finally HF clears the pad oxide. Because no pattern is made by these etching steps, the isotropic nature of wet etching is not detrimental, and wet etching is superior to plasma etching in terms of selectivity.

In the next step, sacrificial oxidation is done. Ca. 80 nm of thermal oxide is grown and immediately etched away in HF. The purpose of this step is to make sure that no nitride remains from the LOCOS process. This residue is known as *white ribbon* because defects at the periphery of the active area are seen as a white ribbon in an optical microscope.

Gate oxidation is preceded by the RCA-cleaning process. Ammonia–peroxide cleaning is for particle removal, HF for native oxide removal and hydrochloric acid–peroxide cleaning for metallic contamination elimination. Dry oxidation at 1050 °C, 65 min, produces ca. 80 nm thick gate oxide.

The third lithography step is used to tailor the threshold voltage of PMOS transistors (Figure 25.2(e)). A dose of $1.2 \times 10^{12}\,\mathrm{cm}^{-2}$ of boron is implanted with energy of 50 keV.

PMOS transistor threshold-current tailoring by implantation is a case where the order of steps can be chosen at will. Two sequences are possible.

Sequence I:	Sequence II: gate oxide first
Lithography	Cleaning
Implantation	Gate oxidation
Resist stripping	Lithography
Cleaning	Implantation
Gate oxidation	Resist stripping
Polysilicon deposition	Cleaning
	Annealing
	Polysilicon deposition

In the first sequence, the implanted dopants diffuse further during gate oxidation and the dopants penetrate deeper than in the 'gate oxide first' option. In the second sequence, the gate oxide experiences implantation and photoresist stripping, both of which are potentially damaging. Cleaning after stripping becomes critical because it determines the oxide–polysilicon interface quality. In the first sequence, polysilicon deposition takes place on the fresh oxide surface, which is very clean (assuming no delay between gate oxidation and polysilicon LPCVD).

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

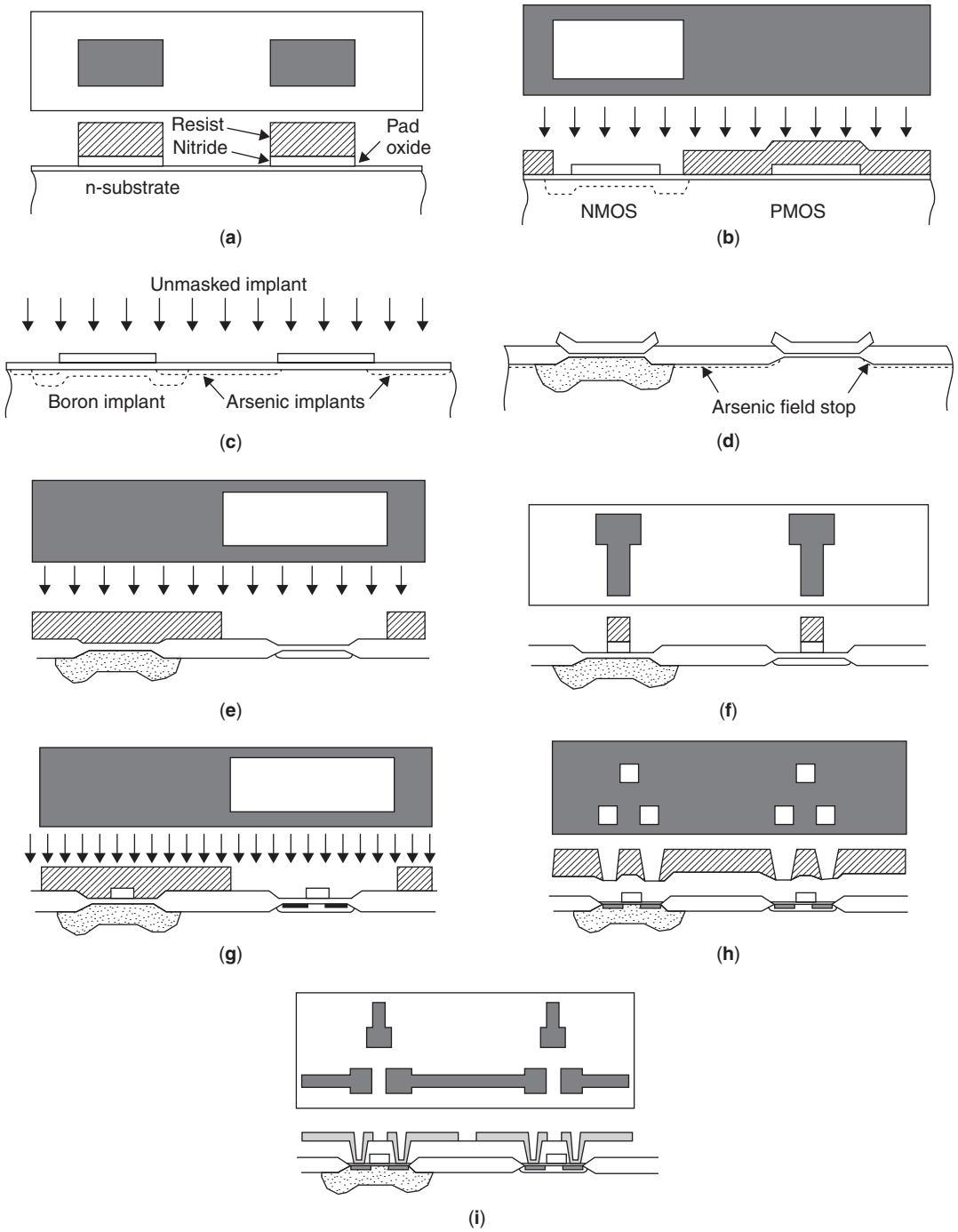


Figure 25.2 (**a**) Active area definition; (**b**) P-well: boron-ion implantation; (**c**) arsenic field stop implantation; (**d**) LOCOS wet oxidation; (**e**) PMOS threshold voltage–tailoring by boron implantation; (**f**) polysilicon gate etching, photoresist still in place; (**g**) self-aligned source/drain high-dose boron implantation. Note that this is the same mask that was used in threshold voltage tailoring; (**h**) contact hole lithography, photoresist pattern before etching and (**i**) finished device with aluminium metallization

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Polysilicon, thickness ca. 500 nm, is deposited undoped. A separate $POCl_3$ gas-phase doping step is performed after deposition, and the resulting poly sheet resistance is ca. 30 ohm/sq. Both NMOS and PMOS gates are made of the same material, the phosphorus-doped poly.

The fourth photomask defines the polysilicon gates. Gate poly etching is done in CF_4/O_2 plasma (Figure 25.2(f)). The selectivity requirement is not very demanding because the gate oxide is fairly thick, so the process can be optimized for sidewall profile, rate and/or uniformity. After photoresist stripping and cleaning, a mild oxidation step (900 °C, 10 min, dry oxidation) is performed, and ca. 50 nm of oxide is grown on polysilicon. This removes plasma etch damage and re-grows gate oxide on source/drain areas a bit.

The fifth photomask is actually the same mask as the third, the PMOS threshold voltage mask: it defines PMOS-transistor area. This time, it protects the NMOS areas from PMOS S/D boron-ion implantation. A high dose $2 \times 10^{15}\,cm^{-2}$ of boron is implanted at 40 keV (Figure 25.2(g)).

The sixth mask is a reverse polarity version of the previous mask: areas that are not PMOS area are either NMOS area or isolation, and can be doped by phosphorus. The sixth mask is thus an automatically generated mask: there is no need to design it once the PMOS mask has been drawn. NMOS S/D implantation with phosphorus is at 120 keV energy with a dose of $3 \times 10^{15}\,cm^{-2}$. After resist stripping and wafer cleaning, a short diffusion/oxidation step is done at 900 °C for 20 min.

CVD oxide (phosphorous-doped silica glass, PSG) of ca. 1 μm thickness is deposited next. PSG is a glassy material and above its glass transition temperature (ca. 1050 °C) it will flow, resulting in beneficial smoothing of the top surface. This is the last high-temperature step, and dopant profiles are now 'frozen'. Junction depths of both PMOS and NMOS transistors are ca. 1 μm (L/5), with source/drain area sheet resistances of ca. 30 ohm/sq for NMOS and ca. 90 ohm/sq for PMOS. The p-well depth is ca. 4 μm and its sheet resistance is ca. 4 kohm/sq. Threshold voltages for NMOS and PMOS are ca. 1.3 V and -1.5 V, respectively.

The seventh mask defines contact holes in the oxide (Figure 25.2(h)). Wet etching in BHF is used to open the contacts. Contact hole–design rules must take into account the fact that there will be ca. 1 μm undercut in this etching step. After photoresist stripping and wafer cleaning, ca. 1 μm of aluminium is sputtered on the wafers.

The eighth mask defines metallization patterns. Aluminium is etched in H_3PO_4-based wet etch. Aluminium lines will be ca. 2 μm narrower than the photoresist pattern, whereas the contact holes will be ca. 2 μm wider than the resist dimensions. Overlap rules must make sure that the metal covers the contact completely (Figure 25.2(i)). After stripping and wafer cleaning, forming gas anneal at 450 °C improves silicon-to-aluminium contact.

Passivation layer of silicon oxynitride is deposited by PECVD. The ninth mask defines bonding pad openings, and plasma etching of oxynitride opens those pads. The wafer-level processing is now complete.

The wafers will be tested electrically, at wafer level, and non-functional chips will be inked. Dicing will separate the chips, and functional chips will proceed to encapsulation and packaging. Many tests cannot be performed at wafer level and more characterization will take place on packaged chips. The cost of testing can be very high if the chips need to be tested for a multitude of parameters.

25.1.1 CMOS variations

A prototypical 5 μm CMOS process has been described. There are many minor variations between different CMOS manufacturers: implant doses and diffusion times differ, oxide thicknesses and junction depths vary, mask compensations can be used, and so on. More variety enters the picture if, for example, analog CMOS is made. Then some of the doping steps will be used to make resistors, and extra lithography masks may be needed. In more advanced analog CMOS processes, an extra polysilicon layer is added for resistor and capacitor fabrication. EEPROM processes also need extra polysilicon for the floating gate. Bipolar transistors can be added to a CMOS process, which will be discussed in Chapter 26.

25.2 MOS TRANSISTOR SCALING

As linewidths were scaled from 5 μm to ca. 1 μm, plasma etching replaced wet etching not only for critical steps but for all patterning etches. Oxidation and diffusion times were scaled down in order to make shallower junctions. Steps such as PSG flow were eliminated because S/D diffusion spreading had to be minimized. We will now discuss some issues relevant to scaling of CMOS, both from device and fabrication point of view.

25.2.1 Lithography scaling

The contribution of lithography to scaling has been constant over the past decades. Resolution of projection optical systems has been pushed down in a seemingly continuous evolutionary process, as discussed in Chapter 9 (Equations (9.4) and (9.5)). Depth of focus (DOF) has dramatically suffered from exposure wavelength reduction and NA improvements, and it is major

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Table 25.2 Lithographic scaling of CMOS

Linewidth (μm)	Wavelength λ (nm)	NA	k_1	DOF (μm)
1	436	0.38	0.8	±1.5
0.5	365	0.48	0.6	±0.8
0.25	248	0.60	0.6	±0.35
0.18	248	0.65	0.5	±0.30

concern. Table 25.2 shows CMOS lithography trends assuming $k_2 = 1$ but letting k_1 evolve.

One approach to better resolution (and smaller linewidths) is by wavelength reduction. This strategy has been steadily used: from 436 nm (g-line from an Hg-lamp) to 365 nm (i-line from an Hg-lamp) to 248 nm (KrF laser) to 193 nm (ArF laser). Should all else be equal, this alone would result in an improvement by a factor of two in resolution and a factor of four in device areal density.

Numerical aperture (NA) enhancement is another clear route that has been used. In 20 years, NA has been increased from ca. 0.15 to 0.7, an improvement by a factor of 4 or 5. Resolution enhancement by NA increase has been dearly paid for on the focus side: DOF is becoming very small indeed

Depth of focus defined above is an optical concept but resist chemistry and resist profile specifications (which depend on subsequent process steps) must be considered. Besides optical DOF, other factors must be accounted for: the wafer is not flat and neither is the wafer chuck, and stepper focus mechanisms are not perfect. All these contribute 0.1 to 0.2 μm to the focus budget. Previous etching and deposition steps can easily create a topography variation of the order of half a micrometre, so planarization is critical for lithography. Fortunately, in the backend of the process linewidths are somewhat larger than in the front end, and this relieves some pressure on DOF.

The 'constant' k_1 has had a major role recently. Scaling down k_1 involves a much higher degree of control over details of the patterning process: photomask dimensions, focussing mechanics, resist thickness, developer concentration, development time, and so on. In research laboratories, k_1 can be as small as 0.3, but then extensive process control measurements must be carried out. In volume manufacturing, k_1 has to be somewhat higher, for example, 0.5, for process robustness.

25.2.2 Transistor scaling

CMOS transistor scaling (Table 25.3) is most often discussed from the lithographic, linewidth-scaling point of view, but vertical scaling is equally important. Source/drain diffusions must be made shallower because they must not extend sideways under the gate. If the diffusions touch, catastrophic failure occurs, but even in the case where they do not touch, they degrade device performance via increased leakage current and parasitic capacitances. Sideways diffusion is kept to a minimum when vertical diffusion, and therefore junction depth x_j, is minimized.

Transit time from source to drain, which is a proxy for device speed, can be calculated as

$$\tau = L/v = L/\mu E = L^2/\mu V_{ds} \qquad (25.1)$$

where L is channel length, v is the velocity and μ the mobility of the electron in electric field $E = V_{ds}/L$. The gate and the substrate form a capacitor, with the gate oxide as the capacitor dielectric of thickness T. The gate capacitance is then

$$C = \varepsilon W L/T \qquad (25.2)$$

where W is the width of the gate and ε is the dielectric constant of oxide. The charge in transit is

$$Q = -C_g(V_{gs} - V_{th}) = -(\varepsilon W L/T)(V_{gs} - V_{th}) \qquad (25.3)$$

and the current

$$I_{ds} = Q/\tau = \mu\varepsilon W/LT(V_{gs} - V_{th})V_{ds} \qquad (25.4)$$

V_{gs} is the gate–source voltage, V_{th} is the threshold voltage where the gate starts controlling the charge carriers and V_{ds} is the drain–source voltage.

Scaling down transistor dimensions (lateral dimensions L and W, and vertical dimensions, oxide thickness T and junction depth x_j), smaller by a factor n ($n > 1$) leads to the following new dimensions:

$$L' = L/nW' = W/nT' = T/n \qquad (25.5)$$

For many CMOS generations, the operating voltage was kept constant at 5 V (Table 25.4), but the electric field cannot be increased without limit because of dielectric breakdown and hot electron considerations,

Table 25.3 CMOS scaling by a constant factor n (>1)

$\tau' = (1/\mu)((L/n)^2/(V/n)) = (1/\mu)(L^2/V) = \tau/n$
$C' = C/n$
$I' = I/n$
$P'_{switch} = C'V'^2/2\tau' = P_{switch}/n^2$
$E'_{switch} = (1/2)C'V'^2 = E_{switch}/n^3$
$P'_{dc} = I'V' = P_{dc}/n^2$

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Table 25.4 Front-end scaling (ca. 1980–1995): supply voltage constant at 5 V

Generation	3 μm	2 μm	1.5 μm	1 μm	0.7 μm	0.5 μm
T_{ox} (nm)	70	40	30	25	20	14
x_j (nm)	600	400	300	250	200	150
Gate delay (ps)	800	350	250	200	160	90

Table 25.5 CMOS front-end scaling at the turn of the millenium

Generation	0.35 μm	0.25 μm	0.18 μm	0.13 μm
T_{ox} (nm)	8	6	4.5	4
Supply (V)	3.3	2.5	1.8	1.5
V_{th} (V)	0.65	0.6	0.5	0.45

which necessitates lower operating voltage, V', given by $V' = V/n$ (Table 25.5). Using shorthand $V \equiv V_{gs} - V_{th}$, we can write the physical parameters for the scaled devices as shown in Table 25.3.

Scaling is mostly beneficial: transistor area scales as $1/n^2$ $(A' = L'W' = LW/n^2 = A/n^2)$, transistor speed increases as $1/n$, switching power decreases as $1/n^2$ and switching energy decreases as $1/n^3$. The power density (P/A) remains constant. Junction depth scaling, x_j, has been mostly in line with oxide thickness scaling, but more recently it has been difficult to keep the pace. This is because ion implantation damage necessitates high-temperature annealing, which inevitably leads to diffusion however shallow the original implantation profile. Linewidth scaling is just one factor in packing density increase: process and device cleverness can contribute amazingly large area reductions.

Note that gate oxide thickness is related to linewidth L roughly as $L/45$ and junction depth is ca. $L/5$.

25.2.3 Front end simulation

The CMOS front end is a transistor parameter optimization. It involves mostly process simulation to produce diffusion profiles and film thicknesses, which are fed into device simulators to obtain transistor characteristics such as threshold voltages and current–voltage characteristics. If a 1D process simulator is used, it feeds 1D device simulation, and similarly 2D for 2D and 3D for 3D. This process development loop is pictured below (Figure 25.3).

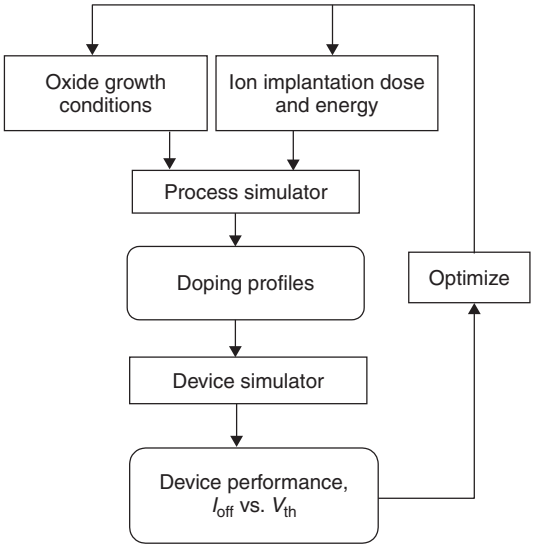


Figure 25.3 Front-end process development loop depends heavily on process simulation

25.3 ADVANCED CMOS ISSUES

The 5 μm CMOS process presented above has main features similar to any modern CMOS process. Over the years, refinements, modifications, materials changes and many other improvements have taken place. The CMOS process of the year 2000 with 0.25 μm linewidth and over 25 mask levels is quite advanced compared to 9 mask levels for 5 μm. We will not discuss changes generation by generation, but rather look at some important trends in processes and structures themselves.

At and below 1 μm, the following features have been implemented in CMOS:

- step-and-repeat 5X reduction lithography with $\lambda = 365$ nm;
- spacers and LDD implants;
- silicides;
- CVD-W plugs;
- planarization.

CMP planarization and shallow trench isolation (STI) in the place of LOCOS become standard for half-micron generations. Deep sub-micron (0.35 μm, 0.25 μm, 0.18 μm, 0.13 μm) generations (Figure 25.4) have taken advantage of many more new techniques and materials:

- DUV-lithography with $\lambda = 248$ nm;
- nitrided oxides instead of pure SiO_2;
- p^+ gate for PMOS and n^+ gate for NMOS;

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

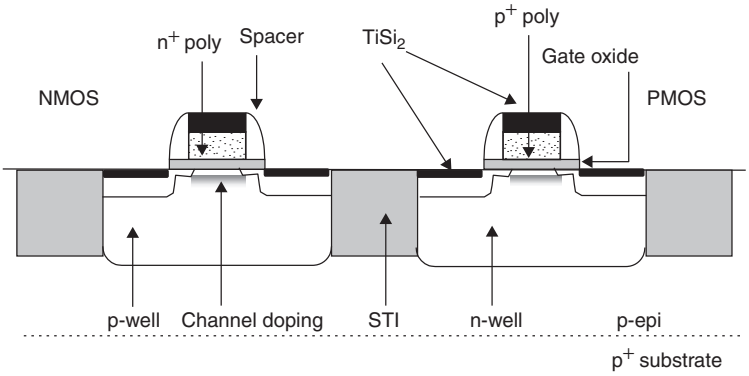


Figure 25.4 Deep sub-micron CMOS: 200 nm gate length, 5 nm gate oxide, 70 nm junction depth. n^+ poly for NMOS and p^+ poly for PMOS. Shallow trench isolation on epitaxial n^+/p^+ wafer

- tilted and halo implants for S/D engineering;
- RTA junction annealing;
- high-density plasmas for etching and deposition.

25.3.1 Wafer selection

CMOS process integration begins, like all other processes, with wafer selection (Table 25.6). Note that the tightening wafer specifications go hand in hand with wafer size via linewidth: 300 mm wafer specs are tighter because 0.13 μm linewidths are made on 300 mm wafers, whereas 0.5 μm to 0.8 μm is typical of 150 mm wafers, and 100 mm wafers are for linewidths above 1 μm.

25.3.2 Wells and isolation

Wells are the deepest diffusions in CMOS, and they must be fabricated early on in the process. There are several ways of making the wells dependent on initial wafer choice and device design requirements: n-well, p-well and twin-well processes are all possible.

The twin-well process requires two lithography steps but both NMOS and PMOS doping levels can be optimized independently. However, as we have seen in Figure 19.2, twin-wells can be made in a self-aligned fashion. Non-self-aligned twin-well structures, however, do not generate surface topography like self-aligned twin-wells.

LOCOS isolation has served CMOS fabrication for 30 years, and it has been scaled to much smaller linewidths than was previously thought possible. Below half-micron technologies, LOCOS was finally replaced: for one thing bird's beak lateral extent wastes area. Second, field oxide growth in narrow spaces is suppressed by compressive stresses, that is, the oxide does not grow to full thickness in narrow spaces. The main

Table 25.6 Wafer specifications for CMOS

Specification	100 mm	125 mm	150 mm	200 mm	300 mm
Thickness	525 ± 20	625 ± 20	675 ± 20	725 ± 20	775 ± 25
TTV (μm)	3	3	2	1.5	1
Warp (μm)	20–30	18–35	20–30	10–30	10–20
Flatness (μm)	<3	<2	<1	0.5–1	0.5–0.8
Oxygen (ppma)	20	17	15	14	12
OISF (cm^{-2})	100–200	100	<10	none	none
Particles (per wafer)	10 @ 0.3 μm	10 @ 0.3 μm	5–10 @ 0.3 μm 100 @ 0.2 μm	10–100 @ 0.16 μm 20–30 @ 0.2 μm	50–100 @ 0.12 μm 10–20 @ 0.16 μm 5–10 @ 0.20 μm
Metals ($atoms/cm^2$)	10^{12}	10^{11}	10^{11}	5×10^{10}	10^9

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

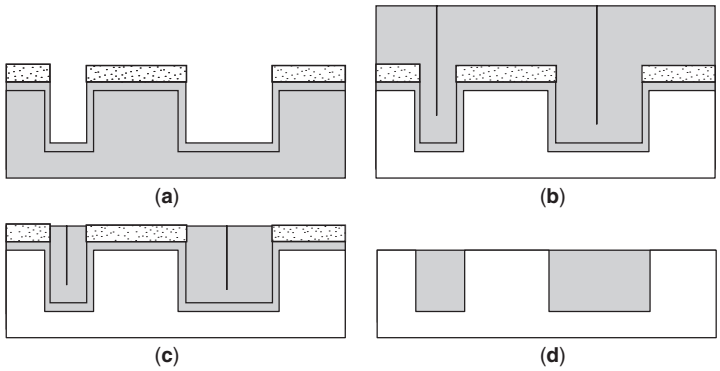


Figure 25.5 Shallow trench isolation, STI: (**a**) trench etching with a oxide/nitride stack followed by liner thermal oxidation; (**b**) CVD oxide deposition; (**c**) CMP polishing until nitride stop layer and (**d**) nitride and oxide etching

isolation method in the deep sub-micron technologies is STI. The process starts very much like recessed LOCOS, but then it takes advantage of CMP, which offers planarity of the final structure. A schematic STI process is described below.

Process flow for shallow trench isolation (STI)

pad oxide (thermal)
pad nitride (LPCVD)
lithography
etching nitride/oxide/silicon (isolation depth determined by etched silicon depth)
resist strip and cleaning
liner oxidation to form a high-quality silicon/oxide interface
CVD oxide deposition (trench overfilling)
CMP planarization of the oxide, polish stop at nitride
etch pad nitride
etch pad oxide.

Note that Figure 25.5 is drawn to scale in x, y and z.

- pad oxide 40 nm
- pad nitride 100 nm
- narrow trench width 250 nm
- trench depth 300 nm
- liner oxide 30 nm
- CVD oxide 500 nm

There are tens of variations of STI, but all of them have to fulfil certain common criteria. Overfill has to fill not only narrow trenches but also larger areas (of course, there can be a design rule limitation on trench widths). CMP planarization has also to be able to polish narrow and large areas at the same rate. If a large area polish rate is higher, planarization will only work for the narrow gaps. Instead of CMP, various etchback processes have also been tried, but they have pattern size and pattern density effects similar or worse than CMP, and the results are therefore no better.

25.4 GATE MODULE

Gate module is critical for transistor action. Gate oxide thickness, channel doping, gate length and source/drain doping profiles determine critical transistor parameters such as threshold voltage, switching speed, leakage current and noise. current and noise. The MOS gate module is very critical with respect to cleaning: as shown in Table 25.7 there are numerous contamination effects.

25.4.1 Gate oxide

Making thin gate oxides is a major wafer cleaning challenge: 100 nm particles are permissible in 0.35 μm technology from a linewidth point of view, but compared to <10 nm oxide thicknesses they are not allowed. Atomic contamination also becomes more crucial as film

Table 25.7 Metal contamination effects in MOS devices. Adapted from ref. Hattori

Metallic species	Contamination effects in MOS
Heavy metals (Cu, Fe, Ni)	Junction leakage current increase Lifetime degradation Oxide dielectric strength failure
Alkali metals (Na, K, Ca, ...)	Threshold voltage shift
Transition metals (Al)	Interface state increase
Noble metals (Au)	Lifetime degradation

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

thicknesses are scaled down. Metals and organics can be removed from the wafers by cleaning, but for very thin oxides, impurities in the gas phase also matter: residual water vapour at 20 ppm concentration level in the oxidation tube will dramatically enhance dry oxidation rate. Surface roughness also affects oxide electrical quality and channel mobility because in the MOS transistor, the current is confined to ca. 10 nm silicon layer underneath the gate oxide.

Silicon dioxide has a lower thickness limit of ca. 2 nm as a CMOS gate oxide because of leakage currents. One problem with ultra-thin gate oxides is boron penetration: boron from the p^+ polysilicon can diffuse through the gate oxide into the channel during thermal treatments and change channel doping, and therefore threshold voltage.

A number of methods and materials have been investigated as replacements for thermal oxide. Nitrided oxide (NO) and oxidation of nitrided oxide (ONO) are evolutionary developments based on thermal oxidation. New alternatives are deposited films, and this is a paradigm shift. Table 25.8 is also a chronological sequence of developments: amorphous and polycrystalline deposited oxides are expected to be the next materials to be implemented; and single-crystal oxides and very high-k materials are still further in the future.

Silicon dioxide is amorphous, and it stays amorphous through the high-temperature steps; single-crystal oxides would also be stable, but most amorphous oxides will crystallize and polycrystalline oxides will exhibit grain growth, both of which lead to problems. Front-end temperatures may have to be limited because of oxides, and not because of junction diffusion.

If, during deposition of the high dielectric–constant material, silicon dioxide is formed at the interface, the system that is formed is a SiO_2/high-ε two-layer structure, which must be analysed as capacitors in series. Interfacial silicon dioxide formation is difficult to avoid because high-ε dielectrics are oxides, and oxygen is present in some form or another during their deposition.

Table 25.8 Gate oxide materials

SiO_2	Thermal oxide, $\varepsilon \approx 4$
NO, ONO	Nitrided oxide, oxidized nitrided oxide, $\varepsilon \approx 6$
Al_2O_3, HfO_2, ZrO_2, Ta_2O_5	Amorphous and polycrystalline deposited oxides, $\varepsilon \approx 10–30$
$<Y_2O_3>$, $<La_2Hf_2O_7>$	Single crystalline deposited oxides, $\varepsilon \approx 10–30$
$Ba_xSr_{1-x}TiO_3$	Very high dielectric constant materials, $\varepsilon \approx 200$

Equivalent oxide thickness, EOT, is often used in describing high-ε dielectrics that replace silicon dioxide. Equivalent oxide thickness is given by

$$EOT = (\varepsilon_{SiO_2}/\varepsilon_{high}) \times t_{high-\varepsilon} + t_{SiO_2} \qquad (25.6)$$

where t_{SiO_2} is the interfacial silicon dioxide thickness, if any.

Zirconium oxide (ZrO_2, $\varepsilon \approx 23$) film of 6 nm thickness has EOT ≈1 nm, under the assumption of no interfacial SiO_2. Even a 1 nm SiO_2 layer will cause a drastic effect on *EOT*. Furthermore, dielectric constants of very thin films are different from bulk values or from values measured for thicker films (recall Figure 5.1). Note that we have used the classical capacitance formula above: in the 3 nm thickness range, a quantum mechanical description should be used for accurate results.

25.4.2 Self-aligned gate

The gate pattern is, together with contact holes, the most demanding lithographic and etching challenge of modern ICs. Gate linewidth scaling is a combined lithography and etching problem: feature size in the resist versus etched feature size. Etching is also related to gate oxide thickness: poly-gate etching has to stop on the thin gate oxide. The length of a gate level conductor is only a few microns, or tens of microns, and low resistivity is not a major requirement. Instead, ease of patterning and thermal stability in the contact with the oxide are primary concerns.

The self-aligned polygate was a major milestone in MOS evolution: source/drain diffusions were automatically aligned to the gate. But as transistor scaling continued, more complex doping patterns were called for. One motivation was to reduce hot electron effects: high electric fields in the channel accelerate electrons to high energies, and these electrons can degrade the gate oxide. In order to reduce these high electric fields, lightly doped drain (LDD) structure was introduced (Figure 25.6). In LDD, source/drain implantation is done in two steps.

After polygate etching, a self-aligned, low-energy, low-dose (ca. $10^{13}\,cm^{-2}$) implant is done, followed by CVD oxide deposition and spacer etching. This spacer shifts the second high dose S/D implant (ca. $5 \times 10^{15}\,cm^{-2}$) further away from gate edge, where the highest electric field occurs. This minimizes hot electron damage to thin gate oxide.

Process flow for LDD structure

implantation for source/drain extension ($10^{13}\,cm^{-2}$)
CVD oxide conformal deposition (thickness similar to junction depth)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

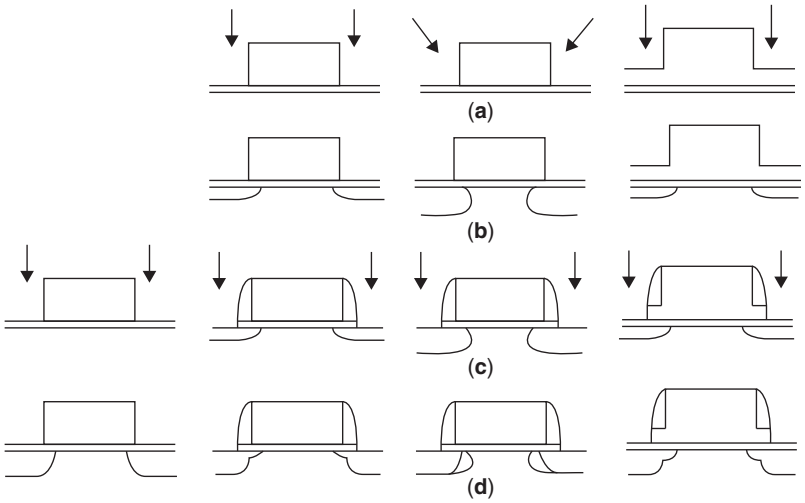


Figure 25.6 Gate-implant possibilities: (**a**) standard; (**b**) lightly doped drain LDD; (**c**) large-angle tilt device (LATID) and (**d**) inverse-T gate. Reproduced from Stinson, M. & Osburn, C.M. (1991), by permission of IEEE

anisotropic oxide plasma etch
etch damage removal/cleaning
implantation for source/drain. ($10^{15}\,cm^{-2}$)

Spacer etching end point is difficult to see because the most abundant material under spacer oxide is thermal oxide, and no selectivity is possible between two oxides. Some field oxide loss is therefore inevitable, and the spacer etch may etch some silicon in S/D areas.

In addition to junction depth, junction profile must be tailored more carefully in deep sub-micron CMOS. Large-angle tilted (halo) implants extend beneath the gate. Various double implant scenarios are depicted in Figure 25.6.

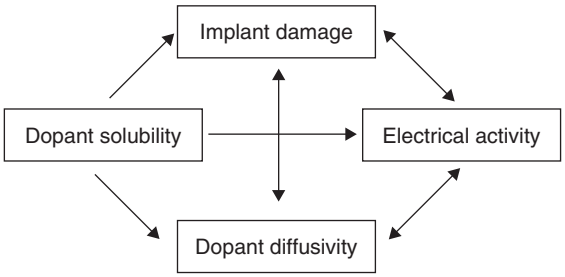


Figure 25.7 Implantation–diffusion interaction matrix. Redrawn from Jones, K.S., Extended defects in from ion implantation and annealing, in R.B. Fair (ed.): *Rapid Thermal Processing: Science and Technology*, Academic Press, 1993

25.4.3 Junction depth

Shallow junction formation is interplay between implantation and annealing. Junction quality means controllable and reproducible junction depth, low leakage current and good (ideal) forward characteristics. Low-sheet resistance requirement necessitates a high degree of electrical activation of dopants. Low leakage current requirement equals efficient damage removal and a low level of contamination. Solid solubility sets limits to activation and plays a role in damage dissolution (Figure 25.7). Clearly the demands are at odds with a typical damage annealing approach.

Point defects are essential for diffusion: vacancies created by the implantation process add to thermally generated vacancies and enhance diffusion. Boron diffusion is dependent on Si self-interstitials that are created, for instance, during thermal oxidation. Boron diffusion under oxidizing atmosphere is thus faster than in an inert atmosphere.

Activation refers to dopant atoms that become electrically active upon annealing. They then occupy lattice sites in the crystal and act as donors or acceptors. A high concentration of active dopants is needed for low resistance, especially at the surface because this affects contact resistance. Dopant atoms above the solid solubility limit do not contribute to electrical properties; they are as interstitial atoms or precipitates.

When two competing processes have different activation energies, we can favour one of the processes by a suitable selection of process conditions. For phosphorus

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

diffusion under normal low concentration conditions, the activation energy is 3.66 eV, but in ion implanted, damaged silicon it is 2.2 eV. Because rate is exponentially related to activation energy (Equation 1.1), dramatic changes in phosphorus diffusion take place. Point defects, interstitials and vacancies, created during implantation, offer fast diffusion paths. This is known as *transient enhanced diffusion* (TED). If defects can be annealed away rapidly, TED is eliminated and thermal diffusion determines doping profiles. Elimination of extended defects, such as dislocation loops, requires 1050 °C anneals.

Rapid thermal annealing (RTA) is a solution to this problem. A short time, high-temperature step (e.g., 1–10 s, 1000–1100 °C) is used to anneal implant damage. Thermal diffusion will be insignificant because the time is very short. Another anneal, at lower temperature but in longer time, will thermally diffuse dopants and activate them. RTA will be further discussed in Chapter 31.

25.4.4 Replacement gate

In order to implement materials that cannot withstand front-end high-temperature steps, dummy structures offer a solution. Replacement gate (dummy gate) of oxide or nitride serves in place of the metal gate during the high-temperature steps (Figure 25.8). After completion of S/D implant activation anneals, the first dielectric layer is deposited and planarized. The dummy gate is etched away, the gate dielectric is grown or deposited, and the final metal gate is deposited (followed by CMP). The replacement gate makes the return of the aluminium gate possible, but refractory metals are more likely candidates. The added process complexity is quite big, and oxidation/oxide deposition into the groove left by dummy gate etching is by no means easy or straightforward.

25.5 CONTACT TO SILICON

Scaling of contact size has rapidly led to problems with contact resistance. Contact resistance is given by Equation 24.1. If 0.4 μm contacts are made only at the bottom of the contact hole, resistance will be 10^{-7} ohm-cm^2/(0.4 μm)^2 = 63 ohm, compared with 16 ohm for 0.8 μm contacts. If, however, the whole source/drain area (1 μm × 1 μm) is silicided, silicon-to-silicide contact resistance will be 10^{-7} ohm-cm^2/1 × 10^{-8} cm^2 = 10 ohm. Metal-to-silicide contact area is 0.4 × 0.4 μm^2, so that will contribute only 1.25 ohm. Total contact resistance is thus only 11.25 ohm, compared with 63 ohm for non-silicided contacts. As shown in Figure 25.9, silicidation helps to increase packing density: signals buses can be routed over transistors if the S/D area is silicided, because then fewer contact holes are needed, saving area.

Contact hole etching–selectivity requirement is related to junction depth. If selectivity between oxide and silicon is poor, oxide etching might reach through the shallow junction. With better selectivity, etching will stop with minimal silicon loss. Etching selectivity of oxide against silicide is much higher than selectivity

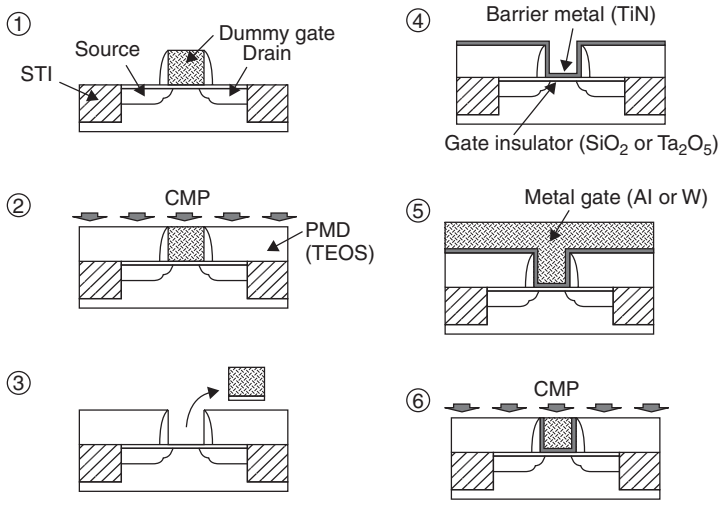


Figure 25.8 Replacement gate process. See text for discussion. Reproduced from Yagishita, A. *et al.* (2001), by permission of IEEE

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

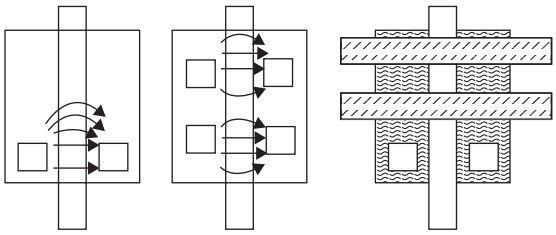


Figure 25.9 (**a**) MOS-transistor current paths in non-silicided contact; (**b**) current paths in multiple contact non-silicided contacts and (**c**) silicided contacts. In the case of silicided contacts, metal lines can run over the transistor, leaving greater freedom for signal routing. Adapted from Liu, R., Metallization, in C.Y. Chang & S.M. Sze (eds.) (1996), by permission of McGraw-Hill

against silicon, which also makes silicided contacts beneficial from the process integration point of view.

25.6 EXERCISES

1. Where in a CMOS would you find the following sheet resistances?

0.05 ohm/sq
0.5 ohm/sq
5 ohm/sq
50 ohm/sq
500 ohm/sq
5000 ohm/sq

2. Silicon dioxide forms readily during Ta_2O_5 deposition because oxygen is present in all oxide deposition processes. What is the effective capacitance of the SiO_2/Ta_2O_5 composite? Ta_2O_5:$\varepsilon = 25$, SiO_2:$\varepsilon = 4$.
3. EOT of 1.9 nm, 2.3 nm and 3.1 nm have been measured for 2 nm, 4 nm and 8 nm thick HfO_2 films, respectively. What is the interfacial SiO_2 thickness when HfO_2 dielectric constant is 20?
4. Design fabrication process for the power-MOSFET shown in Figure 1.6. The hatched structure is the gate oxide, and the source/drain/gate and the cross-hatched backside structures are metallizations.
5. Gate oxide thickness in 1 μm CMOS is 20 nm. On S/D areas, it is thinned during gate poly plasma etching, but re-grown during poly oxidation. Calculate the oxide thickness under the following assumptions:
   - poly etch rate is 250 nm/min;
   - poly thickness is 250 nm;
   - Si:SiO_2 etch selectivity is 20:1;
   - overetch time is 20 s;
   - re-oxidation is 900 °C, 10 min (dry).

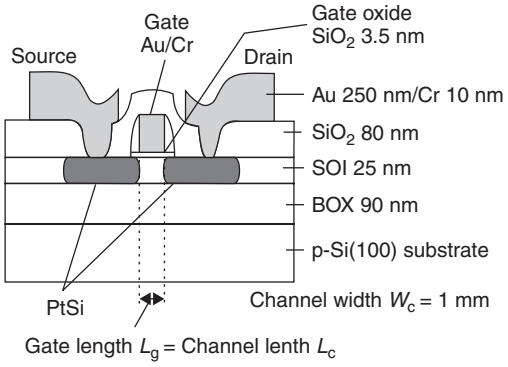
6S. Ion implantation of boron at 40 keV with dose 10^{13} cm^{-2} is done for CMOS p-well formation. The wafers are 4 ohm-cm phosphorus doped. Well depth (position of pn-junction) is designed to be 5 μm. What diffusion times/temperatures should be used?

7S. CMOS S/D implantation is made with arsenic (50 keV, 5×10^{15} cm^{-2}). Designed junction depth is 0.4 μm. Find implant activation conditions when 40 nm of dry oxide forms during activation.

8S. Shallow junctions are needed for advanced CMOS. Compare B-implanted p^+/n and As-implanted n^+/p shallow junctions (5×10^{15} cm^{-2} dose), when substrate doping level is 5×10^{17} cm^{-3}.

9S. Check with your simulator for sheet resistances, junction depths and film thicknesses of the 5 μm CMOS process described in the text. Make sure to select a proper cross section for your 1D simulation.

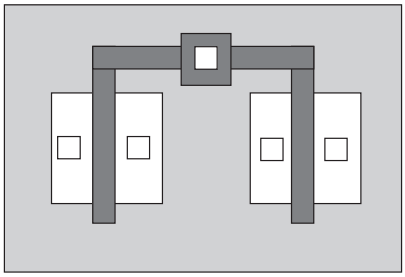
10. Plan a fabrication process for the gold-gate, PtSi S/D MOS-transistor shown below.



From Saitoh, W. *et al.* (1999), by permission of Institute of Pure and Applied Physics.

11. Compare the area of CMOS inverters made by two different lithography tools: (a) 8 μm resolution and 1 μm alignment and (b) 6 μm resolution and 2 μm alignment.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030



12. Compare minimum CMOS inverter area for:
    (a) non-self-aligned Al-gate
    (b) self-aligned polysilicon gate;
    keeping all other factors identical.
13. If NMOS and PMOS gates were fabricated from different metals (optimized for their respective devices), how many process steps would be added compared with n^+/p^+ dual gate (see Figure 25.4).

REFERENCES AND RELATED READINGS

Chesboro, D.G. *et al*: Overview of gate linewidth control in the manufacture of CMOS logic chips, *IBM J. Res. Dev.*, **39** (1995), 189.

Jones, K.S., Extended defects in from ion implantation and annealing, in R.B. Fair (ed.): *Rapid Thermal Processing: Science and Technology*, Academic Press, 1993.

Hori, T. & Sugano, T. (eds.): *Gate Dielectrics and MOS ULSIs: Principles, Technologies and Applications*, Springer, 1997.

Kahng, D.: A historical perspective on the development of MOS transistors and related devices, *IEEE TED*, **23** (1976), 655.

Liu, R., Metallization, in C.Y. Chang & S.M. Sze (eds.): *ULSI Technology*, McGraw-Hill, 1996, p. 400.

Saitoh, W. *et al*: 35 nm metal gate p-type metal oxide semiconductor field-effect transistor with PtSi Schottky source/drain on separation by implanted oxygen substrate, *Jpn. J. Appl. Phys.*, **38** (1999), L629–L631.

Stinson, M. & Osburn, C.M.: Effects of ion implantation on deep-submicrometer, drain-engineered MOSFET technologies, *IEEE TED*, **38** (1991), 487.

Wolf, S.: *Silicon Processing for the VLSI Era*, Vol 2 – Process Integration, Lattice Press, 1990.

Wolf, S.: *Silicon Processing for the VLSI Era*, Vol 3 – The Submicron MOSFET, Lattice Press, 1995.

Yagishita, A. *et al*: Improvement of threshold voltage deviation in damascene metal gate transistors, *IEEE TED*, **48**(8) (2001), 1604, Figure 25.1.

IBM J. Res. Dev., **43**(3) (1999): special issue on Ultrathin dielectric films.

**Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030**

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

26

Bipolar Technology

Both transistors and integrated circuits were initially made by bipolar technologies. The MOS transistor was conceived of and patented in the 1920s, well before the bipolar transistor (1947), but it was not realized until 1960. Bipolar transistors today are used in many specialty applications in which high speed, low noise or high current carrying capability is needed.

Bipolar transistors are traditionally fabricated on <111> because of epitaxial film growth reasons but there is no fundamental reason why they cannot be fabricated on <100> as well. In fact, BiCMOS circuits, which have both bipolar and MOS transistors, are fabricated on <100> wafers because the quality of thin oxide, the MOS gate oxide, is better on <100> orientation silicon. This has to do with the atom arrangement on the silicon surface and the resulting Si–O bonds and their spatial restrictions. Oxide is not a part of the active bipolar device; it has the role of sacrificial and passivation layer. Bipolar transistors are vertical devices, that is, currents are transported perpendicular to the wafer surface, whereas MOS transistors are lateral devices with currents parallel to the wafer surface. The standard buried collector (SBC) bipolar transistor is shown in Figure 26.1. It exemplifies the importance of epitaxy and diffusions in bipolar fabrication.

Bipolar transistor fabrication was already touched upon in Chapter 14, in which the UV photodiode process was described (Figure 14.3). A more detailed outline of the SBC process is given below. Before that, a short excursion to epitaxy on processed wafers is undertaken.

Buried layers are formed either by ion implantation or thermal diffusion. The oxide acts as a mask for thermal diffusion, but it is involved in the implanted process as well: during annealing, a thin thermal oxide is grown to prevent dopant outdiffusion. Before epitaxy, these oxides have to be removed. As a consequence, a step is formed on the wafer surface and this can cause pattern shift and distortion in the growing epitaxial layers (it can also cause growth defects if oxide removal is incomplete or if implant damage is not fully annealed). When the epitaxial-film growth from edges of a pattern is in the same direction, the pattern shifts laterally (Figure 26.2). If the pattern edges are not identical (recall <111> symmetries in Figure 21.19 to understand why rectangular structures on <111> must have different crystal planes at edges), structures can experience a shift in one direction and distortion in the direction orthogonal to the shift. In the extreme case, the epitaxial layer 'planarizes' patterns in what is known as a *wash-out*. Alignment problems will be encountered in all cases.

Buried layers are sources of dopants, and autodoping from buried layers must be considered. An isolated heavily doped region can dope areas many millimetres away in the downstream direction of the epi gas flow. When buried layers are tightly and uniformly spaced, autodoping non-uniformity is reduced, but the doping level change must be accounted for. Buried layers are heavily doped because their role is to minimize collector resistance, but heavy doping will change the lattice constant slightly, and there is a danger of misfit dislocations (as shown in Figure 6.2). Different epitaxial growth conditions (temperature, gases, pressure, reactor design) will result in different shifts, distortions and levels of autodoping.

26.1 FABRICATION PROCESS OF SBC BIPOLAR TRANSISTOR

There are many bipolar technologies but we will discuss a technology known as *standard buried collector* (SBC) bipolar technology, which has been widely used for decades. Even though current bipolars do not resemble it, they share many basic features with SBC.

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

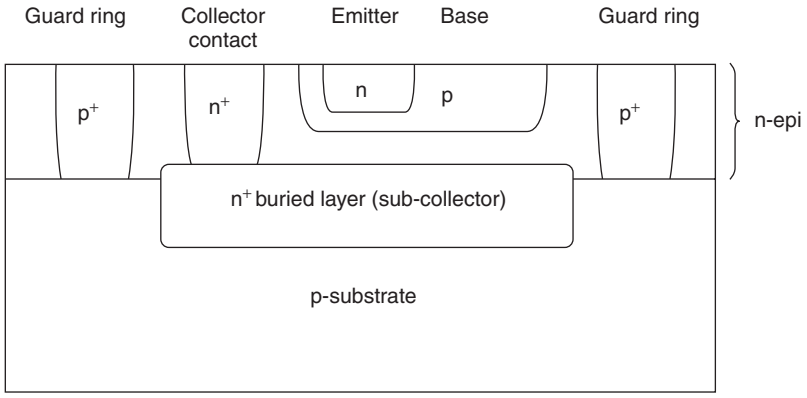


Figure 26.1 Standard buried collector (SBC) bipolar transistor: n-epitaxial layer on p-substrate (note that diffusions are not drawn to scale)

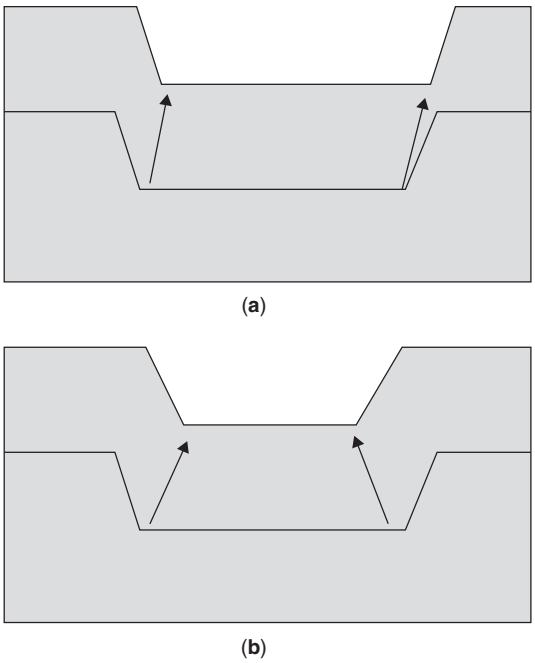


Figure 26.2 (**a**) Pattern shift and (**b**) distortion

The starting wafer is a lightly doped p-type wafer. Photomask 1 defines the area of the buried collector. The buried layer (sub-collector) is doped to a high concentration either by ion implantation or by furnace diffusion (Figure 26.3(a)). If implantation is done, the annealing step must be carried out for damage removal and recovery of a perfect silicon surface for epitaxy. Antimony is often used as the buried layer dopant because of its low vapour pressure, and consequently low evaporative losses during the subsequent epitaxial growth step.

Wafer cleaning after buried collector fabrication is crucially important for the success of epitaxy. A lightly doped epitaxial n-type layer is deposited on top of the sub-collector. Phosphine (PH_3) gas dopes the epilayer n-type during growth.

Photomask 2 defines the guard rings that isolate neighbouring collectors by reverse-biased pn-junctions. Guard rings are formed by boron–ion implantation or diffusion. Photomask 3 defines n^+ contact diffusion (known as *plug* or *sinker*). Phosphorus is implanted. Implantation depths are ca. 200 nm only, whereas epitaxial layer thickness can be up to 10 μm. Both p- and n-type dopants are driven to design depth by a thermal diffusion step at very high temperatures, up to 1200 °C. Deep diffusions must be done early in the process because they require the highest thermal load. A lot of silicon area is used for device isolation in SBC: the p^+ guard ring sideways diffusion distance is equal to the epitaxial layer thickness because diffusion is an isotropic process. The buried collector will experience up-diffusion to a thickness of a micrometre or two, depending on exact conditions during these diffusions.

Photomask 4 defines base areas. Ion implantation is used to introduce the dopants on the wafer because it offers better control of doping concentration. It is crucial to anneal away implant damage quickly so that the base width is controlled by thermal diffusion and not transient enhanced diffusion. It is customary to add to the process, an extra step that will ensure a shallow, high-doping area for good electrical contact to p-base.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

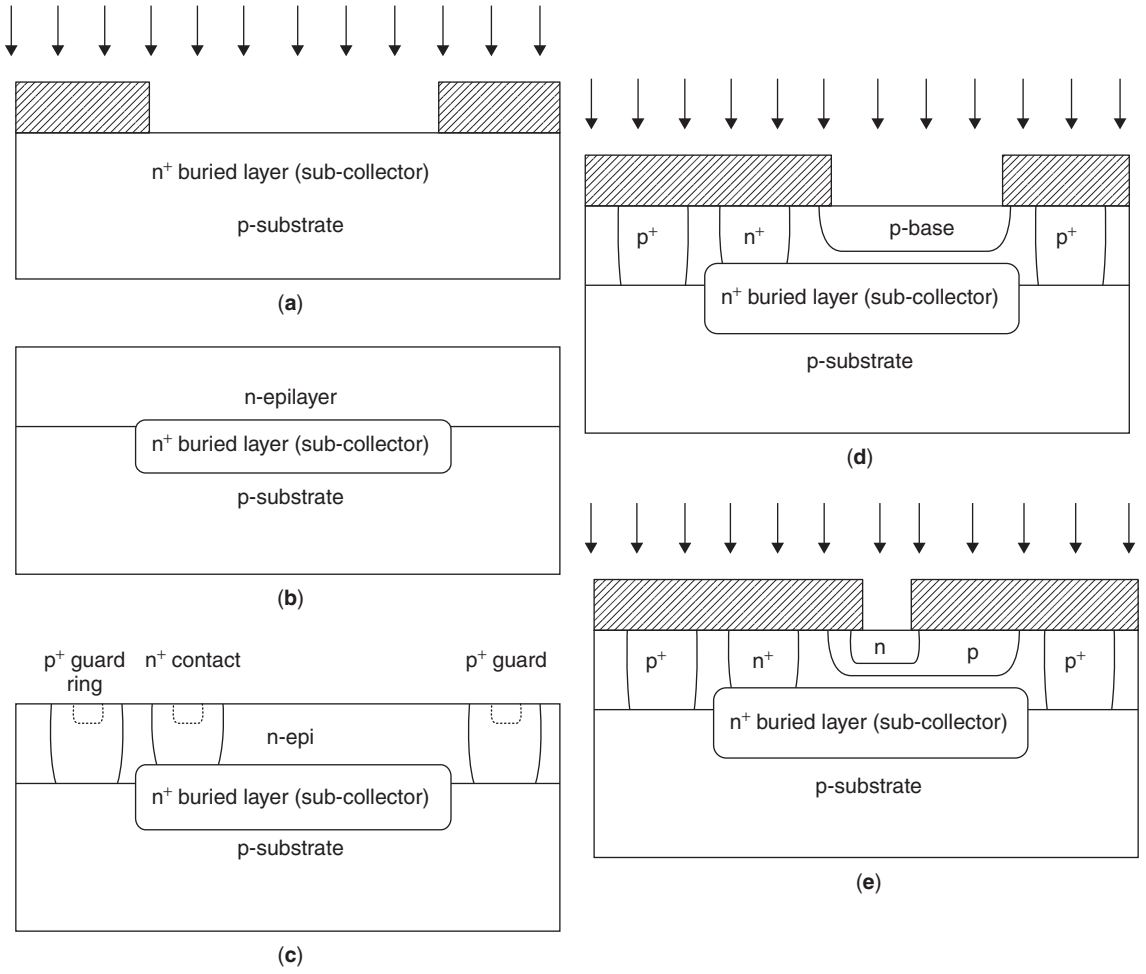


Figure 26.3 Bipolar fabrication steps: (**a**) Photomask 1: buried layer formation by antimony ion implantation; (**b**) growth of epitaxial phosphorous-doped n-type layer; (**c**) photomasks 2 & 3: p^+ guard ring and n^+ sub-collector contact diffusions: lateral spreading of diffusion is approximately equal to epilayer thickness; (**d**) photomask 4: ion implantation for base and (**e**) photomask 5: ion implantation for emitter

The emitter is defined by photomask 5. Emitter implantation and anneal are critical for device speed. Base transit time depends on base width, which is determined by both base and emitter diffusions (transistor speed depends on capacitive charging as well, not just on base transit time). Oxides that have served as diffusion masks are etched away and new thermal oxide is grown.

Contacts to diffusions are defined by photomask 6. Oxide etching is performed either by BHF or by plasma. After photoresist stripping and cleaning, aluminium is sputtered to provide electrical connections. Lithography step 7 defines aluminium wire patterns. After aluminium etching and photoresist stripping, PECVD oxide and/or nitride passivation layer is deposited. The last photomask (8) defines bonding-pad openings in the passivation layer. The wafer is now ready for testing.

Bipolar technologies have evolved over the decades with some familiar general trends: narrower linewidths, smaller vertical dimensions (shallower diffusion depths, thinner epitaxial layer thickness), smaller thermal budget and reduction of the area needed for device isolation. Table 26.1 lists three bipolar technology generations with their main structural features.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Table 26.1 Bipolar transistors, three generations/technologies

Layers (dopants)	Amplifying, junction isolated	Switching, junction isolated	Switching, oxide isolated
Substrate (B)			
Resistivity (ohm-cm)	10	10	5
Orientation	(111)	(111)	(111)
Buried layer (Sb/As)			
R_s (Ω/sq)	20	20	30
Up-diffusion (μm)	2.5	1.4	0.3
Epitaxial film (P)			
Thickness (μm)	10	3	1.2
Resistivity (ohm-cm)	1	0.3–0.8	0.3–0.8
Base (B)			
R_s (ohm/sq)	100	200	600
Diffusion depth (μm)	3.25	1.3	0.5
Emitter (P/As)			
R_s (ohm/sq)	5	12	30
Diffusion depth (μm)	2.5	0.8	0.25

Source: Adapted from Muller, R.S. & T.I. Kamins John Wiley, 1986.

26.2 ADVANCED BIPOLAR STRUCTURES

Bipolar transistor scaling is not as straightforward as in the case of CMOS. The number of transistors per chip is not the main driving force for bipolar technologies, but performance is. Two different aspects of bipolar scaling will be discussed shortly: vertical scaling, which concentrates on base and emitter structures; and lateral scaling, which is related to isolation between transistors. Vertical scaling is related to transistor speed via base transit time: smaller base width leads to faster operation. Lateral scaling is related to transistor speed too, because advanced isolation structures eliminate junction capacitances and allow faster switching. Despite all advanced structures, bipolar device packing density remains very low compared to CMOS.

26.2.1 Polyemitter bipolar transistor

To make a bipolar transistor faster, the base diffusion has to be made shallower. However, base width is determined by two diffusions: both base and emitter diffusion must be considered. A general strategy is to eliminate high-temperature steps. Using polysilicon as an emitter, less silicon is consumed in making the emitter. Dopants diffuse out of the heavily doped polysilicon emitter and reach just the topmost layer of single-crystal silicon, ensuring electrical continuity between polysilicon and single-crystal silicon. This approach has a number of benefits: the single-crystal silicon emitter will not be implanted, and therefore defects from implantation and transient-enhanced diffusion are eliminated. Elimination of implant annealing reduces high-temperature steps and unwanted base diffusion. The polyemitter also eliminates the danger of aluminum spiking: if the emitter is very thin, aluminium might spike through it, destroying the device (recall Figure 7.6(e)). Polysilicon, for example 200 nm thick, between aluminium and the emitter/base junction eliminates the aluminium-spiking problem.

26.2.2 Self-aligned polyemitter bipolar transistor

Bipolar transistor fabrication can utilize the same self-alignment principles as CMOS. One of the many self-aligned polysilicon emitter processes is presented in Figure 26.4. It employs self-alignment to the maximum, with three implants self-aligned to each other. In addition to being a self-aligned transistor, it is also a polyemitter transistor.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

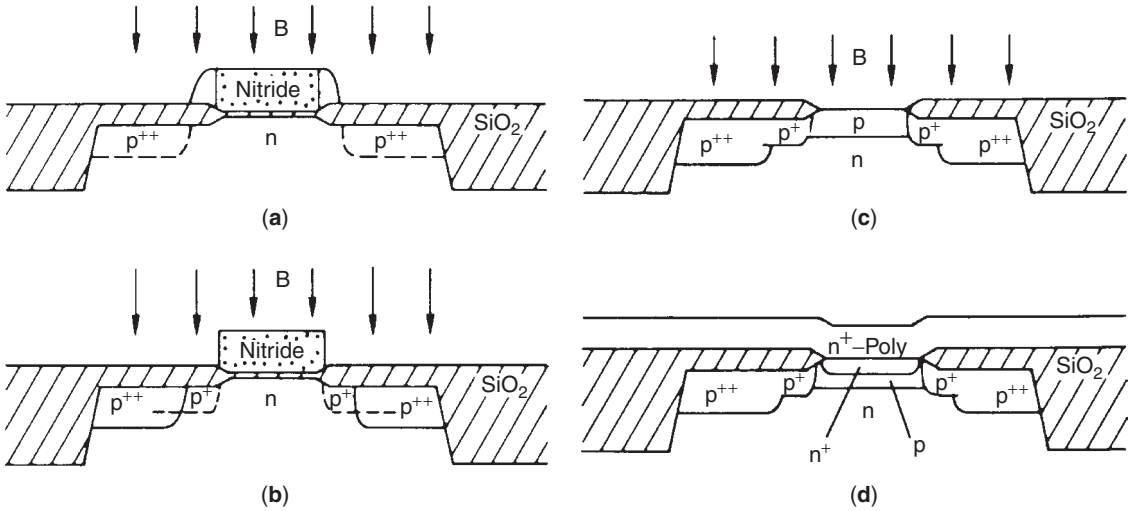


Figure 26.4 Self-aligned single poly bipolar transistor. Reproduced from Chen, T.-C. *et al*. (1988), by permission of IEEE

The thick (600 nm) recessed LOCOS isolation oxide is made first. A thin pad oxide (10 nm) is grown, followed by 75 nm LPCVD nitride. After nitride etching, a second LOCOS oxide is grown, this time 200 nm thick. LOCOS nitride is not removed after field oxidation. Instead, polysilicon spacers are formed on nitride by conformal LPCVD poly deposition and anisotropic etching in chlorine plasma. Boron implantation is carried out to form heavily doped external base (p^{++}), with energy high enough to penetrate the 200 nm thick LOCOS oxide. Polysilicon spacers are etched away, with high selectivity against oxide and nitride. Another boron implantation forms a link (p^+) between external and intrinsic base. The p^+ and p^{++} areas are self-aligned to each other like the source/drain and source/drain extension in an LDD MOS. Nitride is etched away in CF_4 plasma, selectively against oxide. The oxide beneath the nitride protects single-crystal silicon from being etched by fluorine. The oxide is then removed selectively against silicon in HF. The oxide also has, of course, a role as a stress relief layer in LOCOS structure. The third boron implantation forms the shallow active base. Because it is done last, it experiences the least thermal load and consequently the least diffusion. LPCVD polysilicon is deposited for the emitter. It is doped by phosphorous ion implantation. Anneal is required to drive out n-type dopant from the polysilicon emitter into single-crystalline silicon. The emitter reaches into the single-crystal silicon only to a depth of a few tens of nanometres.

26.2.3 Self-aligned double poly bipolar transistor

Phosphorous-doped polysilicon can act as a diffusion source for the emitter, and correspondingly boron-doped poly can act as a doping source for the p-base. This double-poly process (Figure 26.5) offers a different self-alignment scheme from the previous example.

Process flow for self-aligned double poly bipolar transistor

base link poly deposition (undoped)
base link poly doping by boron
CVD oxide-1 deposition
lithography
etching of CVD oxide/base link poly stack
base link diffusion (p^+)
boron implantation (pre-deposition)
intrinsic base diffusion
CVD oxide-2 deposition
oxide spacer etching
emitter poly deposition, *in situ* phosphorous doping
emitter outdiffusion.

The base link doping level is independent of the intrinsic base doping. The base link has to be in electrical contact with the intrinsic base, and the diffusion depth must be similar to the spacer width. CVD oxide is needed on top of the link poly because it will insulate the base link poly and the emitter poly later on. This, of course, adds a little complexity to the etching because a double layer

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

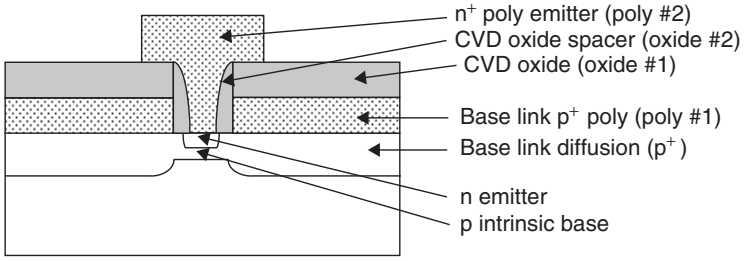


Figure 26.5 Self-aligned double poly bipolar (see text for details)

structure has to be etched. Etching of the base poly leads to some loss of the underlying single-crystal silicon too, but the intrinsic base has not yet been made so its depth is not affected. CVD oxide deposition determines the distance between the link base and the intrinsic base non-lithographically, in a self-aligned manner. The emitter will be automatically aligned to the base, too. Intrinsic base implant dose, energy and annealing are optimized irrespective of link base properties. Emitter poly is doped *in situ* in order to reduce thermal budget: poly LPCVD temperature is ca. 600 °C, as against the ca. 950 °C required for poly doping by thermal diffusion or implantation annealing.

26.2.4 Lateral scaling

In a standard buried collector, bipolar devices are isolated from each other by guard-ring diffusions (Figure 26.1). The diffusion depth has to be equal to the epilayer thickness, and guard rings take up a lot of area. LOCOS isolation, shown in Figure 26.3, becomes possible when epilayer thicknesses become similar to

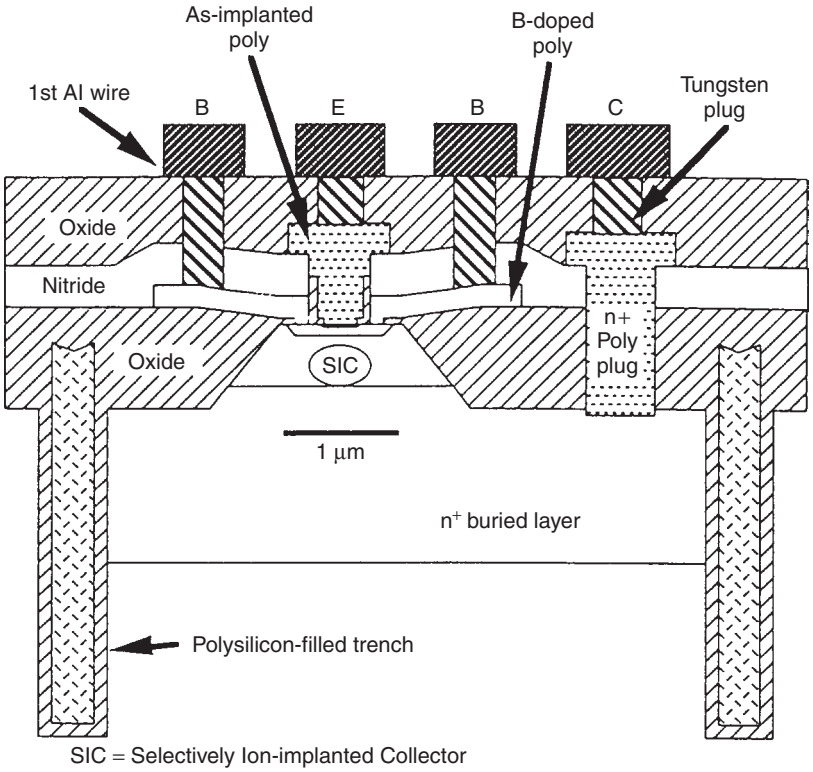


Figure 26.6 Trench isolated bipolar. Reproduced from Ugajin, M. (1995), by permission of IEEE

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

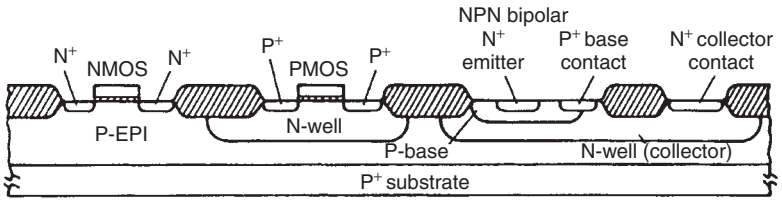


Figure 26.7 Simple BiCMOS technology: triple diffused-type bipolar transistor added to a CMOS-process with minimal extra steps: only p-base diffusion mask is added to CMOS process flow. Reproduced from Alvarez, A.R. (ed.) (1989), by permission of Kluwer

thermal oxide thicknesses. Oxide isolation improves not only area usage but also transistor speed because sidewall capacitances are minimized.

Trench isolation, which is even more area efficient than LOCOS, is used for high-performance bipolars. In bipolar technology, deep trenches of 5 μm are typical, in contrast to CMOS isolation where shallow trenches (ca. 0.3 μm) are used (Figure 25.5). Area usage for isolation becomes independent of epilayer thickness, limited only by lithography and trench etching. Trench filling (Figure 20.7) is usually done in two steps: a thin liner is grown/deposited first, followed by the filling material. For instance, thermal oxidation forms the liner, and TEOS or undoped polysilicon is used to fill up the trench. One variant of many trench-isolated bipolar transistors is shown in Figure 26.6. It makes use of four polysilicon layers: for trench filling, link base doping and emitter and buried layer contact plugs. Some of these layers can be used for resistor structures in analog devices.

26.3 BiCMOS TECHNOLOGY

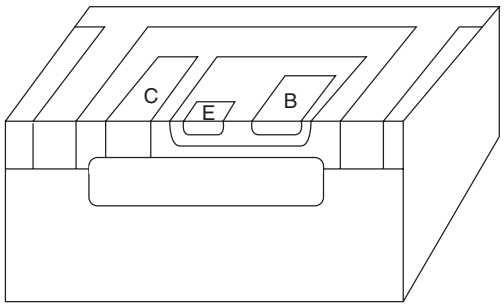
BiCMOS tries to combine the best of both bipolar and CMOS: high speed, low noise and high current-carrying capacity of the former with the integration density and low power consumption of the latter.

BiCMOS has been approached from both directions: taking a full-blooded bipolar process and adding CMOS to that, or taking CMOS as a starting point and adding process modules to create bipolar transistors. The latter approaches are more prevalent but they often fail to take advantage of the best features of bipolars. Unfortunately, the cost would rise too much if all the features of both processes were combined; some performance trade-off has to be accepted. In the BiCMOS shown in Figure 26.7, the n^+ doping step is used to form both NMOS source/drain areas and bipolar emitters and collector contacts; and similarly, the p^+ doping step creates both PMOS S/D and the bipolar base contact. Only the p-base diffusion step is needed in addition to the standard CMOS steps. The elimination of buried layer and epitaxy leads to increased collector resistance and lower operating frequency for bipolars, but the fabrication process is greatly simplified.

As a rule of thumb, the cost is directly related to the number of photolithography steps. The evolution of a 13-photomask, 1 μm CMOS process into a 1 μm BiCMOS process can be done in several ways. In its simplest form, only a base implant photomask is added. If true bipolar performance is needed, buried layer and epitaxy are needed and the collector is made separately from n-well. If analog elements such as resistors are required, the mask count still increases, but this is true for both CMOS and bipolar alike. Analog and high-performance BiCMOS are therefore ca. 20 to 30% more expensive than either pure CMOS or bipolar of the same linewidth.

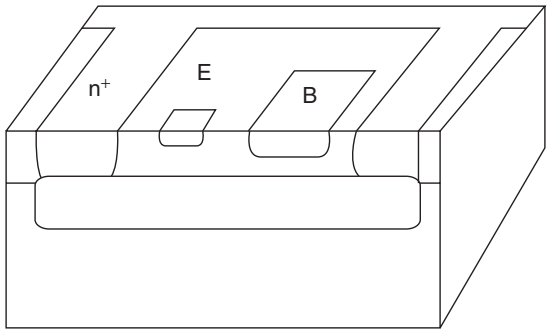
26.4 EXERCISES

1. SBC is pictured below. Calculate the minimum transistor area under the following assumptions:
   - the minimum lithographic linewidth L is 3 μm, and it is the width of E, C and B;
   - the emitter is square; the base length is 2 × width and the collector length is 3 × width;
   - the epilayer thickness is 5 μm;
   - the buried layer up-diffusion is 1 μm;
   - the base diffusion depth is 1.5 μm;
   - the emitter diffusion depth is 0.5 μm.

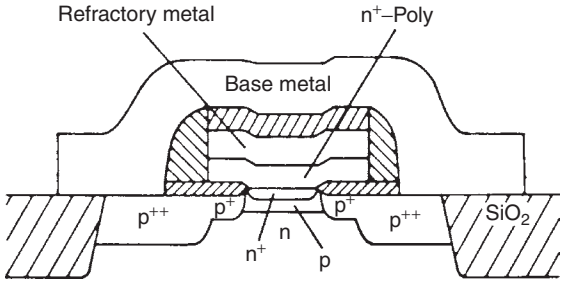


Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

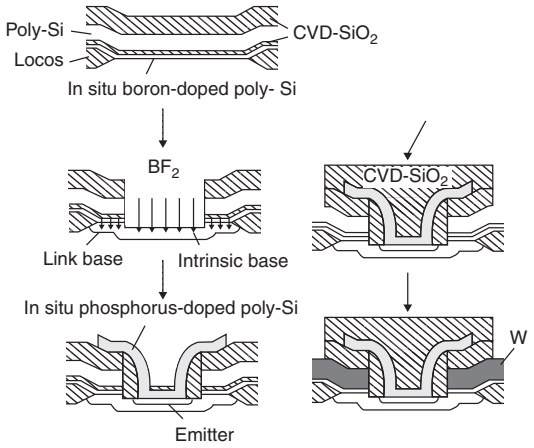
2. What will be the minimum transistor area if the p^+ guard ring isolation of an SBC transistor is replaced by a deep trench isolation?
3. What is the area of a collector diffusion isolation (CDI) transistor when the same baseline process described above is used?



4. Perform the front-end simulations to obtain sheet resistances and diffusion depths of switching for the junction-isolated transistor described in Table 26.1.
5. Design metallization process steps for the polyemitter transistor. This is the same device as shown in Figure 26.4. From Chen, T.-C. *et al*. (1988), by permission of IEEE.



6. Analyse the main fabrication steps of the bipolar transistor shown below. From Onai, T. *et al*. (1997), by permission of IEEE.



REFERENCES AND RELATED READINGS

Alvarez, A.R.: (ed.): *BiCMOS Technology*, Kluwer, 1989.

Chen, T.-C. *et al*: An advanced bipolar transistor with self-aligned ion-implanted base and W/poly emitter, *IEEE TED*, **35** (1988), 1322, Figure 26.1

Muller, R.S. & T.I. Kamins: *Device Electronics for Integrated Circuits*, John Wiley, 1986.

Onai, T. *et al*: 12 ps ECL using low-base-resistance Si bipolar transistor by self-aligned meta/IDP technology, *IEEE TED*, **44** (1997), 2207–2212, Figure 26.2

Reisch, M.: *High-frequency Bipolar Transistors*, Springer, 2003.

Ugajin, M.: Very-high f_t and f_{max} silicon bipolar transistors using ultra-high performance super self-aligned process technology for low energy and ultra-high-speed LSI's, *IEDM*, 1995, p. 735.

Wolf, S.: *Processing for the VLSI Era: Volume 2 – Process Integration*, Lattice Press, 1990.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

27

Multilevel Metallization

Multiple levels of metallization offer possibilities for circuit designers to route signals over transistors, and thus to reduce the area needed for wiring. Multilevel metallization structures for submicron technologies (0.8/0.5/0.35/0.25 μm) are based on aluminium with two process technology innovations: contact and via filling with plugs of tungsten CVD and oxide planarization by CMP (Figure 27.1). Copper metallization emerged in the late 1990s, and more recently low dielectric constant materials (low-k) have been introduced. These are completely new materials, driven by CMOS-metallization time delay concerns.

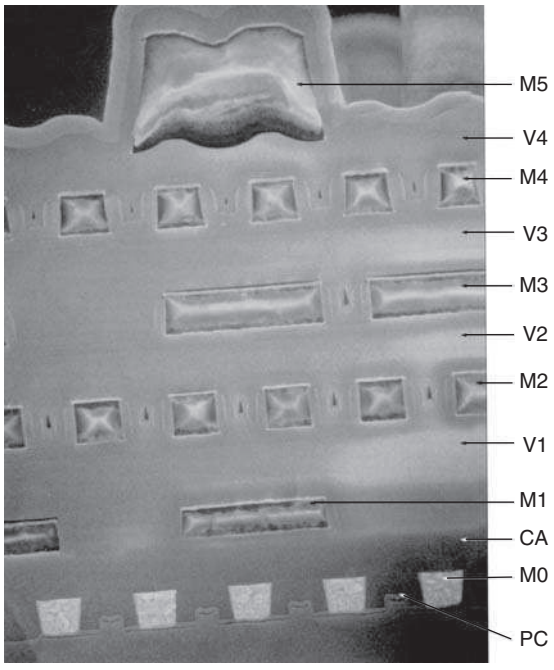


Figure 27.1 Cross-sectional view of six level metal structures (M0 is metal zero). Reproduced from Koburger, C.W. *et al*. (1995), by permission of IBM

27.1 TWO-LEVEL METALLIZATION

Two-level metallizations are extensions of one-level metallizations (see Figure 25.2(i)), with additional dielectric and metal films and only minor conceptual differences. The process continues after first metal as follows:

Process flow for two-level metallization

intermetal dielectric	PECVD oxide
planarization	SOG etchback
via holes	oxide plasma etch
second metal deposition	TiW/Al sputtering
metal etching	Cl_2-based plasma
passivation	PECVD nitride
bonding pad open	CF_4-plasma etch

There are a number of practical aspects in two-level metal processes that demand attention. Each additional (PE)CVD step adds to thermal loads, causes stresses and plasma damage. Silicon/metal interface stability needs to be rechecked and barrier re-evaluated. Stresses from additional layers can cause hillock growth and crack propagation, which must be checked. Hillock sizes are amenable to optical microscope inspection, but electrical data from short/continuity test structures will provide more quantitative data on this and other metallization issues. Second metal step coverage in the

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

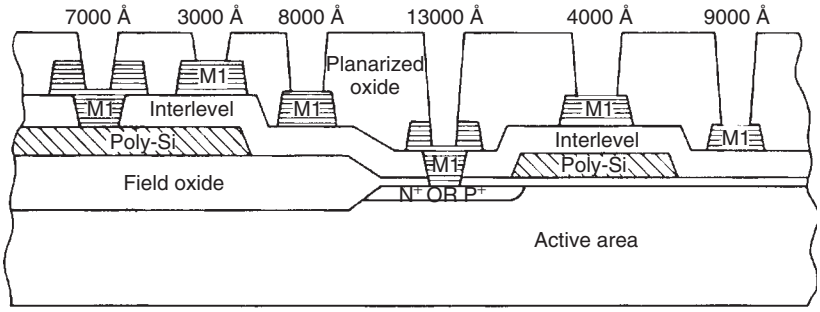


Figure 27.2 Via-depth problem due to planarization. Reproduced from Brown, D. (1986), by permission of IEEE

via hole is often critical. Fortunately, via holes are larger than contact holes, and aspect ratios are therefore smaller (but they need not be, if intermetal dielectric thickness is greater than interpoly dielectric). Via hole etching is similar to contact-hole etching, but the etching needs to be stopped at the top of the first metal, and selectivity between oxide and aluminium is much higher than selectivity between oxide and silicon. However, because there is metal on the wafer, cleaning solutions after via etching are limited.

Two-level metallization cannot be extended to three levels because topography of the wafer gets more pronounced after each level, and gap-filling capability of (PE)CVD dielectric deposition as well as sputtering step coverage in via holes will hit the limits. Planarization helps, but it is no panacea: the surface may become flat, which eliminates optical lithography depth-of-focus problems but, as shown in Figure 27.2, creates problems in via-hole etching and sputtering because holes will be of different depths.

27.2 MULTILEVEL METALLIZATION

True multilevel metallization starts at three levels of metal. Historically, this occurred in the late 1980s when submicron CMOS technologies were introduced. In 0.25 μm technology, up to six levels of metal are used in ASICs and logic chips and three levels in memory chips. It is expected that in 65 nm technology generation, there can be ten levels of metal.

A fully planar structure can be created when contact and via holes are filled by CVD tungsten, and excess tungsten is removed, by etchback or by CMP (Figure 20.7). The number of metal levels can be increased simply by repeating the process over and over again because the topography does not change (Figures 16.1 and 27.3).

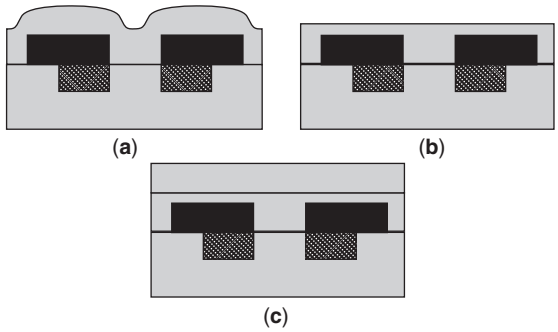


Figure 27.3 Oxide-CMP planarization: (**a**) (PE)CVD oxide fills the gap between aluminium lines; (**b**) blind polishing of oxide (no end point) and (**c**) second CVD oxide deposition

Backend process integration differs from front end in the sense that thermal budget concept has a very different meaning. Whereas front-end thermal budget is about temperature-diffusion relationship, backend thermal budget is about temperature-stress relation. For n-level metallization there will be $2n$ steps at 300 to 400 °C (one CVD tungsten and one PECVD dielectric deposition for each layer), with room temperature steps (etching, spin coating, CMP) in between. Stress, strain, adhesion, hillocks, voids and cracks have to be understood.

27.2.1 Contact/via plug

In order to get planarized metallization, CVD W-plug fill has been adopted (see Figure 20.7). There are many possible routes to achieve the same final structure, and they are pictured in Figure 27.4. Both selective tungsten CVD and contact-hole filling with sputtered aluminium would be advantageous from the process simplicity point of view, but they have proven

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

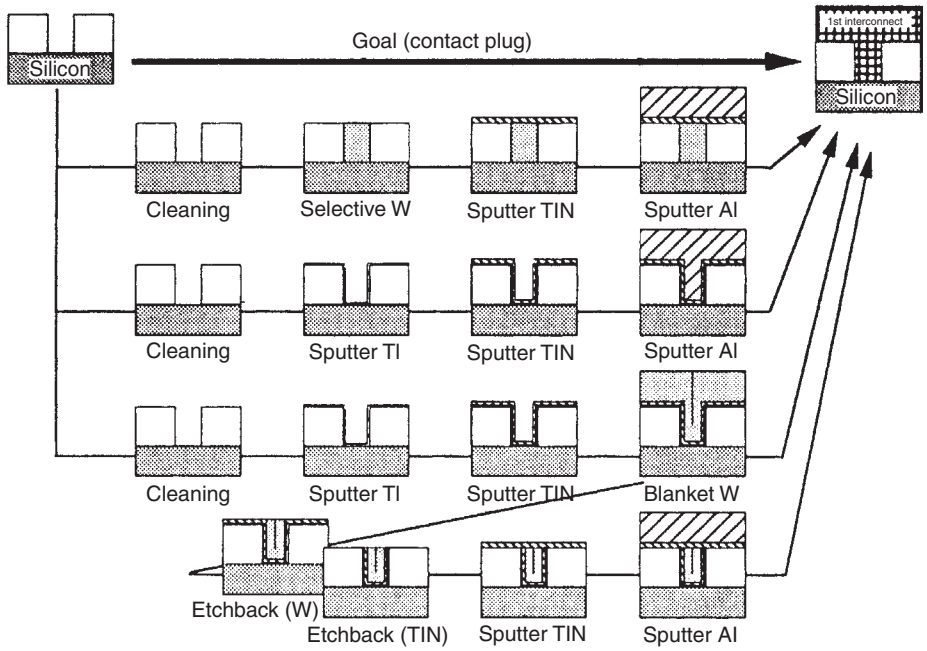


Figure 27.4 Three different routes to Ti/TiN/W/Al contact plug fill. Reproduced from Ohba, T. (1992), by permission of Materials Research Soc

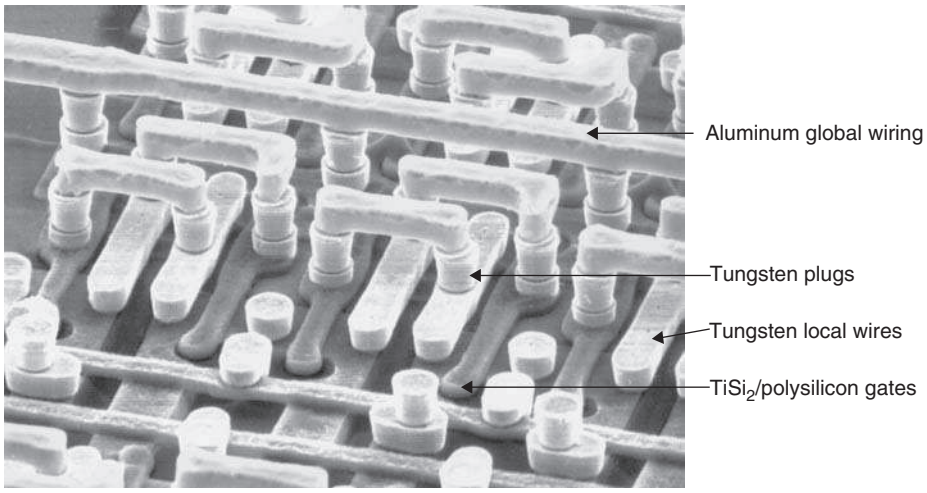


Figure 27.5 Tilted top-view scanning electron micrograph (SEM) of planarized multilevel metallization: all dielectric layers have been etched away to reveal the metal levels. Reproduced from Mann, R.W. *et al*. (1995), by permission of IBM

to be difficult in principle and practice. The blanket tungsten/etchback route has been the most widely adopted one.

The SEM micrograph of Figure 27.5 shows the structure of a planarized multilevel metallization scheme. The top aluminium wiring levels are very planar. Tungsten has been used for local interconnects (in the length scale $\sim 10\,\mu m$). All dielectric layers have been etched away to reveal the metallization for analysis (for example for failure analysis).

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

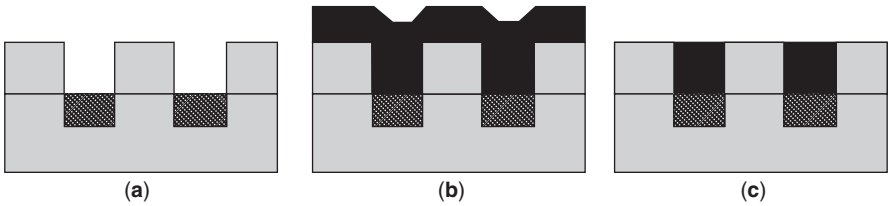


Figure 27.6 Damascene process: (**a**) trenches etched in oxide till underlying metal; (**b**) metal overplating into oxide trenches and (**c**) metal CMP

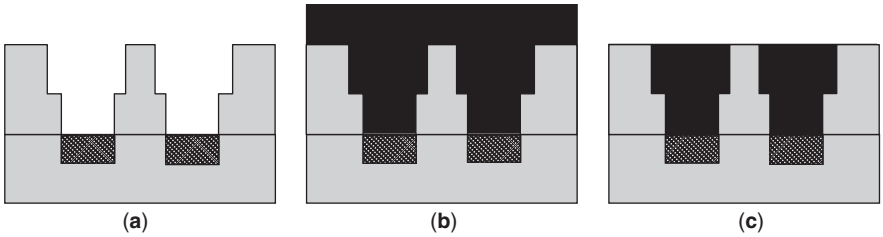


Figure 27.7 Dual damascene metallization: (**a**) two lithography and two etching steps define vias and wires in oxide; (**b**) vias and wire trenches filled by metal in one deposition step and (**c**) metal polishing to yield a planar surface

27.2.2 Stacked vias

When vias can be stacked on top of each other in a multilevel metallization scheme, a lot of area can be saved and freedom of wire routing increases. In Chapter 24, sputtering step coverage was found to be poor for stacked vias (Figure 24.12), but with W-plugs and planarization, stacking becomes natural. In Figure 27.5, tungsten plugs can be seen on top of each other. Misalignment is still there, but because the surfaces are planar, misalignment does not lead to topography build-up.

27.3 DAMASCENE METALLIZATION

Damascene metallization (Figure 27.6) relies on etching trenches in oxide, filling those trenches with metal, and CMP for removal of excess metal. As we have seen in Figure 16.1, this will result in a structure identical to the one made by metal deposition, metal etching and oxide planarization. Oxide etching, which is easy, and copper CMP, which is difficult, are used in damascene. Because copper etching is practically impossible, copper metallization must be implemented in damascene.

The CMP can provide globally planar surface, but if the original topography is not amenable to global planarity, CMP cannot help. If the deposition process leaves voids (Figure 7.17), these can emerge as crevasses after the CMP. This poses reliability problems as residues from processing can accumulate in these pockets. It must be remembered that even though CMP can planarize, the sixth level can never be as smooth as the first level.

27.3.1 Dual damascene

One of the advantages of damascene metallization is its ability to offer even more ingenious multilevel metal fabrication routes. Dual damascene process (Figure 27.7) combines via filling and wire metal deposition into one integrated process step.

In practice, it has been difficult to decide the order of process steps: how should lithography and etching of vias and wire trenches actually be combined for maximum benefit. Dual damascene promises great reductions in the number of process steps, but it is not an easy process. Dual damascene discussion continues in connection with copper/low-k materials towards the end of this chapter.

27.4 METALLIZATION SCALING

In CMOS front-end scaling, vertical parameters: junction depth x_j and oxide thickness t_{ox} are scaled to smaller and smaller values, leading to improved transistor performance. In the backend, however, vertical scaling is detrimental. If metal lines are made thinner, resistance increases and linewidth scaling works in the same

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

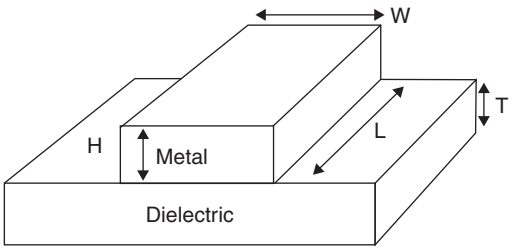


Figure 27.8 Wire geometry for simple RC-time delay model

direction. If the dielectric thickness is scaled down, capacitance between metal layers increases, leading to increased RC-time delays. At 1 μm linewidths, transistor delays are more significant than wiring delays, but the situation changes somewhere around 0.2 μm technology, and below 100 nm wiring delay clearly dominates over transistor delays.

A simple model (Figure 27.8) for backend interconnect wire scaling gives RC-time delay as

$$\tau = RCL^2 \qquad C = \varepsilon WL/T \qquad R = \rho L/HW \tag{27.1}$$

where L is line length and resistance R and capacitance C are per unit length.

Scaled local connection lengths are given by L/n ($n > 1$) because smaller devices are closer to each other. Long distance connections do not scale, however, because chips are not getting any smaller, quite the contrary, in fact, because more and more functions are crammed on a chip. In our simple model, we will assume a constant line length, L. Scaled capacitance and resistance are given by

$$C' = \varepsilon(W/n)L/(T/n) = C \tag{27.2}$$

$$R' = \rho L/(H/n)(W/n) = n^2 R \tag{27.3}$$

RC-time delay τ' is then given by

$$\tau' = R'C' = n^2 RC \tag{27.4}$$

Because scaling factor n is larger than unity, time delays are increasing. When linewidths are scaled down, film thicknesses are scaled down in order to keep aspect ratios about the same (Table 27.1), which is not an unreasonable assumption since very tall but narrow metal lines would be difficult to make. Because chip sizes (L) are increasing, time delays are bound to increase. Historically, RC-time delay has increased 26% per generation.

In order to battle RC-time delay, aluminium ($\rho \approx 3$ μohm-cm) has been replaced by copper ($\rho \approx 1.8$ μohm-cm) and silicon dioxide dielectrics ($\varepsilon \approx 4$) have been replaced by low-k dielectrics ($1 < \varepsilon < 4$).

Table 27.1 Backend scaling trends

CMOS generation	0.35 μm	0.25 μm	0.18 μm	0.13 μm
Min. metal linewidth/μm	0.4	0.3	0.22	0.15
Min. space/μm	0.6	0.45	0.33	0.25
Metal thickness/μm	0.7	0.6	0.4	0.4
Dielectric thickness/μm	1	0.84	0.70	0.6

27.5 COPPER METALLIZATION

All ICs used aluminium for metallization till 1997, and most still do, but copper has been introduced into high-performance applications from 0.25 μm generation on. Resistance reduction is advantageous but copper has many drawbacks and limitations (Table 27.2). Copper diffuses rapidly in both silicon and silicon oxides, and new barrier materials have to be invented: tantalum and its compounds and alloys are prime candidates. Copper has to be chemical–mechanical polished, so CMP is a must. Whereas aluminium deposition is always by sputtering and tungsten is by CVD, there are a number of copper deposition methods available: electroless, electroplating, CVD and sputtering. Sputtering is ruled out because of poor step coverage and inability to fill holes, but it can still be used to deposit a thin seed layer for electrodeposition. Both CVD and electrodeposition methods can fill high-aspect ratios encountered in deep submicron devices.

In aluminium/tungsten metallization, barriers are needed between metals but in copper metallization barriers are required for dielectrics as well (it is of course possible to develop new dielectric materials that would be stable in contact with copper, but currently copper needs to be clad from all four sides, see Figure 27.9).

Table 27.2 Issues in copper metallization

- Adhesion to dielectric
- Diffusion in (and reaction with) dielectric
- Compatibility with tungsten contact plug
- Deposition of seed layer
- Deposition of copper
- Contamination on the chip
- Contamination in the equipment

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

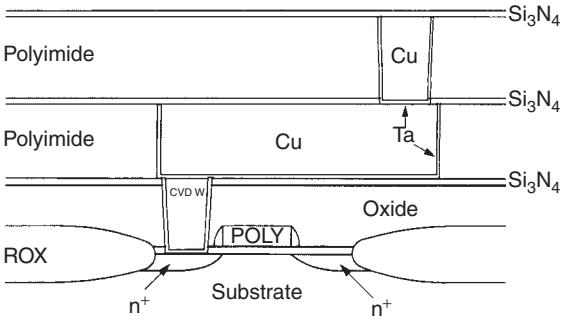


Figure 27.9 Cu/polyimide multilevel metallization with Ta-barriers, W-plugs and silicon nitride polish-stop layers. Reproduced from Small, M.B. & Pearson, D.J. (1990), by permission of IBM

Silicon nitride (PECVD) is stable in contact with copper but nitride has a fairly high dielectric constant (ca. 7), which is disadvantageous for RC-delays. Double layers of low-k material with nitride barrier can be used. Nitride and carbide (PECVD SiC) serve other functions, too: they act as polish-stop layers for CMP, and protect low-k materials that are polished at fairly high rates.

Metallic barriers are thin: below 100nm for 1 μm technology, and thinner for each subsequent generation. For 0.18 μm technology barriers need to be 10 to 20nm; that is, barrier thickness needs to be scaled down because conductor thickness is scaled down. Resistivity of the barrier and plug are not big issues for micron-sized contacts, but they are becoming critical for 0.18 μm technology because the full benefit of the low resistivity of copper cannot be realized if the high-resistivity barrier reduces effective resistivity of the plug.

Copper/polyimide metallization with tantalum barriers and nitride etch-stop layers is shown in Figure 27.9. Copper is completely clad by either tantalum or nitride. Contact with silicon is made by Ti/TiN/W-plug, even in cases where all other levels of metal are copper.

CMP selectivity between copper and tantalum is very high, which means that removal of tantalum leads to long overpolish times (cf. long overetch times). CMP non-idealities dishing and erosion have to be analysed. Dishing is strongly linewidth dependent, but rather insensitive to pattern density, whereas oxide erosion is very strongly pattern density dependent and only mildly linewidth dependent, as shown in Figure 27.10. CMP dishing and erosion in the 20 nm range are targeted for 100 nm technologies. Erosion and copper thinning can somewhat be compensated by using thicker starting layers, but this is a cost issue.

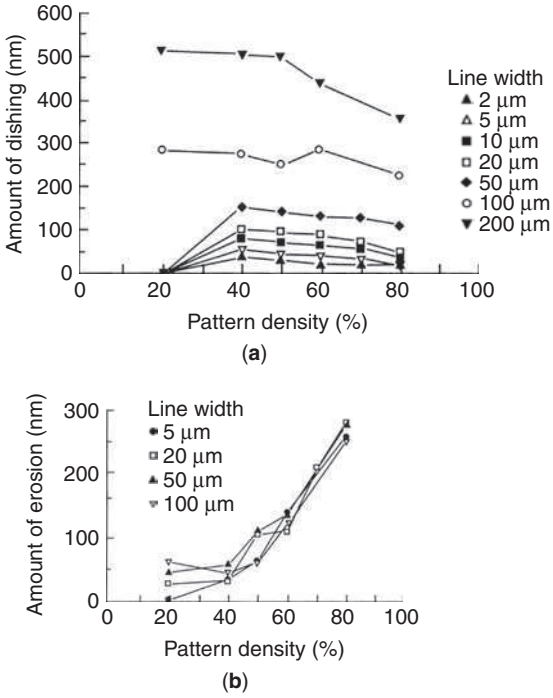


Figure 27.10 Dishing of copper and erosion of oxide. Source: Steigerwald J. M., *et al*, Chemical–Mechanical Planarization of Microelectronic Materials, © Wiley, 1997. This material is used by permission of John Wiley & Sons, Inc

27.6 LOW-K DIELECTRICS

Dielectric constant can be reduced by modifying oxides or by switching to other materials. With SiO_2-based glasses (with $\varepsilon \approx 4$) there is an evolutionary development down to ca. $\varepsilon \approx 2.7$. The first approach is to deposit fluorine-doped oxide by CVD. This will lead down to $\varepsilon \approx 3.6$. Carbon doping, with CH_3-groups in silicon dioxide, designated as SiOC:H, can bring dielectric constant down to ca. 2.7. Composition of SiOC:H films is typically 20 to 25% Si, 30 to 40% O, 15% C, and 20 to 40% hydrogen. These films are well-known, dense, inorganic materials, compatible with existing CVD tools, processes and metrology.

Siloxanes and silsesquioxanes are familiar materials from spin-on planarization, with methyl silsesquioxane (MSQ) ε as low as ≈2.6. In spin-film planarization, the spin-film is most often etched away, but it can be used as a permanent part of the device. This leads to whole new characterization of siloxanes. For instance, during subsequent sputtering step, outgassing from SODs can poison the metal, leading to contact problems.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Switch to polymers is a discontinuous shift: it requires a lot of work in materials science, process technology, metrology, process integration, equipment and reliability. For instance, adhesion and interface stability with metals need to be assessed and etching and polishing processes have to be developed. Sufficient mechanical strength of low-k films is essential for successful CMP. Fluoropolymers, aromatic hydrocarbons, poly (arylene ethers), parylene and PTFE offer dielectric constants down to ≈2.

The next step is to go for porous materials, with $\varepsilon \approx 2$ (also known as ULKs, for ultra-low k). Pores can be made by controlled evaporation, nanophase separation or drying. Aerogels and xerogels, dried silica with 90% air in it, promise further improvements in ε.

The ultimate dielectric is air (or vacuum) with $\varepsilon \approx 1$. There are some practical problems with air, however: mechanical strength is not very good, thermal conductivity is poor and long- term stability is questionable. In spite of these drawbacks, gas-filled and vacuum dielectric structures have been demonstrated.

A wide repertoire of measurements is needed to characterize novel candidate materials (Table 27.3). PECVD boron nitride was measured for some 15 properties (see Table 7.2). New polymeric low-k materials need to be measured for 15 more parameters before they can be accepted in manufacturing.

Modulated photoreflectance methods, already in use in implant-dose monitoring, are useful for multilayer analysis when time-resolved mode is employed. A short laser pulse heats the sample, which then expands locally, giving off sound waves. Reflectivity is modulated by the propagating sound waves, and this can be measured by a probe laser. Time-resolved measurement can distinguish between reflections from various interfaces in the sample, enabling multilayer measurement of both metals and dielectrics. Optical measurements are fast, and amenable to wafer mapping, yielding uniformity maps.

CMP of soft and porous materials with Young's moduli of 1 to 10 GPa is difficult because they are mechanically weak. They are also subject to peeling by shear forces, especially when multiple layers of materials are present (and there can be tens of layers in a multilevel structure). Polymeric abrasives have been tried as replacements of silica and alumina for soft material polishing. Cleaning remains a major problem for low-k materials – post-CMP cleaning, post-etch cleaning and photoresist strip. Many wet chemical cleaning solutions are out of the question because they penetrate pores and cause swelling. Measurement of pore size and porosity is needed for reproducibility of ultra-low k materials. Various methods are being developed: candidates include gas phase, optical, X-ray, positron and neutron methods.

Table 27.3 Characterization needs for new dielectrics

Parameter	Comment
– CMP rate	– Young's modulus 1–10 GPa, high polish rates
– T_g / T_d	– Glass transition and decomposition temperatures (ca. 450 °C)
– Plasma resistance	– Organic materials are etched in oxygen plasma
– Cleaning resistance	– Photoresist removers and solvents
– Shrinkage	– Volume changes upon heat treatment as solvents evaporate
– Adhesion	– Scotch tape test is the first hurdle
– Outgassing	– Even cured films may release gases into sputtering vacuum
– Porosity	– Tightly controlled for reproducible ε
– Pore size	– Oversized pores behave like pinholes
– Shelf life	– Decomposition during storage not unlike photoresists
– Viscosity	– Film thickness depends on viscosity (and spinspeed)
– Impurities	– (Alkali) metals have to be measured
– CTE	– Polymeric materials have a wide range of expansion coefficients
– Loss tangent	– Electrical losses at high frequencies must be understood

When new materials are introduced, they are evaluated in several phases. Initial tests are carried out on planar wafers using blanket films. Basic physical and chemical characteristics are measured: dielectric constant, shrinkage, moisture absorption, uniformity of deposition, blanket etching and polishing. Single-level test structures are then applied to check patterning issues (etch, strip) and interface stability under various process steps (metallization, CMP, etch). Multilevel test structures include electrical tests and more complex interaction tests such as etch and polish stop, adhesion during CMP, and so on.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

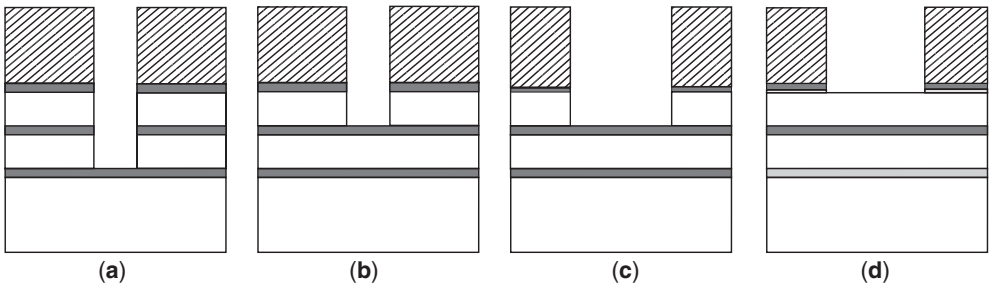


Figure 27.11 Four possible dual damascene processes with etch-stop layers: (**a**) full via first; (**b**) partial via first; (**c**) wire first and (**d**) partial wire first

While thermal oxide serves as a reference material when CVD oxides are evaluated, PECVD oxides serve as references when low-k materials are developed. Leakage current between neighboring lines, interline capacitance, breakdown field between copper lines, metal continuity, metal bridging and line resistance uniformity are compared to oxide reference processes.

Dual damascene copper/low-k dielectric combination introduces novel process integration features: hard mask layers (barriers) that protect (organic) low-k material and act as etch-stop and polish-stop layers. Insulator structure is then either barrier/low-k/barrier (shown in Figure 27.9) or barrier/low-k/barrier/low-k/barrier (shown in Figure 27.11). Order of dual damascene process steps is not clear-cut, and the alternatives are discussed below.

Full via first (Figure 27.11(a)) is problematic because very deep, high- aspect ratio via hole is produced in the first step, making second photoresist spinning difficult. Additionally, the bottom hard mask needs to tolerate two etch steps: it is exposed in the end of the via etch and all the time during trench (wire) etch. One solution is to protect the bottom of a via with undeveloped resist during the second etch step.

In partial via first approach (Figure 27.11(b)), via holes are etched till the mid etch-stop layer in the first step. Wire trench etching is easier than in full-via-first approach. Misalignment can cause a grave error in this structure: if the wire trench is misaligned so much that via is partially photoresist covered, the area of metal contact will be small and erratic.

Wire trenches first (Figure 27.11(c)) approach does not need a top hard mask. Wires are etched down to the middle hard mask. Next, lithography has to be done in a recess, and lithography depth-of-focus may pose problems.

The partial wire trench first approach (Figure 27.11(d)) needs a top hard mask. In the first step, the top hard mask is etched and resist is then stripped. The next lithography step (for via) can now be done on a practically planar surface. After etching the top low-k layer with resist mask, resist is stripped, and the wire trench and the bottom half of the via are etched using hard mask only. Misalignment in the via-lithography step can cause problems similar to 'partial via first' described above.

In the era of 5 μm CMOS, the front-end contributed most of the process steps and most of the cost of processing. Today the backend dominates both the number of steps as well as costs. Back end is also beginning to dominate the time delays of advanced circuits, which means that the backend issues will remain important in the foreseeable future.

27.7 EXERCISES

1. If a 2:1 aspect ratio via plug in 0.25 μm technology has a resistance of 0.4 Ω, is it made of tungsten or copper?
2. What is copper plug resistance in 0.1 μm technology?
3. What is the breakdown field requirement for low-k dielectrics?
4. What is the effective dielectric constant of nitride/BCB/nitride (20 nm/500 nm/20 nm) stack when $\varepsilon = 7$ and 2.5, respectively?
5. What is the etch or polish selectivity needed in a low-k approach that uses 20 nm thick nitride etch/polish-stop layers on 300 nm low-k material?
6. What were the etching processes used to prepare the sample for SEM Figure 27.5? What are the selectivities and other criteria required for those etching processes?
7. Does the simple RC-time delay model described in the next fit with the historical RC-time delay trend of 26% per generation? Use data from Table 27.1.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

REFERENCES AND RELATED READINGS

Anand, M.B. *et al*: Use of gas as low-k interlayer dielectric in LSI's: demonstration of feasibility, *IEEE TED*, **44** (1997), 1965.

Brown, D.: Trends in advanced process technology, *Proc. IEEE*, **74** (1986), 1678 (special issue on integrated circuit technologies of the future).

Chen, W.-C. *et al*: Chemical mechanical polishing of low-dielectric constant polymers: hydrogen silsesquioxane and methyl silsesquioxane, *J. Electrochem. Soc.*, **146** (1999), 3004.

Davis, J.A. *et al*: Interconnect limits on gigascale integration (GSI) in the 21st century, *Proc. IEEE*, **89** (2001), 305 (special issue on limits of semiconductor technology).

Ho, P.S., Lee, W.W. & Leu, J.: *Low Dielectric Constant Materials for IC Applications*, Springer-Verlag, 2002.

Hsu, H.-H. *et al*: Electroless copper deposition for ultralarge-scale integration, *J. Electrochem. Soc.*, **148** (2001), C47.

Koburger, C.W. *et al*: A half-micron CMOS logic generation, *IBM J. Res. Dev.*, **39** (1995), 215.

Mann, R.W. *et al*: Silicides and local interconnections for high performance VLSI applications, *IBM J. Res. Dev.*, **39** (1995), 403.

Murarka, S.P.: *Metallization, Theory and Practice for VLSI and ULSI*, Butterworth-Heinemann, 1993.

Ohba, T.: Multilevel metallization trends in Japan, *Proc. ULSI-VII* (1992), MRS.

Rao, G.K.: *Multilevel Interconnect Technology*, McGraw-Hill, 1993.

Small, M.B. & Pearson, D.J.: On-chip wiring for VLSI, *IBM J. Res. Dev.*, **34** (1990), 858.

Steigerwald, J.M., Murarka, S.P. & Gutman, R.J.: *Chemical Mechanical Planarization of Microelectronic Materials*, John Wiley & Sons, 1997.

Wrschka, P. *et al*: Chemical mechanical planarization of copper damascene structures, *J. Electrochem. Soc.*, **147** (2000), 706.

**Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030**

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

28

MEMS Process Integration

MEMS devices come in a bewildering variety, with regard to structures, materials and functions. Whereas all CMOS technologies are close relatives, MEMS devices are made with a multitude of related, distantly related and unrelated technologies. Pressure sensor operation can be based on piezoresistive, capacitive, thermal conductance or resonance mechanisms; and the first three share some structural features and fabrication steps whereas the fourth bears more resemblance to gyroscopes and RF oscillators.

Identical DRIE fabrication steps are utilized in making microfluidic valves, variable optical attenuators, accelerometers and enzyme microreactors. Anisotropic wet etching is similarly used for a plethora of applications that have nothing in common at the device level, even though they share some of the crucial fabrication steps.

MEMS technologies require new materials: nickel as mechanical material, copper as thick electroplated metal, platinum as chemically inert electrode in microfluidics, palladium as catalyst, gold as low-resistivity metallization, SnO_2 as gas sensitive film, zinc oxide as piezoelectric material, PZT as ferroelectric material, VO_2 as strong temperature coefficient of resistivity material, and the list goes on. Some of these are known materials from other applications: gold is routinely used in GaAs microwave circuits, polyimide films are well-known materials in chip packaging and the printed circuit board industry, and Teflon® coating is widely used in frying pans, but many are new in microdevices or in thin-film form.

MEMS structures have high aspect ratios and highly complex 3D shapes resulting from DRIE or from anisotropic wet etching and wafer bonding. These put new requirements for subsequent lithography, doping and thin-film steps, and introduce novel metrology requirements. The fact that MEMS devices have through-wafer holes limits some process steps: for instance, spinning of resist over holes is out of the question and unconventional patterning approaches are needed. Through-wafer structures require double-sided processing of the wafer, and even without through-holes, there is often a need to align structures on the two sides of the wafer. Double-side alignment is also mandatory for structured wafer bonding.

MEMS devices are not 'solid-state devices' in the sense that they are not solid throughout but have free-standing, moving, rotating, vibrating and sliding parts with air gaps or vacuum cavities. These create additional topology challenges for the following process and packaging steps. Capillary forces in drying, silicon dust and vibrations during dicing or stresses and temperature in encapsulation may damage delicate mechanical structures. Cavities can sometimes be handled without problems, but high temperatures and changing pressures during fabrication can cause some design limitations, especially when the cavity roof is a thin diaphragm.

28.1 DOUBLE-SIDE PROCESSING

Although intricate three-dimensional topography can build up on the wafer surface by etching and deposition processes, utilization of both sides of the wafer leads large-scale 3D structures that pose special problems of their own. Processing must be tailored so that both sides of the wafer are under controlled conditions at all times. Double-side processing is intricately intertwined with process equipment, which has historically been designed for top surface processing only, and therefore processes on wafer backside have been neglected and they depend heavily on particular equipment designs.

Three kinds of processes take place on the wafer backside:

- patterning;
- blanket processing (doping, growth and deposition);
- unintentional processes.

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Many processes take place on all surfaces in the reactor. The films or doping structures on the wafer backside are often of poor quality because most processes are optimized for the front side alone. If single-side polished wafers are used, backside roughness prevents proper film growth. Sometimes, backside films result from front-side processing spillovers: the photoresist covers the wafer edge erratically and some resist is deposited on the wafer backside; or alternatively, material from the wafer chuck or transport system adheres to the wafer back.

Blanket processing involves growth and deposition of films either simultaneously or in sequence on both sides. Thermal diffusion can be done either way, with an oxide film to prevent diffusion on the protected side. Ion-implantation doping is inherently one-sided. Applications of blanket processing include doping for backside metallization for power devices, contact resistance minimization, etch mask formation and gettering treatment (polysilicon film deposition, ion implantation or damage creation).

Some fabrication processes are inherently one-sided, some double-sided, and for yet others the distinction depends on equipment design. All beam-like processes are one-sided: lithography, implantation, evaporation and sputtering. Most thermal processes, such as oxidation, diffusion and anneal, are double-sided (Table 28.1). Wet chemical etch and clean processes are also double-sided. CVD, PECVD and plasma etching processes can be either one-sided or double-sided: if wafers are loaded upright in a wafer boat (Figure 28.1), deposition/etching takes place on both surfaces, but if wafers are loaded flat, or clamped, on an electrode, only the top side is processed, with some unintentional spill-over over the edge. In CVD processes, the backside can be protected to some extent by placing the wafers in the reactor back-to-back: reactant flow is then minimized and unwanted deposition is eliminated. This is of course only a partial solution; some deposition will take place.

Table 28.1 Double-sided and single-sided processes

Double-sided	Single-sided
Furnaces, oxidation	Sputtering
Furnaces, CVD	Evaporation/MBE
Furnaces, PECVD	Ion implantation
Furnaces, diffusion	PECVD
Furnaces, annealing	Epitaxy
Wet etching and cleaning in a tank	CMP
Spray processing	Plasma etching
Resist stripping in barrel plasma	Spin processing
Resist stripping in wet solutions	Lithography

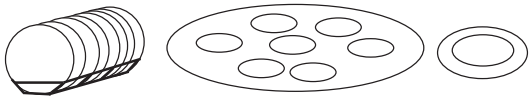


Figure 28.1 A batch of wafers upright in a jig; wafers flat on electrodes

In most equipment, inserting the wafers into the reactor upside down is allowed, but potential damage to the patterns on the front by transport mechanisms, clamping or chucking must be considered. Temperature allowing, photoresist is a quick fix that protects the front side. Sometimes, a film that was deposited on both sides is first patterned on the back, while the front side is under cover.

28.1.1 Double-side polished wafers

In single-side polished (SSP) wafers, the backside is rough with micrometre peak-to-valley heights. Both sides of double-side polished wafers are mirror polished to sub-nanometre RMS roughness. However, the side that was polished last is of better quality than the other side, and double-side polished (DSP) wafers are therefore not fully symmetric. This has implications especially for bonding, which is critically dependent on roughness and flatness.

Wafer thickness refers to centre-point thickness. It is difficult to produce precise thickness specifications because some wafering steps are batch processes for many wafers at a time and some are single-wafer steps; therefore, variations are inevitable. Wafer thicknesses are compromises between material usage and mechanical strength. Mechanical strength is especially important in high-temperature steps as many mechanical properties (for instance yield strength) are strongly temperature dependent. MEMS devices that extend through the whole wafer require exacting thickness control. In crystal plane–dependent wet etching, the 54.7° slanted sidewalls waste area in proportion to wafer thickness, and in plasma etching, thick wafers lead to longer etch times.

Standard wafer thicknesses range from 380 to 770 μm, but 4 to 1500 μm are available. Mechanical stability increases with thickness, and thickness has to increase with wafer size (Table 28.2), therefore extremely thin wafers are limited to small wafer sizes, but handling problems limit their usability. Through-wafer MEMS has not been done on 300 mm so far, and 200 mm is on the fringe, too.

Total thickness variation (TTV) of IC wafers is not of great concern, and 1 to 5 μm is acceptable, but in MEMS, through-wafer etched structures' TTV is of paramount importance. If 10 μm thick beams or diaphragms need to be fabricated, 1 μm TTV results in 10% variation (and possibly much larger variation in device properties,

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Table 28.2 Standard wafer sizes and thicknesses

Wafer diameter	Thickness	Comments
3 in.	380 μm	
100 mm	525 μm	380 μm for MEMS; thinner wafers exist
150 mm	625 μm	380 μm for MEMS; 250 μm minimum
200 mm	725 μm	500 μm minimum
300 mm	770 μm	

which may depend on the square or cube of the thickness). MEMS-wafer TTV values of 1 μm are typical, and 0.5 μm is specified for the most demanding applications.

Double-side polished wafers were first introduced for silicon bulk micromechanics. Double-side lithography, through-wafer etching and anodic bonding were not possible with standard single-side polished wafers. More recently, advanced IC fabrication processes have introduced DSP wafers for twofold reasons: TTV of DSP wafers is less, which relieves the lithography focus budget somewhat. Process cleanliness is also improved because the polished backside minimizes the surface area, which reduces contamination.

28.1.2 Double-sided growth, doping and deposition

Thermal oxidation oxidizes both sides of the wafer, which may or may not be advantageous. Oxide on the backside can be a useful protective layer, for example, to prevent diffusion in the next step. LPCVD nitride masking can be used to protect either side, as in the LOCOS process.

Diffusion from the gas phase will dope both sides of the wafer. Again, oxide or nitride films can prevent unwanted diffusion. Doping by implantation and from thin film sources (e.g., PSG or BSG) are single-sided processes.

Epitaxy presents a special case of backside effects on the front side: if a lightly doped epilayer is grown on a highly doped substrate wafer, evaporated dopant from the substrate will mingle with the source gases and affect epilayer doping. Therefore, CVD oxide is used as a backside-capping layer to prevent dopant outdiffusion from the substrate.

For integrated circuits, backside diffusion is not a problem because diffusion depths are ca. 1% of wafer thickness at maximum and therefore backside diffusions will not interfere with the top surface devices. For volume devices such as power transistors or solar cells, the backside is an active part of the device, and diffusions on the backside are essential for device operation.

Rather thick stacks of films can build up on the wafer backside. Stresses in such film stacks can cause flaking and rupture, which generates particles. Another problem is wafer curvature due to film stresses. For these reasons, backside films are sometimes removed even though no device reason would necessitate it.

28.1.3 Double-side lithography

Double-side lithography comes with three degrees of difficulty:

- arrays without alignment;
- non-critical alignment;
- critical alignment.

Regular array structures on the wafer backside without alignment to the front include, for example, solar-cell back surface field diffusion (Figure 1.6). In non-critical alignment, the major function of the device is determined by structures on one side only, and the coarse auxiliary structures are made on the other side. These include the opening of optical paths and fluidic connections (see Figures 11.14 and 22.11(a)), or the removal of silicon mass for thermal insulation.

Critical alignment involves device functions that are highly dependent on the accuracy of pattern location, for example, symmetric resonating mass or positioning of piezoresistors to the point of maximum deflection of a pressure sensor diaphragm.

Double-side lithography is done on one side at a time: resist application on top, alignment and exposure on top and development, rinsing and drying on top. Then, depending on the device structure, either etching of the front-side or backside lithography is performed.

Backside lithography involves backside resist application, which means that the front side of the wafer is placed in vacuum contact with the spinner chuck. The front side must be protected. Photoresist is often used but it cannot be used for patterning after being vacuum-chucked.

The alignment mechanism in double-sided lithography (Figure 28.2) relies on image processing. The image of the mask alignment marks is stored, the wafer is then inserted between the mask and the alignment microscope, and the alignment marks on the wafer are aligned to the stored mask alignment marks. Alignment accuracy is ca. 1 μm at best, and usually a few microns.

28.1.4 Bond alignment

Anodic bonding alignment resembles standard lithography: the glass wafer with its metal patterns can be

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

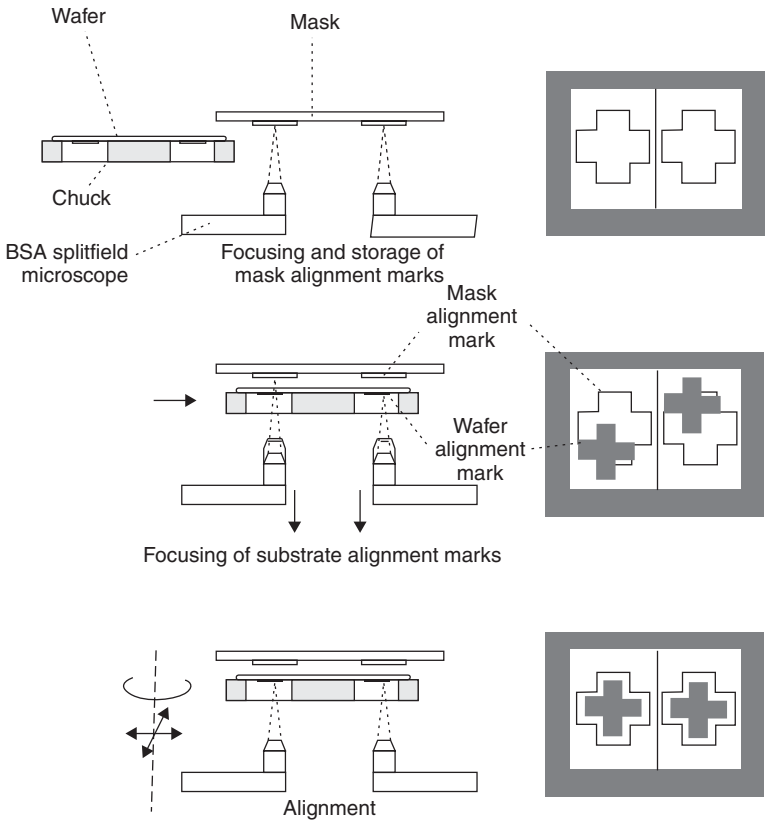


Figure 28.2 Double-side alignment. Figure courtesy of Suss Microtech GmbH

aligned to the bottom silicon wafer (photomasks are glass plates with metal patterns). Bonding of two structured silicon wafers requires a tool similar to the double-side lithography system. Alignment marks on the first wafer are registered, the second wafer is aligned to those marks and the wafers are then brought to contact. The critical step is to maintain the alignment while the wafers are transferred to the bonding equipment. This is accomplished by a special fixture that fits both the aligner and the bonder, and therefore, wafers need not be handled after alignment. Bonding is a process that can be repeated: wafer stacks with up to six wafers have been made, with ca. 1 μm alignment between the wafers.

28.1.5 Etching

Wet etching (and wafer cleaning) in a tank takes place on both sides simultaneously. It may be useful to etch from both sides, either for symmetry reasons, or for doubling the apparent etch rate. If all etching is on one side only, it is mandatory to preserve the protective films on the backside. Single-wafer plasma etching is an obvious choice and if wet etching is preferred (e.g., because of surface quality considerations), the backside must be protected.

Protection by spin-coated polymers is a quick and easy method. Photoresist is suitable for many applications, such as mask oxide etching in BHF, but aggressive etchants like KOH require either inorganic films (oxide, nitride) or more stable polymers. CYTOP (cyclized perfluoropolymer) can tolerate KOH and 49% HF. CYTOP can be removed by oxygen plasma. Blue tape common in wafer dicing can also be used as a protective layer, but removal of the tape can be difficult if fragile freestanding structures are present on the wafer.

A single-wafer holder that exposes only one side of the wafer to the liquid is a universal solution. In electrochemical etching or deposition, this holder also provides the necessary electrical contacts to the wafer. However, some wafer front surface area is covered by the holder, and single-wafer processing is more expensive than batch processing. With a holder, the topside

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

processing and materials can be selected from a device operation point of view, and no extra protective coatings are needed during processing.

28.2 MEMBRANE STRUCTURES

Sometimes, two etchings are needed to define structures. It is important to understand which should be performed first. Three examples are shown in Figure 28.3: a capacitive pressure sensor (with anodic bonding to a glass wafer), a thermally insulated nitride diaphragm with a silicon heat distribution mass and a Weir-type microfluidic particle filter (bonded to a glass wafer).

The pressure sensor gap is very small, of the order of 1 μm. This cannot be considered a topography increase in MEMS even though it would lead to serious depth-of-focus problems in deep sub-micron lithography. Deep etching is done as the second step, just before bonding. After bonding, the mechanical strength of the bonded stack is adequate for further handling without special care, whereas handling of through-etched wafers is a delicate business.

For the thermal equalization mass, a rim is etched first, to a depth that corresponds to the desired thickness of the thermal mass; and a large square pattern defines the isolation nitride membrane size. In the Weir-filter, the shallow etch depth determines the pass size, and the deep V-groove etching defines the flow channels. Shallow etches in the micron range are easy, and shallower ones could be made. However, the anodic bonding process and glass structural stability determine how shallow passages shall remain open (as discussed in Chapter 17). Auxiliary pillars (on the first mask) act as supports for the glass roof.

A pressure sensor can make use of a similar approach as the thermal mass structure: a large boss is left in the middle of the structure, for added mass. This improves capacitor parallelism: due to the added mass, diaphragm movement is much more parallel and less curving. The exact shape of the boss is determined by concave- corner etching of fast etching planes; but in this application, corner rounding is not critical.

28.2.1 Piezoresistive pressure sensor

The piezoresistive pressure sensor is one of the oldest and most widely produced micromechanical devices (Figure 28.4). The simplest version of pressure sensor diaphragm control is the timed etch. Perhaps the dominant method for thickness control is the electrochemical etch stop with n-type epilayer on p-substrate. However, the process flow discussed below is based on an advanced Si:B:Ge etch-stop structure.

The simple p^{++} etch stop does not work for a piezoresistive pressure sensor for two reasons: piezoresistors cannot be fabricated in heavily doped silicon, and the

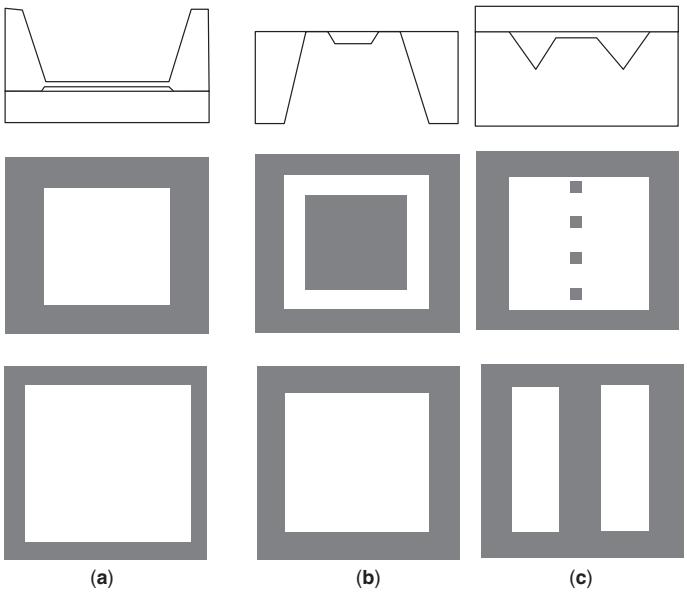


Figure 28.3 (**a**) Pressure sensor (bonded to a glass wafer); (**b**) a thermally isolated nitride membrane with a silicon thermal equalization mass and (**c**) a microfluidic particle filter. The two photomasks are shown (for positive resist patterning of mask oxide)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

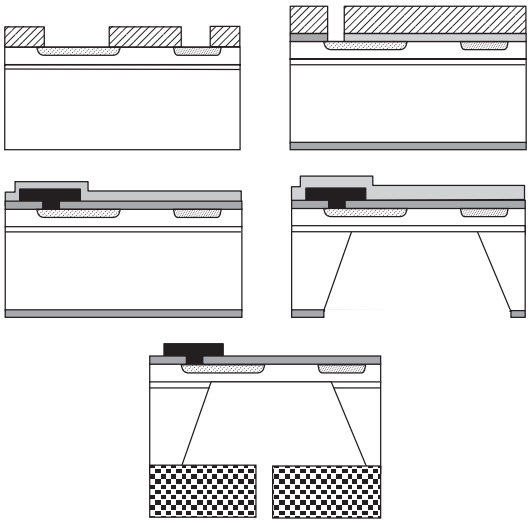


Figure 28.4 Piezoresistive pressure sensor fabrication (see process flow for details)

mechanical properties of highly doped ($>10^{18}\,cm^{-3}$) diaphragms are inferior to low or moderately doped material. An advanced etch-stop structure relies on double epitaxial layer structure: etch-stop layer and a device layer. The first epilayer to be deposited is heavily boron doped, but in order to minimize mechanical stresses from boron doping, the film is compensated by germanium ($10^{21}\,cm^{-3}$ germanium, $10^{20}\,cm^{-3}$ boron). The boron atom is smaller than silicon, and germanium is larger, which prevents stresses from volume mismatch building up. Germanium is a column-IV element beneath silicon and therefore isoelectronic with silicon, so no electrical effects are introduced. The second layer, lightly doped, is deposited on top of the Si:Ge:B etch-stop layer. This second layer is the actual device layer, and we can choose the piezoresistor-doping level freely. Anisotropic etching of silicon stops at the Si:Ge:B layer, which is then removed by a wet etch that etches highly doped silicon but not lightly doped silicon. Lightly doped silicon (>1 ohm-cm) is etched at 1 nm/min in an HF:HNO_3:CH_3COOH (1:3:8) etch, whereas for heavily doped silicon (0.01 ohm-cm), the etch rate is 1000 nm/min. This is an electrochemical effect: there are not enough holes in lightly doped silicon for etching to proceed.

Process flow for piezoresistive pressure sensor

wafer selection: p-type silicon
epitaxy: Si:Ge:B + lightly doped epi
  (front side)
lithography for piezoresistors
  (front side only)
ion implantation for resistors
  (front side only)
photoresist stripping
resistor diffusion in dry oxidation
  (thin pad oxide grown simultaneously)
LPCVD nitride
  (both sides)
lithography for resistor contacts
  (front side)
plasma etching of contacts
  (backside will not be etched)
photoresist stripping
metal sputtering
  (front side only)
lithography for metal
metal etching
photoresist stripping
PECVD nitride protective coating for metallization
  (front side)
photoresist spinning for front side protection
photoresist spinning on backside
lithography for diaphragm release
  (on backside)
nitride + oxide etching; CF_4 plasma
  (front side not etched)
photoresist stripping
  (both sides simultaneously)
KOH etching for bulk silicon removal
  (front side protected by PECVD nitride)
HF:HNO_3 isotropic etching for p^{++} epi removal
  (selective against lightly doped silicon)
plasma-etch nitride + HF-oxide etch
  (to reveal silicon for anodic bonding)
anodic bonding.

The diaphragm thickness is determined by the epitaxial layer thickness. If bulk wafers are used, diaphragm thickness would be determined by wafer thickness and etched depth. Epilayer thickness is independent of wafer specifications (thickness, TTV), enabling a much higher degree of control in diaphragm fabrication.

At first, it might appear that the backside lithography step for a diaphragm-etch is a non-critical lithography step: it merely removes a big block of silicon. But it is, in fact, a critical lithography step: the position of the piezoresistors should coincide with the maximum deflection point of the diaphragm, and therefore alignment is critical.

Even if the double side alignment is perfect, the piezoresistor could be misplaced relative to the diaphragm because of two additional factors:

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

1. If the wafer thickness is not exactly known, the diaphragm size will be wrong (epitaxial layer does not help here). Too thick a wafer will result in a diaphragm smaller than designed, and vice versa. Piezoresistors on the wafer front side will not coincide with mis-sized diaphragm.
2. If the etch selectivity between the (100) and (111) planes is not accurately known and included in the mask design, the size of the diaphragm will be wrong.

28.3 THROUGH-WAFER STRUCTURES

A nozzle is a basic through-wafer structure. It can be done by one-sided lithography and etching: the nozzle size is determined by the mask size (W_{mask}), wafer thickness (t_{wafer}) and silicon crystal geometry (Figure 28.5). The condition for zero nozzle orifice is $W_{\text{mask}} = \sqrt{2}t_{\text{wafer}}$. This simple process has too many limitations that make it impractical.

Double-side processing and boron etch stop eliminate the effects of wafer thickness and TTV from the nozzle fabrication process: the nozzle orifice area is protected by an oxide layer, and the rest of the top surface is p^{++} doped. Backside etching stops at the heavily boron-doped etch-stop layer but continues at orifice sites that did not receive boron doping (Figure 28.5(b)). Alignment between the top and the bottom is not critical because orifice dimensions are determined by top-side processes: lithography, oxide etching and boron diffusion. This approach not only solves thickness and TTV problems but also enables free-form nozzle shapes to be fabricated, whereas simple anisotropic etching results in square and rectangular nozzles only.

Despite all the good features of anisotropic wet etching, through-wafer structures take up a lot of silicon area. Nozzles fabricated by anisotropic through-wafer wet etching cannot be packed close to each other, and for ink-jet printers, other nozzle geometries have been studied. Side-shooting geometries are not limited by wafer thickness or etch geometries. One such design is described in Figure 28.6.

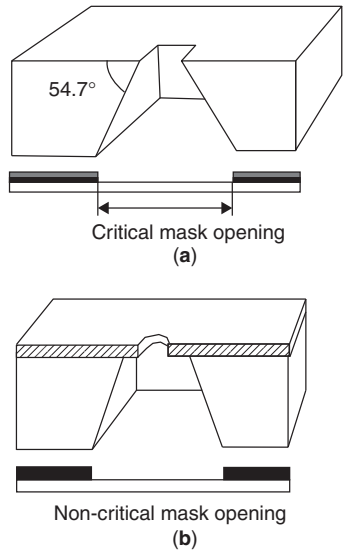


Figure 28.5 (**a**) Nozzles fabricated by simple anisotropic wet etching through the wafer and (**b**) nozzles fabricated by double-side lithography and boron etch stop (shown hatched). See text for details

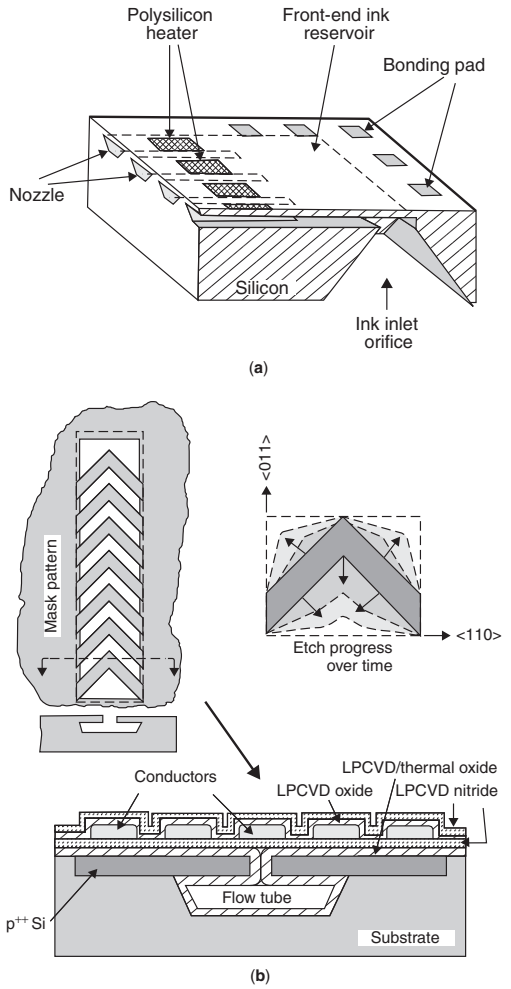


Figure 28.6 Side-shooting ink jet. The chevron structure enables both anisotropic under-etch and roof sealing. Reproduced from Chen, J. & Wise, K.D. (1997), by permission of IEEE

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Process flow for ink jet: (photoresist stripping and cleaning steps omitted)

thermal oxidation, 1 μm thick
lithography step 1: chip area definition
oxide etching
boron diffusion, 2 μm deep
lithography step 2: chevron pattern: 1 μm width
RIE of silicon, 4 μm deep
anisotropic silicon etching to undercut p^{++} chevrons
thermal oxidation, 0.5 μm
LPCVD nitride deposition for chevron roof sealing, 0.6 μm
etchback (or polishing) of nitride
LPCVD polysilicon deposition, 0.8 μm
poly doping, 20 ohm/sq
lithography step 3: poly-heater pattern
polysilicon etching
aluminium sputtering
lithography step 4: metal pads
aluminium etching
passivation: CVD oxide 1 μm + PECVD nitride 0.3 μm
lithography step 5: opening of bonding pads
RIE of nitride and oxide
lithography step 6: pattern for gold lift-off
evaporation of Cr/Au
lift of Cr/Au
lithography step 7: fluidic inlet definition on the backside
anisotropic etching through the wafer from the back.

Boron-doped silicon provides mechanical strength for the structure, as compared to nitride membrane, which can be only hundreds of nanometres thick, versus micrometres for the silicon roof. The chevron patterns open fast etching crystal planes that enable undercutting on <100> wafer. Chevron openings must be as narrow as possible so that flow tube sealing is easy: however, 0.5 μm oxide plus 0.6 μm nitride is much more than the 1 μm chevron opening. This has at least three reasons: RIE etching results in some widening, thermal oxide is ca. 50% inside silicon sidewalls and does not contribute its full thickness to sealing; and LPCVD nitride step coverage can be less than 100%. Figure 23.13 shows what the chevrons look like before and after sealing.

Thinning of nitride/oxide stack is done to improve thermal speed: the closer the heater resistor is to the flow tube, the faster the heating will be. Aluminium is not absolutely required because polysilicon is heavily doped and it can be used for wiring. However, aluminium wiring reduces resistive losses. Gold on bonding pads makes wire bonding easy, and gold protects the front side during backside anisotropic etching (areas that are not gold-covered are either nitride or oxide, which are resistant to alkaline etchants). Through-wafer etching is non-critical because it will stop automatically on the bottom oxide of the flow tube.

28.4 PATTERNING OVER SEVERE TOPOGRAPHY

28.4.1 Resist technology

Spray coating of resist works for wet-etched deep structures with 54.7° angles but exposure focus depth is another issue. Electrochemical coating of resist is a standard technique in the printed circuit board industry and negative working electrodeposited resist can cover sidewalls of vertical holes and cavities. However, electrodeposited resist can be used for many ordinary applications as well. Even though its resolution is not stellar, it can be handy for large structures.

28.4.2 Peeling masks/nested masks

Photoresist coating over severe topography can be eliminated by double masking (peeling masks/nested masks, Figure 28.7): two different mask materials are patterned on a planar wafer, before the first deep etching. The first mask is discarded after the first etching step, and etching continues with the second mask. Combinations of oxide, nitride and silicon carbide have been tried.

28.4.3 Shadow masks

Shadow masks (Figure 28.8) enable metallization of wafers with severe topography or even wafers with through-holes. However, pattern size control over severe topography may not be very good because of flux divergence. It can be improved if the shadow mask itself is a silicon wafer patterned to match the 3D geometry already fabricated, patterning accuracy is regained.

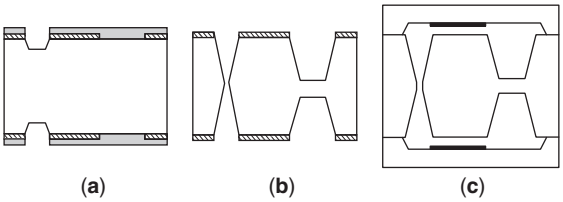


Figure 28.7 Peeling mask/nested mask: (**a**) nitride (hatched) deposition and patterning; oxide (grey) deposition and patterning; first silicon etching; (**b**) oxide etching in HF; second silicon etching with nitride mask and (**c**) capacitive accelerometer by three-wafer bonding

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

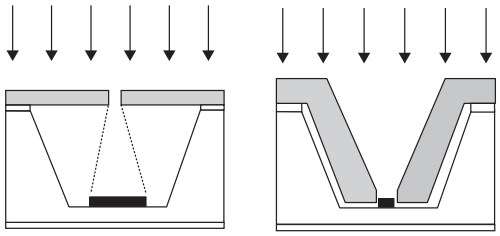


Figure 28.8 Conventional and micromachined 3D silicon shadow masks compared. Redrawn from Brugger, J. *et al*. (1999), by permission of Elsevier

28.5 DRIE VERSUS ANISOTROPIC WET ETCHING

Both plasma etching (RIE/DRIE) and wet etching have their advantages (Tables 28.3 and 28.4), and in many applications, both etching techniques are mandatory. The decision in favour of either technique depends not only on technological factors such as etched shape, sidewall angle or surface quality, but also on practical issues such as etch rate, backside protection or equipment availability.

In the micropipette process shown in Figure 28.9, both DRIE and KOH etching are utilized, in addition to almost all other major microfabrication processes. Flow channels are made in the Pyrex glass wafer by isotropic etching in HF, and aligned to the micronozzles fabricated in silicon. Anodic bonding seals the flow channels.

Process flow for micropipettes

DRIE of nozzles (30 μm deep, 2 μm in diameter);
LPCVD nitride;
KOH etching (nitride masked);
wafer thinning (unmasked KOH etching);
nitride RIE etching;

Table 28.3 Main features of DRIE

– Any shape can be made (RIE lag, ARDE and microloading limitations)
– Tightly spaced structures can be made
– High aspect ratio vertical structures are possible (10:1 to 20:1 AR typical)
– If membrane structures are needed, SOI wafers must be used
– Photoresist masking is possible
– Single-side processing, no backside protection needed
– 1–3 hours for through-wafer etching in single-wafer operation
– 1–3 days to etch a batch of 25 wafers through-the-wafer

Table 28.4 Main features of anisotropic wet etching

– Very accurate dimensional control by crystal plane–dependent etching
– Structural shapes limited by crystal plane–dependent etching
– Accurate 45°, 54.7°, 70.5° or 90° sidewalls
– Smooth and well-defined surfaces
– ca. 4–8 hours for through-wafer etching for a single wafer
– ca. 4–8 hours for through-wafer etching for a batch of 25 wafers
– Etches both sides, protection needed on backside
– Etches both sides, symmetric structures can be made in a single etch step
– Aggressive to metals and many other materials
– Limited selection of mask materials, thick oxide and LPCVD nitride standard
– Many etch-stop mechanisms available: boron p^{++}, pn-junction, SOI BOX

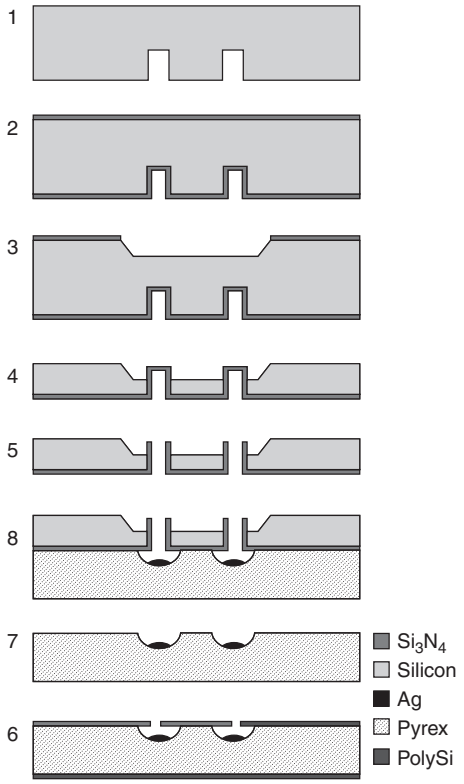


Figure 28.9 Fabrication process for micropipettes: both DRIE, KOH and isotropic HF etching have been used. Reproduced from Guenat, O.T. *et al*. (2003), by permission of IEEE

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

HF etching of Pyrex glass with polysilicon mask;
silver lift-off metallization;
anodic bonding.

28.6 IC–MEMS INTEGRATION

Silicon is just one possible substrate for MEMS, but it is the one that promises integration with electronic (e.g., CMOS circuitry) and optical (e.g., photodiodes) functions that can be fabricated on the same wafer. This section discusses some general integration issues encountered with IC–MEMS integration.

There are three main ways of integrating IC and MEMS devices on a wafer level:

- MEMS before IC;
- MEMS and CMOS interleaved;
- MEMS post processing.

All of these have their strengths and weaknesses, but in all cases, process complexity increases and cases of successful commercialization of monolithic integration remain few. Hybrid integration at chip level is still the norm in the industry: MEMS chip and the accompanying ASIC (for readout, calibration and self-testing) are separate chips. This is partly a commercial (production volume) issue, and partly a technical issue: very few advanced IC fabs are capable of MEMS processing.

IC packaging is generic and simple: both plastic and hermetic packages are independent of chip design and technology. With MEMS, it is a wholly different story: movable structures may stick during the anodic bonding process, even though sticking might have been avoided in release etching. Wafer dicing relies on 20 000 rpm saw blades that might bring MEMS structures to resonance, water cooling may lead to sticking and silicon dust may block cavities and gaps.

Zero-level package is a structure that seals the MEMS part from the ambience. It is preferably applied on the whole wafer, in a manner not unlike passivation nitride deposition in IC industry. Two routes have been explored: deposition and wafer bonding (see Figure 17.2). The former should have zero step coverage for optimum performance, acting as a roof only. The latter has the disadvantage that an additional wafer is required.

In the MEMS-first approach, MEMS devices are processed and covered (e.g., by TEOS), and hopefully, they will not be adversely affected by the hundreds of process steps it takes to complete the IC. IC-process temperatures severely limit the selection of materials for MEMS-first integration: silicon, polysilicon, oxide and nitride are really the only candidates. Connecting the MEMS part to the IC part is preferably done by diffusions because metal–silicon interfaces cannot be made until fairly late in the process. Despite its name, this approach still has some of the MEMS steps to be done after the completion of IC processing: usually the release of freestanding structures and maybe metallization.

The plug-up process shown in Figure 28.10 is an SOI MEMS–IC process that consists of the following main modules:

1. MEMS structure processing and encapsulation;
2. CMOS process;
3. MEMS structure release.

There is no topography increase in SOI MEMS steps, and the sealed cavities do not pose problems for subsequent CMOS processing if the CMOS and MEMS parts are side by side on a wafer.

Interleaved fabrication offers the greatest challenges for process and device designers because there are so many trade-offs to be made. Take polysilicon, for instance: CMOS gate polysilicon is typically 0.25 μm thick, whereas micromechanical poly is ca. 2 μm thick. Gate poly is optimized for poly/SiO_2 interface properties and it is highly doped. Micromechanical poly is designed for minimal stresses and stress gradients. If two separate polysilicon depositions are needed, with two different doping/annealing steps, the benefits from integration start disappearing.

Post-processing of MEMS devices (Table 28.5) includes a great number of choices: micromechanical structures can be made by both subtractive (etching) techniques and additive (deposition) techniques.

Table 28.5 MEMS post-processing

Subtractive	Notes
Bulk silicon backside etching	Wet or DRIE, double-side lithography
Bulk silicon front side etching	Single sided, wet or plasma
Surface; front-side etching	Thin-film mechanical elements only
SOI front/back etching	Buried oxide etch stop for both, wet or DRIE
Additive	**Notes**
Polysilicon/polySiGe (LPCVD)	Thermal limit on poly annealing
Aluminium (sputtering)	Layer thicknesses limited
Nickel (electroplating)	Thick layers possible
Nitride (PECVD)	Stress control

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

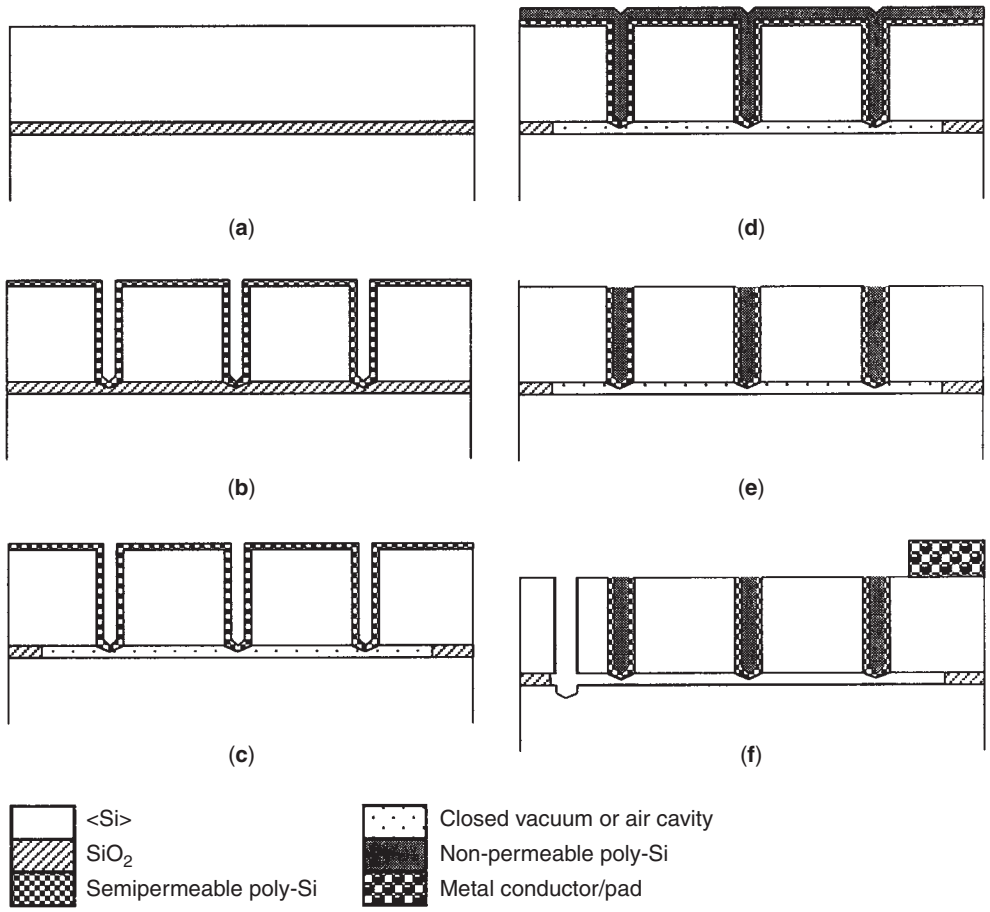


Figure 28.10 Integration of MEMS and CMOS on SOI: (**a**) SOI wafer; (**b**) DRIE of access holes to buried oxide and deposition of semi-permeable polysilicon; (**c**) buried oxide etching through semi-permeable poly; (**d**) refilling the holes with non-permeable polysilicon; (**e**) poly etchback and planarization and (**f**) further IC and/or MEMS processing. Figure courtesy Jyrki Kiihamäki, VTT

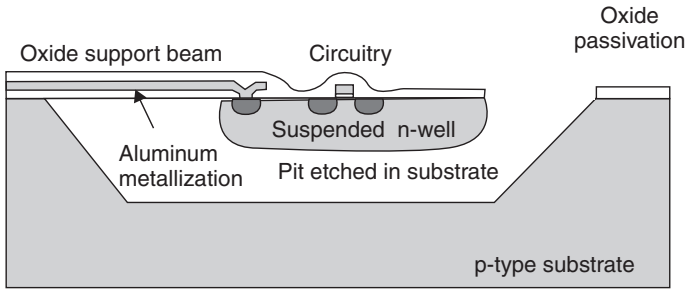


Figure 28.11 Post-CMOS wet etching with electrochemical etch stop to protect n-well of CMOS part. Reproduced from Kovacs, G.T.A. *et al*. (1998), by permission of IEEE

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Another distinction relates to silicon real estate: are the IC and MEMS devices on top of each other, or side by side? This has important implications for etch stop, alignment and device packing density. Bulk silicon removal can also be used to leave n-wells of the CMOS-part intact by electrochemical etch stop, which provides thermal isolation (see Figure 28.11). This offers improved sensitivity for weak thermal signals.

CMOS wafers can be treated as any other substrates, even though they are very expensive: CMOS wafer cost is ca. $500 for a finished 150 mm wafer with 0.8 μm devices on it, versus $20 for a bulk wafer, $50 for an epiwafer and $200 for an SOI wafer. CMOS wafers as substrates have certain limitations: the maximum processing temperature is limited by the silicon–metal interface stability. The standard 450 °C limit has been raised to ca. 700 °C by utilizing tungsten with diffusion barriers. Usually, the topmost metallization layer is not planarized, but CMP is needed when CMOS is used as a substrate. CMOS transistors have to be protected from chemical contamination. This has been done successfully by combined oxide/nitride passivation and polymeric protective coating, and KOH etching can be accomplished without any deleterious effects on the CMOS. Array devices with CMOS transistor drivers include digital micromirror devices (DMD), IR pixel sensors and fingerprint sensors.

28.7 EXERCISES

1. Nozzles are fabricated by etching through a 380 μm thick <100> silicon wafer anisotropically (Figure 28.5). A 540 μm wide mask pattern is used.
   (a) Calculate the size of holes produced by an ideal process.
   (b) Calculate the effect of the following real world uncertainties:
      1. Wafer thickness variation: 380 μm ±5 μm;
      2. Total thickness variation TTV of 1 μm;
      3. <100>:<111> crystal plane selectivity 33:1 versus 30:1;
      4. Mask width +1% narrower than the design value.
2. If a piezoresistive pressure sensor diaphragm is made in an epitaxial layer, and diaphragm etching is stopped by pn-junction etch stop, how do the following affect sensor structure:
   (a) wafer thickness;
   (b) wafer TTV;
   (c) epitaxial layer thickness.
3. If vertical walled through-wafer structures are made, what is the minimum size and space that can be realized by: (a) DRIE, (b) <110> wet etching and (c) <100> wet etching?
4. The deflection of a circular membrane under pressure is given by $h = 0.666\,(r^4 \Delta p/Et)^{1/3}$, where r is the radius, t the thickness and E the Young's modulus of the diaphragm. What is the deflection that corresponds to a pressure difference of 25 mtorr? What is the corresponding capacitance change?
5. Analyse the fabrication process for the nanoholes shown in Figure 13.13.

6S. What is the thickness of beams and membranes that you can make with the p^{++} etch stop technique if diffusion is used to fabricate the p^{++} layer?

7. Calculate the mask dimensions for both masks when 100 μm lateral isolation distance is needed in the thermally isolated structure with silicon heat equalization mass (Figure 28.3(b)).
8. Calculate the mask dimensions and estimate vertical etched depths for the accelerometer shown in Figures 21.10 and 28.7.
9. Design a fabrication process for the 3D silicon shadow mask shown in Figure 28.8
10. What is the linear density of ink channels of technology shown in Figure 28.6?

REFERENCES AND RELATED READINGS

Briand, D. *et al*: Design and fabrication of high-temperature micro-hotplates for drop-coated gas sensors, *Sensors Actuators*, **B68** (2000), 223.

Brugger, J. *et al*: Self-aligned 3D shadow mask technique for patterning deeply recessed surfaces of micro-electro-mechanical systems devices, *Sensors Actuators*, **76** (1999), 329.

Chen, J. & Wise, K.D.: A high-resolution silicon monolithic nozzle array for inkjet printing, *IEEE TED*, **44** (1997), 1401.

de Boer, M.J. *et al*: Micromachining of buried micro channels in silicon, *J. MEMS*, **9** (2000), 94.

Griss, P. *et al*: Development of micromachined hollow tips for protein analysis based on nanoelectrospray ionization mass spectrometry, *J. Micromech. Microeng.*, **12** (2002), 682.

Guenat, O.T. *et al*: Ion-selective microelectrode array for intracellular detection on chip, *Transducers '03* (2003), p. 1063.

Hierlemann, A. *et al*: Microfabrication techniques for chemical/biosensors, *Proc. IEEE*, **91** (2003), 839; special issue on chemical and biological microsensors.

Kovacs, G.T.A. *et al*: Bulk micromachining of silicon, *Proc. IEEE*, **86** (1998), 1543.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Leclerc, S. *et al*: Novel simple and complementary metal-oxide-semiconductor-compatible membrane release design and process for thermal sensors, *J. Vac. Sci. Technol.*, **A16** (1998), 876.

Lin, L. & Pisano, A.P.: Silicon-processed microneedles, *J. MEMS*, **8** (1999), 78.

Mehra, A. *et al*: Microfabrication of high-temperature silicon devices using wafer bonding and deep reactive ion etching, *J. MEMS*, **8** (1999), 152.

Meng, E. *et al*: Silicon couplers for microfluidic applications, *Fresenius; J. Anal. Chem.*, **371** (2001), 270.

Pham, N.P. *et al*: IC-compatible two-level bulk micromachining process module for RF silicon technology, *IEEE TED*, **48** (2001), 1756.

Trimmer, W.S.: *Micromechanics and MEMS, Classic and Seminal Papers to 1990*, IEEE Press, 1997.

Proc. IEEE (1998), special issue on integrated sensors, microactuators and microsystems.

**Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030**

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

29

Processing on Non-silicon Substrates

29.1 SUBSTRATES

We are already familiar with devices made on non-silicon substrates: the acoustic resonator of Figure 7.9 and the passive integrated chip of Figure 24.13 were fabricated on glass/fused silica because substrate capacitances had to be eliminated. The photomask is also a microstructure on glass, even though it is not usually considered one. It shows many of the issues that make non-silicon substrates different: it is square, thick and made of glass, which is not a well-defined material like silicon. The coefficient of thermal expansion (CTE) for soda lime glass is 10 ppm/°C (2.6 ppm/°C for Si), and as a photomask material soda lime glass is limited to applications above 3 μm linewidths in which dimensional control requirements are lax (remember exercise 9.3). The big difference in CTE relative to silicon makes soda lime glass unsuitable for anodic bonding.

Glasses contain, by definition, alkali metals, usually sodium. These alkali ions are essential for some applications, such as anodic bonding even though they are detrimental to electronic devices. Pyrex glass has composition SiO_2:B_2O_3:Al_2O_3:Na_2O in the approximate ratio 80:10:5:5. Pyrex glass is available in round formats and is extensively used in anodic bonding, because its CTE matches that of silicon. In photoactive glasses there are also lithium and other exotic metals, which are major contamination risks. Photoactive glasses have CTEs four times that of silicon, which excludes anodic bonding. Fused silica is 100% SiO_2 and is quite compatible with silicon processes. It is mechanically strong enough to withstand standard high-temperature process steps and it is available up to the 300 mm wafer size, which has made it the material of choice for some silicon-based optical devices. However, because of the lack of mobile ions, it is not amenable to anodic bonding.

The limited temperature range available for processing is a hindrance for processing on glass. This comes from two main factors: glass is mechanically soft and it loses its stiffness above ca. 500 °C (very much dependent on exact composition). Secondly, sodium diffusion at elevated temperatures can be detrimental to electronic devices.

Quartz is pure silicon oxide, just like fused silica, so there is no alkali metal contamination risk. While fused silica is glass in the sense of being amorphous, quartz is crystalline, but the word quartz is often used as shorthand for fused silica. Etching of crystalline quartz. The etching of quartz in HF-based solutions leads to crystal plane-dependent etching, just like silicon etching in alkaline solutions. This crystallinity has important implications for piezoelectric devices, which must be oriented along proper crystal axes.

Flat panel displays (FPDs) are the most important devices fabricated on glass, by sales volume. Radiation detectors and photodetectors of various designs have been made on glass substrates, using a-Si, SiC and diamond as active materials. Glass substrates have several advantages from a manufacturing point of view: they are available in large sizes; 50 × 60 cm is fairly typical, and 140 × 185 cm is available. Secondly, glass is cheap. Thirdly, it is fairly smooth and can be cleaned with RCA-cleans just like silicon wafers; in fact, the RCA-clean was invented for glass cleaning in TV-picture tube manufacturing.

Some problems of non-silicon substrates are related to processing them in a silicon-oriented lab. Even though fused silica wafers are round like silicon, have flats like silicon and are available in the same thicknesses as silicon, complications can still arise, especially in automated tools. The detection of the presence and the movements of wafers are based on either optical or capacitive sensors, and these are fooled by transparent dielectric wafers. Amorphous silicon or polysilicon deposition on the wafer backside can be used as a preventive measure, but the role of this extra film needs to be considered for all process steps and tools.

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Many non-silicon substrates are not round but square. Many substrates are available in both shapes, including glass, quartz and aluminum titanium carbide (which is used in thin-film heads (TFH) for magnetic storage). Exotic materials such as microwave substrates and printed circuit board substrates of glass fibre-filled polymers or alumina are traditionally squares, and plastic and steel come in rolls.

One process step particularly suited for round substrates is photoresist spinning. Square substrates rotating 5000 rpm create turbulence in the corners, and uniformity cannot be obtained. One solution is to use a round carrier with a recess for the square substrate. Another solution is to rotate both the substrate and the bowl in unison, to minimize turbulence.

Not only are the substrates square, the standardization of their sizes is almost non-existent. This is difficult for process tools and tool automations, in particular. What is more, thicknesses are not standardized, either. Add to this the fact that some ceramic substrates have densities three times that of silicon and quartz, and they can be 2 mm thick, which translates to a factor of a 10 mass difference. Thickness also has an effect on thermal equilibrium and the heating of wafers, intentional and unintentional.

Substrates of piezoelectric and ferroelectric materials like $LiNbO_3$ not only pose contamination dangers, but "react" to processes: plasmas cause charging which leads to mechanical volume changes which can relax via unexpected mechanisms. Special material properties like magnetism or superconductivity depend on crystalline structure, and sometimes process temperatures are severely limited. For example, PECVD protective coatings must be deposited at 120 °C, but of course, film quality is not comparable to 300 °C deposition.

29.2 THIN-FILM TRANSISTORS, TFTs

Thin-film transistors (TFTs) are MOS devices with deposited films as channel materials and as gate dielectrics. The most common channel material is amorphous silicon, a-Si:H, and sometimes, temperature allowing, a crystallization process can turn a-Si:H into polysilicon, but there is no need to limit oneself to silicon: conducting polymers such as pentacenes and thiophenes can be used. However, carrier mobilities of these materials are rather different from single-crystal silicon: mobility of SCS is ca. 500 cm^2/Vs, polysilicon ca. 100 cm^2/Vs, a-Si:H ca. 1 cm^2/Vs and organic molecules between 0.001 to 1 cm^2/Vs. Deposited PECVD oxide or nitride are used as gate dielectrics. TFT performance is therefore inherently worse than MOS with thermal gate oxide. Liquid crystal displays (LCDs) use active pixel switching by implementing a transistor for each pixel (AMLCD).

TFTs come in two basic varieties: bottom gate and top gate. Both are MOSFETs but the order of gate versus source/drain is opposite. One of the many bottom-gate versions is described in Figure 29.1, and one top gate TFT is shown in Exercise 29.3.

Process flow for bottom-gate TFT

Process	Function/comment
Cr deposition	Gate metal
Gate lithography and etching	Wet etching
SiN_x deposition	Gate dielectric
Channel a-Si:H deposition	Undoped
SiN_x deposition	S/D separation
SiN_x lithography & etching	Plasma etching
n+ a-Si:H deposition	S/D contact improvement
Cr deposition	S/D metal contact
Lithography	Transistor isolation
Etching Cr/n+ a-Si:H/a-Si:H	Wet etch selective against nitride

Metallization for row and column address electrodes is not shown.

Amorphous silicon is the active material in the channel and its annealing is one of the crucial steps. Amorphous (and polycrystalline silicon) have many dangling bonds, which have to be passivated for long-term stability. Forming gas anneal (H_2/N_2) at ca. 400 °C is a standard procedure.

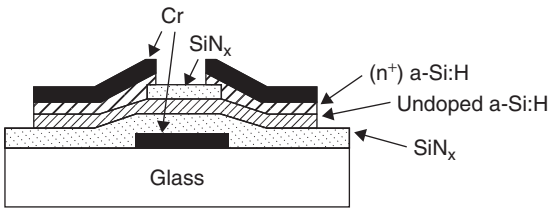


Figure 29.1 Bottom-gate TFT on glass. From Gleskova, H. *et al.* (2001), by permission of The Electrochemical Society

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Thermal oxidation cannot obviously be used but all dielectrics are (PE)CVD or sputter deposited. Ion implantation damage anneal, which is done at 900 °C, cannot be used and implantation is not a very attractive technique for large-area microelectronics because it is a slow, serial process. Other doping processes, such as gas-phase doping during PECVD silicon, must be used. Activation anneal temperatures are so low that we must accept only partial activation of dopants.

TFT performance can be improved by the same techniques used in silicon MOSFETs, but the low-cost/large-area limitations must be borne in mind. Self-aligned structures have been developed for TFTs with spacers, lightly doped drains (LDDs) and self-aligned silicides. CMP cannot be used because of cost considerations and large-area limitations, and plasma-etching uniformity across 50 cm panels can also be problematic. However, because linewidths are of the order of 10 μm, wet etching is suitable for most etching steps.

If alkali glass is used, sodium contamination is a problem: the very first process step must be an ion barrier deposition to isolate the silicon devices from the glass substrate. Aluminum oxide and various other oxides are employed. This barrier must be dielectric, in contrast to diffusion barriers in metallization. The barrier is also part of the optical path of the device, and its influence on display properties, for instance interference colors, must be borne in mind.

In FPDs, depending on the optical design of the display, transparent conductors are used for metallization. Transparent conducting oxides (TCOs) are curious materials, which combine high electrical conductivity (σ) and low optical absorption (α). Transparency and resistivity cannot, of course, be independently optimized because charge carriers are responsible for both optical absorption and electrical conductivity. The figure of merit for TCOs is the ratio of electrical conductivity to optical absorption, and this must be maximized.

Typical TCOs include indium oxide (In_2O_3) and tin oxide (SnO_2) and their alloys, such as SnO_2:F or In_2O_3:Sn, indium tin oxide, known as ITO. Resistivities of transparent conducting oxides are 100 to 500 μohm-cm (a factor of 100 higher than that of true metals), which translates to sheet resistances of a few ohms, and to transmission of over 70% from 400 to 1000 nm (with absorption coefficient $\alpha \approx 0.04\,\mu m^{-1}$).

The yield is paramount because there are usually just a few displays per panel: a 50 cm by 50 cm plate may contain just four displays. Yield statistics are very different from ICs, which have hundreds of devices per wafer (yield will be discussed in Chapter 36 in more detail). Fortunately, linewidths are very relaxed. However, large areas need to be exposed (and still larger ones are required in the future) whereas IC lithography benefits from small area exposure. Film thicknesses are, however, similar to IC fabrication, and particles, pinholes and hillocks are dangerous. Killer defect is half the film thickness, which puts high demands on cleanroom facilities.

29.2.1 Super-self-aligned thin film transistor (TFT)

Fabrication on glass substrates offers intriguing ways of self-alignment in TFT fabrication. A bottom-gate version is described in Figure 29.2. After chromium bottom-gate lithography, etching and stripping, a stack of PECVD oxide (gate oxide), a-Si:H (channel) and nitride are deposited. A photoresist is applied on the top but exposure is made from the backside, with the Cr-gate blocking light (photomasks are glass plates with chromium patterns on them). The resist is then developed and the nitride etched. After resist stripping and wafer cleaning, chromium is deposited. During annealing, chromium silicide will form on the a-Si layers, but not on the nitride.

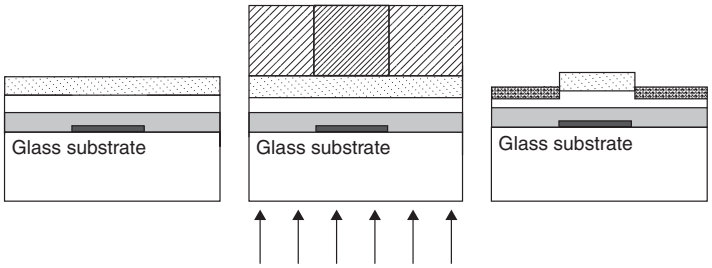


Figure 29.2 (**a**) Cr-gate has been patterned on the glass substrate, and PECVD oxide gate, a-Si:H channel and nitride stopper layers have been deposited; (**b**) topside resist backside exposure and (**c**) nitride etching and resist stripping, plus chromium sputtering and $CrSi_2$ formation. Redrawn after Hirano, N. *et al*. (1996), by permission of The Institute of Electronics, Information and Communication Engineers

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

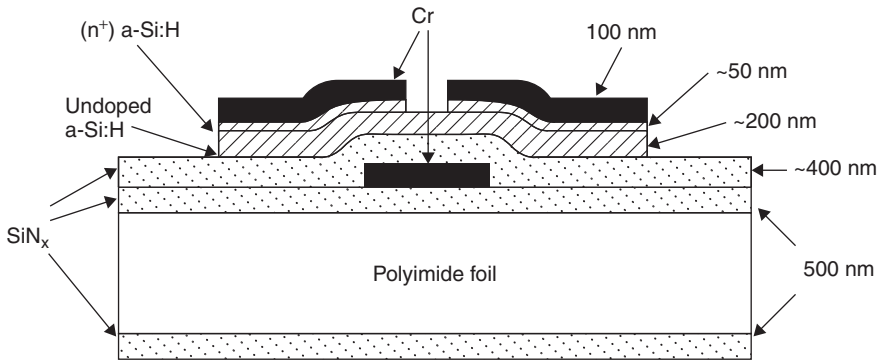


Figure 29.3 TFT on polyimide; the maximum processing temperature is 150 °C. From Gleskova, H. *et al*. (2001), by permission of The Electrochemical Society

29.2.2 TFTs on other substrates

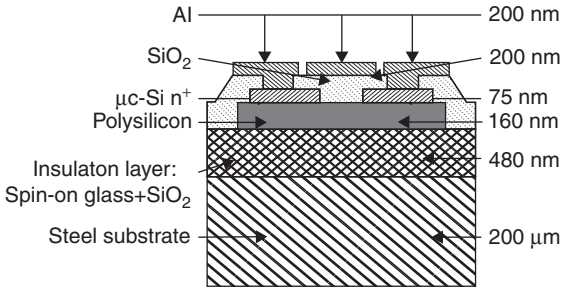
Limitations that hold true for glass plates are also true for TFTs made on steel foils, even though there are some differences. Higher processing temperatures can be used from a mechanical strength point of view, but iron contamination is a concern. Steel is a conducting material and an electrical insulator layer must be deposited on it before any electrical devices. Iron contamination concern replaces the sodium-contamination danger, so an ion barrier is needed. If the same film can act both as electrical insulation, ion barrier and smoothing layer, it is better. Steel surface smoothness is inferior to glass, and planarization may be needed. Spin-on-glass can fulfill all these disparate requirements and is clearly a strong candidate.

Processing TFTs on polymer substrates sets even stricter limits on the thermal budget. Shown in Figure 29.3 is a TFT on polyimide substrate. Maximum processing temperature has been limited to 150 °C (polyimide thin films on silicon wafers can tolerate much higher temperatures, up to 400 °C because conduction to the substrate effectively spreads excess heat). Plasma nitride serves two important functions: it passivates the device from the substrate and it acts as the gate dielectric.

The mechanical strength of polyimide substrates is inferior to both glass and steel, but fortunately low process temperatures are helpful, and due to low temperatures stresses are also minimized.

29.3 EXERCISES

1. If flat-panel lithography is done with a 50 μm proximity gap, what is the smallest possible linewidth on an FPD?
2. Calculate row and column address electrode resistances on a 15 in. TFT display. Compare ITO and real metals.
3. Design a fabrication process for the top gate TFT shown below. The maximum process temperature is 350 °C. From Wu, M. *et al*. (1999), by permission of AIP.



4. TFT itself takes up very little area compared with pixel, and transistor packing density increase offered by self-alignment is not important. What are the benefits of self-alignment in TFT fabrication?
5. What are the integration issues when the RCL passive chip in Figure 24.13 and TFBAR in Figure 7.9 are made on:
   (a) Si
   (b) glass
   (c) fused silica.

REFERENCES AND RELATED READINGS

Becker, H. *et al*: Planar quartz chips with submicron channels for two-dimensional capillary electrophoresis applciations, *J. Micromech. Microeng.* **8** (1998), 24.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Danel, J.S. *et al*: Micromachining of quartz and its application to an acceleration sensor, *Transducers '89* (1989), p. 971.

Gleskova, H. *et al*: 150 °C amorphous silicon thin-film transistor technology for polyimide substrates, *J. Electrochem. Soc.*, **148** (2001), G370.

Hirano, N. *et al*: A 33 cm diagonal high-resolution TFT-LCD with fully self-aligned a-Si TFT, *IEICE Trans. Electron.*, **E79** (1996), 1103.

Kuo, Y. *et al*: Plasma processing in the fabrication of amorphous silicon thin-film-transistor arrays, *IBM J. Res. Dev.*, **43** (1999), 73.

Leech, P.W.: Reactive ion etching of quartz and silica-based glasses in CF_4/CHF_3 plasmas, *Vacuum*, **55** (1999), 191.

Moy, J.-P.: Large area X-ray detectors based on amorphous silicon technology, *Thin Solid Films*, **337** (2000), 213.

Stewart, M. *et al*: Polysilicon TFT technology for active matrix OLED displays, *IEEE TED*, **48** (2001), 845.

Wu, M. *et al*: High electron mobility polycrystalline silicon thin-film transistors on steel-foil substrates, *Appl. Phys. Lett.*, **75** (1999), 2244.

Proc. IEEE, **90** (2002), special issue on flat panel displays.

**Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030**

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Part VI

Tools

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

30

Tools for Microfabrication

The size of the microfabrication tools tends to be inversely proportional to the size of the structures they make. Small tabletop instruments can pattern and etch 3 μm lines, but tools for 100 nm lines require garage-sized behemoths with multimillion-dollar price tags. The analogy with elementary particle physics is obvious: the smaller the objects being studied, the bigger the instruments needed. Price tags for individual tools are up to 10 million dollars today, even though $100 000 can still buy a system suitable for research purposes, be it a mask aligner, a furnace or a plasma etcher.

30.1 BATCH PROCESSING VERSUS SINGLE-WAFER PROCESSING

Microfabrication economies were earlier touted to result from batch processing: tens of wafers with hundreds of chips are processed simultaneously in, for example, a furnace or a wet etch bench. However, the scaling down of linewidths has put increasing demands on process control, and single-wafer tools have superseded batch equipment in many process steps. Besides, batch equipment for large wafers can become prohibitively cumbersome.

Wet processing in a tank is a prototypical batch process: a full cassette of wafers is processed simultaneously (see Figure 12.3). Wafer cleaning and non-patterning etching (e.g., removal of sacrificial oxide by HF) are widely done in batch-mode wet processing, even in the most advanced processes. Wet etching for patterning (e.g., H_3PO_4-based aluminium etching or BHF-etching of oxide) is not an option when linewidths are below 3 μm, because process control is difficult in batch wet processing: no *in situ* monitoring is possible and wafer-to-wafer variations are often encountered. However, model-based control with ionic strength and temperature measurement can be used to improve rate control to some extent.

In batch processing, uniformity over the batch must be added to uniformity across the wafer. Variation comes from wafer position in a batch system: flow patterns of gases and liquids over wafers depend on wafer position, and the thermal environment may also be position dependent: the first and the last wafer have only one neighbour, but the others are sandwiched between two wafers.

During the 3 in. era, most wafer processing was batch processed and the major shift started at the 100 mm wafer size. Robotic loading/unloading is simple in single-wafer systems, and they are more amenable to factory automation, including data gathering. Film thicknesses have been scaled down with linewidths, and thinner films require less process time in deposition and etching, which works in favour of single-wafer processing. However, single-wafer systems rarely even approach batch system throughputs, which can be up to 200 wafers per hour (WPH) and in some simple PECVD applications (in solar cells), even 500 WPH. It may also well be that in the back end of the process, wafers are so expensive that manufacturers do not want to risk a lot by batch processing: 200 mm wafers with 300 chips selling for $10 are worth $2500 (yield is not 100%), or the batch of 25 is worth $60 000. If a batch is lost at the end of the process, it will take time to fabricate the replacement lot, typically three to six weeks. This can be an even greater burden than the money loss if delivery time is used as a criterion for choosing a chip supplier.

In single-wafer processing, wafer-to-wafer repeatability is a major issue. First-wafer effect means that the system has not stabilized, and therefore the first wafer experiences, for example, lower temperature or more concentrated chemicals. In addition to batch and single-wafer processing, various combinations are being used, as shown in Table 30.1.

Single-feature processing is so slow that it is relegated to special applications only. Throughputs of a few

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Table 30.1 Granularity of processing

Single-feature processing
Direct writing for research and pilot production
Mask making by e-beam or laser beam
Mask repair, chip repair, chip customization
Throughputs a few wafers per hour (WPH)
Single-chip processing
Reduction steppers and scanners
Better alignment and resolution
Throughputs up to 100 WPH
Single-wafer processing
Easy automation
In situ monitoring
Throughputs 10–50 WPH
Plasma etching, sputtering, (PE)CVD, medium current implantation (MCI)
Batch processing
Enormous throughputs: up to 200 WPH
Wet cleaning, oxidation, thermal CVD (oxide, poly, nitride)
Combinations
Load multiple wafers but process one wafer at a time (HCI, CVD)

wafers per hour are considered good for direct write processes. Single-chip processing is done only in lithography, using reduction steppers and scanners. They are close to 1X systems in throughputs, with the best systems approaching 100 WPH.

Single-wafer processing benefits from easy process development because fewer wafers are needed and batch effects are eliminated. Robotic handling from cassette-to-cassette and *in situ* monitoring without averaging over a batch enables a much higher degree of process control than in batch systems. There are various combination systems, for instance, high-current ion implanters load a batch of wafers on a rotating holder, but the beam scans one wafer at a time, and the rotation of the holder takes care of the batch processing. In epitaxy, single-wafer and batch tools co-exist, but in plasma etching and sputtering, single-wafer tools are the norm in mainstream IC production.

30.2 EQUIPMENT FIGURES OF MERIT

Equipment figures of merit include various aspects such as process, capital cost, labour, consumables, and so on. Some of the most important ones are briefly discussed below.

30.2.1 Uptime/downtime

Uptime is an overall measure of equipment availability. Uptime is reduced both by scheduled and non-scheduled maintenance. Recalibration/test wafers required to set the process running after a disruption can contribute significantly to downtime. Regular reactor cleaning is mandatory for deposition equipment. Sometimes chamber cleaning is done after every wafer, so that there is no build-up of films on chamber walls (this is plasma cleaning, and not mechanical cleaning which would necessitate chamber opening). Uptime is drastically lower, but yield is higher. Uptimes vary from almost 100% for wet benches to 90% for furnaces and plasma etchers, 80% for implanters and to 40% for PECVD.

30.2.2 Utilization

Utilization is a measure of equipment use: actual productive hours of all available hours. General-purpose tools such as lithography have high utilization while the more dedicated tools have lower utilization. A 10 million dollar lithography tool must not wait for a 1 million dollar resist coater, but the resist coater can sit idle waiting for a stepper. Rapid thermal processor for silicide anneal is used twice during a CMOS process, and its utilization is the lowest of all tools, together with the dedicated wet bench for selective titanium etching.

30.2.3 Throughput

How many wafers per hour can the system handle? Single-wafer tools have throughputs of 25 to 50 WPH, but batch tools can handle up to 200 WPH. This is very much process-dependent: if the LPCVD polysilicon process is run at 635 °C, its rate is four times higher than at 570 °C. Similarly, if film thickness to be deposited is doubled, deposition time is doubled. Throughput, however, might not change much if the overhead (loading, pump down, temperature ramp, etc.) is high relative to deposition time. In etching, throughput can be severely reduced even if film thickness remains unchanged, but overetch requirement changes due to topography (recall page 129).

30.2.4 Footprint

How big is it? The cleanroom space is premium priced: $10 000 per square metre is the price range for a class 1

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

(Fed. Std.) cleanroom. In most cases, just the front panel of the system is in the cleanroom and the rest of the tool is in the service area, which has more relaxed particle cleanliness requirements.

30.2.5 MTTF, MTBA, MTBC

How long will the tool work before failure? Do operators need to interfere with its operation? How often does it have to be cleaned? These questions are operationalized by MTTF (mean time to failure), MTBA (mean time between assists) and MTBC (mean time between cleans).

MTBC is process-dependent: particle counts (on test wafers) are checked regularly, and increased counts indicate a cleaning need. However, the acceptable particle count depends on the chip size, sensitivity of the particular process step to particulate contamination (a subsequent step may be a cleaning step that effectively removes particles) or just an engineering judgement about the acceptable level of particles. Particle counts in individual process steps cannot easily be correlated with process yield, and therefore short loop test runs with specially designed test structures are used to check the effects of individual process steps.

30.3 TOOL LIFE CYCLES

Tool development takes a long time: from the first proof-of-concept tool to multiple orders for volume manufacturing easily takes 10 years. Proof-of-concept tool is a home-built or modified equipment that demonstrates the key features of a new process. For e-beam lithography, it might be a new column design; for a plasma etcher, it might be a new RF-coupling scheme. The alpha tool is a built-to-purpose system that has the new key elements designed in from the beginning. The alpha tool does not have productivity features such as robotics and software, but is designed for the final wafer size. The reliability of the alpha tool is not comparable to production tools; it is a test-bed for process research, not for production. Alpha tools are not shipped to outsiders. The beta tool is a fully equipped version, with essentially all the features that will make the final product distinct. Beta tools are shipped to select customers who are willing to bear part of the burden of testing new equipment in order to benefit from new technology. Beta customers provide productivity-related data that is difficult or even impossible to acquire at the tool-manufacturer site: What is uptime in production-like conditions? Is wafer yield comparable to existing or competing designs? What are the field servicing requirements?

Both academic and industrial labs buy equipment for research and development, but what will happen when a successful new process needs to be scaled up for production? The popular answer today is that the basic design of the process chamber (e.g., spinner bowl geometry, sputter cathode design, etcher gas manifold, RTA lamp configuration) is fixed. Research labs buy the very basic configuration, essentially the process chamber only (obviously this works better for some tools than others and not at all for optical lithography). Later on, when the process is transferred to manufacturing, productivity features such as cassette-to-cassette automation and advanced software can be added. This reduces the risk of new equipment purchase for the industry, and it allows academic labs to do industrially relevant research without the need to invest in volume manufacturing tools.

30.4 PROCESS REGIMES: TEMPERATURE–PRESSURE

Two major process parameters are pressure and temperature. Most microfabrication processes are vacuum/low pressure processes (CVD, etch, sputter, implant), some are room ambient processes (lithography, wet cleaning) and high-pressure oxidation is an exception. The temperature scale extends from 1200 °C diffusions to 850 to 1100 °C oxidation, 300 to 900 °C CVD to room-temperature processes (plasma etch, sputtering, implant, lithography, wet cleaning). Some etch processes use cryogenic cooling down to −100 °C for suppression of spontaneous chemical reactions. Many room-temperature processes can be run at higher temperatures for special purposes: sputtering at 450 °C for aluminium flow, implant at 800 °C for SIMOX wafers or plasma etching at elevated temperatures to reduce residues. Figure 30.1 shows major processes on a temperature–pressure chart. High temperature/high vacuum processes are difficult because of outgassing from vacuum components during high-temperature operation.

There are five main methods that are currently in use to heat wafers, but for example microwaves have been tried (Table 30.2).

The first three methods are used in high-temperature processes and the latter two in low-temperature processes. Some degree of heating and/or temperature control is desirable in almost all tools. In all plasma equipment, there is plasma heating; in ion implantation, the beam flux can heat the wafer considerably; photoresist baking and UV-assisted stabilization depend on hot plate treatments. Whereas older hot plates had no active control of wafer-to-plate contact because there was an

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

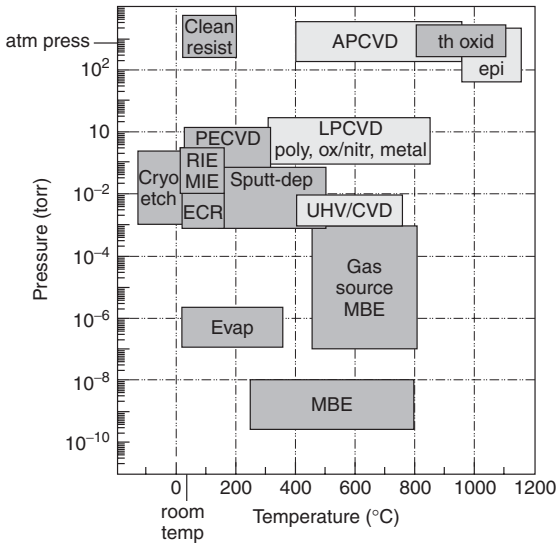


Figure 30.1 Equipment classified on temperature/pressure axes. Reproduced from Rubloff, G.W. & Boronaro, D.T. (1992), by permission of IBM

Table 30.2 Methods for heating

Method	Example
Resistance heating	Furnace
Induction heating	Epitaxial reactor
Photon heating	Rapid thermal processing RTP
Conduction	Horizontal electrodes in PECVD
Convection	Argon backside heating in a sputtering system

inevitable air mattress between the wafer and the hot plate, today the degree of thermal contact can be controlled at will (with hot plate price tags up to $20 000).

In most tools, wafers lie horizontally on electrodes/susceptors, and the electrode or susceptor is heated. Clamping the wafer to the substrate electrode is the simplest way of increasing thermal contact. Both mechanical clamping and electrostatic clamping (ESC) are used. In the former, pins hold the topside of the wafer, which limits usable wafer area, and there is the danger of contamination from the clamp pins. Mechanical clamping is widely used because it is much simpler than ESC. Clamping is essential when wafers are processed in the vertical position (for instance, in ion implanters in which the long acceleration tube (see Figure 15.6) can only be built horizontally) or when wafers are processed face down (as in CMP, Figure 16.2).

Heating (and cooling) can also be affected by direct backside contact with a fluid. Argon is employed in sputtering systems to ramp up wafers to 400 to 500 °C, in a timescale of 10 s. In etchers, the wafer backside is often cooled by helium flow. Some of these gases leak into the process chamber, and the type of heating/cooling gas has to be compatible with the process. In a plasma etcher, energy is supplied to the wafer both from the plasma and from exothermic etching reactions. If no clamping is done, the temperature can easily rise to 80 °C during the first minute of plasma etching, and reach the photoresist glass transition temperature of ca. 120 °C in a few minutes. Steady-state temperatures can be kept below 40 °C indefinitely by backside cooling.

30.5 SIMULATION OF PROCESS EQUIPMENT

Process simulation covers length scales of a few micrometres in both lateral and vertical directions. In process-equipment simulation, the length scale is defined by the tool size, and it can be up to a metre. In practice, this scale difference means that tool simulation is carried out independently of process simulation. In tool simulation, 3D is the norm, but of course, all symmetries in the tool geometry are utilized to reduce computational load.

Typical tool simulation includes temperature distribution, flow patterns and plasma properties. Mass, momentum, energy and charge balances are calculated. Plasma modelling is difficult because it involves so many parameters: collision cross-sections, ionization, attachment, recombination, dissociation, and so on. These plasma reactions must then be combined with surface reactions (deposition or etching). Taken together, these determine, for instance, PECVD film uniformity. For reactors operating in the mass transport–limited regime, flow patterns are of utmost importance. For reactors operating in the surface reaction–limited regime, thermal design is a high priority.

30.6 MEASURING FABRICATION PROCESSES

There are three different aspects that can be measured in a fabrication process: tool, process and wafer. Tool parameters such as RF power, mass flow, process time or electrode temperature are easily measured. Process measurements deal with ionic strength in a cleaning solution, electron and ion energies in plasma or an ion dose. In lithography, exposure time is usually set, but exposure, of course, depends on the UV energy, which drifts with lamp lifetime. Indirect measurements

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

are often much simpler than direct measurements: for example, vacuum chamber base pressure is a good indication of vacuum quality, but mass spectrometry (usually called RGA, for residual gas analysis) can actually identify the residual atoms and molecules, which can be truly significant in understanding vacuum-film interactions. Molecular recognition also helps in trouble-shooting leaks.

Very few measurements are actually done on the wafers during processing. This is understandable because process chamber conditions are often harsh, for example, RF-fields, corrosive gases or high temperatures. Wafer temperature in RTA can be measured by pyrometry during processing. In ultra-high vacuum conditions, surface spectroscopy can be used to monitor deposition processes in real time: reflection high-energy electron diffraction (RHEED) and low-energy electron diffraction (LEED) are routinely employed in MBE systems to check the crystallinity of the growing film. Unfortunately, most deposition processes are operated under conditions in which such systems cannot be used. Film thickness during deposition or etching can be measured by, for example, ellipsometry or interferometry, but such systems are not commonplace.

Measurements can be classified into four categories according to their immediacy:

- *in situ*: during wafer processing in the process chamber
- in-line: after wafer processing in the process tool (e.g., exit load lock)
- on-line: in the wafer fab by wafer fab personnel
- *ex situ*: outside the analytical laboratory by expert users.

In situ resist development monitoring with an interferometric end-point detector can improve linewidth control considerably. It can compensate for changes in exposure dose, resist (de)composition, developer concentration and temperature or resist bake drifts and shifts, which could easily result in 10% development time differences.

Plasma etching is almost always monitored in real time, in order to determine the end point and to prevent excessive etching of the substrate or the underlying film. Optical emission spectroscopy (OES) is commonly used: the intensity of some suitable excited species in the plasma is monitored with optical systems, including a wavelength selective detector. In fluorine plasmas, a signal at $\lambda = 704$ nm (from excited fluorine atoms) can be used. During etching, the signal is small because there is little free fluorine: most of it is bound as reaction products, such as SiF_4 or WF_6. At the etching end-point, free fluorine intensity increases because it is not consumed by the reaction. A more selective method would be the monitoring of reaction products themselves. This must be developed for every process individually. Nitrogen signal (396 nm) is suitable for monitoring nitride etching: there will be a sharp drop in nitrogen signal when all the nitride has been etched away. OES does not, however, measure wafers but, rather, the process.

One of the oldest applications of *in situ* monitoring is the quartz crystal microbalance (QCM) film-thickness control during evaporation and sputtering. The QCM is placed in the same atom flux as the wafers, and therefore it experiences the same film deposition. Mass change is detected as a frequency change and converted to film thickness. The resonance frequency of the QCM is given by

$$f = v_{\mathrm{tr}}/2x \tag{30.1}$$

For quartz wafer of 500 μm thickness with transverse wave velocity of 3340 m/s, this translates to 3.3 MHz. The frequency drop due to thickness increase is given by

$$\Delta f = -2f^2\Delta x/v_{\mathrm{tr}} \tag{30.2}$$

Taking into account the fact that the deposited film density differs from that of quartz (but neglecting that its elastic properties differ), we get the thickness from the frequency change:

$$\Delta x = (v_{\mathrm{tr}}\rho_{\mathrm{quartz}})\Delta f/(-2f^2\rho_{\mathrm{film}}) \tag{30.3}$$

With a 1 ppm frequency shift easily detectable, the minimum thickness change that can be seen is of the order of angstroms. Temperature sensitivity of QCM is 0.5 ppm/°C, which has to be accounted for because deposition is usually accompanied by temperature rise.

In-line tools are located, for example, in load locks or cool down chambers, and they measure wafers immediately after, but not during, processing. Having the instrument outside the process chamber helps because the ambience is usually benign: nitrogen or vacuum atmosphere without RF-fields, plasmas or toxic gases.

On-line measurements constitute the bulk of measurements in wafer fab. These include measurements of standard film-thickness (ellipsometry, reflectometry), sheet resistance, implant damage by thermal waves, step height by profilometer, and so on. Some measurements, such as those for sheet resistance or film thickness, are performed in seconds; while some, such as those for sample preparation or pumpdown (SEM, AFM), require a few minutes.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Ex situ measurements include physical, chemical and structural measurements. Transmission electron microscopy (TEM), secondary ion mass spectrometry (SIMS) and Rutherford backscattering spectrometry (RBS) are also slow methods, and can be bought as services from outside contractors.

Surface analytical methods are problematic because sample transfer from the process chamber to the analytical chamber takes some time and gases and vapours adsorb on the sample surface and disguise the original surface signal. In-line tools do exist for integrated surface analysis, for example, RIE etch chamber connected to an X-ray photoelectron spectrometer (XPS), but such systems are for basic research only.

30.7 EXERCISES

1. By how much will the wafer temperature rise during implantation of arsenic ions of energy 100 keV and dose $10^{15}\,\mathrm{cm}^{-2}$ with a current of 1 mA on a 200 mm wafer? Make simplifying assumptions as needed.
2. In sputtering, ca. 10 to 20 $\mathrm{mW/cm^2}$ of energy is supplied to the surface (heat of condensation, kinetic energy of sputtered particles, ion and electron bombardment and ion neutralization each contribute ca. 2–5 $\mathrm{mW/cm^2}$). How much do wafers heat up during sputtering?
3. If the oxidation furnace is ramped up at 10 °C/min from a stand-by temperature of 800 °C, and ramped down from the process temperature at 5 °C/min, what is the process time for (a) 15 nm dry oxide at 900 °C; (b) for 300 nm wet oxide at 1000 °C?
4. Calculate the minimum deposition rate that can be monitored by a QCMB sensor if the wafers are heated by the deposition process at 3 K/min.

REFERENCES AND RELATED READINGS

Loewenstein, L. *et al*: First-wafer effect in remote plasma processing: the stripping of photoresist, silicon nitride and polysilicon, *J. Vac. Sci. Technol.*, **B12** (1994), 2810.

Moslehi, M.M. *et al*: Single-wafer integrated semiconductor device processing, *IEEE TED*, **39** (1992), 4–32.

Rubloff, G.W. & Boronaro, D.T.: Integrated processing for microelectronics science and technology, *IBM J. Res. Dev.*, **36** (1992), 233.

Schuegraf, K.: Single-wafer process technology: enabling rapid SiGe BiCMOS development, *IEEE TSM*, **16** (2003), 121.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

31

Tools for Hot Processes

Thermal treatments constitute a major fraction of front-end processes. Traditionally, the horizontal tube furnace (see Figure 13.1) has been the workhorse for thermal processing (for oxidation, diffusion, annealing), but more recently, vertical furnaces and rapid-thermal processors (RTP) have entered the scene.

31.1 HIGH TEMPERATURE EQUIPMENT: HOT WALL VERSUS COLD WALL

Two main varieties of high-temperature systems exist: hot wall and cold wall. Hot-wall systems remain hot constantly, usually by resistive heating as in horizontal furnaces. Cold-wall systems heat only the wafers and the actively cooled system walls remain at room temperature. By analogy with kitchen equipment: an oven is a hot-wall system, a microwave oven is a cold-wall system. Warm-wall systems do exist: system walls are heated unintentionally by the process but they remain at a much lower temperature than the wafers.

Large thermal masses in hot-wall systems provide excellent temperature uniformity but very slow temperature ramp rates: 0.1 °C temperature uniformity and 5 to 10 °C/min ramp-up rates, and even slower cooling rates. New vertical furnaces have an order of magnitude higher ramp rates: tens of degrees per minute. Thermocouples are used for temperature monitoring. In hot-wall CVD systems, deposition takes place on all hot walls and successive depositions build up thick films on walls. Film cracking and particle generation are especially probable when two different films are deposited at different temperatures.

In cold-wall systems, only the wafers are heated, and the rest of the system stays cool, which enables faster temperature ramp rates and less deposition on the walls (because chemical reactions are exponentially temperature-dependent). Heating can be achieved by inductive coils (as in epitaxy), by a susceptor/bottom electrode that is kept at a high temperature or by lamps (in rapid-thermal processing, RTP).

31.2 FURNACE PROCESSES

Thermal oxidation is the prototypical hot-wall furnace process. Dry oxidation for a 25 nm oxide is shown in Figure 31.1 and Table 31.1. The process consists of ramp-up, oxidation, post-oxidation anneal (POA) and ramp-down.

Wafer cleaning before all high-temperature processes is essential but in order to also guarantee tube cleanliness, chlorine cleaning can be done prior to thermal oxidation. This process reduces metallic contamination, much like RCA-2 clean, which uses HCl; in fact, HCl has been used as a furnace cleaning agent but today, organic chlorocompounds such as 1,2-dichloroethene (DCE) are used (see Figure 13.1). Alternatively, some chlorine-containing gases can be used during oxidation.

Open-tube furnaces are flushed with nitrogen during wafer loading, and this is usually effective in removing residual water vapour. However, even 100 ppm of residual water vapour will change dry oxidation rates, and 5 ppm of oxygen will lead to titanium silicide deterioration. Double tubing is used if better atmospheric control is required, but loadlocked systems must be used when exact atmospheric control is mandatory. It is useful to have a small, controlled oxygen flow during ramp-up to prevent thermal nitridation of the silicon surface, and accept minor oxidation instead, but of course this is not applicable for very thin oxides.

Actual oxidation time can be a very small fraction of total process time, as in the horizontal tube gate oxidation example in Table 31.1. An optional POA densifies the film, but does not, in the first approximation,

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

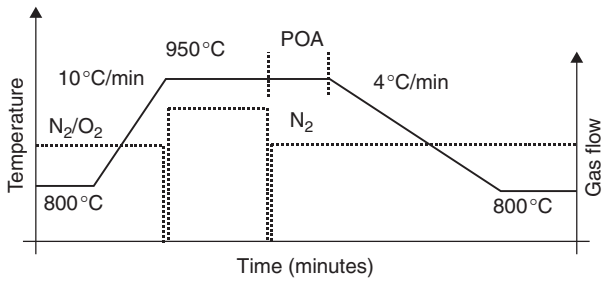


Figure 31.1 Thermal and gas-flow ramping during oxidation in a horizontal furnace

Table 31.1 Gate oxidation (25 nm thick dry oxidation)

Wafer cleaning RCA-1 (NH_4OH:H_2O_2)
organic impurity removal
Wafer cleaning RCA-2 (HCl:H_2O_2)
metallic impurity removal
Dip in dilute HF (1/100; 30 secs)
native oxide removal
Rinse & dry wafers
Boat insertion speed 25 cm/min
(nitrogen flow to prevent oxidation)
Furnace standby temperature 800 °C
Ramp temperature from 800 to 950 °C in N_2/O_2
(15 min, ramp rate 10 °C/min)
Introduce oxygen
(mass flow controlled, 4 slpm)
Oxidize for 35 min at 950 °C
(target thickness 25 nm)
Shut off oxygen flow; introduce nitrogen
Post-oxidation anneal (POA) in nitrogen
(20 minutes at 950 °C)
Cool down to 800 °C
(40 min in nitrogen, ramp rate 4 °C/min)
Unload wafers at 800 °C
(total process time 110 min)
Measurement for thickness and uniformity
Ellipsometry/reflectometry

affect its thickness. POA can also be used to tailor fixed oxide charges (Q_f): while oxidation temperature is, by and large, determined by thickness requirement, POA temperature can be higher, which leads to reduced Q_f density.

31.3 RAPID-THERMAL PROCESSING/ RAPID-THERMAL ANNEALING

Rapid-thermal processors, or RTP systems, have emerged as solutions to some of the difficulties discussed above: in silicide anneal, oxygen must be eliminated and this is easier in a single-wafer tool. RTP emerged early on as an ion implantation–control tool: the implanted wafer was annealed in RTP and measured for sheet resistance in a matter of minutes, as against hours if furnace annealing was used.

Rapid-thermal processing is an alternative to resistively heated tube furnaces. Rapid heating is brought about by either of the following two methods: switching on powerful lamps, or by rapidly transferring the wafer(s) into a hot zone. Three designs for RTP systems are shown in Figure 31.2.

Tungsten halogen lamps deliver a kilowatt or two and a bank of lamps is needed, while a single xenon arc lamp can deliver tens of kilowatts. Ramp rates of the order of 50 to 300 °C/s are used in RTP, a factor of 1000 higher than in horizontal furnaces. The arc-lamp output is in the visible and near infrared, while the tungsten-lamp spectrum extends to 4 μm. This leads to some differences in processes because high-energy photons can contribute to, for example, oxidation.

Lamp geometry is important for uniform processing (Figure 31.3). Large thermal non-uniformities, for example, centre-to-edge temperature differences, may reach 100 °C during ramping, which will result in detrimental crystal slips when the elastic deformation limit is exceeded, as discussed in connection with Equation 4.8. Cooling is usually by natural convection and 50 °C/s is typical. This cannot be affected much.

In addition to annealing, RTP can be used for oxidation (known as RTO) and for CVD (RTCVD). Rapid-thermal oxidation is not significantly faster than furnace oxidation when it comes to oxidation rates, but from the equipment point of view it is: loading-ramping-oxidation-cooling cycle can take a few minutes compared to hours in furnace processing.

Lamp spectrum has implications for temperature measurement: pyrometry is a non-contact method that can monitor wafer temperature in real time, but its operating wavelength must not overlap with that of the heating source. Pyrometry is based on the Stefan–Boltzmann

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

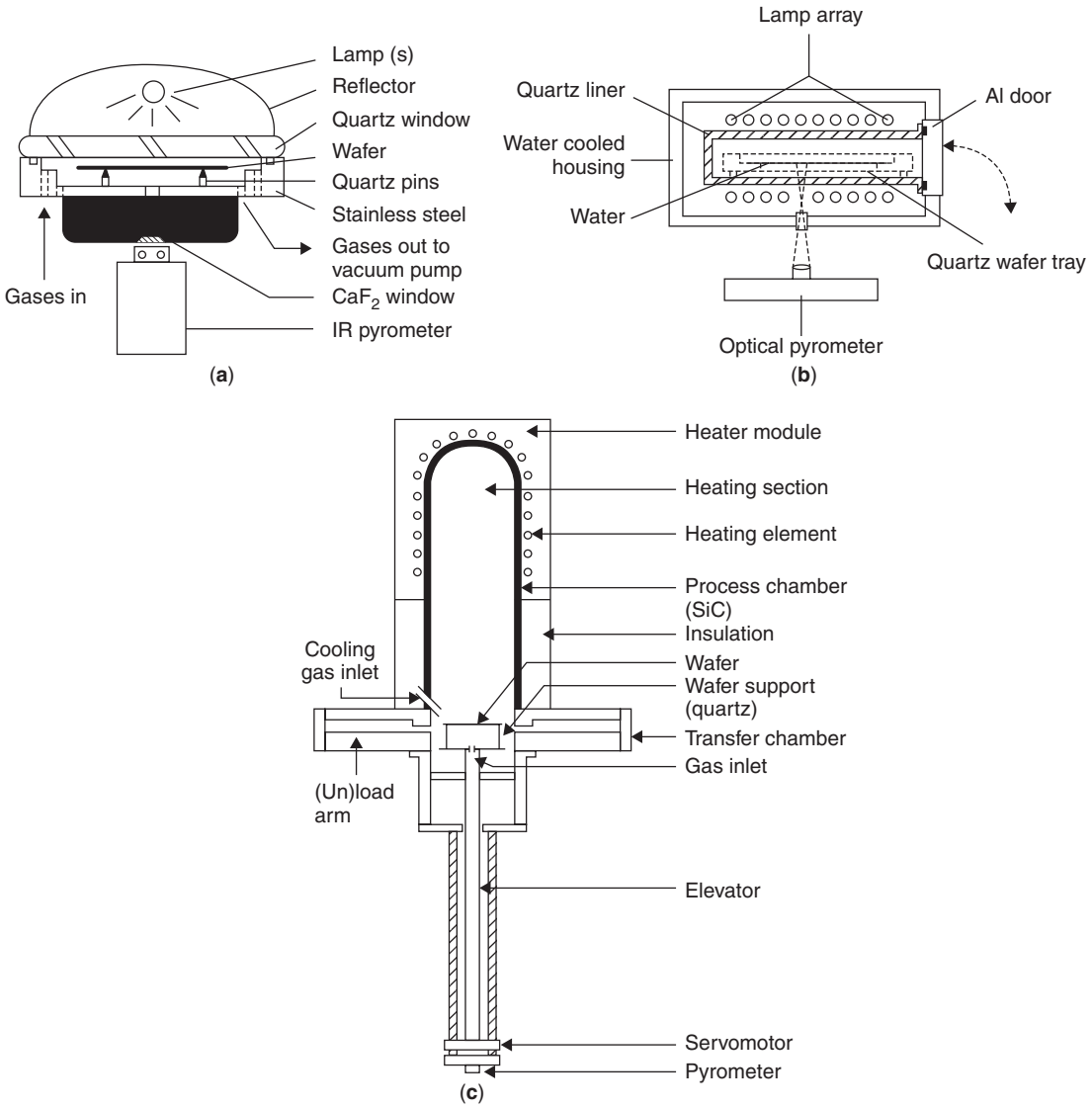


Figure 31.2 RTP systems: (**a**) arc-lamp heated, cold-wall system; (**b**) tungsten-lamp heated, warm-wall system and (**c**) resistively heated fast ramp, hot-wall system. Reproduced from Roozeboom, F. & Parekh, N. (1990), by permission of AIP

law of emitted power

$$P = \varepsilon\sigma T^4$$

where the Stefan–Boltzmann constant is $\sigma = 5.6697 \times 10^{-8}\,\mathrm{W/m^2\,K^4}$.

Emissivity ε ranges from $\varepsilon = 1$ for an ideal black body to $\varepsilon = 0$ for a white body. Silicon emissivity is strongly dependent on charge-carrier density, temperature and wafer thickness in the range up to ca. 600 °C. Above 600 °C, silicon has reasonably constant emissivity of ca. 0.7, but minor changes in emissivity result in large temperature errors. For example, oxide films on silicon act as interference filters and change emissivity from 0.71 to 0.87 when oxide thickness increases from 0 to 400 nm. Below 600 °C, thermocouples are employed. Thermocouples suffer from RTP thermal cycling and contact to silicon is not necessarily reproducible. Metallic contamination from a thermocouple is also an issue.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

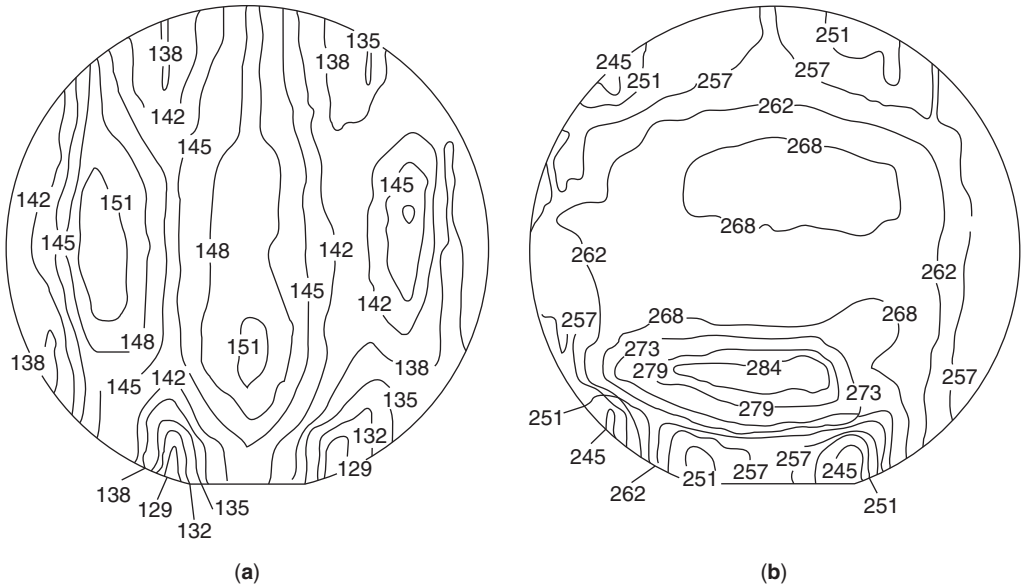


Figure 31.3 Rapid-thermal oxidation uniformity: (**a**) vertical lamp bank geometry can be seen in oxide thickness chart and (**b**) gas-flow patterns are seen in oxide thickness: incoming gas cools the wafer near the flat, and wafer edges are cooler than the centre. Reproduced from Deaton, R. & Massoud, Z. (1992), by permission of IEEE

Metal chamber RTP tools are water-cooled to keep them cold; quartz chambers are allowed to heat up; that is, they are warm-wall systems. System walls do not contribute to contamination because evaporation and desorption of material is minimized by keeping the temperature low.

A hybrid technology between resistively heated furnaces and RTA is the fast ramp furnace. A heater, typically made of silicon carbide, is kept at a very high temperature, and the wafers are rapidly brought to its vicinity. A massive radiation source emits at much longer wavelengths than RTP lamps, and thermal equilibrium is possible. This ramping arrangement can significantly reduce wafer emissivity variation and temperature non-uniformities. Ramp rates for fast-ramping systems are 10 to 100 °C/s, somewhat lower than in RTP systems.

Rapid annealing times are typically tens of seconds (Figure 31.4), very fast compared to 30 to 60 min furnace anneals. In order to reduce unwanted diffusion during annealing, high temperature/short time combination has been refined to zero-time anneal (also known as *spike anneal*): the anneal temperature refers to the highest temperature reached by the system, but power is turned off immediately after reaching that temperature.

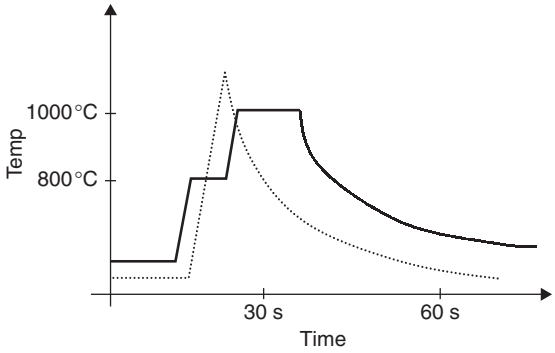


Figure 31.4 Temperature profile in rapid-thermal annealing: solid curve: 1000 °C, 10 s anneal; dashed curve: 1100 °C spike anneal (zero-time anneal)

The main features of furnace and RTP systems are compared in Table 31.2.

When oxide thicknesses are scaled down, rapid-thermal oxidation becomes more competitive but furnaces are still the workhorses of oxidation. In implant activation anneal, RTA is the only choice when shallow junctions are made, as discussed in Chapter 25.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Table 31.2 Comparison of furnace and RTP processes

Furnace	Rapid-thermal processing
Batch	Single wafer
Hot wall	Cold wall
Long time	Short time
Small dT/dt	Large dT/dt
Indirect	Direct temperature measurement

31.4 EXERCISES

1. What should the oxygen flow be in a horizontal batch furnace to make sure that oxidation is not mass transfer–limited? Write out and justify the assumptions you need in your solution.
2. If reproducibility and other uncertainties in a batch-loading furnace limit the shortest practical oxidation time to 15 min, what is the thinnest gate oxide that can be grown at 1000 °C, at 950 °C, at 900 °C and 850 °C? What are the corresponding CMOS linewidths?
3. How rapid is RTP? Calculate how long the heat pulses must be to result in thermal equilibrium of the whole silicon wafer. Thermal diffusivity in silicon is 0.80 cm^2/s at room temperature and 0.1 cm^2/s at 1400 °C.

4S. Rapid-thermal oxidation (RTO) data is given in the table below. How does RTO compare with furnace oxidation? Data from Deaton, R. & Massoud, Z.: Manufacturability of rapid-thermal oxidation of silicon: oxide thickness, oxide thickness variation and system dependency, *IEEE TSM*, **5** (1992), 347.

Constant time 30 s		Constant temperature 1050 °C	
Temp	Thickness	Time	Thickness
950 °C	44 Å	30 s	75 Å
1050 °C	75 Å	150 s	158 Å
1150 °C	145 Å	270 s	240 Å

5. What temperature error does emissivity change from 0.71 to 0.87 cause in rapid-thermal oxidation?
6. What power rating does an RTP system for 300 mm wafers need if its maximum operating temperature is 1200 °C?
7. Anneal time and junction depth are connected as follows: $x_j = k \times (Dt)^{1/3}$. If junction depth is ca. 100 nm in 0.25 μm technology and the corresponding anneal time is 10 s, what is the anneal time for 0.1 μm technology? What is the junction depth?

8S. Typical furnace anneal activation is 950 °C/30 min, but in RTA, a much higher temperature and a much shorter time are used. Compare junction depths that can be made by RTA and FA. Use implant conditions of 20 keV boron, $10^{15}\,cm^{-2}$ into a phosphorous-doped wafer with $10^{15}\,cm^{-3}$.

REFERENCES AND RELATED READINGS

Bensahel, D. *et al*: Front-end, single wafer diffusion processing for advanced 300 mm fabrication line, *Microelectron. Eng.*, **56** (2001), 49.

Bratschun, A.: The application of rapid thermal processing technology to the manufacture of integrated circuits – an overview, *J. Electron. Mater.*, **28**(12) (1999), 1328 (special issue on RTP).

Deaton, R. & Massoud, Z.: Manufacturability of rapid-thermal oxidation of silicon: oxide thickness, oxide thickness variation and system dependency, *IEEE TSM*, **5** (1992), 347.

Endoh, T. *et al*: Influence of silicon wafer loading ambient on chemical composition and thickness uniformity of sub-5 nm thickness oxides, *Jpn. J. Appl. Phys.*, **40** (2001), 7023.

Fair, R.B., *Conventional and rapid thermal processes*, in C.Y. Chang & S.M. Sze (eds.): *ULSI Technology*, McGraw-Hill, 1996.

Roozeboom, F. & Parekh, N. Rapid thermal processing systems: a review with emphasis on temperature control, *J. Vac. Sci. Technol., B*, **8**(6) (1990), 1249.

Saga, K. *et al*: Influence of silicon-wafer loading ambients in an oxidation furnace on the gate oxide degradation due to organic contamination, *Appl. Phys. Lett.*, **71** (1997), 3670.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

32

Vacuum and Plasmas

When we talk about vacuum processes, pressures can be anything from slightly below atmospheric pressure down to 10^{-11} torr. Reduced pressure processes would be a more accurate description, but the word 'vacuum' is handy. In evaporation, a vacuum of 10^{-6} torr is typical; in sputtering, 1 to 10 mtorr is used, depending on system configuration (DC, RF, magnetron). CVD process pressures range from atmospheric to ultra-high vacuum. Units of pressure (and flow) are many, and the reader is referred to conversion tables (Appendix B).

Transport of ejected atoms or ions from the target to substrate requires vacuum to prevent collisions and flux divergence. Mean free path (λ, MFP), or the distance travelled by atoms between collisions, is a useful measure of transport.

$$1/\lambda = \sqrt{2} \times \pi d^2 n \tag{32.1}$$

where n is the atom density and d is the molecule diameter.

This can be approximated for diatomic molecules at around 300 K as λ (m) $\approx 5 \times 10^{-5}$/P (torr), which gives $\lambda \approx 65$ nm for nitrogen ($d = 3.75$ Å) at room temperature and 1 atm (760 torr) pressure, and 5 cm at 1 mtorr pressure.

The Knudsen number, *Kn*, relates mean free path and reactor chamber size:

$$Kn = \lambda / L \tag{32.2}$$

where L is the characteristic dimension of the chamber. Kn > 1 is equivalent to collisionless transport across the vacuum vessel. This regime is known as *molecular flow* and the equipment molecular beam epitaxy (MBE), refers to the molecular flow regime since it is atoms, not molecules, that are transported in MBE. In the regime Kn < 0.01, fluid dynamics has to be taken into account.

32.1 VACUUM-FILM INTERACTIONS

Contamination from the gas phase to the surface can be estimated from kinetic gas theory. The impingement rate of molecules on the surface is given by

$$z = P/\sqrt{2\pi mkT} \tag{32.3}$$

where P is pressure, m is mass and T is absolute temperature.

If the residual gas is assumed to be nitrogen ($m =$ 28 amu), then at 10^{-6} torr (1.33×10^{-4} Pa) $z = 3.8 \times 10^{18}/\mathrm{m}^2\,\mathrm{s}$. A monolayer of residual gases will be adsorbed on sample surface in a timescale:

$$t_{\text{monolayer}} = N_{\text{surf}}/\delta z \tag{32.4}$$

where δ is sticking probability and N_{surf} is the density of surface sites, which can be taken as approximately ${N_{\text{vol}}}^{2/3}$. For silicon, N_{vol} is $5 \times 10^{22}\,\mathrm{cm}^{-3}$, and N_{surf} is ca. $10^{15}\,\mathrm{cm}^{-2}$. Under the conditions described above, monolayer formation time is ca. 1 s under the assumption of unity δ (which gives a shortest possible monolayer formation time) (Figure 32.1). For oxygen, the sticking coefficient is estimated to be ca. 0.1 (but sticking coefficient is strongly temperature-dependent). Residual gases are not similar in their effects: oxygen, water vapour and hydrocarbons are much more problematic than nitrogen, carbon monoxide, carbon dioxide or argon. The sticking coefficient can be tailored by surface preparation: for instance, HF-last treated surfaces are much more resistant to water adsorption than RCA-1 treated surfaces.

Adsorbed species have a characteristic desorption time that is exponentially dependent on activation energy,

$$\tau = (1/\nu)\exp(E_{\text{a}}/kT) \tag{32.5}$$

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

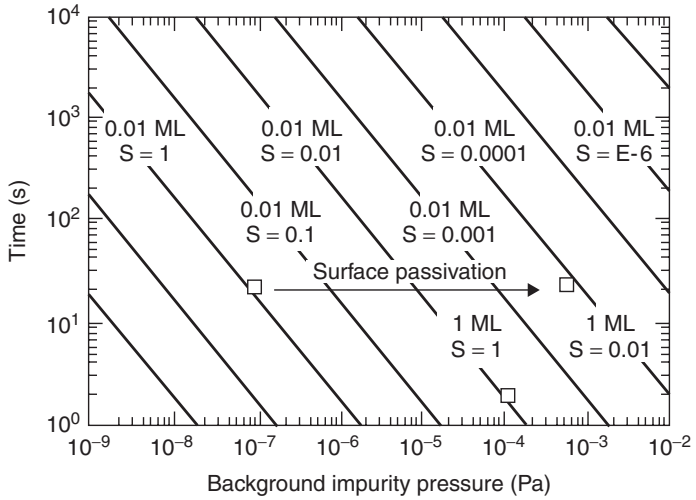


Figure 32.1 Monolayer (ML) and 0.01 ML formation times as a function of pressure and sticking coefficient (S). Surface can be passivated by, for example, HF-treatment. Reproduced from Grannemann, E. (1994), by permission of AIP

The order of magnitude for the frequency factor ν is $10^{13}\,s^{-1}$, which describes a simple harmonic oscillator with frequency kT/h. Chemisorbed species have an E_a of ca. 1 eV and physisorbed species, an E_a of 0.4 eV, which translate roughly, at room temperature, to hours and microseconds, respectively.

Impurities in the vacuum chamber will be incorporated into the growing film. Partial pressure of the impurities must be considered together with the deposition rate in order to determine the concentration of impurities in the film. Table 32.1 shows how gas-phase impurities are incorporated into growing films as a function of residual gas pressure.

At 10^{-6} torr, impurities deposit approximately at a rate of one monolayer per second (~0.1 nm/s). Even the very high rate of 100 nm/s, which corresponds to ca. 1000 atomic layers per second, will result in 0.1% impurity in the film. Purities of typical starting materials for PVD are 99.999%. Poor vacuum can therefore contribute many orders of magnitude more impurities into film than the target materials. Of course, not all impurities are equal: some manifest themselves much more strikingly than others. Unity sticking coefficient presents the worst case. At base pressures of 10^{-9} torr, target purity starts becoming a limiting factor.

Deposition rates in batch systems are usually much slower than in single-wafer systems: an order of magnitude difference is not unusual, and therefore throughput rather than deposition rate is often mentioned for batch systems. But as shown in Table 32.1, film quality is related to deposition rate, not to throughput.

Table 32.1 Fraction of foreign atoms incorporated into growing film (unity sticking coefficient; worst case estimates)

Partial pressure (torr)	Deposition rate (nm/s)			
	0.1	1	10	100
10^{-9}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
10^{-8}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
10^{-7}	10^{-1}	10^{-2}	10^{-3}	10^{-4}
10^{-6}	1	10^{-1}	10^{-2}	10^{-3}
10^{-5}	10	1	0.1	0.01

32.2 VACUUM PRODUCTION

Starting from the ideal gas law

$$p = NkT/V \tag{32.6}$$

we can get a feeling for vacuum production. Vacuum production means a change (decrease) in the number of atoms N over time, dN/dt. We use the following definitions:

Particle density:	$n \equiv N/V$	in units atoms/m^3
Flux:	$J \equiv dN/dt$	in units atoms/s
Pumping speed:	$S \equiv -J/n$	in units m^3/s, a.k.a. volumetric flow rate

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Time evolution of pressure can be written as

$$\mathrm{d}p/\mathrm{d}t = (\mathrm{d}N/\mathrm{d}t)kT/V = -nSkT/V \qquad (32.7)$$

which can be solved to yield

$$p = p_0 \exp(-St/V) \qquad (32.8)$$

Pressure drops exponentially over time with characteristic time τ proportional to V/S.

Low to medium vacuum (10^5–0.1 Pa) can be produced by rotary vane pumps, rotary piston pumps, roots blowers and sorption pumps. High vacuum ($0.1–10^{-4}$ Pa) is produced by capture pumps (cryopumps, getter pumps) and momentum-transfer pumps (turbomolecular pumps, diffusion pumps). Capture pumps capture and hold all the gas and therefore they need forepumps because of limited holding capacity; and they have to be regenerated regularly. Momentum-transfer pumps, on the other hand, require roughing pumps because they cannot start operation at ambient pressure.

Crossover is the pressure at which the high vacuum pump is connected to the chamber. For capture pumps, this is calculated from torr-litre specification (Pa-L/s), by dividing with the chamber volume. Capture pumps hold the pumped material, and therefore knowledge of chamber volume is essential. Capture pumps often bring the pressure down faster than roughing pumps, because the pumping speed of a mechanical roughing pump gets worse at lower pressures.

Ultimate pressure that can be reached by a pumping system is determined by pumping speed and vacuum chamber leak rate. We need the concept of conductance to estimate this: conductance is flow divided by gas density difference on the two sides of the vacuum system. Its unit is thus cubic metre per second. Conductances add like capacitors in series:

$$1/C_{\mathrm{tot}} = (1/C_1) + (1/C_2) \qquad (32.9)$$

Maximum conductance is limited by the orifice opening, and further limited by tube conductance that leads from the orifice.

The number of atoms leaking in from the outside is given by

$$\mathrm{d}N/\mathrm{d}t = J = -C\,\Delta n \qquad (32.10)$$

For high vacuum, Δn is equal to the density of the gas outside the system (approximating high vacuum with $n = 0$), which, for STP conditions, is $n = 2.4 \times 10^{25}\,\mathrm{m}^{-3}$. Identifying flux J as the leak, we get from the ideal gas law (Equation 32.6)

$$pS = kTJ_{\mathrm{leak}} = kTnC \qquad (32.11)$$

and the ultimate pressure that can be reached is then given by

$$p_{\mathrm{ult}} = kTnC/S \qquad (32.12)$$

If the leak rate is $3.8 \times 10^{15}\,\mathrm{s}^{-1}$ and 1000 L/s pump is employed, the base pressure is ca. 1.6×10^{-5} Pa or 1.2×10^{-7} torr. Ultimate base pressures are produced by cryopumps or getter pumps, with values in the range of 10^{-11} torr. MBE systems operate at such base pressures.

The theoretical maximum pumping speed is derived from kinetic theory as

$$S = (A/4)v_{\mathrm{ave}} \qquad (32.13)$$

where A is the inlet area and $v_{\mathrm{ave}} = \sqrt{(8kT/\pi m)}$ is the molecular average speed. This represents the case in which all atoms impinge only in one direction, with no return flux. Real life pumping speeds of diffusion pumps can be 50% of the theoretical maximum value, but rotary pumps fare much worse. Pumping speed is usually specified for nitrogen, and light gases hydrogen and helium are difficult to pump. Water vapour is difficult to remove because its desorption rate is very low.

Gases will adsorb on surfaces when energetically favourable surface sites are available. Adsorbed gases are 'surface gases' as opposed to 'volume gases'. The latter are related to chamber volume; the former to chamber wall area. Large surface area equals large quantity of adsorbed gases. The analogy is with water in a bucket: initially each cup will decrease the water level in the bucket by a cupful until almost all the water is removed. When almost all water has been removed, the remaining water is found in cusps that are smaller than the cup, and therefore each removal cycle removes less than a cupful. This points to the importance of surface finish in vacuum chamber manufacturing. Pumping can be limited by surface gas desorption. It can be helped by heating or UV radiation.

Ultra-high vacuum (UHV) chamber materials and surfaces, valves, and all components must be compatible with baking, which is done to outgas the adsorbed species. UHV systems are baked at elevated temperatures; MBE systems, for instance, are baked at 200 °C for 24 h, every 30 days.

The pressure can be brought down by a multiple-stage vacuum system. The sputtering system may have three levels of vacuum:

- vacuum cassette lock, pumped down to 10 to 100 mtorr by a mechanical pump;
- transfer chamber, pumped down to 0.01 mtorr by a turbopump;
- process chamber, cryopumped to 10^{-6} mtorr.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

If transfer and process chambers take only one wafer at a time, the volume to be pumped can be made very small. In a batch deposition system, the vacuum vessel volume is easily 100 L, and the corresponding pumpdown time is of the order of an hour, or hours, and somewhat less with a loadlock.

Loadlocks come in two varieties: single loadlocks, or separate entry and exit loadlocks. The former loadlocks are used when the process time is long compared to transfer time. Load locks serve many purposes: they protect the main chamber from clean room air, and the clean room air from harmful or toxic gases that have been used in the process. They can also protect the wafers from the atmosphere: for instance, after aluminium plasma etching, chlorine residues remain on the wafer (in the resist and on aluminum surfaces), and if the wafer is taken into cleanroom air with 45% humidity, the chlorine will react with water vapour, and HCl is formed:

$$2AlCl_3 + 3H_2O \longrightarrow Al_2O_3 + 6HCl \qquad (32.14)$$

Hydrogen chloride will etch aluminium locally. This is termed *corrosion*. Exit loadlock can be used to strip the photoresist in oxygen plasma, and to passivate aluminum surfaces to Al_2O_3.

In an evaporator, there is just residual gas to be pumped out; but in sputtering and UHV-CVD systems, we feed in process gases intentionally, and must be able to pump them out. Despite similar base vacuum, the process vacuum in sputtering and UHV-CVD is 1 to 10 mtorr, 3 orders of magnitude higher than the base vacuum, and 10 to 100 Pa-L/s pumps can be used.

32.3 PLASMA ETCHING

Plasma generation has a major role in etching, sputtering, ion implantation, photoresist stripping and PECVD. Plasmas used in microfabrication are low-temperature, low-density plasmas (ca. $10^{10}\,cm^{-3}$ ion density), compared to, for example, welding or fusion plasmas. In microfabrication, high-density plasma (HDP) means ion density in excess of $10^{11}\,cm^{-3}$. The degree of ionization is still fairly low: at 1 mtorr pressure, it is only a fraction of a percent.

Plasma etching has a very high number of parameters that need to be controlled (Figure 32.2). This makes plasma etching difficult, both experimentally and simulation-wise. Furthermore, the machine parameters affect plasma parameters, which, together with surface reactions, determine the final outcome: rate, selectivity and other process responses of interest.

32.3.1 Direct plasmas

Plasma etch reactors can be classified in various ways, and the following is just one. A parallel-plate diode reactor with two electrodes, one powered and one grounded, is a basic construction for an etcher (see Figure 11.9). It is called *RIE* when the wafer(s) is (are) on the biased electrode, or PE when the wafer(s) is (are) on the grounded electrode. Wafers are placed on electrodes that produce the plasma; plasma density, sheath voltage and ion bombardment that hit the wafers are thus dependent on each other, and cannot be controlled independently. Despite this seemingly inconvenient state of affairs, this arrangement is very widely used because of its simplicity. 13.56 MHz RF generators are used to create plasmas of typical density $10^{10}\,cm^{-3}$.

32.3.2 Remote plasmas

In remote plasmas, plasma generation takes place in a region outside the wafers, and the wafers see a

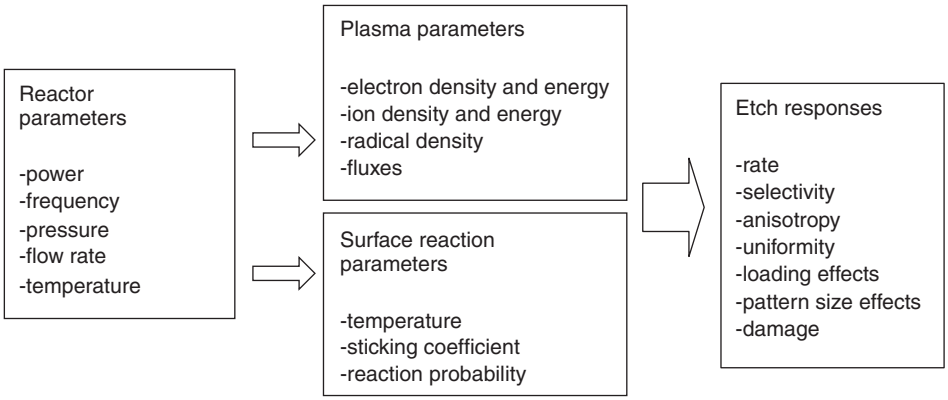


Figure 32.2 Plasma etching parameters and process responses

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

controlled flux by, for example, a separate bias power source. Alternatively, the wafers may be shielded from ions completely by a Faraday cage. Because of this decoupling, high-density plasmas ($10^{11}-10^{12}\,cm^{-3}$) can be achieved, without high sheath voltages or severe ion bombardment on the wafer. Since a high density of ions and radicals means a high concentration of active species, high-density plasmas (HDP) offer higher etch and deposition rates. DRIE reactors use ICP (inductively coupled plasma) and employ 2 to 5 kW power sources for plasma generation.

Higher etch rate, lower damage, easier photoresist removal and higher selectivity favour HDP reactors. Remote plasma reactors are often difficult to scale to large diameters because of the physical separation between plasma and wafer, whereas in parallel-plate reactors, the plasma is naturally 'aligned' to the wafer. But larger wafer sizes make direct plasma reactors less attractive: in order to maintain the same power density, the absolute size of the RF-generator may grow far too big.

32.4 SPUTTERING

The oldest and simplest of sputter deposition systems is the DC-diode system, which consists of a negatively biased plate (target cathode), which is bombarded by argon ions at ca. 100 mtorr pressure (see Figure 5.4). In order to get high deposition rate, high sputtering power has to be used, which leads to high voltage operation. This is undesirable because of damage to thin oxides.

In order to improve DC diodes, RF diode systems were introduced. RF sputtering systems usually work at 13.56 MHz. They can be used to deposit dielectrics, something that is not possible with DC systems because of charging. Electrons oscillating in an RF field couple energy more efficiently to the plasma, and higher deposition rates are possible in RF than in DC, at the same power levels. However, a very high voltage of 2000 V is used.

Magnetron sputtering has emerged as the main configuration. A magnet behind the target creates a field that confines electron movement, and therefore, ionization is much more efficient, leading to high deposition rates at low power (5–20 kW are used, depending on target size). Voltages in magnetron systems are, for example, 500 V (and argon ion energies are 500 eV), clearly lower than in RF diodes. Magnetron sputtering systems work at ca. mtorr pressures (0.1–10 mtorr), with argon flows of 10 to 100 sccm. Impurity-wise, however, sputtering systems are described by their base pressures, which are 10^{-7} to 10^{-9} mtorr because high purity argon sputtering gas (99.9999%) contributes less than background gases.

Sputtering systems have, in addition to plasma generation and vacuum subsystems, many other features: the wafers can be heated, they can be biased and they can be shielded from the plasma by shutters, as shown in Figure 32.3.

32.4.1 Reactive sputtering

Sputtering in a reactive atmosphere, in argon/nitrogen or argon/oxygen mixtures, results in nitride or oxide films, or stuffed films with small amounts of reactive impurities at grain boundaries. Typical applications of reactive sputtering are TiN, Ta_2O_5, ZnO, AlN, TiW:N and WO_3. Often, reactively sputtered films are not stoichiometric, and a (reactive) annealing step (e.g., in oxygen) is needed to improve film quality.

Introduction of small amounts of nitrogen or oxygen into argon plasma does not appreciably change the properties of the discharge or of the growing film, but after a critical partial pressure is reached, the target surface transforms into nitride or oxide, and the plasma discharge is established at another equilibrium. If the reactive gas flow is then reduced, the target remains nitrided/oxidized, and return to initial conditions takes place at much lower partial pressures, that is, reactive sputtering exhibits hysteresis.

32.4.2 Sputter etching and bias sputtering

If the voltages in a sputtering system are switched, and power is applied to the wafer electrode instead of the target, the wafers will experience argon ion bombardment. This is called *sputter etching*. (Sputtering systems can be turned into true plasma etch systems by introducing reactive gases instead of argon. The term *RSE*, for *reactive sputter etching*, was used in the early days of plasma etching.)

If the wafer electrode is biased during sputtering (by a separate power supply), the wafer will experience simultaneous deposition and etching. This will generally densify the film because ion bombardment kicks off loosely bound film atoms, and it also affects film stresses. Geometry of structures is important because argon etching depends on the angle of incidence: convex corners are etched faster, and faceting occurs. This is pictured in Figure 32.4 (PECVD oxide has been etched in argon). Smoothing of sharp corners is beneficial for step coverage in the next deposition step, but such dep-etch (deposition-etch) processes are understandably slow.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

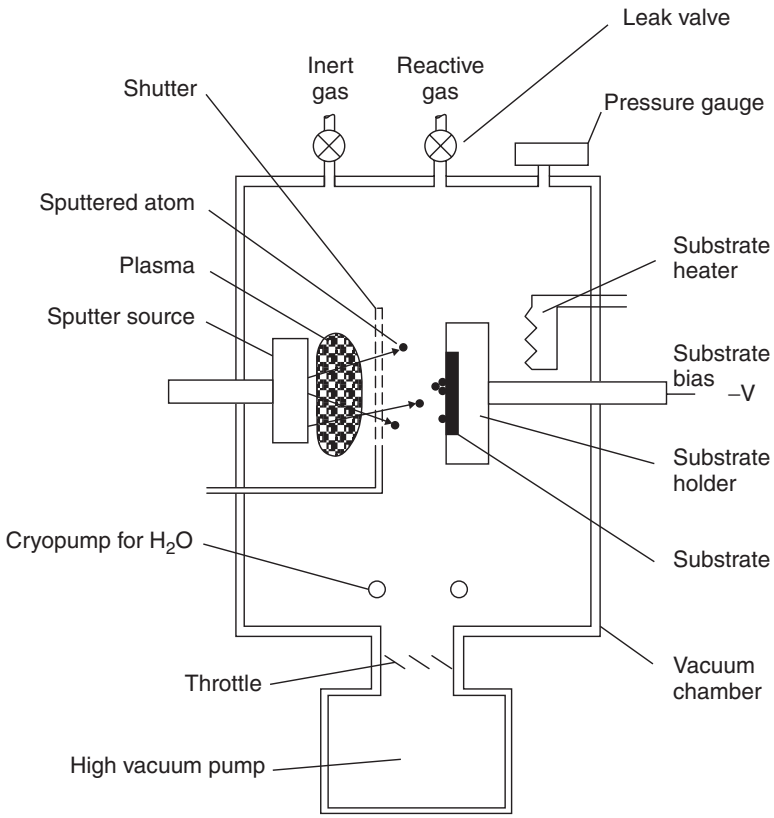
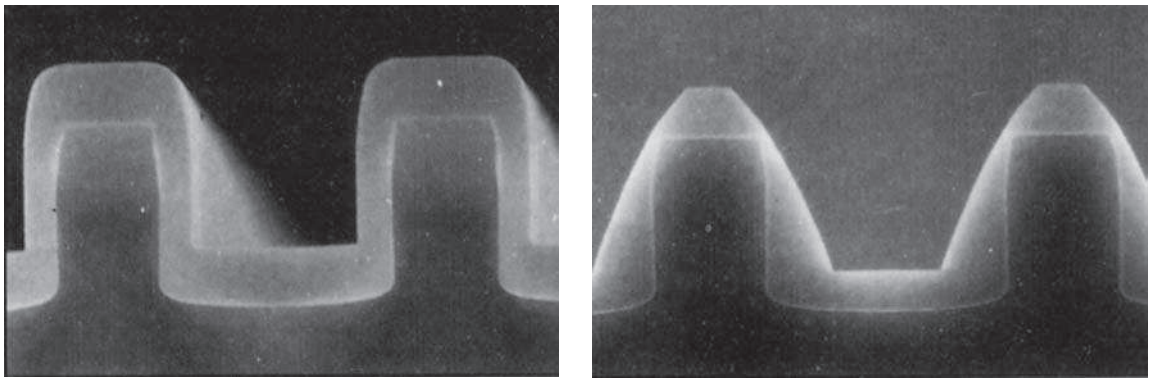


Figure 32.3 Sputtering system. Reproduced from Parsons, R., Sputter deposition processes, in J.L. Vossen & W. Kern (eds.) (1991), by permission of Academic Press



(**a**) (**b**)

Figure 32.4 (**a**) PECVD TEOS oxide profile after deposition and (**b**) after argon sputter etching. Reproduced from Cote, D.R. *et al*. (1995), by permission of IBM

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

32.5 PECVD

PECVD reactors are very much like plasma etchers. From the hardware point of view, the heated electrode is the main difference. Other aspects, such as RF generators, reactive gases and pumping systems, among others, are similar. In etching, high density plasmas (HDP) offer enhanced etch rates; in PECVD, HDP equals enhanced deposition rate and/or improved film quality.

Higher deposition temperature leads to denser, more stable films. This may be useful, but the main advantage of PECVD is low deposition temperature. Typical PECVD temperature is 300 °C, but there is no fundamental lower limit to deposition temperature. Processes at 100 °C have been demonstrated but film properties are strongly temperature-dependent. In particular, hydrogen content of the films increases rapidly as temperature is lowered, and the films become less dense. The above discussion is about first-order effects only: when two reactant gases interact, many things can be different.

Increasing RF power initially increases the deposition rate, because more reactant gases are ionized, fragmented and available for reaction. Further increase in power leads to decreased rate, however: more and more ion bombardment causes sputtering of the growing film.

Utilization is a measure for reactant usage. It is the ratio of atoms incorporated into the film to atoms in incoming gases. Utilization cannot even approach 100% because flow patterns in a reactor cannot be optimized for such a high efficiency. Some metal–organic precursor molecules undergo disproportionation reaction, and only 50% of source gas atoms are available for deposition in the best case.

Deposition takes place not only on the wafers but also on the reactor walls and the electrodes. It is standard procedure to etch these deposited layers away at regular intervals, for example, after every wafer, after a certain thickness has been deposited, when deposition temperature is changed or when the material to be deposited is changed. The similarity of PECVD to RIE is evident from the fact that introduction of CF_4 or NF_3 gas into a PECVD reactor chamber turns it into an etch system. *In situ* cleaning of the PECVD chamber can thus be accomplished easily. NF_3 gas has a nice feature in that it decomposes into gaseous products only, whereas CF_4 or SF_6 are potential sources of carbon and sulphur residues. NF_3 is, however, toxic and hard to handle. It is also a greenhouse gas just like fluorinated hydrocarbons.

32.6 RESIDENCE TIME

The effects of pressure and flow can be deduced from residence time τ (for PECVD and other processes alike):

$$\tau = (p/p_0)(V/F)(273/T) \tag{32.15}$$

where p_0 is a reference pressure of 1 atm.

Residence time is the characteristic time that a molecule spends in the reactor before being pumped away. Increasing the pressure leads to increased residence time, which translates to higher deposition rate: the molecules have a higher probability of being incorporated into the film if they spend more time in the reactor. Increasing the flow will sweep the molecules away faster, leading to smaller τ and lower deposition rate.

32.7 EXERCISES

1. What is the Knudsen number in
   (a) sputtering;
   (b) evaporation;
   (c) MBE;
   (d) RIE.
2. What is the maximum theoretical pumping speed of a diffusion pump with vacuum flange of diameter 10 cm?
3. If the sticking coefficient of a water molecule is 0.01 and the partial pressure of water is 10^{-4} Pa, how long will it take to form a monolayer?
4. What must the leak rate be in an MBE system in order to achieve a base pressure of 10^{-11} torr?
5. What would the crossover pressure be for film purity to become dependent on target purity when a 99.9999% pure target (6N) is used?
6. How deep into aluminium sputtering target will 500 eV argon ions penetrate?
7. Pulsed (Bosch) process DRIE chamber volume is 50 L, flow rate is 200 sccm and operating pressure is 20 mtorr. What is the shortest possible pulsing period?
8. If 5-kW power is applied to aluminium sputtering target of 200 mm diameter, what is the maximum possible deposition rate?
9. XPS measurement takes 15 min. What is the pressure in a XPS chamber?

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

REFERENCES AND RELATED READINGS

Cote, D.R. *et al*: Low-temperature CVD processes and dielectrics, *IBM J. Res. Dev.*, **39** (1995), 437.

Hess, D.W.: Plasma-material interactions, *J. Vac. Sci. Technol.*, **A8** (1990), 1677.

Mahan, J.E.: *Physical Vapor Deposition of Thin Films*, John Wiley & Sons, 2000.

Nguyen, S.V.: High-density plasma chemical vapor deposition of silicon-based dielectric films for integrated circuits, *IBM J. Res. Dev.*, **43**(1–2) (1999), 109 (special issue on plasma processing).

Rossnagel, S.M.: Sputter deposition for semiconductor manufacturing, *IBM J. Res. Dev.*, **43**(1–2) (1999), 163.

Lee, J.T.C. *et al*: Plasma etching process development using *in situ* optical emission and ellipsometry, *J. Vac. Sci. Technol., B*, **14** (1996), 3283.

Loewenhardt, P. *et al*: Plasma diagnostics: use and justification in an industrial environment, *Jpn. J. Appl. Phys.*, **38** (1999), 4362.

Parsons, R., Sputter deposition processes, in J.L. Vossen & W. Kern (eds.): *Thin Film Processes II*, Academic Press, 1991, p. 179.

Somorjai, G.A.: From surface materials to surface technologies, *MRS Bulletin*, **23**(5) (1998), 11.

IBM J. Res. Dev., **43**(1–2) (1999); special issue on plasma processing.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

33

Tools for CVD and Epitaxy

Thermal CVD processes share many equipment features with oxidation and diffusion furnace processes, whereas PECVD is more akin to plasma etching. The epitaxial processes to be discussed here are limited to flow-type silicon CVD epitaxy processes, which share many features with thermal CVD.

CVD reactors are classified by their operating pressure range:

- atmospheric pressure APCVD;
- sub-atmospheric SACVD 10 to 100 torr;
- low-pressure, LPCVD at ~torr;
- ultra-high vacuum, UHV-CVD, 10^{-6} torr (base pressure), 1 to 10 mtorr (operating pressure).

In UHV reactors, the actual process pressures are 1 to 10 mtorr when gases are flowing, much like magnetron-sputtering systems. In both cases, a good base vacuum (of 10^{-6}–10^{-9} torr level) is mandatory for the removal of residual gases from the chamber.

The pressure range has profound effects on the mechanism of film deposition. While temperature affects the rate in a predictable manner (Arrhenius behaviour), pressure has subtler effects: the rate-limiting step can change from surface reaction-limited to transport-limited by a pressure change. Depending on application and reactor design, it may be advantageous to operate in a transport-limited regime in which the temperature dependence is small, but flow control must be accurate. On the other hand, in the surface reaction-limited regime, uniformity of deposition becomes independent of fluid dynamics, but critically temperature-dependent.

33.1 CVD RATE MODELLING

CVD can be modelled with a simple model that bears resemblance to the Deal–Grove model of thermal oxidation. Flux of reactants from the gas flow to the surface is controlled by diffusion through the boundary layer, and film deposition takes place at the wafer surface (Figure 33.1). Flux from the gas phase to the surface is given by

$$J_{\text{gas-to-surface}} = h_g(C_g - C_s) \tag{33.1}$$

where h_g is the gas-phase transport coefficient, C_g is the gas-phase concentration and C_s the surface concentration of reactants. The surface-reaction rate is assumed to be directly proportional to reactant concentration:

$$J_{\text{surface reaction}} = k_s C_s \tag{33.2}$$

Under steady-state conditions, the fluxes are equal

$$J_{gs} = J_s, \text{ or } C_s = C_g/(1 + (k_s/h_g)) \tag{33.3}$$

Conversion from fluxes to rate is given by $R = J_s/n$ where n is atom density in the film.

From the above formula we can recognize two familiar regimes (recall Figure 5.6):

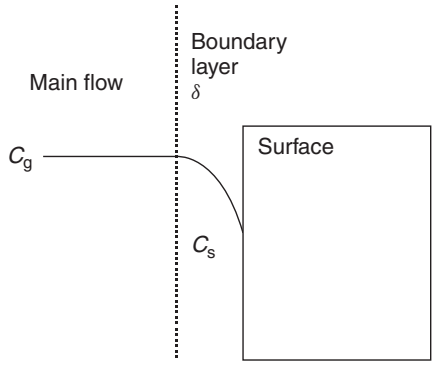


Figure 33.1 Model of gas-phase deposition

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

1. transport-limited deposition, $k_s \gg h_g$; $C_s = (h_g/k_s)C_g$;
2. surface reaction-limited deposition, $k_s \ll h_g$; $C_s = C_g$.

In the former, the reaction rate at the surface is very high and leads to local depletion of reactants. Supply of reactants by the gas flow or their diffusion through the boundary layer is then the rate-limiting step. In the latter case, an oversupply of reactants is brought to the vicinity of the surface, but the surface reaction cannot consume all of them.

The gas-phase transport coefficient h_g, can be gauged as follows: in Fick's law $J = -D(dC/dx)$ we identify (dx) with the boundary layer thickness δ and get

$$J_{\text{gas-to-surface}} = -(D/\delta)C_g \tag{33.4}$$

Boundary layer is the region of fluid where wall friction is important. Boundary-layer thickness δ is given by

$$\delta = (\eta L/v\rho)^{1/2} \tag{33.5}$$

where η is viscosity, v is fluid velocity, ρ is its density and L is the characteristic dimension of the system. Boundary-layer thickness increases along the flow and is thicker in the exhaust end of the reactor compared with the inlet end.

For atmospheric system at ca. 1000 °C, the values are $D \approx 10\,\text{cm}^2/\text{s}$, $L \approx 100\,\text{cm}$, $\eta \approx 10^{-4}$ poise (g/cm-s) and $\rho \approx 10^{-4}\,\text{g/cm}^3$ ($\rho \propto (1/T)$) we get an approximate boundary-layer thickness of 3 cm, which is close to values found in real systems. Gas-phase transfer coefficient h is then ≈3 cm/s.

If we lower the operating pressure by a factor of 1000, diffusivity increases thousand-fold because D changes as a function of pressure and temperature roughly as

$$D \propto T^{3/2}/P \tag{33.6}$$

There is an opposing trend of boundary-layer thickness increase because density decreases and flow velocity increases, but because of square root dependence (Equation 33.5), this opposing trend is ca. one order of magnitude only. Diffusivity increase clearly dominates, and gas-phase transport of reactants to the surface is greatly enhanced. A reaction that was transport-limited at higher pressure can be turned to surface reaction controlled, by operating at reduced pressure.

In order to get a feeling for temperature dependence, we have to compare k_s and h_g as a function of temperature. Chemical reactions obey Arrhenius behaviour with exponential dependence, and thus, surface reaction-limited deposition is strongly temperature dependent (high E_a). The gas-phase transport coefficient h_g is proportional to D, which has $T^{3/2}$ temperature dependence. This explains the shallower slope in the transport-limited regime of Figure 5.6.

33.2 CVD REACTORS

APCVD reactors operate in a transport-limited mode and flow geometries are important for film uniformity. LPCVD reactors operate in a surface reaction-controlled regime and wafers can be packed closely, which increases system throughput. LPCVD reactors are similar to oxidation tubes (Figure 13.1), and both

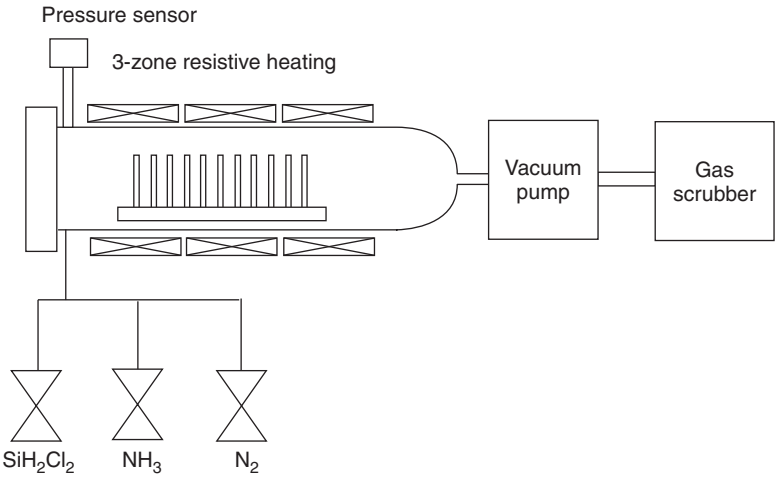


Figure 33.2 LPCVD nitride batch furnace (thermal CVD). Compare with Figure 13.1

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Table 33.1 LPCVD of silicon nitride (Si_3N_4)

If wafers come directly from another furnace operation (e.g., LOCOS pad oxide growth), no cleaning is required. Time limit for a new clean can be set, for example, at 2 h.
Load the wafers in the boat, fill with dummy wafers to equalize load and flow patterns.
Ramp temperature from 500 to 750 °C under nitrogen flow, 50 min (5 °C/min).
Pump to vacuum and perform leak check, 2 min.
Introduce ammonia NH_3, stabilize flow at 30 sccm, for 1 min.
Introduce dichlorosilane SiH_2Cl_2, flow 120 sccm, deposition starts.
Deposit at 300 mtorr for 25 min (thickness 100 nm, or 4 nm/min deposition rate).
Cool down to 700 °C (10 min).
Boat out.
Measurement: film thickness and refractive index monitoring by the ellipsometer.

LPCVD (Figure 33.2) and oxidation tubes can be fitted to the same furnace stack. A process for LPCVD silicon nitride (Table 33.1) bears similarity to oxidation process (Table 31.1).

Flow, temperature and pressure are important CVD reactor design criteria. Practically all CVD processes use toxic, corrosive and flammable fluids such as ammonia, silane, dichlorosilane, hydrides and metal organics. Reactor designs include double piping, inert gas flushing and venting and other safety features. Some of the reaction byproducts are harmful to pumps and mechanical constructions, which translates to special care in materials selection. Environmental, safety and health issues will be discussed further in Chapter 35.

CVD furnace systems are hot-wall systems, meaning that deposition also takes place on the walls. This leads to film build-up and flaking problems.

Gases are introduced in one end of the tube. Deposition leads to reactant gas depletion towards the end of the tube, and boundary-layer thickness increase also reduces deposition rate. However, this is compensated by increased temperature (=increased rate of chemical reaction). Heating elements are arranged in three zones: for example, T1: 747 °C, T2: 750 °C and T3: 753 °C for LPCVD silicon nitride (Figure 33.2). This temperature ramp along the tube helps to keep deposition rate constant.

In polysilicon LPCVD, this three-zone system results in grain size gradient along the length of the tube. In so-called flat-poly systems, the temperature is kept constant and gas introduction is made uniform by an elaborate distribution system. Alternatively 'poly' can be deposited in amorphous state at 570 °C to eliminate grain size gradients.

33.3 ALD (ATOMIC LAYER DEPOSITION)

Surface-controlled reactions result in better step coverage (microscale phenomenon) and uniformity across the wafer (macroscale phenomenon) compared to transport-limited reactions. ALD (which is also known as *atomic layer CVD*) is the ultimate surface-reaction limited case: one atomic layer is deposited in a single pulse of reactant gases. The first layer to react at the surface (AB) is chemisorbed with bond energies of the order of 1 eV, while additional layers are physisorbed with bond energies of the order of 0.4 eV. By selecting temperature and flush-gas pulses suitably, it can be arranged so that chemisorbed species are stable and physisorbed species and the excess precursor are flushed away. With the desorption time for the chemisorbed species at least of the order of seconds and residence time for physisorbed species a fraction of second, only the chemisorbed layer will remain. A second pulse of a different precursor (CD) is then introduced and allowed to react with the adsorbed species AB to form solid film according to

$$\text{AB (adsorbed)} + \text{CD (adsorbed)} \longrightarrow \text{AD (solid)} + \text{BC (gas)} \qquad (33.7)$$

$$ZrCl_4\,(\text{ad}) + 2H_2O\,(\text{ad}) \longrightarrow ZrO_2\,(\text{s}) + 4HCl\,(\text{g}) \qquad (33.8)$$

Repeated cycles of pulses of precursors AB and CD lead to the growth of solid film AD. Layer thickness is given by the number of pulses multiplied by monolayer thickness. In theory, one monolayer per pulse is deposited, but in many cases a sub-monolayer growth is seen.

In both cases, however, growth is self-limiting. Practical growth rates range around 1 Å/cycle: for Al_2O_3 deposition, it is 1.1 Å/cycle and for TiN, it is 0.2 Å/cycle (for other precursor gases this can, of course, be very different). When thickness/cycle numbers are translated into deposition rates, one has to take into account the flushing cycles between the pulses. Overall rates of a few nanometres per minute are typical for ALD, similar to LPCVD nitride or polysilicon, which are much higher temperature processes. ALD is a slow process, but there are many applications in which very thin films are needed, and step coverage requirements are strict: for example, diffusion barrier deposition into a high aspect

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

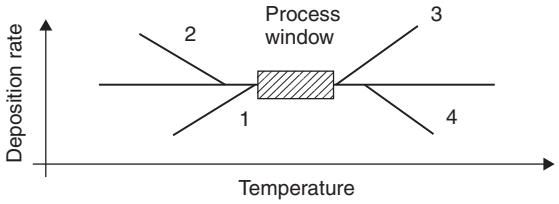


Figure 33.3 Process window for ALD (see text for details)

ratio contact hole, or scaled down gate oxides. In both cases, a few nanometres are enough.

ALD operating temperature is limited from below by two mechanisms (numbers refer to Figure 33.3): low temperature leads to a low reaction rate (1), and precursor condensation on the surface leads to excessive deposition (2). The former leads to less than the monolayer deposition, and the latter to non-self-limiting deposition of unwanted composition. Upper operating temperature is also limited by two mechanisms: thermal decomposition of the precursors, which results in deposition in the normal CVD fashion (3), and high re-evaporation rate, which leads to sub-monolayer growth per cycle (4). Under the right conditions, a uniform monolayer (or sub-monolayer) formation is observed.

ALD is a variant of CVD, but its deposition mechanism is definitely different: in CVD, the deposition rate is strongly temperature dependent, but in ALD there is a (wide) process window in which the rate is independent of temperature. For example, the rate for $SrTiO_3$ has been measured as 0.3 Å/cycle from 225 to 325 °C. Uniformity of ALD is exceptionally good, with <1% uniformities reported for both within wafer and wafer-to-wafer.

ALD results in very conformal films, as shown in Figure 33.4. The nanolaminate of aluminium and tantalum oxides covers the oxide step 100%, whereas the sputtered metal shows only ca. 50% step coverage.

ALD is free of one of the main mechanisms of irreproducibility in CVD: homogeneous gas-phase reactions, which make, for instance, reaction $SiH_4 + O_2 \rightarrow SiO_2 + 2H_2$ prone to gas-phase SiO_2 particle generation. Because only one gas is introduced at a time, there cannot be gas-phase reactions between precursors.

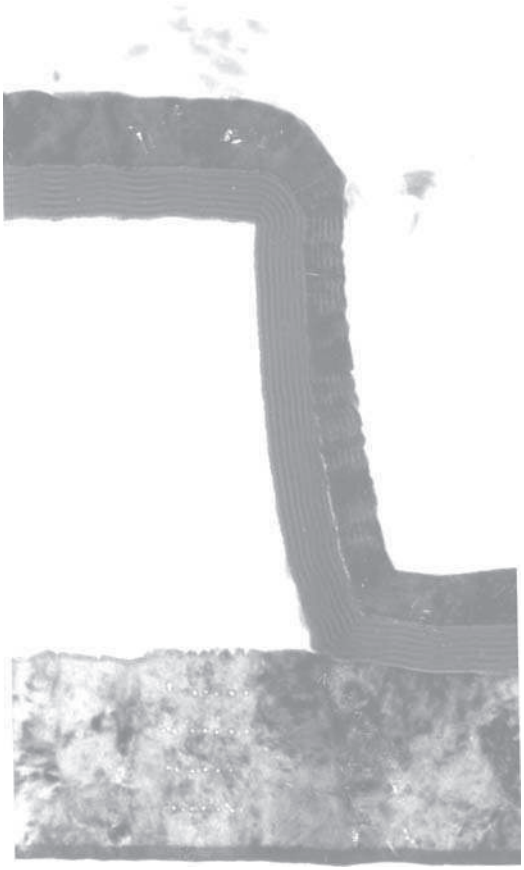


Figure 33.4 ALD nanolaminate (Al_2O_3 and HfO_2) step coverage over an oxide step is fully conformal, whereas the sputtered metal step coverage is ca. 50% only. TEM courtesy Hannu Kattelus, VTT

33.4 MOCVD

Most CVD processes use simple source gases such as silane and hydrides but there is the possibility of using liquid precursors. A widely used liquid source for CVD is TEOS (tetraethoxysilane) for oxide deposition. Liquid is heated in a container to increase its vapour pressure, and then a carrier gas, nitrogen, helium or hydrogen, is bubbled through the liquid and the precursor vapours are carried away by the carrier gas stream. The same method is also applied in gas-phase diffusion: dopants such as $POCl_3$ are introduced with bubbling and wet oxidation can be done by bubbling nitrogen carrier gas through water.

When the precursors are metal-organic compounds (MOs), the technique is termed MOCVD. It is widely used in III-V compound semiconductor epitaxy, with group III elements supplied as metal organics, such as trimethyl gallium $Ga(CH_3)_3$ or triethyl aluminium $Al(C_2H_5)_3$, while group III precursors are usually hydrides, AsH_3 and PH_3.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

MOCVD has also been studied for metal deposition. Copper has been deposited from precursors such as vinyltrimethylsilane hexafluoroacetylacetonate, VTMSCu(hfac), or Cu(I)-β-diketonate. Conformal deposition is possible and filling of high aspect ratio holes has been demonstrated. Trimethyl aluminium source gas has been used for MOCVD of aluminium. It would be beneficial to deposit aluminium films with copper alloying (0.5–4%), but this complicates MOCVD even further. MOCVD and ALD are methods of choice for new gate oxides such as HfO_2 and Ta_2O_5. Because of oxidizing atmosphere in CVD oxide deposition, the dielectric films are actually SiO_2/HfO_2 film stacks. SiO_2 formation is, in fact, beneficial because Si/SiO_2 interface is good and well known; the problem is in limiting and controlling the silicon dioxide thickness to keep the EOT low.

The problems with MOCVD are both practical and fundamental. The vapour pressure has to be right, the precursor must not react with other gases or materials present in the system, and its decomposition reactions must be reproducible. There is always the danger of carbon incorporation into the film when MOs are used as source materials. On the practical side, purity must be high, and this is difficult for complex compounds such as metal organics. Many MOs are extremely reactive with oxygen, and premature contact with oxygen will destroy the precursors.

33.5 SILICON CVD EPITAXY

Silane gases (SiH_xCl_{4-x}, $x = 0, \ldots, 4$) can all be used for epitaxy, but the temperature regimes are different (Figure 33.5). Growth temperature is a compromise between rate (thickness) and thermal budget (dopant diffusion during growth). Temperature is closely related to substrate/epi interface steepness: higher deposition temperature offers higher growth rate but at the expense of more thermal diffusion. Other factors that must be considered are autodoping from the substrate and from buried layers, pattern shifts and distortions (see chapters 6 and 26).

Because silicon homoepitaxy is a CVD reaction, the same laws about mass transport and surface-reaction limited growth apply to it. At high temperatures, all arriving source gas atoms react at the surface and the growth is limited by the arrival rate of atoms; at low temperatures an abundance of reactants wait to react. Different source gases have different useful temperature

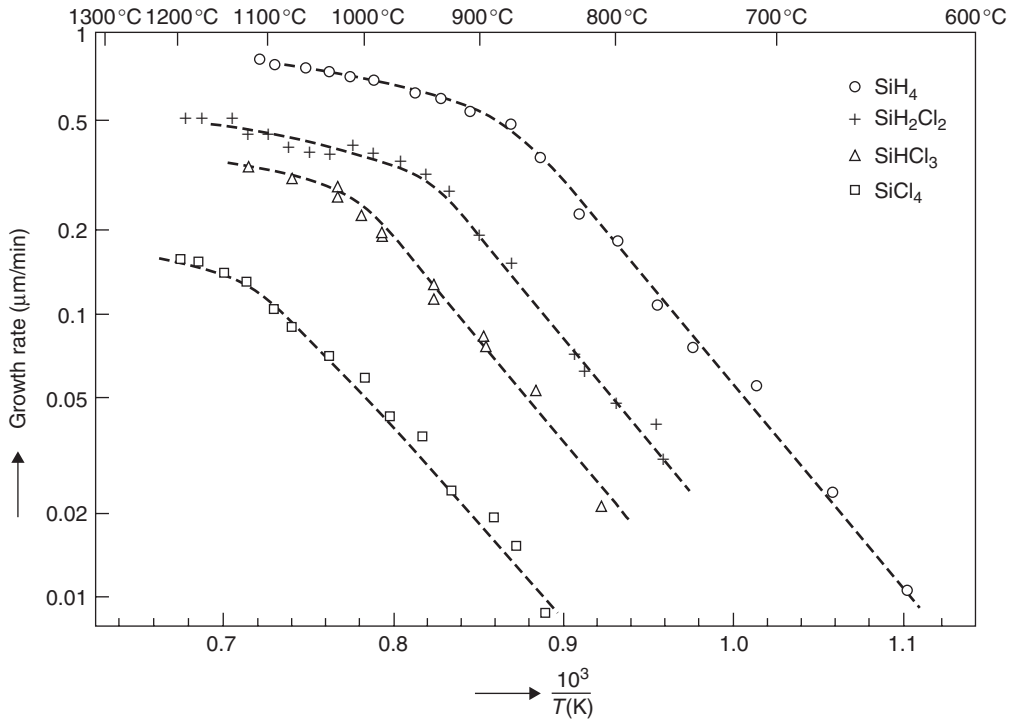


Figure 33.5 Epitaxial growth for different SiH_xCl_{4-x} source gases. Reproduced from Everstyen, F.C. (1967), by permission of Philips

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

ranges but practically identical activation energies in the surface reaction limited regime. Most epitaxy reactors, however, operate in the transport-limited regime, and gas-flow design in the reactor is crucial.

Epitaxy is not necessarily a high-temperature process. It has traditionally been so, but epitaxy as such can be carried out at any temperature. *In situ* cleaning of the wafer has been a factor for high temperatures: HCl or H_2 gas-phase cleaning processes worked better at elevated temperatures. Surface composition, however, is also dependent on the preceding cleaning step, and if that can be modified to reduce native oxide growth, *in situ* cleaning temperature can be lowered.

33.6 EPITAXIAL REACTORS

Reactors can be classified according to gas-flow patterns: gas flow parallel to the wafer surface is used in barrel (aka hexode) reactors where the wafers are vertically placed, and also in single-wafer reactors where the wafer is horizontally placed. In vertical reactors, wafers are flat on a susceptor but gases flow vertically perpendicular to wafers; vertical reactors are known as pancake (disk) reactors (Figure 33.6).

Two wafer heating methods; induction (RF coils) and lamp heating; are used. Lamp heating can be used in all major reactor types. The wafer surface is hotter than the backside because lamps heat the wafers from top, and the wafers are bowed up at the centre. Induction heating heats the graphite susceptor, and wafers bow up at edges, which is countered by designing curved wafer recesses in the susceptor. Induction heating is more suited for sustained high temperatures, and lamp heating to short depositions/thin layers.

There are both batch and single-wafer reactors on the market. Both designs coexist because they have different strengths as regards film thickness, growth rate, interface abruptness or doping uniformity. Batch reactors typically have ca. 1 μm/min growth rates, and they are preferred for thick-layer applications (up to 200 μm in some power devices) in which interface sharpness is not an issue. Batch-loading reactors can take, for instance, 30 wafers of 100 mm diameter or 12 wafers of 200 mm diameter.

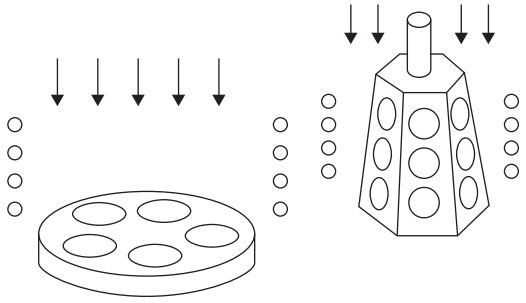


Figure 33.6 Pancake and barrel reactors. Lamps or RF coils for heating are shown, the reactor chamber is not

Single-wafer reactors offer high growth rates, for example, 5 μm/min at 1120 °C, using trichlorosilane. In addition to steep interface due to short deposition time, single-wafer reactors are superior with respect to film uniformity: 1% across the wafer for thickness, 4% for resistivity. Rotating susceptor, which comes naturally in a single-wafer reactor, is responsible for the uniformity, and also for a wider operating window because gas-flow rate, velocity and boundary-layer thickness can be varied over a wider range. A thinner boundary layer, for example, means that evaporated dopants from buried layers will rapidly diffuse to the main gas flow and be swept away.

Epi reactors operate either at atmospheric pressure but reduced pressure, typically 50 to 100 torr, can also be used. Reduced pressure operation adds to equipment complexity, and it is used for demanding applications only, including SiGe epitaxy (which differs from silicon epitaxy in regard to process temperatures, which is only ca. 700 °C vs. 1100 °C).

Reactor chambers are made of either quartz or stainless steel. Of course, metal chambers pose metal contamination dangers, especially because HCl and other chlorine gases can etch metals. Quartz chambers are not mechanically very strong at high temperatures, and they must be air cooled. Wafer susceptors are made of graphite. However, graphite itself is not very pure; it is porous and might trap source gases or reaction products, or it might react, and then carbon-containing species might be incorporated into epi film. Therefore, silicon carbide (SiC) coating is applied on graphite parts.

Gases used in epitaxy are extremely pure: carrier hydrogen must be free of oxygen and water below 100 ppb level. Silane purity is measured by resistivity: >3000 ohm-cm. Dopant gases are very dilute: 100 ppm phosphine or diborane in hydrogen is typical. All piping for process gases must be made of stainless steel because chlorosilanes and HCl are aggressive gases. Electropolishing, down to nanometre-surface roughness, is used in piping to eliminate particle contamination.

Epi reactors are power hungry: keeping wafers at ca. 1100° consumes hundreds of kilowatts, which must be removed: 80 to 90% of it into cooling water and the rest, mainly to hot exhaust gases. These gases are unused silanes (typical utilization is 10–30%) and hydrogen,

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

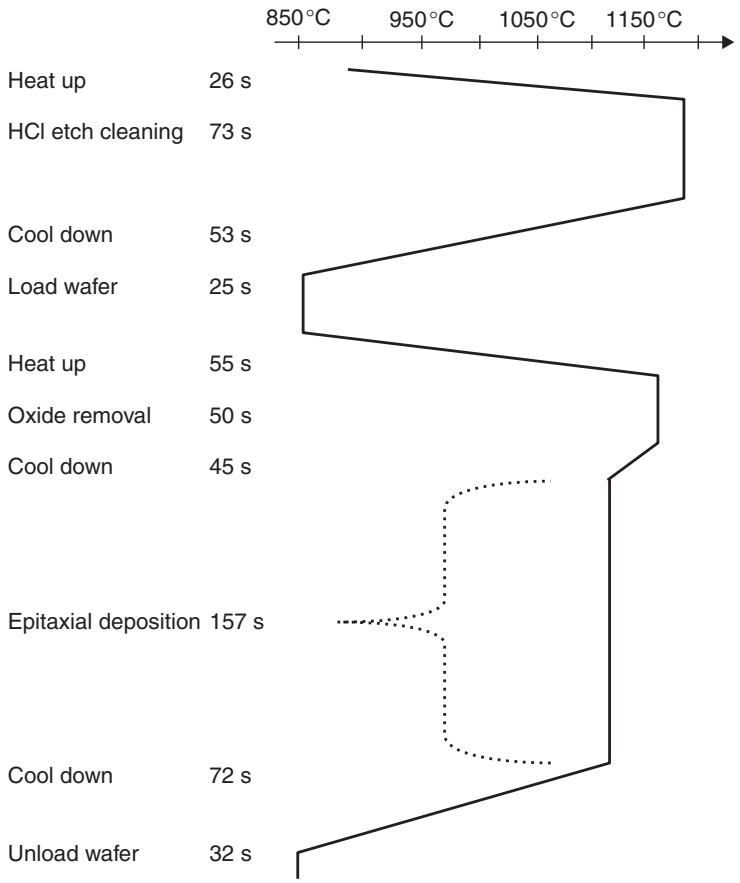


Figure 33.7 Single-wafer epitaxy reactor running $SiHCl_3$ process. Actual deposition time is 30% of the total time. Deposition rate is ca. 5 μm/min, or the film thickness is 13 μm

which can account to 99% of flow. Gas treatment is done by burn systems, wet scrubbers or by thermal decomposition.

A growth process 13 μm thick epilayer in a single-wafer reactor is shown in Figure 33.7. As can be seen, the actual deposition is just a fraction of total process time; the remainder is spent on heating, cooling and cleaning. These steps are essential for epitaxial film quality. Pre-bake has many effects: native oxide is removed (according to Equation 6.2), dopants and oxygen outdiffuse from the surface layer, and damage from preceding implantation step is annealed away. This results in higher crystalline quality and reduced autodoping.

In some reactors, wafers are loaded upright (akin to Figure 33.2), and their backsides are exposed to gas flows, and substrate autodoping can be significant. Backsides of heavily doped wafers are usually protected by, for example, CVD oxide film to prevent the evaporation of the dopant into the reactor. In addition to intentional and autodoping, films on reactor walls release some dopants. This is known as reactor memory effect.

Even though silicon growth in epi reactors is typically in the transport-limited regime, dopant incorporation can be in the surface-reaction limited regime, which necessitates accurate temperature control. Temperature uniformity is also very important because even minor temperature differences lead to crystal slips when silicon yield strength is exceeded (Equation 4.8).

33.7 EXERCISES

1. What is the Knudsen number in
   (a) APCVD
   (b) LPCVD
   (c) UHV-CVD?

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

2. Polysilicon LPCVD activation energy E_a is 1.7 eV. What happens to the deposition rate if, instead of standard 630 °C deposition, 570 °C is used?
3. If the gas-phase transfer coefficient h is 3 cm/s, and the surface reaction coefficient $k = 5 \times 10^7 \exp(-1.7\,\text{eV/kT})$ (in cm/s), at what temperature does the reaction turn from transport-controlled to surface-controlled?
4. What is the cost of a 150 mm diameter epiwafer if the single-wafer epireactor described in Figure 33.7 costs $2 million, running costs are $800 000/year (gas and graphite costs are dominating) and starting wafer cost is $20?
5. What is the utilization of silane in oxide CVD if the flow is 15 sccm silane with overabundance of N_2O in a single-wafer reactor, with 150 mm wafer size and deposition rate of 50 nm/min.
6. Nitride LPCVD is done nominally at 750°. What thickness difference does 6 °C temperature difference indicate if $E_a = 1.9$ eV?
7. What is the thinnest layer that could reasonably be deposited using PECVD parameters of Table 7.2, assuming a single-wafer reactor volume of 5 liters?
8. What is the total gas flow in the process shown in Figure 33.7?

REFERENCES AND RELATED READINGS

Cote, D.R. *et al*: Low-temperature chemical vapour deposition processes and dielectrics for microelectronic circuit manufacturing at IBM, *IBM J. Res. Dev.*, **39** (1995), 437.

Crippa, D., D.R. Rode & M. Masi: Silicon epitaxy, in *Semiconductors and Semimetals*, Vol. 72, Academic Press, 2001.

Everstyen, F. C.: Chemical-reaction engineering in the semiconductor industry, *Philips Tech. Rep.*, **29** (1967), 45.

Leskelä, M. & M. Ritala: Atomic layer deposition (ALD): from precursors to thin film structures, *Thin Solid Films*, **409** (2002), 138.

Ohring, M.: *The Materials Science of Thin Films*, Academic Press, 1992.

Vossen, J. & W. Kern: *Thin Film Processes, II*, Academic Press, 1991.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

34

Integrated Processing

Integrated processing involves the chaining of process steps into longer sequences. Process integration is also about chaining process steps into sequences but in a different sense: process integration is device-related, whereas integrated processing is a tool-view of step chaining.

34.1 AMBIENT CONTROL

In integrated processing, steps follow each other under strictly controlled conditions either in vacuum, inert gas or some other well-known ambient (Figure 34.1). This principle has been used in epitaxial silicon deposition for a long time: surface cleaning by HCl or H_2 gas is done in the same reactor chamber as the deposition itself to guarantee oxide-free surface. The titanium adhesion layer below platinum is another old example of integrated processing: the titanium surface is kept clean under vacuum, and platinum, which is deposited immediately after titanium, adheres to it well, whereas platinum would not adhere to an oxidized titanium surface, which would result immediately if a titanium wafer was transferred from one deposition system to another.

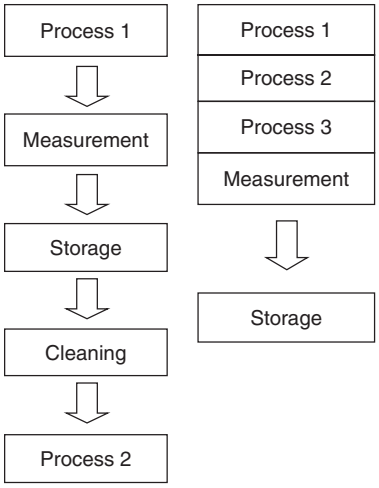


Figure 34.1 Conventional step-by-step process compared with an integrated sequence

Integrated processing has both scientific and manufacturing benefits. It enables a much higher degree of control over materials, interfaces and surfaces. This helps us to understand what is really going on in our processes. In manufacturing, it brings savings via several ways: cleaning steps can be minimized because wafer conditions are known all the time; wait and storage steps are eliminated and cycle time is reduced.

Integrated processing can be applied to any process sequence in principle, but in practice, similar processes are integrated: similar temperature, similar vacuum or similar ambient in general. In epireactor, both cleaning and deposition steps are at ca. 1000 °C, and both use not too different gases. Titanium and platinum are both deposited in the same vacuum at the same temperature. Integration of thermal oxidation with sputtering or CMP with PECVD would be awkward, but PECVD and plasma etching, or RTO and RTCVD can be combined fairly easily.

There are two main approaches to integrated processing (when we leave wet processing aside): vacuum clusters and mini-environments. In vacuum clusters, several process chambers are connected to each other, either serially or by means of a central transfer chamber. In Figure 34.2, a PVD multichamber system is shown. It has a pre-clean chamber, multiple deposition chambers and a cool-down chamber, all connected to a central handler chamber. Multiple identical reactor modules enable increased throughput, or alternatively two different processes can be run without the risk of cross-contamination. The central handler reliability is crucial for cluster operation.

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

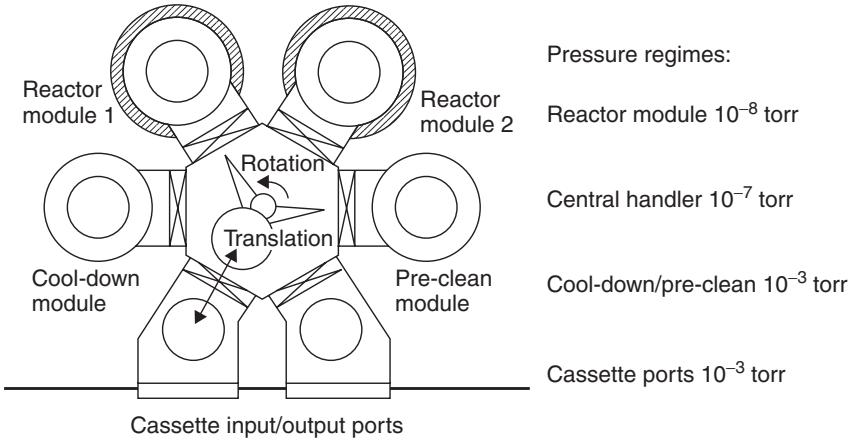


Figure 34.2 Multichamber vacuum cluster for PVD. Reproduced from Grannemann, E. (1994), by permission of AIP

Integrated vacuum tools are single-wafer tools for ease of automation. In the titanium/platinum example, the two steps were carried out in one chamber, sometimes called *multiprocessing*, but most integrated processing tools have separate chambers for each process. This enables a much tighter ambient control, and it enables chemically different steps to be integrated. If Ti/TiN/Al/TiN sputtering would be carried out in a single chamber, nitrogen carryover from TiN step would contaminate aluminum films.

In a mini-environment approach, a small cleanroom is built locally around the tools or the wafers. It is easier to keep a high purity level locally over a small area, than in the whole room. In one extreme, the wafer box is the cleanroom, filled with high purity nitrogen. Compared to the cleanroom, it has two benefits: nitrogen is inert, so reactive impurities from the atmosphere are eliminated, and the gas is stagnant in the box and particles do not move, as they do in the laminar airflow of the cleanroom.

Integrated processing has two major sources of variation under control: particle cleanliness and ambient chemical environment (Figure 34.3). Elimination of the cleanroom itself has been toyed with: if all tools would use a standard interface, wafers could be carried in mini-environment boxes from tool to tool, and they would never see the cleanroom air, in which case the cleanroom would become redundant. Wafer fabs with such standard mechanical interfaces (SMIF) have been built, but cleanrooms have not been made redundant because the conversion of all process and measurement tools has been elusive. This topic will be touched upon again in Chapter 35.

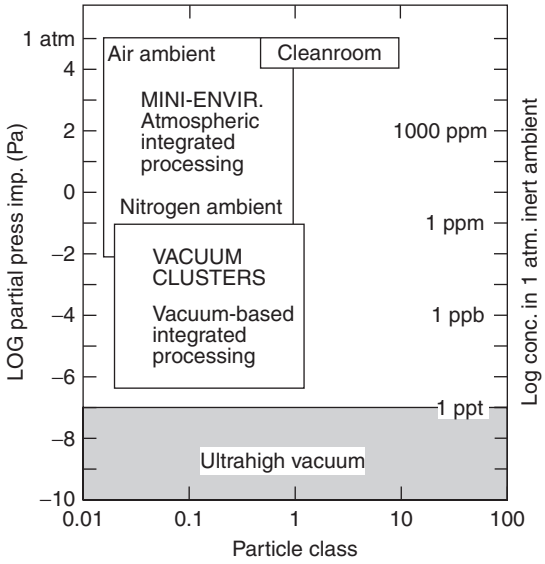


Figure 34.3 Environmental control: chemical/reactive contaminants and particles in vacuum clusters vs. mini-environments. Reproduced from Grannemann, E. (1994), by permission of AIP

34.2 DRY CLEANING

Because it is easy to integrate process modules with similar pressure and temperature regimes, dry cleaning methods are attractive in vacuum integrated cluster tools. Reduced pressure dry cleaning modules could fit into plasma etchers, sputters, PECVD, RTP and single-wafer epitaxial reactors.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Table 34.1 Dry cleaning agents

Vapours	Anhydrous HF
Gases	H_2, HCl
Ions	Ar^+
Atoms	Si
Photons	UV (plus some chemicals like Cl_2 or O_3)
Plasmas	CF_4

Compared to wet cleaning, dry cleaning has the following advantageous features:

- no surface tension effects in small structures
- reaction products are removed efficiently
- no drying necessary.

UV-ozone has been tried for organics removal, UV-Cl_2 for metal removal and HF-vapour for native oxides. Argon and H_2 plasmas have also been utilized, in sputtering systems, to improve contact by etching oxide just prior to metal deposition (Table 34.1). Dry cleaning has a central role in epitaxial systems in which utmost surface cleanliness is mandatory. Thin oxides can be desorbed by a hydrogen bake. The exact temperatures depend on surface termination: hydrogen-terminated surfaces can be baked at temperatures as low as 700 °C to reveal a perfect surface for epitaxy. To date, however, dry cleaning has remained a special method, especially because it is difficult to remove particle contamination with dry methods.

34.3 INTEGRATED TOOLS

Ti/TiN/Al/TiN multilayer stack poses some interesting etch problems. If top TiN is etched with a fluorine plasma, there is the danger that involatile AlF_3 is formed and aluminium will be etched non-uniformly. If top TiN is etched in chlorine plasma, aluminium etching can continue immediately, without the difficult native oxide removal step (when TiN has been deposited on aluminum without vacuum break). If the bottom TiN/Ti is etched in fluorine plasma, AlF_3 will passivate the sidewalls of aluminium lines. This is a desired side effect because otherwise post-etch corrosion from HCl attack would corrode aluminum lines (Equation 32.14). Hydrogen chloride is formed in reaction between chlorine residues on the wafer and water vapour in the air. If the bottom TiN/Ti is etched with chlorine chemistry, a separate passivation/chlorine removal step is needed. Photoresist plasma stripping can provide this passivation through the formation of aluminium oxide. Immediate wet rinsing to remove any HCl formed is also possible, but then the vacuum/plasma tool needs to be integrated with a wet process tool, which is not straightforward.

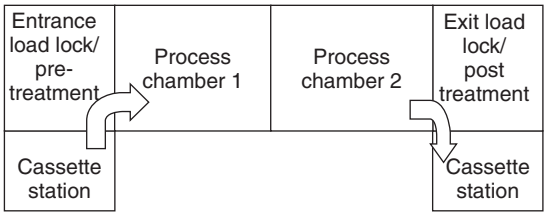


Figure 34.4 Sequential multichamber tool with cassette-to-cassette operation

A sequential multichamber tool is shown in Figure 34.4. If it is used as a TiW/Al etcher, a chlorine plasma process for aluminium etching would run in process chamber 1, and process chamber 2 would accommodate TiW etch process, fluorine or chlorine-based. Exit load lock could be used for photoresist stripping.

If the tool of Figure 34.4 is configured as a gate-module tool, its configuration is as follows:

• entrance load lock:	HF-vapour cleaning
• process chamber 1:	RTO of gate oxide
• process chamber 2:	polysilicon CVD
• exit load lock:	ellipsometry

34.4 EXERCISES

1. What is the throughput of an aluminium etcher as shown in Figure 34.4 for (a) TiW/Al (0.1 μm/1 μm) and (b) for 50/400 nm film stack, if entrance load lock pump-down time is 20 s, aluminium etch rate in process chamber 1 is 500 nm/min, TiW etch rate in chamber 2 is 200 nm/min, and exit load lock purge/pumptime is 30 s?
2. What would be the maximum throughput of a cluster tool of Figure 34.2 if metal deposition rate is 10 nm/s, and 0.5 μm thick films are made?
3. How could metallization be monitored in exit load lock of a sputtering system?

REFERENCES AND RELATED READINGS

Barna, G.G. *et al*: MMST manufacturing technology – hardware, sensors and processes, *IEEE TSM*, **7** (1994), 149.

Grannemann, E.: Film interface control, *J. Vac. Sci. Technol.*, **B12** (1994), 2741.

Rubloff, G.W. & Boronaro, D.T.: Integrated processing for microelectronics science and technology, *IBM J. Res. Dev.*, **36** (1992), 233.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Part VII

Manufacturing

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

35

Cleanrooms

Particle size distributions in cleanroom air, process gases, DI-water and wet chemicals all have the same basic characteristics: four to eight times more particles are detected if the detection threshold is halved. Therefore, if the minimum linewidth is halved, the number of particles that are potential killers increases by four to eight times.

Cleanrooms were initially a solution to particle contamination reduction (cleanrooms were not invented for microelectronics, but for delicate mechanical assembly). Later on, temperature and humidity control for improved reproducibility in lithography was recognized. Other features have been added over the years, and a modern cleanroom is a system of facilities that ensure contamination-free processing under very stable environmental conditions (Figure 35.1).

The main features of cleanrooms are:

- overpressure (50 Pa) for keeping particles outside;
- filtered air (99.9995% at 0.15 μm particle size);
- heating/cooling/humidification/drying of incoming air;
- laminar (unidirectional) air flow in the working areas;
- materials compatibility;
- mechanical and electrical interference minimization;
- working procedures.

35.1 CLEANROOM STANDARDS

Cleanrooms are classified mainly on the basis of particle counts. Older specifications such as Fed. Std. 209 (Table 35.1) specify particles per cubic foot. Newer ISO standards (Table 35.2) employ units of particles per cubic metre (conversion factor: $1\,m^3 = 35.3\,ft^3$). ISO standard cleanliness class N with particle concentration C_n (particles/m^3) is calculated as

$$C_n = 10^N \times (0.1\,\mu m/D)^{2.08} \qquad (35.1)$$

where D is particle size in micrometres.

Table 35.1 Simplified Fed. Std. 209D airborne particle cleanliness classes (particles/ft^3)

Class	1	10	100	1000	10 000
No. of particles 0.5 μm	1	10	100	1000	10 000
No. of particles 0.1 μm	35	350	3500	35 000	350 000

Table 35.2 ISO standard airborne particle cleanliness classes (/m^3)

	0.1 μm	0.2 μm	0.3 μm	0.5 μm	1 μm	5 μm
ISO class 1	10	2				
ISO class 2	100	24	10	4		
ISO class 3	1000	237	102	35	8	
ISO class 4	10 000	2370	1020	352	83	
ISO class 5	100 000	23 700	10 200	3520	832	29

The proper way to specify cleanroom cleanliness is therefore: Class X (at Y μm particle size).

The example in Table 35.3 shows that there are a multitude of cleanroom features in addition to particle specifications. These are related to air quality plus mechanical and electrical environment.

Cleanliness is defined for three different stages of cleanroom construction:

1. as-built: cleanroom construction is finished, but no tools installed;
2. static: with process tools installed and running, but no personnel;
3. operational: with people working in the cleanroom.

As-built tests should indicate around one class better cleanliness than the designed operational class. Laser scattering of sampled air is used to measure particle counts. There are some methodological problems in the

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

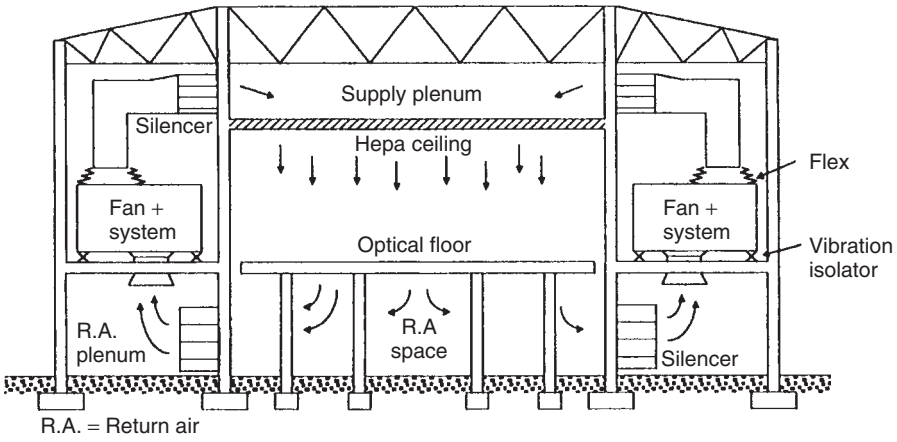


Figure 35.1 Cleanroom: fans generate unilateral airflow from HEPA (high efficiency particle) filter ceiling. Air is highly purified and temperature- and humidity-controlled. Optical floor, isolated from the rest of the building, prevents vibrations that would destabilize microlithography and microscopy operations. Source: Cleanroom Design, W. Whyte, 1999, © John Wiley & Sons, Ltd

Table 35.3 Fed. Std. class 1 cleanroom

Feature	Values
Cleanliness, process area	<35 particles/m^3, >0.10 μm
Temperature, lithography	22 °C ± 0.5
Temperature, other areas	22 °C ± 1.0
Humidity, lithography	43 ± 2%
Humidity, other	45 ± 5%
Air quality	
Total hydrocarbons	<100 ppb
NO_x	<0.5 ppb
SO_2	<0.5 ppb
Envelope outgassing	6.3×10^8 Torr L/cm^2/s
Pressure	typical 30 Pa relative to outside
Acoustic noise	<60 dB
Vibration	<3 μm/s (8–100 Hz)
Grounding resistance	1 Mohm
Magnetic field variation	< ±1 mG
Charging voltage	< ±50 V

Source: Cheng, H.P. & R. Jansen (1996)

best cleanrooms: there are simply too few particles to get good statistics.

The cleanroom must include not only the structure itself and airflows, but also procedures for transfer of people and materials. Cleanrooms are built with stages of increasing cleanliness: at the heart of the cleanroom is the process area, which is surrounded by the service area (known as gray area), which is clean compared to the outside world but which does not have unilateral air flow. People enter cleanrooms in stages of increasing cleanliness: at the entrance, footwear is changed into cleanroom shoes and hair is covered. In the next stage, an overall is put on. Depending on the cleanliness class, further protective garments are added: a mouthpiece, a second layer of headgear and cleanroom boots to cover the shoes. Finally, gloves are put on (Figure 35.2). A similar, but somewhat reverse, procedure of increasing cleanliness is applied when new tools, wafer boxes, sputtering targets or any other material is transported into the cleanroom: in the anteroom, the outermost layer of packaging is removed and the gadgets are



Figure 35.2 Fed. std. Class 100 cleanroom with wet benches. Photo courtesy Ulrika Gyllenberg, VTT Microelectronics Centre

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

taken into an airlock where the inner packing material (which was wrapped in the cleanroom of wafer, target or tool manufacturer) is removed. Depending on the item, manual cleaning with isopropyl alcohol may be undertaken.

As discussed in Chapter 34, cleanrooms need not be large halls or rooms; mini-environments are locally clean areas around critical process tools. If wafers are enclosed in portable mini-environments, they will never experience cleanroom air, which can then be orders of magnitude less clean, as shown in Figure 35.3.

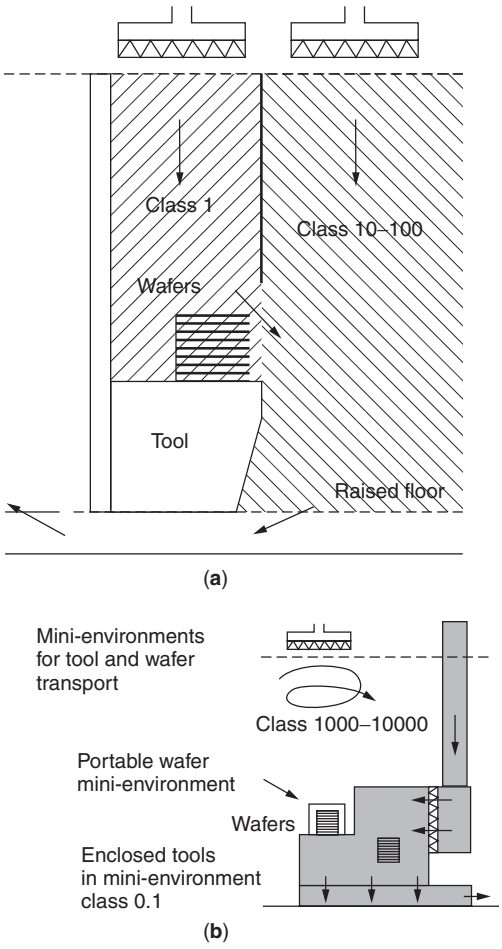


Figure 35.3 (**a**) Cleanroom versus (**b**) mini-environment. In a mini-environment, wafers are processed, transferred and stored in tight, portable containers; a cleanroom is four orders of magnitude dirtier, for example, class 0.1 mini-environments in a class 1000 cleanroom. Reproduced from Rubloff, G.W. & D.T. Boronaro (1992), by permission of IBM

35.2 CLEANROOM SUBSYSTEMS

35.2.1 Construction

Cleanroom envelopes – walls, floor, ceiling, and so on – need to be made of materials compatible with the overall objective of environmental control. The walls must not outgas, they must be easy to clean and they must be easily removable for equipment installation. They must also be tight because cleanliness is partly ensured by slight overpressure, which prevents outside air from entering. (In a virus research laboratory, cleanliness must be achieved even though underpressure must be applied in order to prevent samples from escaping.) The ceiling consists of blank elements and filter elements. The higher the proportion of filter elements, the better the cleanroom class.

A raised, perforated floor is essential for unidirectional (laminar) flow conditions: air from ceiling filters can travel unidirectionally. If particles are generated in the cleanroom, they will be transported away directly through the floor, hopefully not interfering with the wafers. Return air will travel laterally under the raised floor, and return either in the service aisles or in separate return air ducts. If service aisles are used as the return path for the air, there will be turbulent upstream flow, and even though the particle counts are low, the service area is not suitable for wafer processing.

Vibration isolation is important for lithography and microscopy. Massive air-handling units generate vibrations, and therefore mechanical separation of air circulation fans from other parts of the building is needed. Sensitive process areas for lithography can be established on isolated concrete slabs extending down to bedrock.

35.2.2 Air

Air handling consists of four major blocks:

- extraction unit
- make-up air unit
- recirculation unit
- filter fan units.

In the first phase, the air is filtered from coarse objects, humidification or dehumidification is performed, and airborne pollutants such as SO_x, NO_x and ammonia are removed by activated carbon filters. Cooling coils and heaters are used to stabilize air temperature. Successive stages of filtration remove finer particles. The final filter is called HEPA (high efficiency particle) or ULPA (ultra-low penetration air); it is installed in the cleanroom ceiling. ULPA filters have 99.9995% filtration

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

efficiency at particle size $>0.12\,\mu m$. Filter efficiencies can also be classified according to most penetrating particle size (MPPS). Filter defects (pinholes) are also a major concern. Air velocity in the cleanroom is usually ca. 0.35 to 0.45 m/s; and air circulation takes place 50 to 500 times/h, depending on cleanliness requirements. Once the air has been processed, it is re-circulated, with only 10% of replacement air introduced in each cycle.

Many types of process equipment produce excessive heat loads, for example, furnaces in the range of 100 kW, and this heat has to be removed in order to maintain constant temperature in the cleanroom. Most of the excess heat is taken away by cooling water. The design of a cleanroom must, therefore, include knowledge of the processes and tools that are going to be employed.

35.2.3 DI-water

De-ionized water (DI-water), also known as *ultra-pure water* (UPW), is a major sub-system because of enormous water consumption in modern IC fabrication. A big fab uses a million cubic metres of ultra-pure water a year.

Water is treated in many steps as follows:

- sand filter;
- active carbon filter;
- particle filtering at $3\,\mu m$;
- softening of water;
- RO: reverse osmosis;
- CEDI: continuous electrical de-ionization;
- UV treatment;
- ion exchangers;
- particle filtering at $0.2\,\mu m$;
- storage tank;
- continuous DI-water circulation in the cleanroom loop.

Reverse osmosis is a process in which water molecules diffuse through a porous membrane, while microorganisms, particles and ions are rejected. UV treatment kills bacteria and reduces total carbon content. Both RO and UV treatment can be repeated for improved performance. DI-water quality is monitored by resistivity measurements: 18 Mohm-cm is required. Regular bacteria checks as well as particle tests are performed.

35.2.4 Gas systems

Gas system requirements include particle specifications (which set limits to the choice of materials for piping, valves, regulators, mass flow controllers, etc.), leak rates (static leak test, helium leak test) and gas impurity tests.

Bulk gases (also known as *line gases* or *house gases*) are gases shared by many tools. These include nitrogen, oxygen, hydrogen, argon and compressed air. Nitrogen is especially widely used, both in processes and as an inert protective gas. Four purity classes of nitrogen can be offered for different applications:

- process nitrogen: furnace annealing or reactive sputtering, 7N purity;
- dry nitrogen: venting and flushing of process chambers, 5N purity;
- pistol nitrogen: for drying;
- pump nitrogen: as ballast for pumps.

Specialty gases are used by dedicated equipment, and they are supplied from gas bottles in a one-to-one distribution topology. These include, for example, SF_6 and Cl_2 for etchers, SiH_2Cl_2 and NH_3 for nitride LPCVD, SiH_4 and N_2O for PECVD oxide, PH_3 for doped polysilicon LPCVD and WF_6 for tungsten CVD. Ion implanter gas consumption is very small, and AsH_3, PH_3 and BF_3 mini-bottles are usually located inside the implanter cabinet. Implanter gases can also be supplied from safe delivery system (SDS) sources: the dopant gases are absorbed in solid absorber material in the bottle, and released by application of temperature or underpressure.

35.3 ENVIRONMENT, SAFETY AND HEALTH (ESH) ASPECTS

Various gases, chemicals and tools are sources of potential health hazards to cleanroom personnel. Ion implanters operate at 200 kV and they are sources of X-rays (and gamma rays may be emitted in hydrogen implantation); plasma systems may leak microwave energy and UV radiation, and wet etch and plating baths may contain cyanides. These hazards are dealt with in different ways.

Strong mineral acids such as H_2SO_4, HNO_3, H_3PO_4 and HCl are routinely used. Normal burn hazards are associated with them and they must be neutralized after use. HF is different because its effect is not immediate but delayed, and it does not attack skin but bone. Special care is needed for all HF-containing liquids and separate disposal of HF is required.

Solvents and organics come from various sources: HMDS, which is used as a priming agent before photoresist coating, is released into cleanroom air (HMDS

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

is the main airborne pollutant in many cleanrooms), solvents are released from resists upon baking and IPA and acetone are used for drying and cleaning. Solvents are major reasons for wafer fab fires.

Process exhausts remove unwanted thermal and mass flows from the cleanroom. Acid vapours from wet benches are removed and safely disposed of in plastic ducts while solvent exhausts are removed in stainless steel ducts. Separate piping is required not only because of materials issues but also to prevent explosive mixing. In most cases, cleanroom systems protect wafers from humans, but in wet benches, the protection of humans from chemicals is required (this is the usual concern in e.g., pharmaceutical cleanrooms). Acid vapours are cleaned by gas-abatement systems (solid absorber, combustion system and/or gas effluent washing machine, aka wet scrubber) before release into the air.

In many processes, the utilization of source gases is very low and the outpumped flow consists mostly of unused source gas. These gases, for example, SiH_4 from an LPCVD system, may be incinerated or diluted. Silane is spontaneously flammable. It is used at 100% concentration in LPCVD polysilicon, but in PECVD systems it is usually diluted, 1 to 5% SiH_4 in nitrogen, argon or helium. Wet oxidation is usually done by *in situ* generated water from H_2 and O_2 gases (see Figure 13.1). Hydrogen/oxygen mixtures are flammable between 4 and 75% hydrogen, and hydrogen content in exhaust gases needs to be controlled by combustors or by other gas-abatement systems.

A toxic-gas alarm system is required because many of the gases used in semiconductor processing are extremely toxic (Table 35.4): hydrides, PH_3, AsH_3 and B_2H_6 are lethal in low parts per million concentrations. Chlorine was used as a battle field gas in World War I. Many chlorine-containing gases react with humid air to form HCl, which is similarly toxic and corrosive.

Pumps and pump oils can accumulate considerable amounts of unknown compounds: for example, products from reactions between etch gases and photoresist. Pumping oxygen is a safety concern: oxygen can explode if it reacts with pump oil. Therefore, most plasma and CVD equipment use either inert perfluorinated pump oils (Fomblin®, Krytox®) or else dry pumps are employed. Dry pumps are also beneficial because they tolerate more corrosive and abrasive chemicals than standard mechanical pumps.

Fire detection in a cleanroom cannot be done similar to normal office rooms because high cleanliness prevents particle-based detection and ionization detectors in the ceiling would see nothing because of unidirectional downflow. Local sampling and thermal detection are used. Fire extinguishing must be accomplished without generating particles because damage from extinguishing might be intolerable to the cleanroom as a whole. Carbon monoxide or water-mist systems are used.

Table 35.4 Toxic gases in semiconductor manufacturing

	TLV (ppm)	IDLH (ppm)	Other properties
NH_3	25	300	DO: 0.04–50 ppm
Cl_2	0.5	10	DO: 0.03–0.4 ppm
HCl	5	50	
HF	3	30	
BF_3	1	25	*
SiH_4	5	N/A	ER: 1.37–96%
GeH_4	0.2	N/A	
$SiCl_2H_2$	**	N/A	ER: 4.1–99%
AsH_3	0.05	3	DO: 0.5–4 ppm, garlic
PH_3	0.3	50	DO: 0.01–5 ppm, fishy
B_2H_6	0.1	15	DO: 1.8–3.5 ppm, sweet

*Reacts to form HF upon contact with moisture.
**Reacts to form HCl.
TLV – threshold limit value: no adverse effects for prolonged exposure.
IDLH – immediately dangerous to life and health: 30 minutes escape time to ensure no permanent health effects.
ER – explosive range (% by volume in air).
DO – detectable odour.
N/A – not applicable.

Alarm strategies in a microfabrication cleanroom need to be carefully planned. In the case of a toxic-gas alarm, the personnel need to be evacuated, but it does not necessarily mean that oxidation furnaces have to be shut down. If a lot of 200 wafers is lost in a case of unplanned shutdown, huge damages will be incurred. In the case of fire alarm, air circulation needs to be closed down as otherwise it would spread the fire efficiently, but it is important to keep the exhausts operational. If the fire originated from a wet bench (which is usually the case), then the wet bench exhaust will at least remove hot acid and/or solvent vapours but there is the danger that the fire will spread along the exhaust ducts.

Static electricity elimination, acid neutralization, acid regeneration, waste chemical storage, particle counters, air quality monitors and various other systems are required to operate a cleanroom. The cleanroom can be regarded as a single big instrument because proper

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

cleanroom conditions can only be fulfilled when all subsystems are running.

35.4 EXERCISES

1. What ISO class corresponds to Fed. Std. 209 class 100 cleanroom and class 1, respectively?
2. Make a graphical plot of ISO cleanliness classes 1 to 4 for particle sizes 0.1 to 1 μm.
3. What class of cleanroom would be suitable for (a) 1 μm and (b) 0.1 μm CMOS production?
4. If a 0.5 L bottle (under 50 bar pressure) of boron trifluoride (BF_3) leaks into a 1000 m^2 cleanroom, will it be immediately dangerous to health?
5. Particle deposition rate J on a wafer that is parallel to airflow is given by $J = nu$, where n is the particle density and u is the sum of gravitational and diffusive settling velocities, ca. 5×10^{-4} cm/s for 0.1 to 0.5 μm particles. How many particles will deposit on a 200 mm wafer in an ISO class 2 cleanroom in an hour?

REFERENCES AND RELATED READINGS

Baldwin, D.G., M. Williams & P.L. Murphy: *Chemical Safety Handbook for the Semiconductor/Electronics Industry*, 3rd ed., OEM Press, Beverly Farms, 2002.

Cheng, H.P. & R. Jansen: Cleanroom technology, in C.Y. Chang & S.M. Sze (eds.), *ULSI Technology*, McGraw-Hill, 1996.

Middleman, S. & A.K. Hochberg: *Process Engineering Analysis in Semiconductor Device Fabrication*, McGraw-Hill, 1993.

Misra, A., J.D. Hogan & R.A. Chorush: *Handbook of Chemicals and Gases for the Semiconductor Industry*, John Wiley & Sons, 2002.

Rubloff, G.W. & D.T. Boronaro: Integrated processing for microelectronics science and technology, *IBM J. Res. Dev.*, **36** (1992), 233.

Whyte, W.: (ed.): *Cleanroom Design*, Wiley, 1999.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

36

Yield

Understanding yield loss is a life and death issue in wafer fabs. Yield loss is inevitable, and it is important to understand the factors behind it. Microfabrication is a statistical business: some devices always fail, and usually no repair is available or feasible. There are a few exceptions: big memory arrays with redundant cell blocks can be repaired by disconnecting malfunctional blocks and connecting redundant blocks; and defective photomasks are usually repaired because writing is very slow and expensive.

Yield can be calculated at different points of process and different yield numbers obtained. In all cases, yield is a quotient of 'good outcomes/total'. Fab yield takes into account the number of wafers completing the process, divided by wafer starts. However, note that, it is typical that 20 to 30% of wafers circulating in a fab are for monitoring and testing and do not contribute to saleable chips, even in theory. Fabrication yield for prime wafers approaches 99%.

Die yield, also known as chip yield, is the fraction of functional chips on a wafer. In a 1997 survey, die yields ranged from 46 to 92% for 0.5 cm^2 devices. Again, not all chips on the wafer are product chips: some chips are dedicated to process-monitoring test structures (identical in all products, to gather statistical data on the process) and some are product-specific test structures.

Yield is a product of different yield loss mechanisms

$$Y = \Pi Y_i \tag{36.1}$$

Total yield can never be better than the yield of the lowest yielding step. Yield is a product of process steps Y_i (and processes with lots of steps tend to have low yields) but it can also be viewed as a product of systematic and random components

$$Y_{\text{total}} = Y_{\text{systematic}} * Y_{\text{random}} \tag{36.2}$$

Systematic yield loss comes from process errors and equipment malfunctioning, and from process capability limitations. All processes have variation (across the wafer, wafer-to-wafer and lot-to-lot), and devices cannot be designed to tolerate tails of statistical distributions. The fishbone diagram in Figure 36.1 depicts contributors to die-yield loss. As can be seen, the yield-loss causes can be difficult to pinpoint.

Table 36.1 Yields of IC fabrication at different stages of maturity

	Y_{random}	$Y_{\text{systematic}}$	Y_{total}
Introduction	20%	80%	16%
Ramp-up phase	80%	90%	72%
Mature	90%	95%	86%

SRAM is the prototypical test vehicle for process development: in a regular memory array of transistors, it is easy to locate the electrical fault and to investigate it by optical, physical and chemical means, and to correlate it with a physical defect, a particle, a residue, corrosion or linewidth change.

Yield is related to a particular process, characterized by its linewidth or process-technology generation. It is not constant over a device lifecycle: at product introduction, yield is low and it rises with production volumes. Some schematic values for processes in different stages of process maturity are shown in Table 36.1.

36.1 YIELD MODELS

The random-yield loss has been described by many models. Poisson distribution (Equation 36.1) is the simplest model: defect density D and chip area A determine yield. This holds fairly well for small chips and/or low defect densities (Figure 36.2).

$$Y = e^{(-DA)} \tag{36.3}$$

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

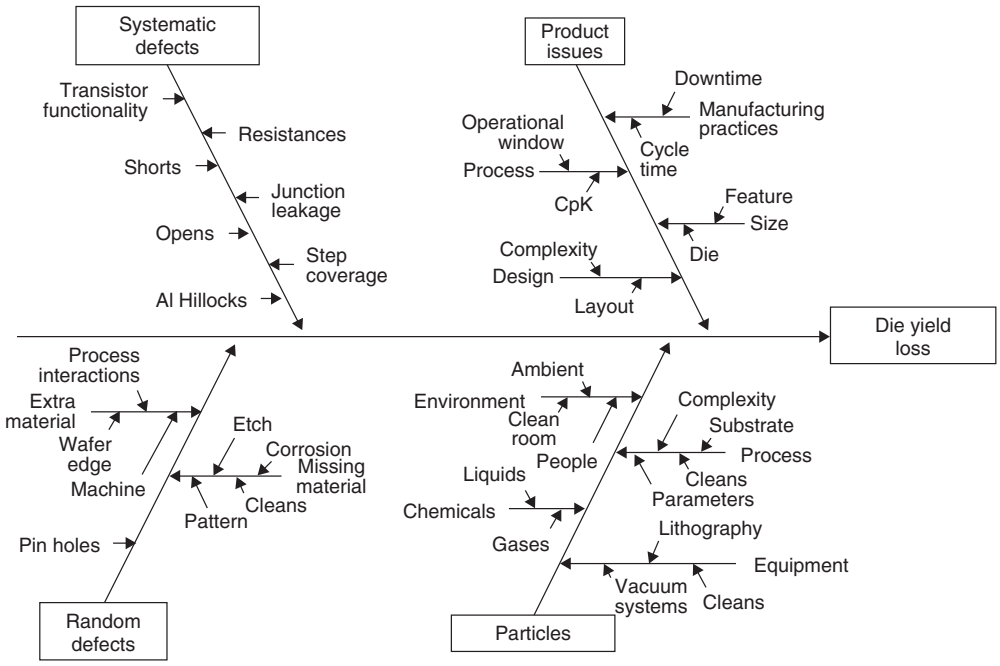


Figure 36.1 Factors influencing die-yield loss. Reproduced from Rao, G.P. (1993), by permission of McGraw-Hill

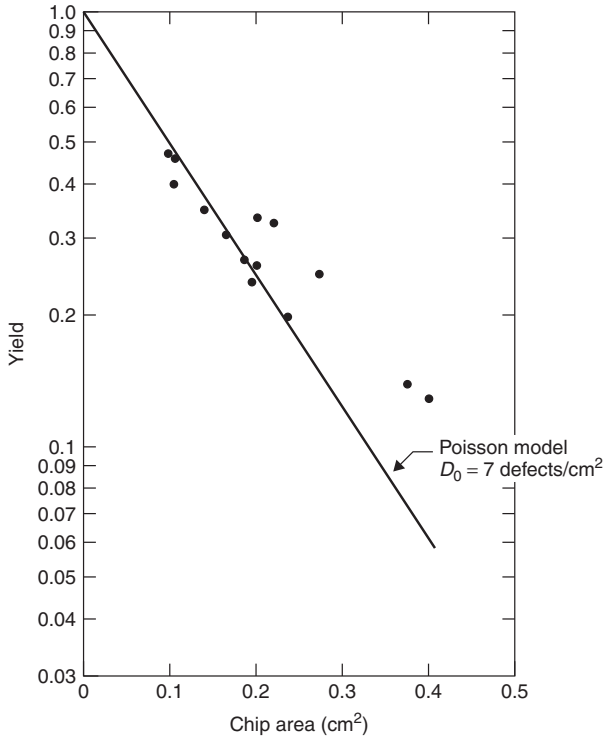


Figure 36.2 Poisson distribution of chip yield: good fit for small chips. Reproduced from Cunningham, J.A. (1990), by permission of IEEE

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

A more general model takes defect clustering into account and models the yield as

$$Y_{\text{random}} = (1 + (AD_o/\alpha))^{-\alpha} \tag{36.4}$$

where α = cluster factor (Figure 36.3).

Cluster factor α presents the tendency of defects to cluster; that is, they are not randomly distributed but tend to concentrate. The values of α are usually considered trade secrets, and companies are very reluctant to reveal their yield statistics. Cluster factor $\alpha = \infty$ corresponds to Poisson distribution, and $\alpha = 1$ results in Seeds model:

$$Y = (1 + AD)^{-1} \tag{36.5}$$

Another yield model is known as *Murphy's*

$$Y = ((1 - \exp(-DA))/DA)^2 \tag{36.6}$$

Chip size A is a result of two opposing trends: as linewidths are scaled down, chip area should decrease; but because more logic functions and more memory capacity is added, the number of transistors on a chip increases so fast that the chip area, in fact, is constantly increasing. Defect density D is not an unambiguous concept, as shown in Figure 36.4. Particles

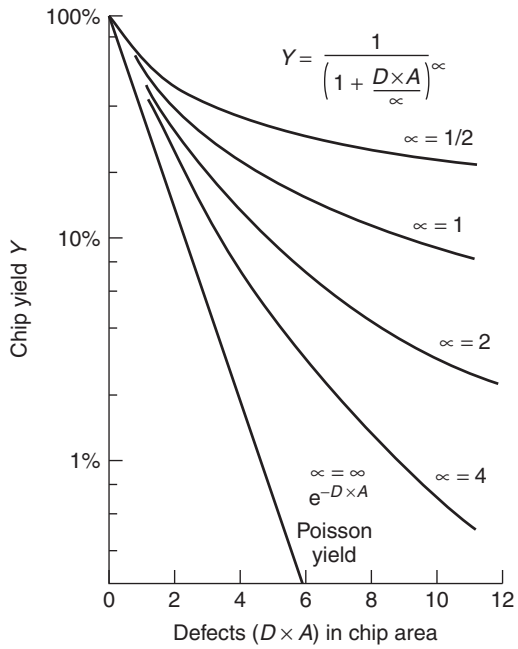


Figure 36.3 Yield models compared: cluster factor α ranges from 0.5 to infinity. Reproduced from Carlson, R.O. & Neugebauer, C.A. (1986), by permission of IEEE

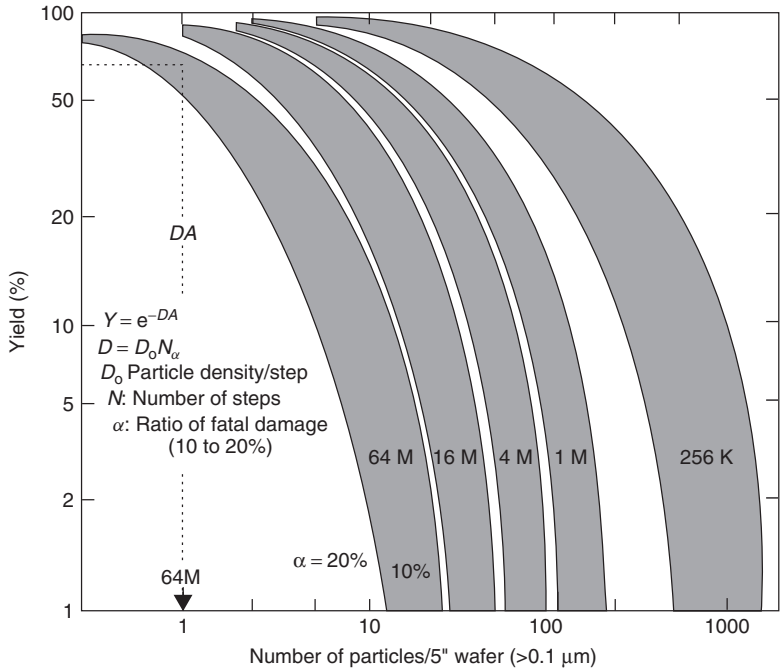


Figure 36.4 Particle-induced yield loss in DRAMs according to Poisson model. Note that only 10 to 20% of particles are assumed to cause fatal damage to chips. Source: Hattori, T. (ed.) (1998)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

are prospective killer defects, but only statistically. Fatal damage proportion has been set to range from 10 to 20% in the DRAM yield model, to give a range of yields.

36.2 PROCESS STEP EFFECT

As the number of process steps goes up, the requirements for yield in each individual step increases asymptotically. In a 100-step process, individual-step yield of 99% results in 37% total yield (0.99^{100}), but in a 500-step process it would yield <1%. Step yield of 99.99 yields 95% total. However, one single, badly yielding step, with say 70% yield, will limit the total yield to less than 70%; therefore, a process-development effort must be carried out in all process steps.

36.3 YIELD RAMPING

Process research for a new generation of chips should start around 10 years before commercial introduction. It involves exploration of new technologies and materials, and novel device structures. Around five years before introduction, the equipment should be available in single units, and two-to-three years before introduction, pilot production quantities of equipment should be purchased, say five units in a major company.

Complete circuits should be functional ca. three years before introduction. This implies device and equipment readiness, but does not give an indication of systematic or random yield. Depending on device type and company culture, 10 to 20 lots, each taking one to three months (running partly in parallel) are fabricated and analysed. Production start is the date when every lot produces functioning devices.

The yield-ramp phase often determines commercial success or failure. Commodity devices such as DRAMs have a market price, and because fab investments are similar for the same generation technology, the difference in revenue comes mostly from the yield in the early phase. The IC industry has been able to prosper in spite of dire predictions about yield-limited economics. In fact, statistics show that yield-ramp rates have been steeper for new, small linewidth processes (Figure 36.5). This is partly due to the policy of building multiple identical fabs, where everything is copied from an existing fab, and data cumulates much faster than in one-of-kind fabs.

Yield stability during ramp-up and production is mandatory, as otherwise there is no yardstick for process-development efforts. Gross variations in the yield would mean that even major process improvements might be rejected because the effects of yield variation and process improvement have opposite signs. Similarly, cosmetic improvements might get an approval even though the effect came from normal yield variation. Yield decrease in the end of the lifecycle is real: it is caused by process phase-out and decreased engineering effort.

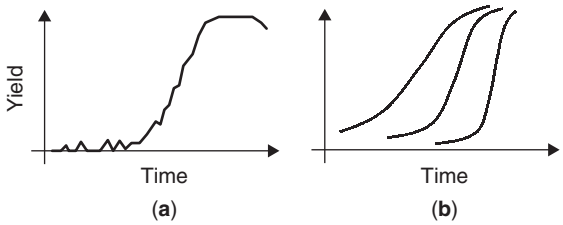


Figure 36.5 Yield over time: (**a**) yield along the life cycle of a device and (**b**) yield-ramp rates of succeeding generations. Ramp rates have become steeper in recent years

36.4 EXERCISES

1. Compare the number of $0.5\,cm^2$ chips on 100 mm and 150 mm wafers with 6 mm edge exclusion rule. Repeat for $2\,cm^2$ chips on 200 mm and 300 mm wafers with 3 mm edge exclusion.
2. If linewidth is halved but the same old cleanroom is used, what will happen to the yield?
3. Use Minesweeper (XMine for UNIX or Minesweeper for Windows) as a tool to simulate the fabrication yield: chips are 1×1, 2×2, 3×3, 4×4, 5×5 or 6×6 areas on the grid. Vary defect density (= the number of mines) and check how defect density and chip size are related.
4. What is the extrapolated yield of a new $2\,cm^2$ chip if $D = 2\,cm^{-2}$ using a model $Y = \exp(-DA)$, measured from a large sample of small chips ($<0.6\,cm^{-2}$). What is the yield if Murphy's model is used instead? How about Seeds model?
5. If 64 Mbit DRAM chips are $2\,cm^2$, what will the fabrication defect density be?

REFERENCES AND RELATED READINGS

Carlson, R.O. & Neugebauer, C.A.: Future trends in wafer scale integration, *Proc. IEEE*, **74** (1986), 1741.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Cunningham, J.A.: The use and evaluation of yield models in integrated circuit manufacturing, *IEEE TSM*, **3** (1990), 60.

Hattori, T. (ed.): *Ultraclean Surface Processing of Silicon Wafers*, Springer, 1998.

Leachman, R.C. & Hodges, D.A.: Benchmarking semiconductor manufacturing, *ESSDERC 1997* (1997).

Rao, G.P.: *Multilevel Interconnect Technology*, McGraw-Hill, 1993.

Stapper, C.H. & Rosner, R.J.: Integrated circuit yield management and yield analysis: development and implementation, *IEEE TSM*, **8** (1995), 95.

Micro Magazine, http://www.micromagazine.com/.

**Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030**

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

37

Wafer Fab

This chapter deals with high-volume IC manufacturing: MEMS fabs and niche IC fabs are considerably smaller, and more diverse than the leading edge CMOS fabs. There are some 1000 IC and 300 MEMS fabs in the world, the latter being mostly very small. Flat-panel display fabs are usually big, but they are different because of large plate size and large 'chip' size, and the lack of high-temperature processes on glass substrates.

Wafer fab cost has increased exponentially with decreasing linewidth. Cleanrooms have become more expensive as the size of a killer particle has gone down but equipment is the most expensive part of a fab. A recent estimate stated that the capital investment in tools is equivalent to 80% of the revenue that the fab is going to generate in its lifetime. All dollar values in this, and the following chapters, are bound to be crude approximations because exact numbers are not revealed by companies and because there are great variations in prices as the market fluctuates heavily (but costs tend to be quite constant). In the IC industry, both 30% annual increases and 20% decreases in production values are common (even though production volumes do not fluctuate that much). In the long run, costs and prices do follow some predictable trends, like cost per bit falling at regular rate, the cost of a processed square centimetre of silicon being constant and the cost of lithography tools and wafer fabs going up exponentially (Table 37.1).

Wafer fabs can be classified into four size categories according to their wafer starts per month (WPM):

High volume	>20 000 WPM
Medium volume	10 000 WPM
Low volume	5000 WPM
Pilot/R&D	500 WPM

In a high volume fab, there are always multiple tools for each and every process (Table 37.2) but there is also a "division of labour" between the tools: there are tubes separately for gate oxidation, other dry oxides, wet oxides, and polysilicon oxides; in a smaller fab or lab the division might be gate oxide versus other oxides, or dry oxides versus wet oxides. Megafabs have plasma etchers dedicated to oxide, poly, aluminium and tungsten. In a university lab with two plasma etchers, the division is based on fluorine- as against chlorine-based processes (or between clean and not-so-clean processes). LPCVD processes have dedicated tubes for poly, nitride and oxides, and this holds for small fabs and labs alike because thin-film interactions would ruin reproducibility. In a research lab, one sputtering system can take care of all metal depositions, but production

Table 37.1 Fab investment for volume manufacturing (top fab of its day)

1957	$0.2 million
1967	$2.5 million
1977	$10 million
1987	$100 million
1997	$1000 million
2007	$3000 million (estimated)

Table 37.2 Equipment numbers for a 25 000 WPM fab

Lithography tools	35
Wet stations	70
Oxidation/diffusion tubes	30
Ion implanters	15
LPCVD tubes	10
PECVD reactors	40
Plasma etchers	50
Metal deposition systems	40
CMP tools	60

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

sputters are dedicated to certain films or film stacks exclusively.

37.1 HISTORICAL DEVELOPMENT OF IC MANUFACTURING

In addition to the scaling of lateral and vertical dimensions, a multitude of other refinements has taken place in IC manufacturing during the last 40 years. These involve new materials for metallization as well as dielectrics, new equipment designs, new control measurements and inspections tools, new contamination control strategies as well as new devices (Table 37.3).

Lithography has evolved from 1X contact/proximity printers to 4X step-and-scan machines. Batch wet etching has been replaced by single-wafer plasma etching. Furnace diffusion has been replaced by ion implantation. Some processes, such as wet cleaning and thermal oxidation have remained unchanged. The industry has been quite conservative, with very few radical changes in any one technology generation.

Early transistors could be made with just five elements: Si, B, P, O and Al; the fabrication of 0.18 μm CMOS uses 14 elements: in addition to the aforementioned, N, As, Ti, W, Co, Ta, Cu, C and F are used. Polysilicon, tungsten, copper and low-*k* dielectrics have been major shifts and the new gate dielectrics HfO_2, ZrO_2 and $BaSrTiO_3$ will present a major shift because they are deposited films, unlike thermal oxides, which are grown.

Plasma etching, wafer steppers, CMP and electroplating have been major tool changes, but the shift from batch to single-wafer processing has been equally important. Sometimes, new materials can be introduced without new tools: diffusion barriers are sputtered films, and aluminium alloying for EM resistance did not affect sputter systems. However, silicides necessitated RTP, and tungsten required CVD. LOCOS, self-aligned polysilicon gate, LDDs and STI have been major shifts in MOS device structures. Taken together, these developments, both revolutionary and evolutionary, have contributed to the transistor number going from one per chip to 100 000 000 in 40 years.

Thin-film head (TFH) fabrication for magnetic data storage, surprisingly, shares many aspects with IC fabrication, especially the steady growth in the number of process steps, the number of thin films (up to 20) and the steady (and very steep) decrease in linewidths: from 1990 to 2000, the minimum linewidth in TFH fabrication came down from 5 to 0.5 μm, and by 2010 it is speculated to be equal to IC linewidths. This means

Table 37.3 Historical development of IC processes

1960 to 70s processes
- 30 to 3 μm linewidths
- proximity and projection 1X lithography at $\lambda = 436$ nm
- fewer than 10 lithography steps
- wet etching
- doping by furnace diffusion
- batch processing
- (pure) aluminium metallization; one level of metal
- Si, O, N, P, B, Al needed
- wafer size increase from 1" to 3"

1980s processes
- 3 to 1 μm linewidths
- step-and-repeat lithography at $\lambda = 365$ nm introduced at 1.2 μm
- 10–15 lithography steps
- plasma etching replaces wet etching for critical steps
- ion implantation for doping
- single-wafer equipment emerging, first in plasma etching
- two levels of metallization
- SOG and resist etchback planarization
- silicides introduced
- new elements: As (n-doping), Cu (in Al-alloy), Ti, W (in TiW barrier)
- 100/125/150 mm wafer size

1990s processes
- linewidths 1 to 0.25 μm
- 20–25 lithography steps for advanced CMOS
- high density plasma (HDP) equipment for etching and deposition
- W-plugs by CVD with TiN barrier
- CMP oxide planarization
- Cu metallization introduced in damascene structure
- number of metal levels increasing up to seven in logic circuits
- 150–200 mm wafer size

2000s processes
- linewidths 0.25 μm and smaller
- 30 lithography steps for advanced CMOS
- step-and-scan lithography with $\lambda = 248$ nm introduced at 0.25 μm
- phase shift masks (PSM) adopted at 0.18 μm
- new elements: Co (in $CoSi_2$), F (in SiOF), Ta (in TaNSi barrier for Cu)
- copper becoming standard for high-performance circuits
- low-*k* dielectrics introduced in multilevel metallization
- 300 mm wafer size emerging

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

that hard disk drive memory density increases faster than semiconductor memory density.

37.2 MANUFACTURING CHALLENGES

The IC industry is faced with a number of challenging issues in fab economics, device structures and packaging. Fab cost is not only high, but the amortization times are also very short, five to seven years only. Lithography cost, especially, is rising very fast, with 20- to 30-million-dollar pricetags for lithography tools in sight. Wafer size transition from 200 to 300 mm introduces additional costs because all tooling has to be upgraded, not just process tools but metrology and test tools as well. Most of the 300 mm tools for the 0.13 μm generation can later be upgraded for the 90 nm generation, and a few are going to be useful even in the 65 nm generation. In 2003, there were 30 fabs running 300 mm wafers.

With 100 million transistors on a 0.13 μm logic chip (which translates to some 20 to 30 million devices per square centimetre), design complexity is enormous, and the same applies to device testing. CMOS was originally a solution to power consumption: CMOS logic consumes energy only during switching, but the sheer number of devices means that excessive amounts of waste heat are generated in advanced chips. Chip cooling has two elements: hot spot cooling and overall cooling. Power consumption of 100 W is becoming typical in high-performance processors (power densities 30 W/cm^2), whereas processors for battery-powered devices consume only a fraction of a watt. Connections from the chip to the outside world require some advanced solutions: attaching lead to just chip periphery is not enough when 1000 connections need to be made. Various ball grid and bump-metallization schemes have been introduced. In these approaches, the traditional division of labour between wafer fab and the packaging house is shifting; a packaging house can do wafer processing – lithography, electrodeposition of bump metal and bump anneal – before the usual steps of testing, dicing and assembly.

Because photomask cost is rapidly rising, it is becoming increasingly difficult to make small series production. A photomask set for advanced CMOS can cost $500 000, and if a wafer sells for $10 000, anything below 50 wafers does not cover even the non-recurring starting costs. Semi-custom chips solve this problem, at least partially: front-end processing, and therefore the transistors, is identical in all products, and chips are customized by a few customer-specific photomasking steps later in the process. In the best case, only one mask is product-specific, and all the other masks are shared between many products. Of course, semi-custom chips cannot use silicon area very efficiently, but the cost reduction relative to full custom design is significant.

37.3 CYCLE TIME

Cycle time (CT) is the number of days it takes to complete a lot. Process time (PT) is the actual time it takes for the wafer to be processed. Process time is the total time when processes act on the wafers, while cycle time includes idle time, like queuing. The ratio of cycle time to process time, CT/PT, is a measure of fab efficiency. For standard processing, CT/PT is about 2; wafers spend half the time in queue and storage.

Cycle time and process time are intimately coupled to batch versus single-wafer tool combination in a fab. Most front-end processes are batch, and most backend processes, single-wafer. For batch processes, process time is 'overhead + batch time', which is fairly constant; but for single-wafer processes process time is 'overhead + lot size × single-wafer time', and lot size has a major effect. All-single-wafer fabs have been experimented with, and record cycle times of three days have been demonstrated for 0.25 μm CMOS. There are no single-wafer fabs running volume production, but in order to reduce risks associated with billion-dollar fabs, the *minifab* concept has been created. Minifabs are low-volume fabs with mostly single-wafer and some small-batch equipment (batch size of 25 wafers in thermal processes, versus 200 wafer batches in high volume fabs). Such minifabs are expected to be more agile because the cycle times will be shorter, and production scheduling is going to be more flexible. There will be little equipment duplication, and only some dedicated equipment for certain process steps. One thermal processor might be running various processes, maybe with only front-end versus backend separation, which is for keeping metallic contamination at bay.

Other ways to reduce cycle time include lot status and priority classification schemes. Hot lots (aka *rush lots*) are priority lots that receive preferential treatment in the fab. When a hot lot arrives at a process tool, it is processed in front of the queue. Hot-lot cycle time may be 30% less than that of a regular lot. 'Super hot' lots (aka *bullet lots*) are even more prioritized: process equipment is reserved for the super-hot lot so that it can be processed as soon as it arrives. For a super-hot lot, CT/PT is thus 1, but there is a way to reduce

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

CT/PT even further: in the backend of the process the lot is made smaller; for instance, only three wafers will be processed to completion and CT/PT can be as low as 0.5. There can be only a limited number of hot lots running simultaneously because they disturb the normal fab operations.

Yields of hot lots tend to be consistently better than those of standard lots. This can be explained by a simple particle deposition model: hot lots spend less time in the wafer fab, and there is less time available for particles to deposit on the wafers.

Split lots, which have process variations designed in them (e.g., wafers having different implant doses but otherwise identical processing), carry a wealth of information, but at the enormous cost of experimentation. In split lot experiments, it is important to understand which process steps are single-wafer and which are batch, because running split lots in batch processes is time-consuming.

Regular wafers are run in lots of 25 or 50 wafers. For batch processes such as oxidation, many batches are combined, which leads to higher CT/PT. Sometimes, a lot is made up of 24 wafers plus a monitor wafer. The monitor wafer is not physically one and the same wafer but an allocation only: in gate oxidation, it is a prime wafer that then continues to polysilicon deposition, poly doping and polysilicon etching, and exits after that. A new monitor wafer starts at first inter-level dielectric deposition, and is then used as a contact hole etch monitor and as first metal resistance and step coverage monitor. This monitor is not a prime wafer, but a monitor-quality wafer.

In addition to device and process-specific monitor wafers that run with the product wafers, a lot of other monitor wafers run in a wafer fab. These are used for

- equipment qualification, for example, after maintenance;
- regular monitoring, for example, particle tests, film thickness/uniformity;
- process development, for example, modifying an existing process step;
- short loop test wafers, for example, via-chain test.

In the start-up phase of a new fab, product wafers may in fact represent less than half of all the wafers. Test/monitor wafers are often re-claim wafers. Reclaim wafers are wafers that have been "reconditioned" after processing. Thin films have been etched away, and the wafers may have been re-polished and inspected. Re-claim wafers have been through various process steps, especially thermal processes, which affect the properties of the wafer bulk, for example, oxygen precipitation and wafer curvature. Re-claim wafers are cheaper choices for non-critical tests: as thin-film thickness monitors, as equipment qualification wafers or as regular particle-test wafers.

37.4 COST-OF-OWNERSHIP (CoO)

Difficulties in tool performance assessment have led to the introduction of a new figure-of-merit, the cost-of-ownership, CoO, which tries to put all tools on equal footing, calculated over the lifetime of the tool. Equipment capital investment has very little meaning in IC cost calculations if other major factors such as yield and throughput are neglected. CoO is an estimate of all costs associated with a certain piece of equipment, and it can be used to compare different mixes of fixed and running costs. Yield, or alternatively cost per good chip, is of paramount importance, and therefore CoO-models are rather 'personal': equipment maintenance, process specification tightness/looseness, the number of monitor wafers, all affect the yield, and the yield has often the biggest contribution to CoO.

$$\text{Cost/wafer} = (\text{tool cost/throughput}) + \text{process cost} \tag{37.1}$$

Process cost includes chemicals, targets, water, labour, electricity, administration, and so on. Wafer cost can be added, or treated separately. Cost-of-ownership (CoO) is defined as

$$\text{CoO} = \frac{\text{equipment} + \text{labour} + \text{consumables} + \text{operation} + \text{yield loss}}{\text{equipment life} \times \text{throughput} \times \text{utilization} \times \text{rework rate}} \tag{37.2}$$

The following calculation (from Moritz, H.: *Professional i-line Lithography, Lecture Notes*, IBM, 1993) shows how different components relate to lithography cost.

Equipment cost	\$3 500 000
Equipment life	5 years
Utilization	85%
Throughput	25 wafers/hour
Rework rate	0.90 (= 10% of wafers reworked)

This translates to 826 000 wafers processed during equipment lifetime, or investment cost of \$4.7 for a lithography step. Process cost is estimated as follows:

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Labour	\$1.7/wafer
Consumables (resist, etc.)	\$2/wafer
Operation (electricity, etc.)	\$0.15/wafer

Total lithography cost is then \$8.55/wafer. So far, 100% yield has been assumed but in real life, yield loss severely affects the actual number of good chips. Assumptions for yield loss calculation are

200 mm wafer size;
350 chips/wafer (0.85 cm^2);
0.01 defects/cm^2 from lithography;
cost of good chip \$3.

Systematic loss comes from tails of statistical distributions: 3σ process capability in both alignment and in linewidth yields 99.4% good chips, or 346 good chips ($0.994 \times 0.994 \times 350$), with four scrap chips. Stochastic losses are calculated from defect density: 0.01 defects/cm^2 translates to three defective chips per wafer. Cost of scrap chips is then \$21, or two and a half times the cost of equipment and its operation. Therefore, even minor improvements in yield will contribute enormously to the bottom line.

37.5 COST OF PROCESSED SILICON

Looking at the cost structure a bit further, the cost of silicon chips can be seen to consist of three elements (after Warwick, C. & A. Ourmazd):

- cost of wafer processing (both capital and running costs);
- cost of scrap (yield loss);
- cost of assembly.

The cost of processed, untested silicon is k_1 \$/cm^2 (all costs in the calculation are normalized to square centimetre of silicon area).

Scrap cost depends on yield according to k_1/Y where Y is modelled by

$$Y = (1 + (1/2D_oA))^{-2} \tag{37.3}$$

Rent's rule assumes that a chip is divided into $n \times n$ circuit blocks with inter-block spacing of b (Figure 37.1). This chip can then be accessed via $4n$ pins at the chip periphery. The number of pins P required for chip area A is

$$P = 4\sqrt{\frac{A}{b}} \tag{37.4}$$

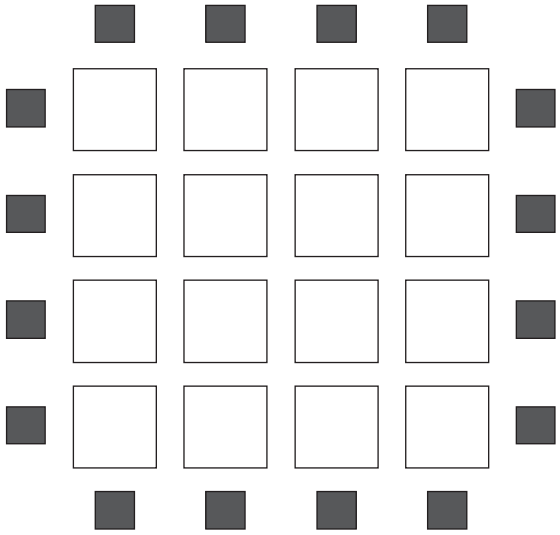


Figure 37.1 Rent's rule: $n \times n$ array can be accessed from the edges via $4n$ pins

Cost of off-chip connection via a pin is experimentally estimated to be 10 cents/pin. The assembly cost per area is $k_2/\sqrt{A}$ \$/cm^2. A chip with 1 cm^2 area and 400 μm inter-block distance has $4\sqrt{1\,cm^2}/0.04\,cm =$ 100 pins, or 10 \$/chip assembly cost. Total cost is thus

$$k_1((1 + (1/2D_oA))^{-2} + k_2/\sqrt{A}\ \$/cm^2 \tag{37.5}$$

If the chip size increases, the assembly cost is reduced because fewer chips need to be assembled, but the scrap cost increases with chip size. Assuming defect density of 0.3/ cm^2 and cost of processing \$10/cm^2, the minimum cost point is at 1.3 cm^2 chip size (Figure 37.2(a)).

The cost of processing has remained more or less constant over 30 years, which is remarkable considering the growth in complexity of fabrication processes. This cost always refers to the most advanced, yet established, process technology of its day; older technologies are cheaper. In 2000, fabless companies paid approximately \$8/cm^2 for 0.25 μm CMOS on 200 mm wafers, and \$2.6/cm^2 for 0.8 μm CMOS on 150 mm wafers.

Defect-density scaling can be estimated from historical trends: there has been a constant 20% per year reduction in defect density. In the year 2010, D_o will then be 0.01 cm^{-2}, a factor of 30 improvement. However, the optimum chip size increases only by a factor of 10 to 13 cm^2 (Figure 37.2(b)).

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

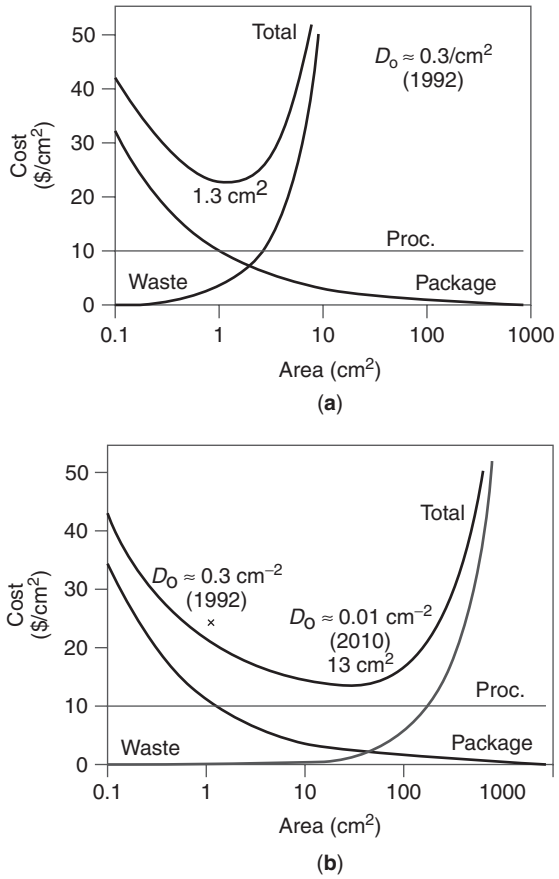


Figure 37.2 Optimum chip size with defect density 0.3 defects/cm^2. Reproduced from Warwick, C. & A. Ourmazd (1993), by permission of IEEE

37.6 EXERCISES

1. The investment for a large-volume wafer fab is \$1 billion (year 2000, 0.25 μm technology, 200 mm wafer size). The fab running costs are \$1 million/day. Assuming 30 000 wafer starts per month (WPM), what will be the cost of finished silicon?
2. Calculate the mask-cost contribution to silicon area price if 0.25 μm CMOS with 25 photomasks at \$3000/mask plate are used, and each mask set is used to fabricate 50/500/5000/50 000 wafers?
3. Maskless lithography by direct writing is expensive because it is very slow, but there is no photomask cost. Assuming identical capital investment (\$6 million) and running costs (\$0.5 million/year) for both optical and direct write lithography systems (a very crude approximation), and 100 WPH for optical and 2 WPH for DW on 300 mm wafers, what would be the number of wafers at which DW becomes competitive with optical lithography for 0.1 μm CMOS if the mask set cost is assumed to be \$500 000?
4. If photoresist stripping in a 30 000 WPM fab is 50/50 between wet tanks and single-wafer plasma strippers, how many wet benches and plasma strip tools are needed? Make assumptions about throughputs based on similar processes/tools.
5. If a 30 000 WPM fab has four gate oxidation tubes, what is their average utilization?
6. Under the conditions of $10^{15}\ cm^{-2}$ phosphorus implant dose, 200 mm wafer size, PH_3 bottle volume 3 L (STP), how many wafers can be implanted? If ion current is 1 mA, what is the interval for bottle changing?
7. If a 2 cm^2 chip has 1000 output pins, what would be the pin pitch at the chip periphery if an arrangement such as the one in Figure 37.1 was employed?
8. How many 30 000 WPM fabs are there in the world?

REFERENCES AND RELATED READINGS

Diebold, A.C.: Materials and failure analysis methods and systems used in the development and manufacture of silicon integrated circuits, *J. Vac. Sci. Technol.*, **B12** (1994), 2768.

Doering, R. & Y. Nishi: Limits of integrated circuit manufacturing, *Proc. IEEE*, **89**(3) (2001), 375.

Leonovich, G.A. *et al*: Integrated cost and productivity learning in CMOS semiconductor manufacturing, *IBM J. Res. Dev.*, **39** (1995), 201.

Liehr, M. & G.W. Rubloff: Concepts in competitive microelectronics manufacturing, *J. Vac. Sci. Technol., B*, **12** (1994), 2727.

Moritz, H.: *Professional i-line Lithography, Lecture Notes*, IBM, 1993.

Spanos, C.J.: Statistical process control in semiconductor manufacturing, *Proc. IEEE*, **80** (1992), 819.

Warwick, C. & A. Ourmazd: Trends and limits in monolithic integration by increasing the die area, *IEEE TSM*, **6**(3) (1993), 284.

Wood, S.C.: Cost and cycle time performance of fabs based on integrated single-wafer processing, *IEEE TSM*, **10** (1997), 98.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Part VIII

Future

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

38

Moore's Law

This chapter deals with the past, present and future of integrated circuits, concentrating on CMOS, which is driving scaling into smaller linewidths and higher device densities. Devices, fabrication processes and industrial issues are discussed with future trends, limits, opportunities and threats to continued scaling.

38.1 FROM TRANSISTOR TO INTEGRATED CIRCUIT

Transistor fabrication in the 1950s was crystallography and metallurgy, not microfabrication. Junction formation was an alloying process that did not share many features with modern transistor fabrication. Pallets of indium, a p-type dopant, were attached to both sides of an n-type semiconductor piece, the diffusion step was performed and metal wires were attached to the two p-type and one n-type region and *voilá*, the pnp-transistor was ready.

The modern key concepts of microfabrication: diffusion masking by an oxide layer, photolithographic patterning, wet etching of the oxide and the use of evaporated aluminium as a conductor emerged in the mid-1950s mostly at Bell Laboratories and at Fairchild Semiconductor. These techniques were put together by Jean Hoerni, in what is known as the *planar process* for transistor fabrication.

The integrated circuit (IC) was invented twice, simultaneously and independently. Jack Kilby of Texas Instruments demonstrated ICs in 1958 and filed for a patent in early 1959. However, Kilby used germanium transistors and gold wires bonds for connecting the devices. Robert Noyce at Fairchild based his invention on the planar process, using evaporated aluminium for metallization and silicon dioxide as an insulator, and created the first device that became the forefather of current ICs.

Integration of transistors was only part of the story: integration of analog elements, resistors and capacitors was also open to new vistas. Because resistances and capacitances are not very accurate, it is useful to use ratios of these rather than the absolute values. Integration of analog elements on the same chip resulted in major improvement in ratios compared to discrete components.

There were many objections to ICs in the beginning of the 1960s, as Jack Kilby reminisces:

1. Electronics designs would become hard to change once the circuits had been etched onto silicon.
2. Electronics engineers would be out of jobs because all design would shift to IC manufacturers.
3. Transistors are low-power devices that are suitable only for some special applications.
4. ICs do not use optimum materials: NiCr resistors are better than silicon resistors, and Mylar capacitors are superior to oxide capacitors.
5. Yield of transistors is low, for example, 80%, and if, say, 20 of them are made on a single chip, the combined yield will be miniscule.

Argument number one still holds today: especially, custom circuits take a long time to design and to fabricate, and changes are hard to make. This is, however, a small price to pay for the enormous gains in speed and functionality. We now know that argument number two was groundless as ICs propelled the electronics industry into super growth. Argument number three was wrong, and some people had already seen it in the 1950s: Bob Wallace of Bell Labs stressed,

"Gentlemen, you've got it all wrong! The advantage of the transistor is that it is inherently a small-size and low-power device. This means that you can pack a large number of them in a small space without excessive heat generation and achieve low propagation delays. And that's what we need for logic applications. The significance of the transistor is not that it can replace the

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

vacuum tube but that it can do things that the vacuum tube could never do!" (from reference Ross).

Many MEMS and nanodevices today are miniaturized versions of existing devices. Sometimes, a smaller size is useful because it results in, for example, smaller power consumption or higher speed. However, it is equally important to look for new applications in which new physical phenomena, new combinations of speed and power can be utilized, or where macroscopic counterparts do not exist, or where the scale economies of microfabrication have not yet been utilized.

The whole can be more than the sum of its parts. The very concept of integration seems to have escaped the attention of the supporters of argument number four. Integrated circuits paved the way for more powerful electronic systems. And the savings in assembly costs quickly more than compensated the higher cost of ICs.

Argument five was mathematically valid, but it was based on the technology of its day, and it did not anticipate the tremendous strides in microfabrication technologies. The success of ICs has been dependent on the fact that in spite of continuous miniaturization and complexity of the manufacturing process, the yield of individual transistors on ICs has improved dramatically. In 1960, the yield of 50% for individual devices resulted in a 3% yield for a five-transistor IC (today, integrated microfluidic systems face a similar situation: while pumps, valves and mixers may have reasonable yields, systems consisting of many such devices have rather low yields). In the year 2000, 64 Mbit DRAMs with some 130 million transistors and capacitors were manufactured with ca. 90% yields, which translates to practically a 100% yield for individual devices, for 0.25 μm devices, compared to the ca. 25 μm devices of 1960.

The early proponents of the IC had to balance between two options:

1. Suitable only for price-insensitive applications like military or space technology.
2. Will be cheap in the future once technology matures.

Early growth was, of course, along the first argument because somebody had to pay for the chips but at the end of the 1960s the second argument was finally realized, and the IC became a household term.

38.2 MOORE'S LAW

The development of ICs seemed to follow a regular pattern: doubling the number of devices on the chip every year. In 1965, Gordon Moore spoke about this pattern. The observation was based on few data points, but the conclusion became famous. Later, the prediction was revised to doubling every 18 months, and this version has been especially long lasting. It has been dubbed Moore's law, even though it is only an empirical pattern without fundamental justification (Table 38.1).

Moore's 1965 prediction extended till 1975 and his extrapolation was quite accurate. The trend has continued approximately at the predicted speed, give or take some fluctuations. At the turn of the millennium, the pace has been even faster than that predicted by Moore's law. DRAM memory chips are best suited for Moore's law studies because the law is about production economics: chip size and cost minimization. Processors are governed by quite different laws: they are design-heavy, rather than manufacturing-driven, and proprietary architectures are not subject to ultimate cost reductions. One gigabit DRAM circuits were unveiled with 0.18 μm geometries as predicted, and 4 Gbit DRAM memory circuits with 0.10 μm dimensions are being made in 2003. It should be borne in mind that sometimes the product demonstration date is used (when the first fully functional chips are fabricated), sometimes the production start date is used and sometimes the peak production year is stated.

Shrink versions make the situation more complex: the first functional 1 Gbit DRAMs were demonstrated using 0.18 μm technology, but production versions have been made at smaller linewidths: 0.13 to 0.10 μm. Add to this the minor differences between companies, and it is fair to accept discrepancies of a few years in Moore's law data.

Moore's law was originally proposed in the era of bipolar transistors and it has held well in the era of PMOS, NMOS and CMOS, and it seems to hold for the next decade of strained silicon and SOI-CMOS and other evolutionary MOS technologies. Moore's law is about device-packing density and cost, and not about any particular technology. There have been a number of dubious extensions of Moore's law: it has been said to apply to computing power, which is not true because computer architecture is not part of Moore's law. Despite its non-fundamental nature, it is one of the few predictions about future technology that has held for 40 years.

Linewidth scaling has been very predictable. Junction depths were scaled with linewidths as L/5 for decades, but more recently it has been difficult to scale x_j down as aggressively as linewidth. Gate oxide thickness used to scale as L/45 for a long time, but with oxide thickness now approaching one nanometre EOT, it is not possible

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Table 38.1 Moore's law

Year	Transistors/chip	DRAM	Linewidth	Wafer size
1959	1		30 μm	0.5″
1960	2			
1961	4			
1962	8			1″
1963	16			
1964	32		20 μm	1.5″
1965	64			
1968	256		12 μm	2″
1970	1024	1 k	8 μm	
1973	4096	4 k	5 μm	
1975	16 384	16 k	3 μm	3″
1979	65 536	64 k	2 μm	
1983	262 144	256 k	1.5 μm	100 mm
1986	1 048 576	1 M	1.2 μm	125 mm
1989	4 194 304	4 M	0.8 μm	150 mm
1992	16 777 216	16 M	0.5 μm	
1995	67 108 864	64 M	0.35 μm	200 mm
1998	268 435 456	256 M	0.25 μm	
2000	536 870 912	512 M	0.18 μm	
2002	1 073 741 824	1 G	0.13 μm	300 mm
2004	2 147 483 648	2 G	90 nm	
2006	4 294 967 296	4 G	65 nm	
2008	8 589 934 592	8 G	45 nm	
2010	17 179 869 184	16 G	32 nm	

to continue at a regular pace. Devices are now being designed with different criteria according to their power consumption: in high performance (HP) systems, gate oxide is aggressively scaled down and leakage currents are allowed to increase, but in low power (LP) portable electronics, leakage currents are minimized by using 'thicker' oxides: 2.4 nm versus 1.3 nm for HP. There are a number of demanding scaling issues as we go from established 130 nm technology to 65 nm technology. Some of these are collected in Table 38.2.

The death of Moore's law has been much discussed but newer predictions of IC scaling have often proven inaccurate, even in a quite short term: in 1994, it was predicted that 0.1 μm technology would become available in 2007, microprocessor chips would have 350 million transistors and operate at 1 GHz with 1.2 V, which was wrong with the date, too high on the transistor count and too pessimistic on the speed. In 1986, it was predicted that 16 Mbit DRAMs would be available at the turn of the millennium, but 256 Mbit was available. Around 1980, the prediction was that optical lithography could not print lines smaller than 1 μm and in 1989, the end of optical lithography was predicted for 1997. Quite regularly, the end of optical lithography has been predicted to be 10 years into the future, and this same prediction holds true even today. In 1989, it was also assumed that silicon dioxide as the gate oxide would be replaced by high-k dielectrics starting from 1993, but in 2003 high-k is still in the development phase. Long-term predictions have been off by a far wider margin: in 1984, linewidth predictions for 2007 were 0.1 μm (optimistic case) and 0.5 μm (pessimistic case).

How long can this scaling continue? If all goes as predicted by Moore's law, in 2059, the 100th birthday of the IC, we will have:

- 2.5 Å minimum linewidth;
- 0.04 Å gate oxide thickness;
- 2 mV operating voltage;
- 64 exabit DRAMs (exa = 10^{18}).

Obviously, a scaled version of the current MOS transistor cannot be the device described above. However, remember that Moore's law is independent of device technology. The first working 1 μm MOSFET was reported in 1974, and ca. 15 years later 1 μm devices entered mass production. The first 100 nm

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Table 38.2 Scaling trends from 130 to 65 nm. Adapted from ITRS Technology roadmap

Technology generation	130 nm	65 nm	
Half pitch (DRAM)	130 nm	65 nm	
Half pitch (processor)	150 nm	65 nm	
Physical gate length L_g	65–100 nm	25–37	(HP vs. LP)
L_g variation (3σ)	6 nm	2.5 nm	
Gate oxide thickness	1.3–2.4 nm	0.6–1.4 nm	(HP vs. LP)
Drain extension	27–45 nm	12–19 nm	
Contact junction x_j	48–95 nm	18–37 nm	
Spacer thickness	48–95 nm	18–37 nm	
Drain extension junction abruptness	7.2 nm/dec	2.8 nm/dec	
R_s drain extension, PMOS	400 ohm/sq	760 ohm/sq	
R_s drain extension, NMOS	190 ohm/sq	360 ohm/sq	
Silicide thickness	36 nm	14 nm	
Silicide sheet resistance	4.2 ohm/sq	10.5 ohm/sq	
Channel doping	$4 \times 10^{18}/\text{cm}^3$	$2.3 \times 10^{19}/\text{cm}^3$	
Number of metal levels	8	10	
Local wiring pitch	350 nm	150 nm	
Aspect ratio of copper	1.6	1.7	
RC-time delay (1 mm line)	86 ps	198 ps	
Copper barrier thickness	16 nm	7 nm	
Dielectric constant, effective	3–3.6	2.3–2.7	
Dielectric constant, unclad	2.7	2.1	
Wafer size	300 mm	300 mm	
Particles/wafer	<123	<77	
Site flatness	130 nm	65 nm	
OSF	$<2.8/\text{cm}^2$	$<1/\text{cm}^2$	

device was unveiled in 1987, and ca. 15 years later 100 nm devices are being mass produced. At the beginning of the third millennium, 10 nm devices exist in laboratories, and they are extrapolated to enter production before 2020. Extrapolation, however, is a tricky business.

Linewidth and gate oxide scaling are the most visible parts of scaling, but there are many other parameters that are continuously being pushed forward. The energy consumption of a logic operation was 10 nJ in 1960, 1 pJ in 1980 and only 1 fJ in 2000. Operating voltage, which was 5 V for many generations (5–0.8 μm), is now being reduced rather regularly, and 1 V operation will soon be usual for non-battery powered devices too. The number of metallization levels for logic is rapidly going up. Since 0.5 μm generation, when three levels of metals was standard, one level of metallization has been added in almost every generation, leading to 8 levels in 0.1 μm technology. The corollary trend is that of output pin-count increase, to thousands, which has led to various ball-grid like packaging solutions.

38.3 EXTENDING OPTICAL LITHOGRAPHY: PHASE-SHIFT MASKS (PSM)

In order to push for smaller linewidths, simple chrome-on-quartz binary masks put the pressure of linewidth scaling on optical lithography tools and resist chemistries. The alternative approach is to tailor the mask. This is now being introduced at 0.18 μm linewidth and smaller.

Phase-shift masks (PSM) consist of three areas, chrome, quartz and the phase shifter, a structure that produces 180° phase shift in the transmitted light (Figure 38.1). Light along the shifted path will be out-of-phase with the light going through the non-shifted part, and the amplitude will go through a zero. Intensity, which is amplitude squared, will be much steeper compared to a binary mask, which improves both resolution and edge contrast, Figure 38.3. There are many variants of PSMs, such as attenuation phase-shift masks (AttPSM) and alternating PSM (altPSM). Embedded amplitude masks (EAM) and light guiding masks (LCM) are not unlike PSMs.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

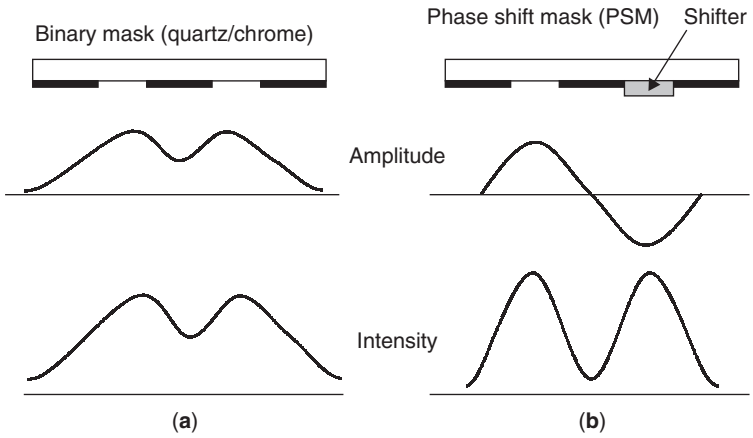


Figure 38.1 Binary mask (**a**) and alternating phase-shift mask (**b**) compared: amplitude goes through zero for PSM, and intensity (= amplitude squared) is steep

Phase shift for light travelling in the air for a distance L is $\Phi = 2\pi L/\lambda$, and for light travelling in the phase shifter material with index of refraction 'n', $\Phi = 2\pi nL/\lambda$. For a 180 degree-phase shift, $\Delta\Phi = 180^{\circ}$, the condition for shifter thickness is given by

$$L(n-1) = \lambda/2 \tag{38.1}$$

For $\lambda = 193\,\text{nm}$ (ArF laser) and $n = 1.6$, shifter thickness is ca. 200 nm, which is not unlike 100 nm chrome thickness in binary masks.

In an alternating PSM, a shifter is either etched or deposited for every second feature which limits altPSM applications to regular arrays. A rim shifter (see Figure 38.2) utilizes undercut and it can be applied to any pattern, shape and size.

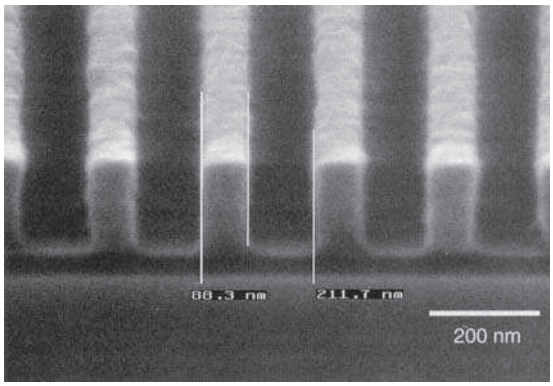


Figure 38.2 PSM enables $\lambda/2$ lines to be printed: 100 nm lines with 193 nm light source. Reproduced from Fritze, M. *et al*. (2003), by permission of IEEE

The rim-PSM fabrication makes use of ingenious self-alignment with backside illumination: an ordinary binary mask is fabricated first, with chrome patterns on a quartz plate. The shifter material is then deposited all over the plate, and the photoresist is spun. The structure is then exposed from the opposite side of the mask plate and the chrome acts as a self-aligned mask for the shifters. The shifters are then etched, followed by chrome undercutting in a second etching step.

Standard single-exposure process flow

chrome deposition
shifter deposition
photoresist application
pattern generation
shifter etching
chrome etching and underetching
photoresist stripping.

Chrome undercutting in both methods results in exactly the same degree of dimensional control. The difference is in mask inspection and repair: in the self-aligned method, the chrome pattern can be inspected and repaired before shifter fabrication. Lack of inspection and repair for PSMs has been the main factor holding back their adoption. Because of complexities in both design and fabrication of PSMs, they have not been widely used. At 0.18 μm and below, PSM has been adopted (Figure 38.3). Estimates put PSM prices at \$10 000 per mask level and \$20 000/level are seen for future reticles.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

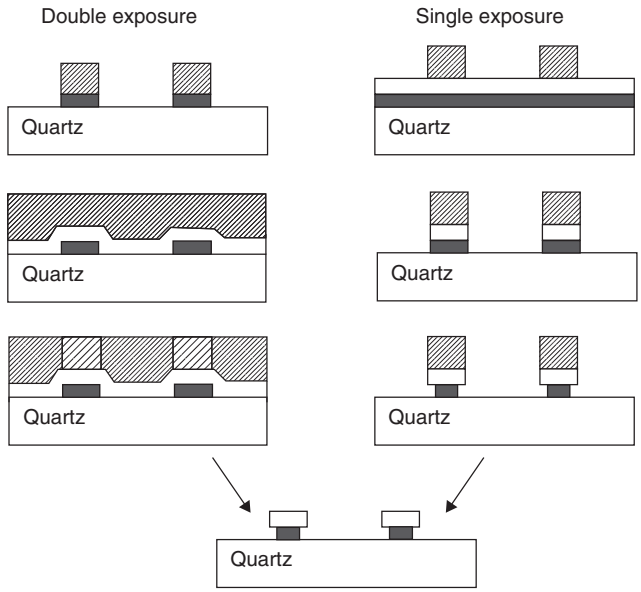


Figure 38.3 Two schemes for fabrication of rim-PSMs: double exposure self-aligned on the left; standard single exposure on the right side. Both processes result in an identical mask plate. See text for details

38.4 ALTERNATIVES TO OPTICAL LITHOGRAPHY

38.4.1 Extreme ultraviolet lithography (EUVL)

Extending optical lithography from DUV to extreme UV involves more changes than previous wavelength reductions. A new light source needs to be developed: at 157 nm, F_2 laser is a candidate but at 126 nm, the choices are open. Below 193 nm, lenses and masks need to be fabricated out of CaF_2 instead of quartz because quartz absorption becomes too high at 157 nm. The high thermal expansion coefficient of 19 ppm/°C of CaF_2 presents major problems with thermal control. A shift from refractive optics to reflective optics would present an even greater paradigm shift. Resist absorption is high at 157 nm and it is not clear if evolutionary approaches in resist chemistry are feasible.

38.4.2 X-ray lithography (XRL)

X-ray reduction optics do not exist, which means that 1X photomasks have to be used, in contrast to optical lithography which relies on 4X reduction masks. In addition to this, the blocking layers need to be thick to effectively block x-rays: heavy elements such as tungsten or gold are used. Aspect ratios of chrome lines on an optical reticle for 0.13 μm linewidths on a wafer are 1:5, whereas in XRL it is 8:1, a factor of 40 difference. XRL has many advantages over optical lithography: the exposure field is large and XRL is relatively insensitive to small particles because, for example, 0.5 μm silicon particles are relatively transparent to X-rays. Traditional X-ray sources are not bright enough to produce reasonable throughputs, so new sources have been developed: synchrotron radiation storage rings and laser plasmas. This leads to enormous starting costs for XRL systems.

38.4.3 Electron and ion projection lithographies

Because direct writing with electron or ion beams is slow, masked versions have been sought after. In electron- and ion- projection lithographies (EPL, IPL), a broad beam illuminates the mask, and the main problem again is the mask: electrons and ions need to be admitted through the mask at selected sites, and blocked elsewhere. This leads to masks with thick (blocking) areas and thin or open (transparent) areas. Thin areas need to be made of low atomic weight materials for good transmission, with thickness of the order of 1 μm. And they must, preferably, be several square centimetres across for large chips to fit in a single-exposure field. Thick blocking layers on these thin membranes cause stresses and pattern distortions. Shadow mask-like structures with open areas are excluded because making doughnut-shaped objects would require two masks and exposures. The mask will be heated by the incoming beam, just like the photomask in optical lithography, but

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

additionally, ions or electrons lead to mask charging and damage. Electron scattering masks, instead of absorbing masks, have been developed for EPL. This eliminates many of the thickness, stress and heating problems. Still, at an estimated 15 million-dollar price tag, EPL systems will only write 15 wafers per hour.

38.5 FUNDAMENTAL AND PRACTICAL LIMITS

38.5.1 Linewidth and film thickness

Nominal or design width is just an idealization of a microstructure. The physical structure in silicon or in thin-film material adds its own features. These effects are more pronounced the narrower the linewidth or the thinner the film. The smaller the details we study, the more are the effects that come into play. Line edge roughness can become significant when compared with linewidth. In the extreme, it is partly a materials limitation: chrome, photoresist and thin film on wafer are granular to some extent, and for instance, polycrystalline materials may be etched at slightly different etch rates for different crystal orientations, and this preferential etching contributes to line-edge roughness.

In $TiSi_2$ formation on polysilicon, three-grain boundaries are crucial for nucleation of the C54 phase, but if the linewidth is narrow and the grain boundaries sparse, nucleation is retarded. This can be battled by increasing the annealing temperature, but this is at odds with diffusion goals, and it will also change the relative rates of silicidation and surface nitridation. Polysilicon grain-size tailoring by ion implantation before titanium deposition can be performed or alternatively the titanium deposition process can be modified by, for example, heating, or a thin (∼nanometre) intermediate layer of molybdenum can be deposited between titanium and polysilicon to modify nucleation kinetics. Yet another method is ion beam mixing: the interface between the poly and titanium is modified by ion implantation after metal deposition. The maximum projected range should coincide with the film interface for maximum modification.

When linewidth scaling is continued, the relative importance of physical effects changes. Current conduction in a $1 \times 1\,\mu m$ cross-sectional conductor line is fully characterized by classical ohmic description. Narrower lines and thinner films reach a limit at which the surface scattering contribution to resistance becomes important, and in the 10 nm-size range, quantum effects come into play and single electron conduction can be seen. The characteristic scale for non-classical effects is the mean free path, which is 40 nm for copper and 15 nm for aluminium. However, some deviation from classical behaviour has been seen even at 500 nm, probably due to grain boundary reflections, and at 100 nm linewidths, copper resistivity has been reported to increase to $4\,\mu$ohm-cm.

Film thickness downscaling at the back end is driven by the need to keep aspect ratios reasonable, even though RC-time delays inevitably increase as resistance increases in thinner wires, and capacitance increases when dielectrics are scaled down. Ultimate limits are fairly close in back-end scaling: copper is as close to minimum resistivity as any metal can practically be, and with dielectrics, $\varepsilon = 1$ (vacuum) is not so far away, with $\varepsilon = 2$ materials being introduced. Superconducting wiring was touted in the early 1990s as a solution to the resistance problem, but enthusiasm waned rapidly when the difficulties of a high-T_c superconductor deposition and structural control became apparent.

Scaling to atomic dimensions leads to inevitable limitations. Gate oxide thickness is approaching such limits: because atoms are discrete, gate oxide thickness is 'quantized' (Figure 38.4): we cannot have any gate oxide thickness, only integral multiples of atomic dimensions. Putting it another way, each transistor will have its own microscopic oxide thickness pattern, and consequently idiosyncratic microroughness that affects channel mobility and tunnelling currents.

38.5.2 Device considerations

When MOS transistors are made extremely small, the ability of the gate to control the current in the channel is diminished. This can be overcome if two (or more) gates are to be used instead of one, as shown in Figure 38.5. Fabrication of these devices is not obvious, and the two-gate version can exist in various configurations, with the gates parallel to the silicon surface or vertical.

So far, very little attention has been paid to the MOSFET channel, but of course, the channel can be improved and tailored just like gate oxide or junctions. Strained silicon is an actively studied channel material. As discussed in connection with thin-film stresses, $Si_{1-x}Ge_x$ alloys have lattice constants larger than silicon and they are under compressive stress, and consequently the silicon on $Si_{1-x}Ge_x$ will be under tensile stress. This tensile stress introduces energy split in the conduction band of silicon, which leads to mobility enhancement, for electrons by a factor of 2 and for holes by a factor or 4 (depending on germanium content, doping level and field strength). Higher operating frequency could be obtained from MOSFETs without lithographic scaling (Figure 38.6).

The smallest MOSFETs fabricated to date, have 6 nm gate lengths, and simple ring-oscillator circuits with

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

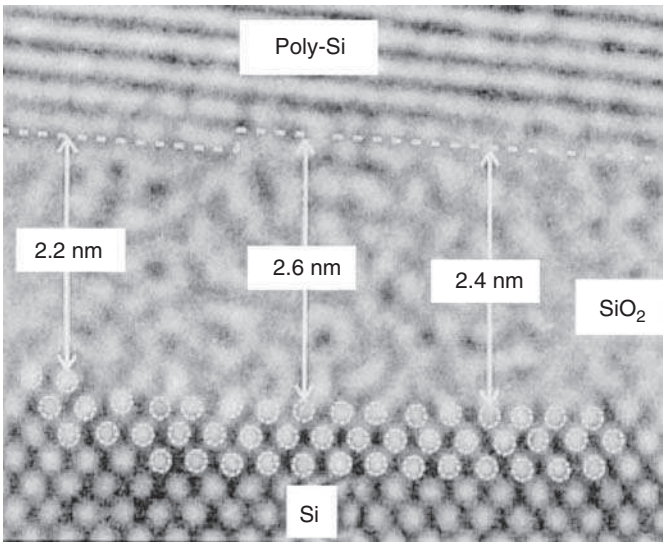


Figure 38.4 Quantized gate oxide thickness: 2.2 nm, 2.4 nm and 2.6 nm represent possible thicknesses. Reproduced from Buchanan, M. (1999), by permission of IBM

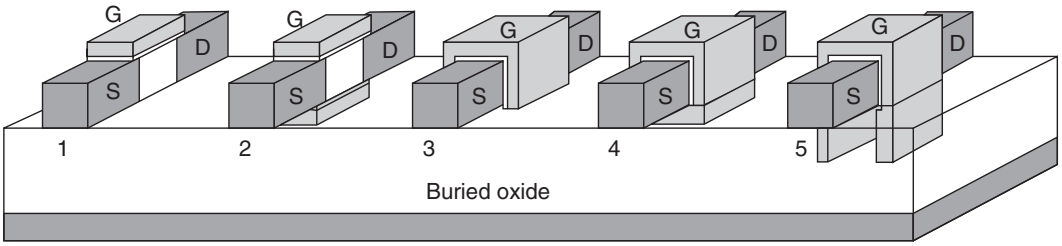


Figure 38.5 SOI MOSFETs with 1) one gate; 2) two gates; 3) three gates; 4) four gates and 5) extended three gates. Reproduced from Park, J.-T. & Colinge, J.-P. (2002), by permission of IEEE

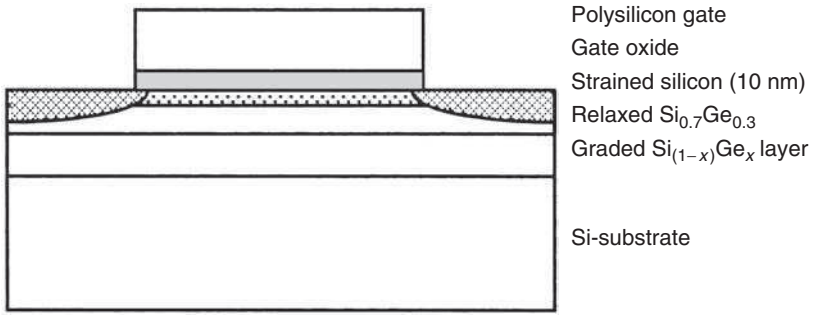


Figure 38.6 Strained silicon n-MOSFET. Silicon, with a lattice constant of 5.43 Å, experiences tensile stress on $Si_{0.7}Ge_{0.3}$, which has a lattice constant of 5.50 Å. Reproduced from Hoyt, J.L. *et al.* (2002), by permission of IEEE

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

26 nm gates have been made, too. The process used SOI wafers with 6 nm ±2 nm thick device silicon layer, and 150 nm buried oxide. Gate oxide EOT was 1.2 nm. The gate was defined by optical lithography at $\lambda = 248$ nm, using resist trimming technique, similar to the one described in Figure 10.8.

38.5.3 Statistics and yield

Yield is tied to the number of process steps, which have been increasing constantly. With 25 lithography steps, and ca. 500 steps altogether, individual step yield has to be very high. This is putting more and more demands on metrology: process monitoring precision and speed have to be increased so that more wafers can be checked. However, scaling also introduces new aspects that need to be measured: for example, junction depth is a too simple one-dimensional measure; it needs to be complemented by the junction abruptness yardstick. With ultra low-k films, film thickness and density are not enough, the pore size and pore size distribution must be known.

Despite aggressive linewidth scaling, the chip area keeps increasing. The number of defects per chip has to remain constant or decrease, which means that defect density has to be scaled down more aggressively than linewidth. The chip area increases because of the economic incentive to integrate as many functions as possible on the chip, in order to reduce packaging and assembly costs (as discussed in Chapter 37). At the moment, it seems that lithographic lenses are limiting chip size increase: it has not been possible to simultaneously improve resolution and to increase lens field size at the same pace. This, of course, applies mostly to evolutionary scaling of refractive optical systems; reflective optics, X-ray lithography or EPL have their own scaling trends.

Chemicals, DI-water, process gases and targets have been 'scaled' to higher and higher purity levels. Metal impurity levels have been reduced by a factor of 100 in four technology generations. Measurement of minutiae impurities must be available for gases, liquids and solids. Cleanrooms have been 'scaled' to higher and higher standards of purity. Cleanliness today is so high that particle measurements have hit the barrier: there are simply not enough particles to statistically assess particle purity. With increasing cleanroom cost, there has been an incentive to find alternative operation modes. Integrated processing is one such approach, keeping the wafers under controlled ambient at all times.

Statistics with extremely large or extremely small quantities can have some surprises even before ultimate limits. In a circuit with 1 000 000 000 devices, tails of statistical distributions can easily cause circuits to fail: there are 20 devices that have variations larger than six standard deviations. In very small volumes, distribution of atoms becomes a source of variation: in a 100 nm linewidth MOS transistor, the volume under the gate is ca. 100 nm × 500 nm × 10 nm ($L_{eff} \times W_{eff} \times$ inversion layer thickness), and the channel-doping level is $N_A \approx 10^{18}/\mathrm{cm}^3$, which translates to ca. 500 dopant atoms only. The small number of dopants in itself leads to detectable fluctuations in the threshold voltage, but the random positions of dopant atoms also must be considered. Standard deviation of the threshold voltage V_T is given by

$$\sigma V_T = 3.19 \times 10^{-8} t_{ox} N_A^{0.4} / \sqrt{L_{eff} W_{eff}} [V] \qquad (38.2)$$

Continued scaling to smaller dimensions together with the increase in the number of devices per chip rapidly leads to situations in which not all devices switch.

38.6 IC INDUSTRY

The IC industry has been growing at 17% annually for over 30 years, whereas the electronics industry as a whole grows only 7% annually. For the IC industry to keep growing at its historical rate, the IC content of electronics has to rise at the expense of discrete devices, circuit boards, connectors, displays, switches and keyboards, or else IC growth will slow down. ICs now account for 15% of the value of electronics. Is it reasonable to expect it to rise to 30 or to 50%, like it is in portable electronics?

Mainframe computers (1980s)	8–10% of the value consists of ICs
Personal computers (1990s)	25–33% of the value consists of ICs
Handheld devices (2000s)	40–50% of the value consists of ICs

Measures from IC manufacturing can be used to check if the rate of introduction of novel devices is slowing down. The ramp rate of production to high volumes is one measure. There are some hints that this might be slowing down. The cost of a fab compared to the revenue it is assumed to generate during its lifetime is another measure. Obviously, the former must be kept to a fraction of the latter but recently the cost of the fab has been rising faster than the revenue. Both these measures are tricky because the IC industry is very

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

cyclical, and long-term trends are easily camouflaged by annual or quarterly fluctuations.

More complex devices are introduced at regular intervals, which means that the R&D effort must grow for each successive device generation: development of the 1 Mbit DRAM has been estimated to have cost $200 million, for 1 Gbit DRAM it is estimated at $1.5 billion. So far the market size has grown steadily, which means that there have always been customers for more memory and more processing power, and therefore, the interval between introductions of new generations has been steady.

38.7 EXERCISES

1. The price per bit has been scaled down at a rate of ca. 30%/year. If 512 Mbyte of DRAM memory cost ca. $100 in 2003, how much will it cost in 10 years?
2. How far from fundamental limits are metallization RC time delays?
3. Given the scaling trend predicted by Moore's law, when will CMOS gate oxide be one atomic diameter thick?
4. The price of the refractive lens used in a wafer stepper has increased rapidly over the years: $25 000 in 1986, $102 000 in 1989, $294 000 in 1992, $670 000 in 1995 and $1.5 million in 1998. What is the price of a stepper lens today? Data from Jeong, H. *et al*: Optical projection system for gigabit DRAM, *J. Vac. Sci. Technol.*, **B11** (1993), 2675.
5. The DRAM memory cell for one bit takes up an area of $8F^2$, where F is the lithographic pitch. What is the chip size of a 1 Gbit DRAM?

REFERENCES AND RELATED READINGS

Anand, M.B. *et al*: Use of gas as a low-k interlayer dielectric in LSI's: demonstration of feasibility, *IEEE TED*, **44** (1997), 1965.

Asenov, A. *et al*: Simulation of intrinsic parameter fluctuations in decananometer and nanometer-scale MOSFETs, *IEEE TED*, **50** (2003), 1837 (special issue on nanoelectronics).

Buchanan, M.: Scaling the gate dielectric: materials, integration and reliability, *IBM J. Res. Dev.*, **43** (1999), 245.

Bühling, S. *et al*: Resolution enhanced proximity printing by phase and amplitude modulating masks, *J. Micromech. Microeng.*, **11** (2001), 603.

Chang, L. *et al*: Moore's law lives on, *IEEE C & D*, **1** (2003), 35.

Doris, B. *et al*: Extreme scaling with ultra-thin Si channel MOSFETs, *IEDM 2002* (2002), p. 267.

Fritze, M. *et al*: Enhanced resolution for future fabrication, IEEE C & D Mag., **1** (2003), p. 43.

Henderson, R.: Of life cycles real and imaginary: the unexpectedly long old age of optical lithography, *Res. Policy*, **24** (1995), 631.

Hisamoto, D. *et al*: FinFET- a self-aligned double-gate MOSFET scalable to 20 nm, *IEEE TED*, **47** (2000), 2320.

Hoyt, J.L. *et al*: Strained silicon MOSFET technology, *IEDM 2002* (2002), p. 23.

Huff, H.R.: From the lab to the fab: transistors to integrated circuits, *Electrochemical Society Proceedings ULSI Process Integration III* (2003), p. 15.

ITRS, International Technology Roadmap for Semiconductors, http://public.itrs.net/HomeStart.htm.

Iwai, H.: Outlook of MOS devices into next century, *Microelectron. Eng.*, **48** (1999), 7.

Jeong, H. *et al*: Optical projection system for gigabit DRAM, *J. Vac. Sci. Technol.*, **B11** (1993), 2675.

Keyes, R.W.: Fundamental limits of silicon technology, *Proc. IEEE*, **89** (2001), 305 (special issue on limits of semiconductor technology).

Kilby, J.: The invention of the integrated circuit, *IEEE TED*, **23** (1976), 648.

Moore, G.: Cramming more components onto integrated circuits, *Electronics*, **38** (1965) (available at http://www.intel.com/research/silicon/mooreslaw.htm).

Park, J.-T. & Colinge, J.-P.: Multiple-gate SOI MOSFETs: device design guidelines, *IEEE TED*, **49** (2002), 2222.

Ross, I.: The foundations of the silicon age, *Bell Labs Tech. J.*, **2** (1997), 3 (50th anniversary issue of the invention of the transistor).

Tuomi, I.: *The Lives and Death of Moore's Law*, http://firstmonday.org/issues/issue7_11/tuomi/index.html.

Wagner, C. *et al*: The technical considerations of extending optical lithography, *Solid State Technol.* (2000), 97.

Wong, H.-S.P.: Beyond the conventional transistor, *IBM J. Res. Dev.*, **46** (2002), 133 (special issue "Scaling CMOS to the limit").

Proc. IEEE, **74**(12) (1986), special issue on integrated circuit technologies of the future.

Giesecke+Devrient GmbH, Giesecke+Devrient Currency Technology GmbH, and Papierfabrik Louisenthal GmbH. - Ex. 1030

39

Microfabrication at Large

Integration of different technologies is a mega trend all over microfabrication. Analog–digital (mixed signal) ICs integrate resistors and capacitors with digital MOS or bipolar transistors; BiCMOS integrates bipolars and CMOS; and microprocessors integrate more and more SRAM memory (which, in fact, takes up most of the silicon area in processors). MEMS, microelectromechanical systems, integrate mechanical and electrical functions. Microsensors for mechanical, optical, chemical and magnetic quantities most often produce an electrical output signal that opens up possibilities to process, store and transmit those signals with microelectronics, which may be integrated on the same chip.

Microfabricated devices have a number of benefits compared to classic or macroscopic devices: small size, low-cost, high speed (of electron transit time across bipolar base, or of microreactor thermal ramp time), low-power consumption (and low-reagent consumption in chemical microsystems) and high device-packing density (of DRAM memory cells or attached DNA strands) all relate to the exceptional possibilities offered by microfabrication. One of the special benefits of microfabrication is the completely different cost structure compared to macroworld manufacturing. Material usage is minuscule and almost any material can be used if it can be micromachined, because material price is not a limiting factor. We will next discuss some novel materials that are being introduced in microfabrication.

39.1 NEW MATERIALS

New materials are being introduced regularly for functionality, ease of fabrication, better compatibility or just curiosity. Recent demonstrations include negative thermal-expansion coefficient material ZrW_20_8, photopatternable electrically conducting polymer (by silver nanoparticle inclusion) and similar magnetic material, photoactive siloxanes that can be patterned like resists but with oxide-like properties, or iridium and ruthenium with good interface properties with high-k dielectric $BaSrTiO_3$.

39.1.1 Silicon carbide and diamond

Both silicon carbide (SiC) and diamond are wide bandgap semiconductors, and transistors, diodes, thyristors and other semiconductor devices can be made on them. Wide bandgap equals low noise and/or high operating temperature. Single-crystal SiC wafers are available in sizes of up to 2 in., with price tags of ca. $1000 for a prime quality wafer. Crystalline SiC comes in many polytypes: 3C-SiC, 4H-SiC and 6H-SiC, which are slightly different with respect to physical, mechanical and electrical properties. Diamond wafers are available with 1 cm diameter, but most diamond devices fabricated so far have been processed on gemstones. Double heteroepitaxy of diamond shows promise: on a sapphire starting wafer, a layer of epitaxial iridium is grown, and a single-crystal diamond is grown on iridium.

In the thin-film form, SiC and diamond are deposited by CVD, and the basic features of their deposition are not unlike oxide or polysilicon deposition. For example, boron addition to PECVD chamber during deposition leads to p-type doped diamond. In the thin-film form, diamond costs about the same as other thin film materials: capital cost and operating costs are similar for (PE)CVD reactors, and methane (CH_4) source gas is even cheaper than silane (purity levels strongly affect prices). However, as always with thin films, the resulting materials differ from bulk materials. Instead of diamond, people prefer to talk about diamond-like carbon, DLC.

Microfabrication applications for diamond/DLC and SiC are based on their very special combination of

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Table 39.1 Diamond and SiC properties

	Diamond	3C-SiC
Melting point (°C)	3550	2830
Thermal conductivity (W/cm K)	20	5
Coefficient of thermal expansion (ppm/°C)	1	3
Young's modulus (GPa)	1200	700
Poisson ratio	0.2	
Yield strength (GPa)	53	21
Friction coefficient	0.03	
Sound velocity (m/s)	18 000	15 000
Resistivity (ohm-cm)	$<10^{16}$	
Bandgap (eV)	5.45	2.2
Mobility (cm^2/Vs)	4500	
Dielectric constant	5.5	9.72
Optical transparency (nm)	225– > IR	
Refractive index (at 591 nm)	2.41	

thermal, mechanical and optical properties (Table 39.1). They are used as protective coatings in high-temperature devices and in aggressive chemical environments. Exceptional abrasion resistance and low friction are useful in fluidic and mechanical devices, and superior mechanical properties combined with special surface properties make them interesting candidates for microswitches. As passive films, DLC coatings are routinely used to protect moving mechanical parts from contact. Diamond is an insulator, but it has exceptionally high thermal conductivity. Optical transparency of diamond from UV to IR combined with electrical insulation is useful for a number of optoelectronic and microfluidic applications.

39.1.2 Active materials

Many sensors and actuators require active materials, for example, piezoelectric (ZnO, AlN), pyroelectric ($LiTaO_3$) or magnetostrictive (FeCoSiB) materials. Future memories (magnetic RAM, ferroelectric RAM) might be made of ferroelectrics ($SrBi_2Ta_2O_9$, SBT and $PbZr_xTi_{1-x}O_3$, PZT). Spintronic devices are made in GaAs:Mn (a few per cent manganese) and GaN:Mn. In magnetic shape-memory alloy, Ni_2GaMn, a difference of 2% in nickel content changes the Curie temperature by 50 °C. Similar operating-temperature changes can be brought about in TiNi shape-memory alloys by palladium doping. Superconductors with perovskite structure ($YBa_2Cu_3O_{7-\delta}$, YBCO), are quaternary compounds that need careful control of oxygen concentration. The control of composition and structure is inherently more difficult for multi-component films than for binary materials. In addition, in active materials, the correlation between deposition process and material properties is more important because, for example, in ZnO or AlN piezoelectrics, the crystal structure strongly influences the electrical-to-mechanical energy coupling. This cannot be compromised while optimizing the usual film properties such as stress and uniformity. Further processing with these films also entails limitations; for example, ferroelectric films must be processed below their Curie temperature.

39.2 HIGH ASPECT RATIO STRUCTURES

Early silicon IC processes were dubbed 'planar processes' because the structures were essentially flat on a wafer surface, whereas older transistor technologies were 'mesa' technologies with large step height differences. Today, deep-trench isolation in bipolars, DRAM trench capacitors and deep sub-micron contact holes are common in ICs, making them all but planar. Similar high aspect ratio structures are found in DRIE micromechanics, in LIGA and in thin-film head fabrication.

Film deposition into high aspect ratio microstructures (HARMS) is difficult. As aspect ratio increases, maintaining good step coverage becomes even more difficult. A few CVD films (TEOS oxide, LPCVD nitride, LPCVD polysilicon) and some electrodeposited films (Cu, Ni) have the gap-filling capability needed to fill aspect ratios up to 100:1. Deposition into any structure usually involves deposition on two or more different materials: for instance, the bottom and sidewalls are usually made of different materials. PVD, CVD and ECD processes are independent of underlying material only in the first approximation: nucleation processes are influenced by both the chemical and the physical nature of the surfaces in question (roughness, texture, bonds, etc.), and film growth rate, grain size and roughness will vary depending on underlying films.

Metrology of HARMS is difficult: even simple measurements, such as step height or film thickness on the sidewall, pose major problems. Scanning probe methods would require needles with even higher aspect ratios than the structures to be measured, and such needles would be mechanically weak. Optical beams (e.g., in interferometry) necessitate beam diameters smaller than the structures, and small beam divergences. Destructive methods such as cross-sectional SEM and TEM must be used quite often.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

39.3 TOOLS OF MICROFABRICATION

Because of ever increasing metrology needs, microsystems have many possible applications in the microfabrication industries. Residual gas analysis (RGA) in vacuum chambers is one application in which microsystems have already been commercialized. Instead of bulky traditional mass spectrometers, vacuum residual gases are analysed by microfabricated mass spectrometers. Their performance does not match that of traditional instruments, perfected over decades, but the lower price makes it possible to install residual gas analysers in every vacuum equipment, for routine monitoring. In the past, RGA was a special tool that was used in troubleshooting and system check-ups by professionals. Another microfabricated tool that is useful for microstructure characterization is the near field scanning optical microscope (NSOM). The resolution of NSOM is determined by the microfabricated aperture size, not by the wavelength of light (see Figure 13.13 for one NSOM aperture fabrication process).

Until now microfabrication tools have become larger and larger even though the structures on the wafer have simultaneously become smaller and smaller. Could it be that some day micromachines could fabricate microstructures? One such tool candidate is the AFM. The equipment exists, but the writing speeds are orders of magnitude too slow for production. However, if millions of AFM tips could be fabricated on a single chip and individually addressed, then the writing speed limitation would be removed. In optical lithography, reticles can be replaced by micromirror arrays, which can be treated as programmable reticles, offering enormous savings in mask costs. In both cases, the data transfer rates easily become bottlenecks: existing optical steppers and scanners expose gigapixels per second.

SOI wafers offer process simplifications in MEMS as in CMOS. A thermomechanical cantilever tip device for data storage is illustrated in Figure 39.1.

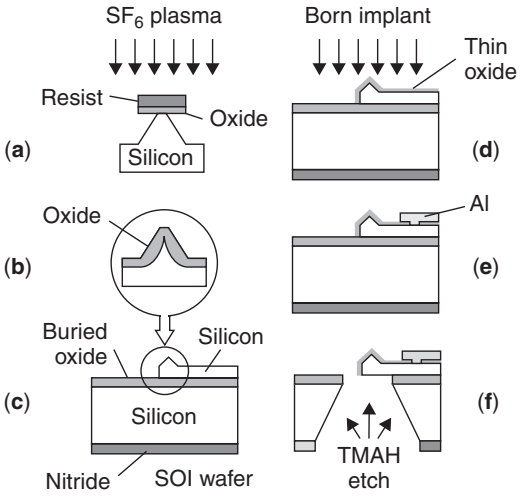


Figure 39.1 Silicon cantilever with a tip. Reproduced from Chui, B.W. *et al*. (1998), by permission of IEEE

Process flow for cantilever and tip on SOI

wafer selection: SOI with 5 μm thick device layer
isotropic silicon etching in SF_6 plasma
(to form a blunt tip)
thermal oxidation for tip-sharpening
cantilever patterning
thermal oxidation for passivation
boron implantation to form piezoresistors
($40\,\text{keV}$, $5 \times 10^{14}\,\text{cm}^{-2}$)
boron implantation for contact improvement
($40\,\text{keV}$, $5 \times 10^{15}\,\text{cm}^{-2}$)
implant activation in RTA
(10 s at 1000 °C; 0.4 μm diffusion depth)
contact opening
aluminum metallization
polyimide protective coating on front side
backside oxide patterning
backside TMAH anisotropic etch
(in a single-wafer holder for front side protection)
etch-stop at buried oxide
buried oxide etching
polyimide plasma removal.

SOI enables precise and easy control of silicon cantilever thickness: this is essential for mechanical devices in order to control cantilever resonance frequency and stiffness. Sharp tips could, of course, be fabricated by anisotropic wet etching of SOI device layer too, but oxidation leads to sharper tips, and the process is better controlled by oxidation time than by etch timing. Boron implantation and RTA are used in piezoresistor formation because piezoresistors should be thin compared to cantilever thickness. Because the wafers become fragile after through-wafer etching, all processing on the front side is completed and the front side is covered by a protective polyimide coating before backside etching. After TMAH etching, the only steps that need to be done are wet etching (of buried oxide) and plasma-etching (of imide), which do not require lithography.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

39.4 BONDING AND LAYER TRANSFER

Silicon wafers used to be made of silicon, but today, wafers are more complex objects. Layer-transfer techniques enable thin layers of expensive or hard-to-make materials to be transferred on common substrates, such as SiC on Si, silicon on quartz and germanium on oxidized silicon, which results in GeOI, germanium on insulator. Bonded wafers with NiSi interlayer have been demonstrated for RF circuits and double-bonded starting wafers have been described for MOEMS (micro-opto-electro-mechanical systems). Layer transfer often necessitates temporary bonding: the thin layers need a support wafer for transfer or for processing, and it must be debonded easily (Figure 39.2). This is obviously quite a departure from traditional bonding, which aims at permanent (and often hermetic) bonding.

An alternative way to increase transistor-packing density without resorting to smaller linewidths is to stack wafers on top of each other (Figure 39.3). 3D integration has been around for decades because it is such an attractive idea. It is possible to thin CMOS wafers down after processing, and align those thinned wafers on top of other CMOS wafers to realize 3D integration. In addition to mechanical joining of the wafers (bonding), the wafers have to be joined electrically too. Metal deposition into vias that extend through the top wafer has been successfully demonstrated.

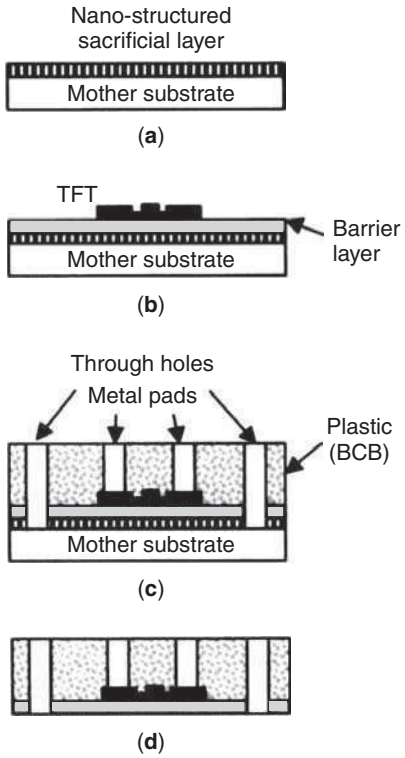


Figure 39.2 Transfer bonding: (**a**) deposition of porous sacrificial layer; (**b**) barrier deposition and TFT processing; (**c**) BCB polymer carrier spinning, exposure and development, followed by etching through the barrier and (**d**) sacrificial layer removal etch. TFT can now be bonded to any substrate. From ref. Lee, Y

39.5 DEVICES

New classes of devices are being introduced in microfabricated versions, as are novel devices with no macroscopic counterparts. New names for devices and categories are popping up, such as nanoelectromechanical systems (NEMS), nanofluidics, biophotonics, adaptive optics (see Figure 17.8), immunosensors, microacoustics (Figure 7.6), micro power systems (turbine in Figure 1.10), pyrotechnical microsystems or DNA-CMOS hybrids. Applications such as CMOS and DNA arrays have small interaction, but if integration is desired, it necessitates a common technology base, which, in most cases, is silicon.

Chemical microreactors form a broad class on microfabricated devices not necessarily related in operation or structure. A hydrogen separation device shown in Figure 39.4 is one example of microfabrication benefits in microreactors. Higher separation selectivity between hydrogen and other gases is possible because thin, yet defect-free membranes do not leak, and only hydrogen can cross the palladium membrane by diffusion. It is fabricated on <110> silicon, and the large structures on the backside are made by KOH wet etching. The 5 μm sieves in top silicon nitride are plasma-etched. Palladium–silver active membrane is sputter-deposited (with titanium adhesion layer) into etched <110> grooves, and the flow channels are made by anodic bonding to a glass wafer. Microfabrication offers benefits in manufacturing: defect-free thin metal membranes can be made reproducibly because fabrication takes place in a cleanroom, and because silicon dioxide surface is extremely flat and smooth. Moreover, the membranes tolerate high pressures because the device geometries and materials in microfabrication allow a lot of design freedom, and higher pressures enable higher gas fluxes. Microfabrication possibilities are everywhere: LIGA and injection moulding have been applied to polyester fibre spinnerets in the textile industry; a micromachined interferometer (Figure 1.8) measures carbon dioxide concentration for heating, ventilation and air conditioning

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

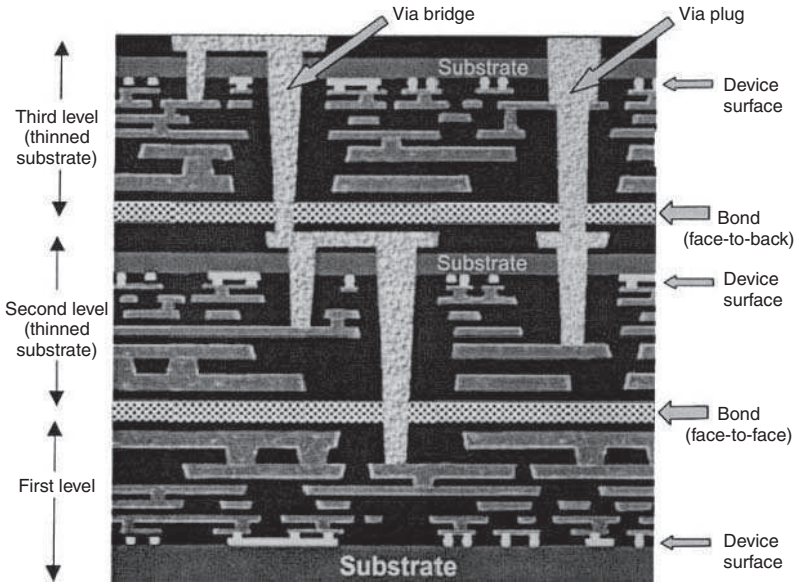


Figure 39.3 Chip stacking by wafer thinning and adhesive bonding. Reproduced from Lu, J.-Q. *et al.* (2000), by permission of Materials Research Society

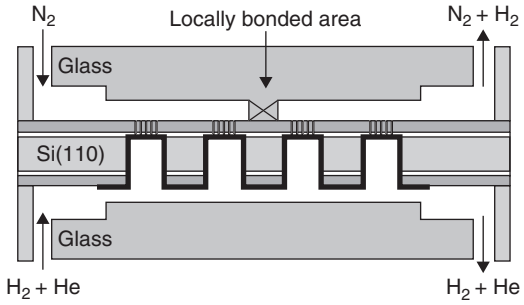


Figure 39.4 A microreactor for hydrogen separation. See text for details. Reproduced from Tong, H.D. *et al.* (2003), by permission of IEEE

applications, and microfabricated superconducting quantum interference devices are measuring weak magnetic fields generated in the human brain.

A wafer with CMOS circuits is usually diced and packaged, after first being electrically tested. This, however, need not be the case. CMOS wafers can be used as substrates for microfabrication. Classes of devices taking the most benefit from CMOS integration include various array devices, which use CMOS for readout: photodetectors, infrared imagers and thermal scanners are typical applications. Displays have been made by many approaches, including LCD on top of CMOS, and micromechanical mirrors on CMOS. In all cases, CMOS provides individually addressable pixels. Fingerprint detectors with pressure-sensitive microstructures have been demonstrated for a variety of applications.

A digital micromirror device is shown in Figure 39.5. It uses standard CMOS wafers as substrates, and builds micromechanical structures on top of that. Mirrors are made of sputtered aluminium, with photoresist as the sacrificial material. Three metal layers form the hinge, yoke and mirror, and this leads to a six-photomask post-CMOS processing. PECVD oxides act as additional protective layers so that the sacrificial resist is not removed when the patterning resist is stripped after metal etching.

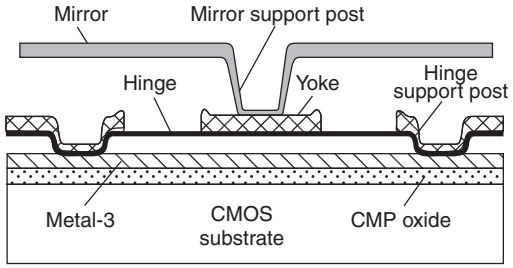


Figure 39.5 Digital micromirror on a CMOS wafer: yoke, hinge and mirror are sputtered metals; photoresist is used for sacrificial layers. Reproduced from van Kessel, P.F. *et al.* (1998), by permission of IEEE

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

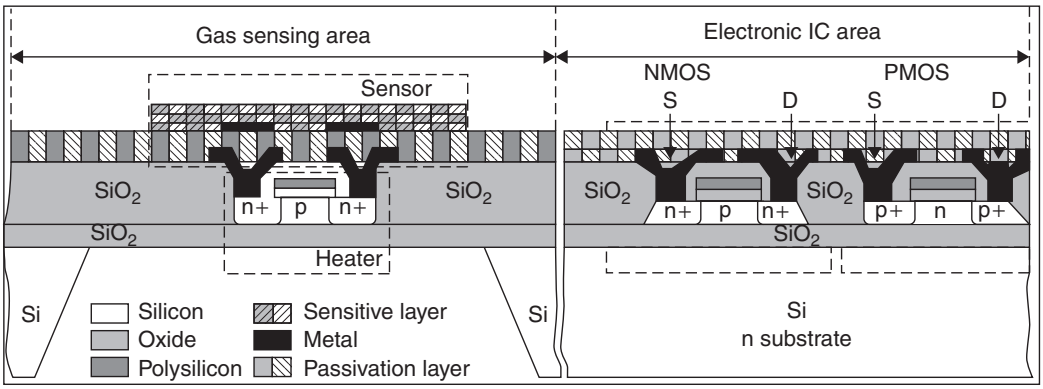


Figure 39.6 Integrated SOI–CMOS micro-hot-plate resistive gas sensor. The MOS transistor below the sensor is for heating; the readout electronics is situated beside the sensor for thermal isolation. Source: Microsensors, MEMS and Smart Devices, J.W. Gardner *et al*, 2001, © John Wiley & Sons, Limited

More than two technologies can be combined, but, of course, at the expense of increased mask count. The integrated micro-hot-plate gas sensor pictured in Figure 39.6 combines bulk silicon micromachining, chemically sensitive resistors and SOI–CMOS technologies. However, simple SOI wafers were not usable in this application because device silicon thickness needs to be ca. 1.5 μm, therefore, epitaxial deposition of silicon on top of a SIMOX SOI wafer was used. Anisotropic wet etching of silicon was used for vertical thermal isolation, with SOI buried oxide as an etch-stop layer. However, the sensor, operating at 350 °C, also has to be laterally thermally isolated from readout electronics. This is achieved by trench isolation, a technique borrowed from advanced IC technologies. CMOS circuitry on the SOI device layer takes care of signal processing but MOSFETs are also used as heaters. This was done in order to simplify the process: platinum heaters would have added a new material, and new cleaning and contamination concerns. Contacting, however, introduces exotic materials: the sensor material, porous palladium-doped SnO_2, makes contact with gold electrodes, which make contact with electronics. In order not to contaminate the SOI CMOS part, Au, Pd and SnO_2 depositions have to be made as post-processing steps, and they put extra demands on barriers.

39.6 MICROFABRICATION INDUSTRIES

Integrated circuits account for a majority of microfabrication turnover, but when different device types are counted, MEMS devices outnumber electronic devices. This is, of course, partly because the MEMS field is new and a lot of experimentation is ongoing, and most of the devices will never enter volume manufacturing, whereas we only see the surviving electronic devices. Linewidth scaling is not an issue in most microfabrication industries: microsystems aim at new functionalities. Costs can be brought down not only by linewidth scaling, but also by device cleverness, new materials and completely new fabrication technologies.

While ICs are being pushed for even smaller operating voltages, down to 0.35 V for digital parts, the power semiconductor industry is hoping to achieve even higher operating voltages, or rather, kilovoltages. Semiconductor power devices have many special features that make them a rather separate entity among the microfabrication industries. Starting wafers are float zone <111> wafers with high resistivities. Power device fabrication is dominated by high-temperature steps and deep diffusions of up to 100 μm. This means, for example, 100 h at 1200 °C. Alternatively, thick epitaxial layers or wafer bonding and thinning can be done. One wafer is one thyristor, which makes yield statistics very different from those of the IC industry. Traditional power devices had very relaxed linewidths, but modern devices are being integrated with CMOS drive electronics (Smart power), and therefore the processes increasingly resemble CMOS, with sub-micron linewidths in MOS-controlled thyristors.

Flat-panel display fabrication deals with large substrate sizes, up to a square metre. Yield models, then, are again different from ICs. Compared to IC memories with regular structure, displays have yield disadvantage because repair is much more difficult: in a memory array, a reserve block can be wired and the faulty block disconnected, but in displays, repair has to be at the very pixel (the same applies to CCD and CMOS image

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

sensors, of course). Large square substrates mean that the FPD industry cannot rely on CMOS for its tools, unlike most other microfabrication industries.

Solar cells are large-area devices like FPDs, but their cost models are completely different: solar cells are the ultimate low-cost devices. Cost reduction starts at starting wafer cost: one dollar would be typical for solar grade silicon, an order of magnitude less than for IC grade wafer. Linewidths are relaxed, in the 10 to 100 μm range. Microfabrication technologies are used for performance, like PECVD nitride antireflective coatings, but traditional techniques such as screen-printing of conductive pastes are used for cost reduction.

The microsystems industry is very fragmented compared to ICs or FPDs. Technologies differ: polysilicon surface micromachining, bulk silicon, DRIE single-crystal silicon (bulk and SOI), thick poly surface micromachining, LIGA and polymer imprint structures share the basic principles of microfabrication but differ in many critical parts. Micropatterning, thin films and etching are core concepts in all microsystems. Polysilicon micromachining applies many of the features of IC fabrication, such as reduction steppers for lithography and plasma-etching for pattern delineation, and the number of photolithographic steps is quite similar to ICs: 10 masks is usual for polysilicon surface micromechanics, whereas most other microsystems are made with four to six masks, and sometimes a single patterning step is enough, as for simple fluidic systems. Microsystems use 100 and 150 mm wafer sizes, and for bulk micromechanics, scaling to 200 mm is not an option because wafer-thickness increases wastes area in through-wafer etching. Waveguide optical microsystems are fabricated on 200 mm wafers because the chips are large due to large radii of curvature.

Integration of two technologies adds to process complexity: roughly speaking, a 20% mask count increase leads to ca. 20% cost increase. A surface micromachined airbag accelerometer, integrated with BiCMOS readout electronics, has been commercialized and is being manufactured in significant volumes. In many sensor applications, extremely advanced readout ICs are required. Processing will then require an advanced CMOS fabrication line, which is often overkill for the MEMS/sensor part.

RF-MEMS devices are close to ICs in many respects: they are mostly planar (or at least not highly 3D) devices, they are internal to the system (unlike sensors and actuators) and reusable blocks and hierarchical design may be amenable to RF MEMS. There is a potentially large market for RF MEMS, in the billion devices per year level, whereas even the most successful MEMS devices sell in tens of million pieces, and more complex ones considerably less, and annual volumes in the 10 000 range are common (corresponding to 1% of monthly production of an IC megafab). Microfluidic/BioMEMS devices have potentially large markets if they can be made cheap enough for disposable applications such as point-of-care measurements in health monitoring where $10 might be a reasonable price, which translates to the cost of ca. $1 for the silicon part.

Microsystems and nanotechnology are still in a nascent state, and there are many contenders for main devices and device classes. Some of them will reach IC-like volumes and markets, some will remain niche applications, and most will never enter the manufacturing stage. However, that is how evolution in technology imitates natural evolution: the more variation and experiments you conduct, the more likely it is that some viable applications and technologies will emerge and will reproduce into many future generations.

39.7 EXERCISES

1. What is the kg price of a CMOS wafer at the end of the fabrication process?
2. What is the kg price (or carat price) of a thin-film diamond if the PECVD capital cost is $500 000 and the running costs are $100 000/year? Take 10 nm/min as deposition rate on a 150 mm wafer size in a single-wafer system.
3. The solar cell cost can be lowered by direct writing because the mask cost is eliminated. If laser direct writing for top metallization is done at a speed of 1 m/s and metal pitch is 200 μm (see Figure 24.1), what is the throughput of such a direct write system?
4. How many metres of wiring is there on a 0.13 μm technology CMOS wafer? What would be the throughput if direct writing at 1 m/s was used?
5. What is the density of AFM tips that could be fabricated on a 1 cm^2 area by the process described in Figure 39.1?
6. Design a DRIE version of the AFM tip array of Figure 39.1 and calculate the tip areal density.
7. Kilogram is defined as the mass of platinum–iridium cylinder that is held at BIPM (Bureau International de Poids et Mesures) in Sevres, near Paris. It has been suggested that a new standard should be made of silicon because silicon is an extremely well-characterized material. What uncertainties can you name for a silicon kilogram standard piece?

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

REFERENCES AND RELATED READINGS

Baliga, J.B.: The future of power semiconductor device technology, *Proc. IEEE*, **89** (2001), 822; (special issue on power electronics technology).

Baltes, H. & O. Brand: CMOS-based microsensors and packaging, *Sensors Actuators*, **A92** (2001), 1.

Chui, B.W. *et al*: Low-stiffness silicon cantilevers with integrated heaters and piezoresistive sensors for high density AFM thermomechanical data storage, *J. MEMS*, **7** (1998), 69.

Gardner, J.W., V.V. Varadan & O.O. Awadelkarim: *Microsensors, MEMS and Smart Devices*, John Wiley & Sons, 2001.

Kärkkäinen A.H.O. *et al*: Photolithographic processing of hybrid glasses for microoptics, *Journal of Lightwave Technology*, **21** (2003), 614.

Lee, S.T. & Y. Lifshitz: The road to diamond wafers, *Nature*, **31**(7) 2003, p. 500.

Lee, Y. *et al*: High-performance poly-Si TFTs on plastic substrates using a nano-structured separation layer approach, *IEEE EDL*, **24** (2003), 19.

Lin, T.-Y. *et al*: Fabrication of low-stress plasma enhanced chemical vapor deposition silicon carbide films, *Jpn. J. Appl. Phys.*, **39** (2000), 6663.

Lu, J.-Q. *et al*: 3D integration using wafer bonding, *Advanced Metallization Conference 2000* (3–5 October 2000, San Diego), paper V3.

Machida, K. *et al*: A novel semiconductor capacitive sensor for a single-chip fingerprint sensor/identifier LSI, *IEEE TED*, **48** (2001), 2273.

Railkar, T.A. *et al*: A critical review of CVD diamond for electronic applications, *Crit. Rev. Solid State Mater. Sci.*, **25** (2000), 163–277.

Renaud, Ph.: Composite photopolymer microstructures: from planar to 3D devices, *Transducers '03* (2003), p. 991.

Sutton M.S. & J. Talghader: Micromachined negative thermal expansion films, *Proc. Transducers '03* (2003), 1148.

Tong, H.D. *et al*: A hydrogen separation module, *Transducers '03* (2003), p. 1742.

Tzeng, S.C. & W.P. Ma: Study of flow and heat transfer characteristics and LIGA fabrication of microspinnerets, *J. Micromech. Microeng.*, **13** (2003), 670.

van Kessel, P.F. *et al*: A MEMS-based projection display, *Proc. IEEE*, **86** (1998), 1687.

Xue, M. *et al*: A self-assembled conductive device for direct DNA identification in integrated microarray based system, *IEDM*, 2002, 207.

Page 401

Giesecke+Devrient GmbH, Giesecke+Devrient Currency Technology GmbH, and Papierfabrik Louisenthal GmbH. - Ex. 1030

Appendix A

Comments and Hints to Selected Problems

1. Introduction

1. How does this value compare with atomic sizes?
4. One mole of gas equals 22.4 liters.
5. Chips can fail by many different mechanisms; $E_a = 0.7\,\text{eV}$ is the activation energy for just one common mechanism.
7. Extrapolation is a dangerous business: how far do you expect scaling to continue?

2. Micrometrology

1. Be careful with units: resistivity is usually given in μohm-cm ($=10^{-8}$ ohm-m).
4. In Equation 2.9, kilovolts and tens of kilovolts are typical.
6. Calculate the volume that is being probed. Express your answer in atomic %.
8. Acceleration voltage and electron wavelength are related, as are wavelength and resolution.
9. TaN_x resistivity is not given: you could try the following: 1) ignore it completely; 2) assume the same resistivity as Ta; and then see how much your result is affected. If you are going to surf the Internet to find TaN_x resistivity, you will find a bewildering range of values, so you are no better off.

3. Simulation

1S. Differences will become apparent at high doping levels.
3S. What is your criterion for penetration?

4. Silicon

1. The unit cell of silicon consists of 8 silicon atoms.
2. This is an order of magnitude question: you just guestimate (guess and estimate) the dimensions for the pool and the balls. See Figure 4.1 for concentration-to-resistivity conversion.
3. Into which direction will segregation work?
4. What if 0.01 ohm-cm boron-doped silicon is used as a boron source, instead of pure boron?
5. When $X = 0$, 10 ohm-cm material will be pulled.
6. This is an order of magnitude question. Yield strength is strongly temperature dependent: at the end of crystal-pulling, neck temperature can be, for instance 600 °C, and yield strength is of the order of 1 GPa.
7. COP size and wafer thickness must be considered.

5. Thin-film Materials and Processes

1. You have four degrees of freedom to work with: width, length, thickness and resistivity.
2. You can only calculate a lower-limit value based on polysilicon minimum resistivity because doping changes poly resistivity by orders of magnitude.
3. This is an order of magnitude question. One significant digit is enough. $C = Q/V$, $Q = Ne$, N is the number of electrons and e is the elementary charge.
4. See Table 5.7.
5. Which fraction of silicon atoms in the flow would you expect to end up on the wafer as an a-Si film?
6. Molar volumes (m_{mol}/ρ) are useful; or you may assume $1\,\text{cm}^2$ area and calculate via the number of atoms.
8. Ni^{2+}, $M = 58.71$ g, and set alpha = 1 for the maximum possible rate. (≈100 μm/hr)

6. Epitaxy

1. Table 4.1.
2. Yes, and very accurately, if there is no spurious deposition over the edges or on the wafer backside.

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

3. See Figure 5.6. The answer will inevitably be a rough estimate only.

4S. Keep the deposition time constant.

7. Thin-film Growth and Structure

1. Layer thicknesses and wavelengths are coupled.
2. Radius of curvature relates to bow. Typical wafer-thickness specification is $\pm 25\,\mu m$; how does this affect your result?
3. $E = h\nu$ and $c = \lambda\nu$
7. R. Shohji *et al*: High-reliability tungsten-stacked via process with fully converted $TiAl_3$ formation annealing, IEEE TSM 12 (1999), p. 302

8. Pattern Generation

1. Equation 2.9 gives the electron stopping range in solids. Two significant digits suffice.
2. Plot this as a function of resist thickness. What do you expect to be the practical minimum resist thickness?
3. Which process is limiting the writing time?
4. See T.R. Groves: Theory of beam-induced substrate heating, J.Vac.Sci.Tech. B 14 (6) (1996)
6. Compare raster scan and vector scan.

9. Optical Lithography

1. How far down can you reduce the wavelength and still call it optical? What is the minimum resist thickness conceivable?
2. Plot this as a function of gap: $20\,\mu m$ is typical of X-ray lithography systems.
3. All misalignment tolerance cannot be used to compensate for thermal expansion: some have to be reserved for mechanical warpage and optical aberrations.
5. Equation 9.1 gives an estimate for resolution, but thick resist profile may be far from ideal.

10. Lithographic Patterns

1. Choose a few wafer sizes and resist thicknesses, and then calculate some rough estimates.
3. You can have a high absorption coefficient in a very thin imaging layer.
6. You can calculate resist thickness from the standing wave period.

11. Etching

1. Look for volatile etch products. Remember that mask selectivity depends heavily on the etched depth: if 100 nm is to be etched, even 1:1 mask selectivity is more than enough for a $1\,\mu m$ thick mask.
2. Graphical solution will reveal something.
3. Remember Arrhenius behaviour.
4. Over-etch time is determined mainly by the step height; not by film or etch non-uniformity.
6. Calculate via masses: before porosification, after porosification and after etching the porous layer away.
7. Pore size and resistivity are connected, and you can come to a range of possible resistivities.
9. Typical chrome thickness is 100 nm.

12. Cleaning

1. 10^{-5} monolayers are difficult to visualize: try another viewpoint: how far are the iron atoms from each other?
2. Use a simple model molecule, like Si-O-Si$(CH_3)_3$ for organic contamination.
3. This is an order of magnitude question; try a few values for phosphorous content (say, 1%, 0.1%, 0.01%) to get a feeling if there is any problem at all.
4. Table 12.1 helps, and tank volumes can be guestimated.
5. How do CVD and evaporation compare?
6. An order of magnitude question. Even if you can find a literature value for sweat salt content, you have to guestimate the droplet size all by yourself. This 0.1 ppb specification is for $0.13\,\mu m$ CMOS.

13. Oxidation

1. The original $1\,\mu m$ oxide will also grow during oxidation.
2. Quadrupling time will result in double thickness, in the parabolic regime.
4. This is a rather heavily doped poly.

7S. Is segregation similar in dry and wet oxidation?

14. Diffusion

1. Graphical approximation according to Equation 14.7.
4. Equations 14.3 and 14.4 can be used to get a feeling for the order of magnitude.

5S. Explore this over a range of temperatures.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

7S. Try different parameters to see which ones are important.

15. Implantation

1. How long will low-dose implantations take with this machine?
2. What issues need to be considered if $^{11}B^+$ ions are replaced by $^{49}BF_2{}^+$?
3. Oxide volumes are calculated on page 146.
4. Germanium mass $M = 72$; interpolation between phosphorus and arsenic gives a good guestimate.

5S. What is your criterion for masking?

16. CMP

1. Young's modulus gives only a very rough estimate. How does it compare with the experimentally determined value?
2. An order of magnitude question.
3. Take a concrete example and figure out which measurements must be carried out to obtain those rates.

17. Bonding and Layer Transfer

1. Assume 100 mJ/m^2 surface energy.
2. Calculate what will happen to an non-bonded area at h_{crit}.
3. The velocity of sound in water is 1.5 km/s.
4. You have to fix wafer thicknesses. Closing a cavity does not depend on its origin: natural and synthetic cavities obey the same laws.
6. Ion ranges, Equations 15.1 and 15.2

18. Moulding and Stamping

1. If cleanroom thermal control is $\pm 1\,^\circ$C, calculate thermal exapnsion over 100 mm stamp.
2. A backing layer is needed to make the channel.
3. Go back to Chapter 9 to find resolution formulas.

19. Self-aligned Structures

1. Recall Exercises 5.5 and 5.6.
2. Use 15 μohm-cm for C54 phase resistivity (Table 5.8).
3. Nordström, H. *et al*: A refined polycide gate process with silicided diffusions for sub-micron MOS applications, J.Electrochem.Soc. 136 (1989), p. 805
4. Guess TiN thickness and estimate etch selectivities.

20. Plasma-etched Structures

1. What selectivity is needed if oxide loss is to be limited to 5 nm, and molybdenum film goes over 300 nm steps?
2. What value should you report: maximum value, average value, instantaneous value?
4. Aspect ratio is different, and pattern size effects may arise.
5. Assume 0 to 2 nm native oxide and use some representative selectivity values.
6. Silicon is etched as SiF_4. Each SF_6 molecule contributes maybe two to three fluorine radicals.

21. Wet-etched Structures

1. Does etching at lower temperatures result in better control?
2. Plot logarithm of rate against $1/T$.
3. The 54.7° sidewall wastes quite a bit of area. Remember edge exclusion. Additionally, some area must be reserved between the chips to ensure bonding and allow for wafer dicing, too.
4. How can you use the concept of pitch to improve such a measurement?
6. G. Vdovin & S. Middlehook: Technology and applications of micromachined adaptive mirrors, J.Micromech.Microeng. 9 (1999), p. R8
8. Estimate dimensional tolerances for it.

22. Sacrificial Structures

1. Silicon-rich nitride is more resistant in HF than stoichiometric nitride, ca. 1000:1 versus 100:1.
3. Take 10% as litho/etch tolerance, and 3% as deposition tolerance.

23. Structures by Deposition

1. Evaporation is the most-often used process with shadow masks: it is highly directional.
3. Cylinder wall area must be calculated but how high can you make the cylinder?
5. T. Shibata *et al*: Stencil Mask Ion Implantation Technology, IEEE TSM 15 (2002), p. 183

24. Process Integration

3. Use either implantation or diffusion, and stick to your choice. Remember to include the cleaning steps.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

4. Consider both layout rules and electrical rules.
5. $\varepsilon = 7$ is the relative permittivity; you have to multiply by vacuum permittivity to get a numerical value for capacitance. Remember the four degrees of freedom in resistor design.
7. Capacitor area is ca. $5\,\mu m^2$.
10. Use 150 μohm-cm for TiW resistivity (very much deposition dependent)
11. Arrhenius behaviour between operating and test temperature.
12. How does this displacement relate to gap height? To surface roughness? To atomic dimensions?

25. CMOS

1. Not all of them are encountered in any one CMOS process: you have to think of different CMOS generations and/or technologies.
2. Explore this for SiO_2 and Ta_2O_5 thicknesses relevant to sub−0.20 μm CMOS.
3. Plot EOT versus physical thickness.
5. This thickness must be added to the first CVD oxide thickness when contact-hole etching is considered.
8S. Make a time/temperature/junction depth/sheet resistance plot of your results.
9S. Are there 2D effects in a 5 μm CMOS process?
10. Linewidth is 35 nm. This will constrain many things.
11. Remember that alignment is not necessarily to the previous layer, but to some important lower layer.

26. Bipolar

1. Assume isotropic diffusion profiles. Include half of the guard ring area.
2. What will happen to the number of process steps if trench isolation is adopted?
5. What is the total mask count of this process, if two levels of aluminium metallization and passivation follow?

27. Multilevel Metallization

1. Remember that 0.25 μm is the gate poly linewidth; other dimensions are larger.
2. How does it compare with silicon-to-metal contact resistance of similar size contacts?
3. What are the dimensions and voltages in technologies that employ low-k?
4. Why can't thicker nitrides be used?
6. This kind of etching is standard procedure in reverse engineering.

28. MEMS Integration

3. Fix wafer thickness. Note that the limiting factors are very different in each case. Figure 21.13 is useful.
4. You have to fix the diaphragm size and thickness.
6. In order to have etch-stop concentration deep inside silicon, surface concentration must be very high. What consequences will that have?
9. How can you implement profile for good lift-off, depicted in Figure 23.7?

29. Processing on Non-silicon Substrates

1. Refer to Equation 9.2 again. What would you expect for warp and bow of 50 cm glass plates?
5. Consider for example, thermal and contamination issues, and give some thought to transparency as well.

30. Tools for Microfabrication

1. Assume no heat dissipation, calculate just the theoretical maximum value.
2. How much will the wafers heat up in evaporation?
3. When wafer-cleaning time is added to furnace time, thermal oxidation is a really slow process.

31. Tools Hot

1. Consider maximum batch size. What is the proper time frame to consider?
2. In addition to temperature, what other parameters affect oxidation rate?
3. $\sqrt{4Dt}$ is the characteristic diffusion distance.
4S. Check whether your simulator understands RTO. If not, you can compare it to the data in Exercise 13.3.
5. This change could be due to the interference effects from thin films on silicon. What difference to oxide thickness would that temperature difference imply?
6. Consider radiation; and remember that radiation comes from both sides of the wafer.
7. How about 0.050 μm technology?

32. Vacuum and Plasmas

2. Find out product specs for some real diffusion pumps and compare your result.
4. You can use a pumping speed of 5000 l/s (which is high).
6. Equation 15.1.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

7. Pumps set one limit via residence time. How about mass flow controllers?
8. Conservation of energy and sputtering yield: divide input energy to 500 eV argon ions.
9. Surface-sensitive measurement must be faster than monolayer formation, otherwise it would not probe the surface of interest.

33. Tools CVD

2. If you cannot find absolute rates, you can calculate relative rates.
3. How sharp do you expect the transition to be?
4. How does epilayer thickness affect your result? How do uptime and yield affect your answer?
5. $SiH_4\,(g) + 2N_2O\,(g) \rightarrow SiO_2\,(s) + 2N_2\,(g) + 2H_2\,(g)$.
6. In three-zone CVD, furnace temperature is used to compensate for reactant depletion along the tube.
7. Calculate residence time and compare it with deposition rate.
8. Remember that the silane flow is just a few percent of the total flow, and that utilization of silane is well below 100%.

34. Integrated Processing

1. Find out which chamber is limiting the process.
3. Barna, G.G. *et al*: MMST manufacturing technology – hardware, sensors and processes, IEEE TSM 7 (1994), p. 149

35. Cleanrooms

2. Fed.ST. 209 uses 0.5 μm particle size as the yardstick.
4. Leak is initially very local because the laminar flow effectively prevents spreading, but once the air is circulated, gases spread uniformly.
5. Compare your result with Table 25.6 which gives prime wafer 'as received' particle specifications.

36. Yield

1. 50% diameter increase results in 125% raw area increase, but how much is chip count increasing?
2. Particle count goes up as the third power of particle size (a worst case assumption).
3. Remember that for small defect densities and small chip sizes the differences between yield models are small or negligible.
4. Which model should you use as the basis for pricing the chip?
5. Graphical solution; Figure 36.4 and Table 36.1.

37. Wafer Fab

1. Take 5 years for fab amortization time.
2. Compare with the price calculated in the previous problem.
3. Calculate the lithography cost per wafer over a five-year life span of the system. Take 25 lithography steps as a baseline CMOS process.
4. Assume batch size 25 for wet tanks, and use a 20 min process time. 50 WPH can be used as a baseline for a single-wafer stripper. Does your answer bear any resemblance to the tool numbers given in Table 37.2?
5. Oxidation time is just a fraction of the total process time, as shown in Table 31.1.
6. Maybe only 25% of the gas introduced into the ion source ends up on the wafers.
8. You can start from silicon-wafer consumption, which is ca. 3 000 000 m^2 per year and then make some assumptions about the size distribution of wafer fabs.

38. Moore's Law

1. 512 Mbyte memory modules will have been phased out long before 2013; the calculation refers to the memory capacity only.
2. One fundamental limit is set by the speed of light.
4. The lithography tool cost in 2005 is projected to be ca. $25 million for 90 nm technology, but this includes everything, not just the lens.
5. You have to add, for example, 10% area for peripheral circuits like sense amplifiers. The first product is much larger than the subsequent shrink versions. SRAM cell takes ca. 30–100F^2 area, which explains why SRAM capacities are lagging DRAM by ca. two generations. Flash-memory cell area can be <2F^2.

39. Microfabrication at Large

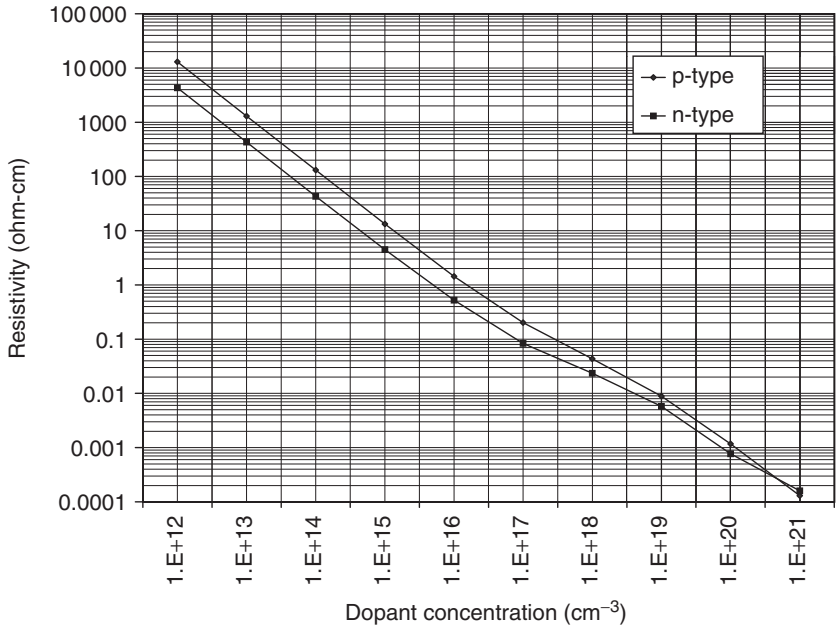
1. For comparison, gold sells for ca. $10 000/kg.
4. How many layers of metallization are there in a 0.13 μm technology?
6. What would be the throughput of such an AFM-writer in the metallization application of Exercise 39.4?

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Appendix B

Constants and Conversion Factors



Atomic mass unit amu	1.66×10^{-27} kg
Electron charge e	1.602×10^{-19} C
Avogadro's constant N_A	6.022×10^{23}/mol
Boltzmann constant k	1.38066×10^{-23} J/K $= 8.6544 \times 10^{-5}$ eV/K
Faraday constant F	96 500 As/mol ($F = e \times N_A$)
Gas constant R	8.3144 J/Kmol ($R = k \times N_A$)
Gas molar standard volume	22.4 l/mol ($V_m = RT_0/p_0$)
Permittivity of vacuum ε_0	8.854×10^{-12} F/m
Speed of light c	2.9979×10^8 m/s
Stefan–Boltzmann constant σ	5.67×10^{-8} W/m^{2}K^4

Introduction to Microfabrication Sami Franssila
© 2004 John Wiley & Sons, Ltd ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Conversion factors

$T/K = 273.15 + t/°C$
$1\,eV = 1.6 \times 10^{-19}\,J$
$1\,eV \times N_A = 96.5\,J/mol = 23.06\,kcal/mol$
$1\,cal = 4.184\,J$
$1\,N = 10^5\,dyne$
$1\,Pa = 1\,N/m^2 = 10\,dyne/cm^2$
$1\,\mu m = 10^{-6}\,m = 1000\,nm = 0.001\,mm$
$1\,Å = 0.1\,nm = 1 \times 10^{-10}\,m$
$1\,mil = (1/1000)\,inch = 25.4\,\mu m$

Pressure conversion

From \ To	Pa	Torr	atm	mbar
	multiply by			
Pascal (Pa)	1	7.5×10^{-3}	9.87×10^{-6}	10^{-2}
Torr (mmHg)	133	1	1.316×10^{-3}	1.33
atm	1.013×10^5	760	1	1013
mbar	100	0.75	9.87×10^{-4}	1

Flow conversion

	Pa m^3/s	Torr l/s	sccm
Pa m^3/s	1	7.5	592
Torr l/s	0.133	1	78.9
sccm	1.69×10^{-3}	1.27×10^{-2}	1

Molarity to weight%

PROPERTIES OF SILICON AT 300 K

Structural and mechanical

Atomic weight	28.09
Atoms, total (cm^{-3})	4.995×10^{22}
Crystal structure	Diamond (fcc)
Lattice constant (Å)	5.43
Density (g/cm^3)	2.33
Density of surface atoms (cm^{-2})	(100) 6.78×10^{14}
	(110) 9.59×10^{14}
	(111) 7.83×10^{14}
Young's modulus (GPa)	190 (111) crystal orientation
Yield strength (GPa)	7
Fracture strain	4%
Poisson ratio, ν	0.27
Knoop hardness (kg/mm^2)	850

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Electrical

Energy gap (eV)	1.12
Intrinsic carrier concentration (cm^{-3})	1.38×10^{10}
Intrinsic resistivity (Ω-cm)	2.3×10^{5}
Dielectric constant	11.8
Intrinsic Debye length (nm)	24
Mobility (drift) (cm^2/Vs)	1500 (electrons) 475 (holes)
Temperature coeff. of resistivity (K^{-1})	0.0017

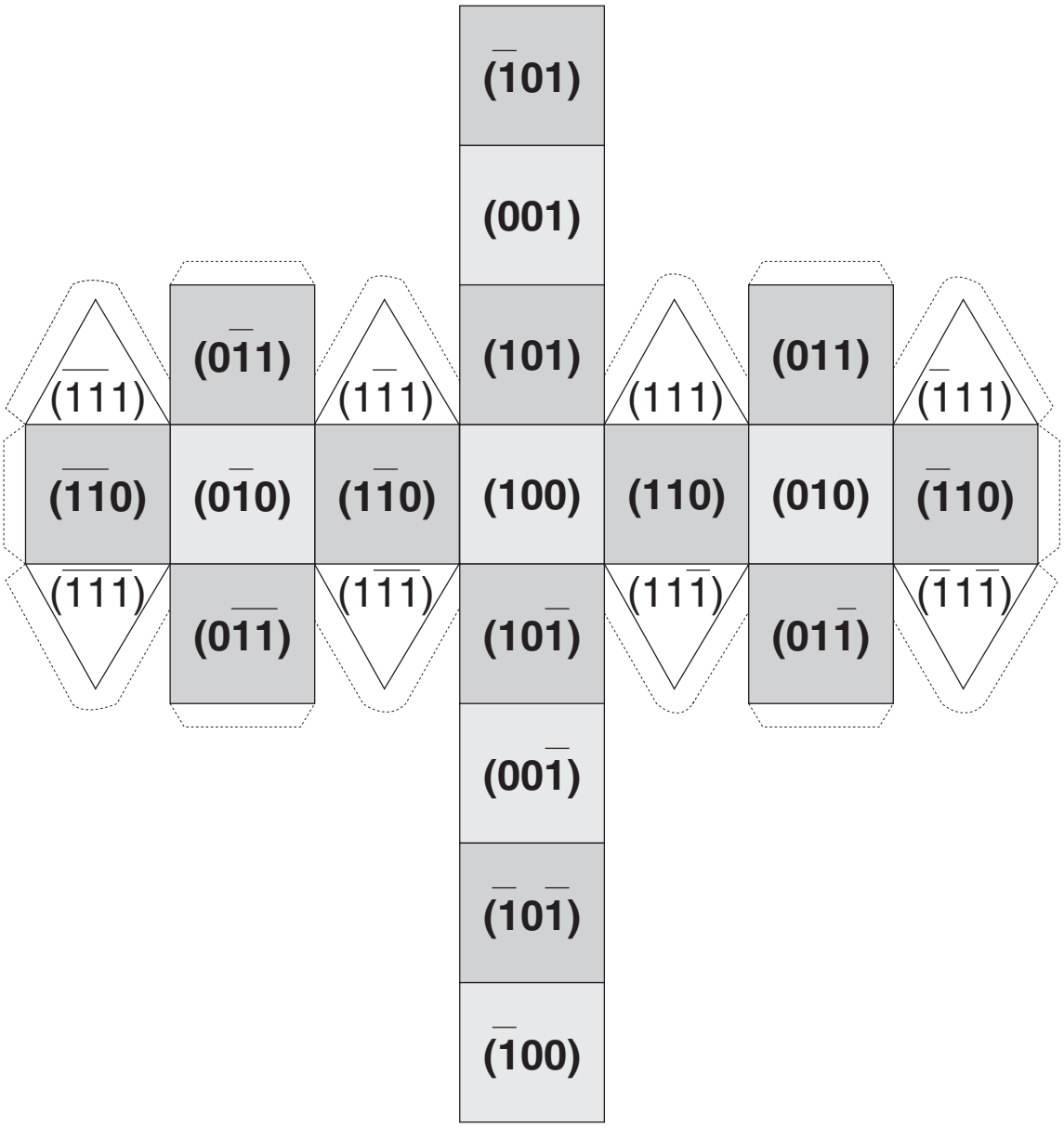
Thermal

Coefficient of thermal expansion ($^{\circ}C^{-1}$)	2.6×10^{-6}
Melting point ($^{\circ}C$)	1414
Specific heat (J/kg K)	700
Thermal conductivity (W/m K)	150
Thermal diffusivity	$0.8\,cm^2/s$

Optical

Index of refraction	3.42	$\lambda = 632\,nm$
	3.48	$\lambda = 1550\,nm$
Energy gap wavelength	$1.1\,\mu m$	(transparent at larger wavelengths)
Absorption	$>10^6\,cm^{-1}$	$\lambda = 200–360\,nm$
	$10^5\,cm^{-1}$	$\lambda = 420\,nm$
	$10^4\,cm^{-1}$	$\lambda = 550\,nm$
	$10^3\,cm^{-1}$	$\lambda = 800\,nm$
	$<0.01\,cm^{-1}$	$\lambda = 1550\,nm$

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030



Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

Index

100 silicon, 40, 205
110 silicon, 40, 212
111 silicon, 40, 213
1D, one-dimensional
  simulation, 28
2D, two-dimensional
  simulation, 29
  growth, same as layer-by-layer growth, 74
3D, three-dimensional
  growth, same as island growth, 74
  simulation, 30
4PP, four-point probe, 19
5N (99.999 % purity), 133

abrasive, 165
absorption, 37
accelerometer, 174, 211, 294
acoustic microscope, 179
acoustic resonator, 83
activation energy, 7, 52, 120, 154, 251, 330
adatoms, 74
adhesion, 81
adhesion promotion, same as priming, 107
adhesive bonding, 177
adsorption, 321, 323, 331
aerial image, 114
aerogel, 56
AES, Auger electron spectroscopy, 22, 82
AFM, atomic force microscope, 17, 19, 44, 171, 375
agglomeration, 196
ALCVD, atomic layer CVD, same as ALD, 339
ALD, atomic layer deposition, 339
alignment, 94, 103, 245
alpha-tool, 311
aluminum, 49, 58, 80, 120, 129, 258, 277, 324, 339, 377
aluminum gate, 193, 255
aluminum nitride, 123, 374
ambient control, 337
amorphization, 161–162
amorphous, 4, 47, 75, 77–79
  silicon, 63
anisotropic plasma etching, 125
anisotropic wet etching, 125, 205–216, 290–298
annealing, 73, 146, 161, 174, 248, 318
anodic bonding, 176
APCVD, atmospheric pressure CVD, 329
APM, ammonia-peroxide mixture, 135
ARC, antireflection coating, 111, 238
arc-lamp, 317
ARDE, aspect ratio–dependent etching, 202
Arrhenius behaviour, 7, see also activation energy
arsine, AsH_3, 164, 347
ashing, same as photoresist stripping, 116
ASIC, application specific integrated circuit, 202
aspect ratio, 7, 86, 202, 374
aspect ratio dependent etching, ARDE, 202
assembly cost, 359
atomic force microscope, AFM, 17, 19, 44, 171
atomic layer CVD, ALCVD, 331
atomic layer deposition, ALD, 331
Auger electron spectroscopy, AES, 22, 82
autodoping, 68

back-end, 6, 255
ball-up, 196, same as agglomeration
bamboo structure, 76
BARC, bottom antireflection coating, 111
barrel reactor, 334
barrier, 81, 281
base (of a bipolar transistor), 270
batch processing, 309
BCB, benzocyclobutadiene, 60, 284
BEOL, back-end of the line, 6, 255
BESOI, Bond-etchback SOI, 180
beta-tool, 311
BHF, buffered HF, 121

Introduction to Microfabrication Sami Franssila

© 2004 John Wiley & Sons, Ltd

ISBNs: 0-470-85105-8 (HB); 0-470-85106-6 (PB)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

BiCMOS, 275
bipolar transistors, 28, 154, 269–276
binary mask, 95. See photomask
bird's beak, 148
blanket deposition, 230
BMD, bulk microdefects, 44, 240
Boltzman's constant, 7, see also activation energy
bond alignment, 289
bonding, 173–182, 289, 376
bond energies, 126, 175, 251
boron etch stop, 185, 207, 293
boron nitride, 79
Bosch process, 127, 203, see also DRIE
bottom gate TFT, 302–304
boundary layer, 329
bow, 239
BOX, buried oxide in SOI, see SOI
BPSG boron–phosphorous-doped silica glass, 52, 249
Bragg-Brentano stress formula, 86
BST, barium strontium titanate $BaSrTiO_3$, 263
buffering, 121, 166
bulk microdefects, BMD, 44, 240
buried layer, 28, 269
buried oxide, BOX, 164, 241, see also SOI

CA chemical amplification (resist), 109
CaF_2, 368
cantilever, 218, 375
capacitor, 57, 244
capillary forces, 219
casting, 183
cavity, 173–174, 178, 180, 232
CD, compact disc, 183
CD, critical dimension, 17, see also linewidth
CD gain, 201
CDI, collector diffusion isolation, 276
channelling, 161–162
chemical amplification (resist), CA, 109
chemical mechanical polishing, see CMP, 165–172
chemical shrink, 113
chemical vapor deposition, see CVD
chip, same as die, 14, 360
chrome, 96
chromium, 120
cleaning, 133–140, 338
cleanroom, 12, 133, 338, 343
cluster tool, 338
CMOS, 11, 255–267, see also transistor, MOSFET
CMOS-MEMS, 296–298, 378
CMOS, as substrate, 298, 377
CMP, chemical–mechanical polishing, 165–172, 265, 282
cobalt silicide, 194
coefficient of thermal expansion, CTE
  anodic bonding, 176
  stamping, 188
  stresses, 83–84
  thin films, 58–60
cold wall reactor, 315, 318
collar, 245
collector, 270
collimated sputtering, 76
comb-drive, 219–221
concave corner, 149, 209
conductance, 323
conformal, 86–87, 129–130, 230
contact angle, 134
contact/via hole, 246, 249, 265
  plug, 278
  stacked, 246, 280
contact lithography, 100
contamination, 133–141, 247
contamination standard, 137
contrast, 110, 188
convex corner, 149, 209
CoO, cost of ownership, 358
COP, crystal originated particle, 44, 137
copper, 49, 54, 58, 129, 168, 281
corner compensation, 212
corrosion, 324
cost of ownership, CoO, 358
cost of processed silicon, 359
cost of testing, 25
critical dimension, CD, 17
cross-contamination, 337
crucible, 38, 49
cryogenic etching, 127
crystal originated particle, COP, 44, 137
crystal structure, 39–41
CTE, coefficient of thermal expansion
CVD, chemical vapor deposition, 51–53, 73, 83
  equipment, 329–331
  rate models, 52, 329
  step coverage, 86
cycle time, 357
CYTOP, 61
Czochralski silicon, CZ, 36–39, 239

damage
  anneal, 249, 264
  implant, 161–162, 264
  plasma etch, 203
damascene, 165, 280
dangling bonds, 146
dark field, 97, 103, 113
dark field microscopy, 17
Dash defect etch, 121
DCE, 1,2-dichloroethene, 144, 315
DCS, dichlorosilane SiH_2Cl_2, 51, 67
Deal–Grove oxidation model, 143
de-embossing, 188

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

deep submicron, <0.5 μm, 260, 366
deep trench isolation, 274
deep UV, 109
defects,
  crystalline, 43
  epitaxial, 65, 69
  oxide, 250
defect density, 351, 359
demoulding, 186
denuded zone, DZ, 240
depth of focus, DOF, 102, 259
design rules, 242–247
desorption, 321, 331
detection limits, 20
development (of photoresist), 108
DHF, dilute HF, 135
diamond, 185, 373–374
diamond-like carbon, DLC, 53, 373
diaphragm, 208, same as membrane
diazonapthoquinine, DNQ, novolak, 108
die, same as chip (pl. dice), 14, 360
dicing, 287, 296
dichlorosilane SiH_2Cl_2, 51, 67
dielectrics, 58–63, 283
diffusion, 153–158
diffusion barrier, 81, 282
diffusivity, thermal, 319
Dill parameters, 108
direct bonding, 173
direct writing, 93, 231, 360
dishing, 171, 282
dislocation, 43–44
disposable mold, 184
dissolved wafer process, 185
DIW, de-ionized water, 12, 140, 346
DLC, diamond-like carbon, 53, 373
DNA chip, 64
DNQ, diazonapthoquinine, novolak, 108
DOF, depth of focus, 102, 259
dogbone, 245
dopant, 153, 159
double poly (bipolar), 273
double side lithography, 289
double side polished wafers, DSP, 42, 288
down force, 165
down-time, 310
drain (of MOS), 11, 255–267
DRAM, dynamic random access memory, 12, 230, 233, 351, 365,
DRIE, deep reactive ion etching, 127, 130, 185, 199, 203, 214
drive-in, 153
dry cleaning, 338
dry etching, 119, see also RIE, wet etching
dry oxidation, 143
drying, 140, 219–220
DSP, double side polished wafers, 42, 288
dual-damascene, 280–282
dummy gate, same as replacement gate, 265
DUV, deep ultra violet, 109
DZ, denuded zone, 240

EBL, electron beam lithography, 93–95
EBR, edge bead removal, 108
edge bead removal, EBR, 108
edge rounding, 41
EDP, ethylene diamine pyrocathecol, same as EPW, 205
EDX, energy dispersive X-ray analysis, 23
EPW, ethylene diamine pyrocathecol water, same as EDP, 205
EGS, electronic grade silicon, 36
electrochemical deposition, ECD, see electroplating, 54
electrochemical etching, 123, 208
electrochemical etch stop, 208
electrodeposited resist, 107
electroless deposition, 54
electromigration, 58, 251
electron beam lithography, EBL, 93–95
electron microprobe analysis, EMPA, 23
electron projection lithography, EPL, 368
electroplating, 54, 219, 227
electropolishing, 123
electronic stopping, 160
ellipsometry, 19, 61
ELO, epitaxial lateral overgrowth, 70
EM, electromigration, 58, 251
embedded amplitude mask, 366
embossing, 183, 187
emissivity, 317
emitter
  bipolar transistor, 269–276
  push, 158
  tip, 217, 233
EMPA, electron microprobe analysis, same as EDX, 23
end point, 129, 199, 278, 313
energy dispersive X-ray analysis, EDX, 23
EOR, end of range damage, 161
EOT, equivalent oxide thickness, 263, 266
epipoly, 66
epitaxial lateral overgrowth, ELO, 70
epitaxial wafers, 240–241, 291–292
epitaxy, 65–71, 333–335
EPL, electron projection lithography, 368
equipment, 237, 309, 355, same as tool
equivalent oxide thickness, EOT, 263, 266
erosion, 171
ERR, etch rate ratio, same as selectivity, 128
ESCA (electron spectroscopy for chemical analysis), same as XPS, 22
ESH, environment, safety and health, 346

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

etchback, 202
etch bias, see undercut, 121
etch damage, 203
etch products, 127
etch rate, 119, 121
etch rate ratio (same as selectivity), 128
etch residues, 203, 324
etch stop, 185, 207, 293
EUV, extreme ultra violet (lithography), 368
evaporation, 49, 229
exposure, 100, 108, 115
exposure field, 100

FA, furnace annealing, 158, 318
fab (IC manufacturing facility), 355
Fabry–Perot interferometer, 10
failure analysis, 26
fast ramp furnace, 318
Fed. standard, 343
FEOL, front-end of the line, 7, 255
FGA, forming gas anneal (N_2/H_2), 248, 302
FIB, focussed ion beam, 96, 231
Fick's law, 155
field stop implant, 256
filter, 217
flat (wafer), 40–42
flat panel display, FPD, 301, 378
float zone silicon, FZ, 39
focal plane deviation, FPD, 239
focus depth, 102, 259
focussed ion beam, FIB, 96, 231
footprint, 310
Fomblin®, 347
forming gas (N_2/H_2) anneal, FGA
four-point probe, 4PP, 19
FPD, focal plane deviation, 239
front-end, 7, 255
front-side micromachining, 211
FSG, fluorine doped silica glass, 53, 283
FTIR, Fourier-transform infrared spectroscopy, 24, 139
furnace, 144
fuse, 231
fused silica, 4, 241, 301
fusion bonding, 173–176, 179
FZ, float zone silicon, 39

galvanic deposition, see electroplating, 54
gap fill, 86
gases, 346
gas-phase transport coefficient, 329
gas sensor, 378
gate (MOS), 62, 193, 255, 262, 265
gate oxide, 143, 262, 369
GDSII design data format, 242
generation (of CMOS technology), 12, 260, 365
Gerber design data format, 242
gettering, 240
glaa-frit, 177
glass, 301
glass transition temperature, 113, 116, 188, 312
g-line, Hg-lamp $\lambda = 436$ nm, 102, 251
global planarization, 169
grain boundary diffusion, 145, 251
grain size, 62, 74–78
grinding, 165
guard ring, 270

h-line, Hg-lamp $\lambda = 405$ nm,
handle wafer, 164, 181
hard mask, 123, 200
HAR, high aspect ratio microstructures, also HARMS, 7, 184, 203, 227, 374
haze, 44
HCI, high current implantation, 162
HDP, high density plasma, 324
HEI, high energy implantation, 162
HEL, hot embossing lithography, 183, 187
HEPA, High Efficiency Particulate Air filter, 13, 344–346
heteroepitaxy, 66
hexafluoroacetylacetonate, hfac, 129
HexSil molding, 185–186
hfac, hexafluoroacetylacetonate, 129
Hg-lamp, 259
high-current implanters HCI, 162
high density plasma (HDP), 324
high energy implantation, 162
high index planes, 40
high-k dielectric materials, 263
high vacuum, 49, 323
hillock, 252
hinged structures, 221
HIPOX, high pressure oxidation, 143, 150
HMDS, hexamethyl disilazane, 107, 134
Hoerni, Jean, 363
homoepitaxy, 66–67
horizontal furnace, 144, 315, 330
hot embossing, HE, 187
hot lot, 357
hot plate, 311
hot wall reactor, 315, 318
HPGL design data format, 242
HPM, hydrochloric acid-peroxide mixture, 135
HV, high vacuum, 49, 323
hydrochloric acid, 135
hydrofluoric acid, 135
hydrophilic, 134, 175–176
hydrophobic, 134, 175–176

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

IBE, ion beam etching, 119
IC, integrated circuit, 11–14, see also Moore's law, scaling
  history, 356, 363
  industry, 12, 357, 371
  manufacturing, 355–360
  wafer fab, 355
  yield, 349–353
ICECREM simulator, 23, 69, 147, 157, 160, 162
ICP, inductively coupled plasma, 325
IDHL, immediately dangerous to health and life, 347
IG, internal gettering, 240
I/I, ion implantation, 159–164, 263–265
i-line, Hg-lamp λ =365 nm, 258
imprinting, 188
in situ monitoring, 313
inductor, 96, 222, 244
inert ambient, 338
ingot, 38
injection molding, 183
inking, 14, 184
ink jet, 130, 293–294
in situ cleaning, 67, 337
impingement rate, 321
indium tin oxide, ITO, 303
insulator, 4, 48, see also dielectric
integrated circuit, see IC
integrated processing, 337–339
integrated tool, 339
interconnect, see multilevel metallization
interfaces, 5, 79–81, 250
interfacial oxide, 79
interference, 110
interlevel dielectrics, 58
International Technology Roadmap for Semiconductors, ITRS, 366
interstitial diffusion, 154
interstitialcy diffusion, 154
ion beam etching, IBE, 119
ion implantation, 159–164, 263–265
ion milling, 119
ion projection lithography, IPL, 368
IPA, isopropyl alcohol, 2-propanol, 140, 206
IR, infra red, 24
island growth, 74
ISO standard, 343
isotropy, 121, 159, 243
ITO, indium tin oxide, 303
ITRS, International Technology Roadmap for Semiconductors, 366

junction anneal, 264
junction depth x_j, 258, 264, 319

Kilby, Jack, 363
killer defect, same as fatal defect, 97, 351
Knudsen cell, 49
Knudsen number, 321
KOH, potassium hydroxide, 205
Krytox®, 347

Lambert's law, 49
laminar flow, same as unidirectional flow, 13, 344
lapping, 41
laser-CVD, 231
latent image, 114
latex sphere equivalent, LSE, 137
LATID, large angle tilt device, 264
lattice constant, 39, 66
layer transfer, 180
layout rules, 243
LCM, light coupling mask, 366
LDD, lightly doped drain, 263
L_{eff}, effective gate length, 371
LER, line edge roughness, 369
lift-off, 228, 294
LIGA, LIthografie, Galvanoformung, Abformung, 184, 218, 228, see also X-ray lithography
light coupling masks, LCM, 366
light field, 97, 103, 112
lightly doped drain, LDD, 266
limited source diffusion, 156
line edge roughness, LER, 369
linewidth, 95, 101, 128, 259–261, 281, 365
liner, 281, see also barrier
Linhard solution, 159
load lock, 339
loading effect, 202
local oxidation of silicon, LOCOS, 147
localized processing, 231
LOCOS, local oxidation of silicon, 147
lot, 358
low-k dielectric materials, 283
low-pressure CVD, LPCVD, 53, 329
LPCVD, low-pressure CVD, 53, 239
LSE, latex sphere equivalent particle size, 137

magnetron (sputtering), 325
Marangoni, 140
mask
  lithography, 94–97, 241–242, see also photomask, reticle
  etching, 116, 123, 200–201, 206
  oxidation, 147
  diffusion, 153–154
  implantation, 159
mass transport limited, 53, 329–333
master, 183–189
MBE, molecular beam epitaxy, 49

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

MC, Monte Carlo simulation, 88, 162
MCI, medium current implanter, 162
MCZ, magnetic Czochralski silicon, 39
mean free path, 321
measurements, 17–26, 312–314
medium-current implanters MCI, 162
megasonic cleaning, 141
membrane, same as diaphragm, 291
membrane devices, 10
MEMS, microelectromechanical systems, 3, 205, 287, 379
mesoporous, 124
metal contamination, 138
metal gate, 265
metal micromechanics, 218
metal organic CVD, MOCVD, 332
Metal Organic Vapor Phase Epitaxy, MOVPE, same as MOCVD, 332
metallic thin films, 56–58, 74–77
metallization, 56, 249, 265
metal-semiconductor contacts, 81, 249, 265
MFP, mean free path, 321
MGS, metallurgical grade silicon, 36
microcontact printing, μCP, 187
microcrystalline, 47, 78
microelectromechanical systems, MEMS, 3, 205, 287, 379
microhotplate, 378
microloading effect, 202
micromirror, 178, 377
micron, same as micrometer
microphotoinductive decay, μPCD, 140
micropipette, 295
microporous, 124
microreactor, 377
microriveting, 180
microroughness, 69, 133
microsystems, same as MEMS, 3, 205, 287, 379
microturbine, 11
microwave photoconductive decay, 24, 140
Miller index, 39
mini-environment, 337, 345
minifab, 357
misalignment, 245, see also alignment
miscut, 41, 68, 239
mix-and-match (lithography), 242
ML, monolayer, 322
MLM, multilevel metallization, 277–285
MLR, multilayer resist, 112
mobility, 35, 37, 155
MOCVD, Metal Organic CVD, 332
modulated photoreflectance, same as thermal waves, 25, 161, 283
MOEMS, microoptoelectromechanical systems, 376
molding, same as moulding, 183
MOSFET, Metal Oxide Semiconductor Field Effect Transistor
  devices, 9, 11, 30, 266
  fabrication, 193–197, 255–267
  scaling trends, 258–260, 364–366
MOVPE, Metal Organic Vapor Phase Epitaxy, 332
molecular flow, 321
molybdenum, 48, 58, 120, 129, 203
monocrystalline, same as single crystalline, 4, 36, 65
monolayer, ML, 322
Monte Carlo simulation, 88, 162
Moore's law, 12, 363–366
MPC, Multi Project Chip, 242
MPPS, most penetrating particle size, 346
MPW, Multi Project Wafer, 242
MTBA, mean time between assists, 311
MTBC, mean time between cleans, 311
MTTF, mean time to failure, 311
multicrystalline, 47
multilayer resist (MLR), 112
Multi Project Chip, MPC, 242
Multi Project Wafer, MPW, 242
Murphy's yield model, 351

NA, numerical aperture, 100
nanocrystalline, 47
nanofluidic filter, 217
nanolaminate, 82, 332
nanotechnology, 3
nanowire, 148
native oxide, 80–81, 133–134
negative resist, 109
nested mask, same as peeling mask, 294
neutron transmutation doping, NTD, 39, 153
nickel, 57–58, 222, 228
nickel silicide, 64
NIL, nanoimprint lithography, 188
NIST, National Institute of Standards and Technology, 25
NO, nitrided oxide, 263
node, same as CMOS technology generation
Nomarski interference microscopy, 17
non-conformal step coverage, 86
non-uniformity, 25
novolak resist, 108
Noyce, Robert, 363
nozzle, 293
NSOM, near-field scanning optical microscope, 150, 375,
n-type silicon, 38–42
NTD, neutron transmutation doping, 39, 153
nuclear stopping, 159
nucleation, 75, 369
numerical aperture NA, 100, 259

O_2P, oxygen precipitate, 44, 240
oblique angle evaporation, 229
OES, 313
ohmic contact, 57, 249

**Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030**

OISF, oxidation induced stacking fault, also OSF, 44, 261
ONO, oxidized nitrided oxide, 263
OPC, optical proximity correction, 97, 242
optical emission spectroscopy, OES, 313
optical lithography, 99–117, 366–368
optical microscopy, 17–18
optical proximity correction, OPC, 97, 242
organic contamination, 138
oven, 107, see also furnace
overetch, 129
overlay, see alignment, 103
overplating, 228
overpolishing, 167
oxidation, 143–151
oxidation enhanced diffusion, OED, 158
oxidation induced stacking fault, OISF, also OSF, 44, 261
oxidation sharpening, 150
oxide breakdown, 146, 250
oxide defects, 250
oxide stress, 148
oxidized nitrided oxide, ONO, 263
oxynitride, 83, 248
ozone, 116

packaging, 296
PACVD, plasma assisted CVD, same as PECVD, 53
PAG, photoacid generator, 109
parabolic growth, 144
particle contamination, 136
parylene, 61, 218
passivation, 250, 258
pattern density effects, 202
pattern generation, 93
PDMS, poly(dimethyl)siloxane, 183, 186
peak anneal, 318
PEB, post exposure bake, 109,
PECVD (plasma-enhanced CVD), 52–53, 327
peeling mask, same as nested mask, 294
pellicle, 101
phase diagram, 80
phase shift mask, PSM, 95, 366. See also photomask
phosphoric acid, 120
phosphorus doped glass, PSG, 52, 218
photoacid generator, PAG, 109
photodiode, 154
photolithography, see lithography, 99
photomask, 94–97, 241–242, see also reticle
  compensation, 242
  count, 275, 379
  defects, 97
  writing, 94
photonic crystal, 122, 170
photoresist, 107–110
photoresist stripping, 116
physical cleaning, 140
physical vapor deposition, PVD, 49–51, 74–77
piezoelectric, 374
piezoresistance, 35, 291
pinhole, 56, 59, 97
Piranha, sulphuric acid peroxide mixture, 135
pitch, 103, 113
pitting, 80
planarization, 169, 202, CH 27
plasma,
  chemistry, 126
  CVD (PECVD), 52, 327
  equipment, 324–327
  etching, 125–130, 199–204
plating, 54–55, 221, 227–228
platinum, 63, 201, 252, 266
plug, 202, 278
PMMA, polymethyl methacrylate (resist), 95, 183
POA, post oxidation anneal, 316
point defect, 43, 154, 161–162
poisoning, 283
Poisson distribution, 351
polarity (of photomask), 102, 241, 258
polishing, 42, 165–172, 280
poly (polycrystalline silicon)
  gate, 193–194, 255–267
  LPCVD, 62, 331
  oxidation, 144, 217
  poly emitter, 272–274
  properties, 62–63
  trench filling, 274
polycide, polysilicon plus silicide, 200
polydimethyl siloxane, PDMS,
polyimide, 61, 223, 304, 375
polymers
  bonding, 177
  embossing, 186
  properties, 60, 283
porous silicon, 123–125, 223–224
post oxidation anneal, POA, 316
positive resist, 108
post exposure bake, PEB, 114
power-supply voltage, 260
power transistor, 9, 378
ppb, parts per billion, 20
ppm, parts per million, 20
ppma, parts per million atoms, 20
ppt, parts per trillion, 20
precipitate, 44, 161, 240
predeposition, 153
pressure sensor, 253, 291
Preston equation, 167
prime wafers, 349, 358. See also reclaim wafers
priming, same as adhesion promotion, 107
process integration, 237–253
process simulation, 27

**Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030**

profile,
  diffusion, lateral, 159
  diffusion, vertical, 155–157
  etch, 121, 125, 128, 199–203
profiler, 17
projected range, 160
projection optics, 100
proximity
  lithography, 99
  effect, 94. See also OPC
PSG, phosphorous doped silica glass, 52, 218
PSi, porous silicon, 123–125, 223–224
PSL, polystyrene latex sphere, 137
PSM, phase shift mask, 366–367
p-type silicon, 38–42
PTFE, polytetrafluoroethylene, 61
pumping speed, 322
PVD, physical vapor deposition, 49–51, 74–77
Pyrex glass, 176, 301
pyrolysis, 51
pyrometer, 316
PZT, $Pb(Zr,Ti)O_3$, 56, 184

QCM, quartz crystal microbalance, 313
quartz, 4, 241, 301, 368
quartz crystal microbalance, QCM, 313

range, 153
rapid thermal processing, RTP, 158, 316–319
rate limiting step, 52, 120, 329
RBS, Rutherford backscattering spectrometry, 22, 195
RCA clean, 135, 175, 256
RC-delay, 281
RCL-chip, 247
reactive ion etching, see RIE, 119
recessed LOCOS, 148
reclaim wafers, 358
reduction stepper, 99–101
reflective notching, 112
reflectometry, 19, 61
reflow, 258
refractive index, 48, 60, 137
refractory crucibles, 50
refractory metals, 129
release etch, 122, 218
remote plasma, 324
Rent's rule, 359
repair, 96, 349, 378
replacement gate, same as dummy gate, 265
replication, 184, 189
residence time, 83, 327
residual gas analyzer, RGA, 313
resist, see photoresist, 108
resistivity, 19
  diffused layers, 155
  DI-water, 346
  metals, 48, 58
  polysilicon, 62
  silicon, 36
resistor, 57, 245
resolution, 102, 112, 258
resonance, 224
reticle, 100, see also photomask
retrograde profile,
  etching, 128
  implantation, 162
reverse engineering, 26
rework, 119, 358
RF-MEMS, 220, 379
RGA, residual gas analyzer, 313
RIE, reactive ion etching, 119, same as plasma etching
  DRIE, 127, 130, 185, 199, 203, 214
  reactor, 324
  RIE lag, 202
rotating structures, 222
RT, room temperature,
RTA, rapid thermal annealing, 158, 316
RTO, rapid thermal oxidation, 316
RTP, rapid thermal processing, 158, 316

sacrificial etching, 122, 218
sacrificial layer, 217
Sailor defect etch, 121
salicide, short for self-aligned silicide, 194
SAM, self-assembled monolayer, 107, 187
SAMPLE simulator, 30, 89, 114
SBC, standard buried collector (bipolar transistor), 269–272
SC, standard clean, 135
scaling, see also Moore's law
  bipolar, 272–274
  CMOS, 258–260
  multilevel metallization, 280
Scanning Electron Microscope, SEM, 17
scatterometry, 44, 137
sccm, standard cubic centimeters per minute, 15, 331, see also STP
Schottky contact, 57
scribeline, 14
scrubber, 330, see also gas abatement, 347
SCS, single crystal silicon, 4, 36
sealing, 232
Secco defect etch, 121
secondary ion mass spectroscopy, SIMS, 21, 69
seed layer, 55, 228
SEG, selective epitaxial growth, 70
segregation, 38, 147
selective deposition, 230
selective epitaxial growth, SEG, 70
selectivity, 128, 170, 201, 207

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

self-align(ment), 193
  bipolar, 273–274
  MOS gate, 193, 258, 263
  phase shift mask, 367
  rotor, 222
  silicide (salicide), 194
  TFT, 303
  Wells, 194
self-interstitial, 44, 158
self-limiting depth,
  <100>, 205
  <110>, 213
SEM, scanning electron microscope, 17
shadow mask, 229, 294
shallow trench isolation, STI, 262
sheet resistance, 19, 26, 64, 155, 246, 258, 266
shrink
  resist, 113
  version, 364
sidewall spacer, 129, 201, 230, 264, 274
SiGe, silicon–germanium, 66, 296, 370
silane, SiH_4, 51–62, 67, 347
silicate, 250
silicide, 63, 194–196
silicon, 35
  crystal growth, 36
  epitaxy, 65–71, 333–335
  plasma etching, 128
  properties, 35–37
  wafers, 40, 238–241, 261, 288–289
  wet etching,
silicon carbide, SiC, 35, 47, 53, 145, 373
silicon dioxide, SiO_2,
  CVD, 51
  etching, 120–121
  properties, 59–62
  reliability, 250–251
  structure, 145
  thermal, 143
silicon monoxide, SiO, 38, 251
silicon nitride, Si_3N_4, 51, 129, 206–209
silicon on insulator, SOI, see this
silicon on sapphire, SOS, 43
siloxane, 60, 373
silsesquioxane, 164
SIMOX, Separation by Implantation of Oxygen, 164. See also SOI.
SIMS, secondary ion mass spectrometer, 21
simulation, 27–32
  anisotropic wet etching, 211
  deposition, 31, 88
  diffusion, 156
  epitaxy, 69
  equipment, 312
  etching, 211
  ion implantation, 29, 161
  lithography, 114–115
  oxidation, 29, 146
single-wafer processing, 309
SiO_2, silicon dioxide, 59–62, 143–151
Sirtl defect etch, 121
slip, 44
slope etching, 122
slpm, standard liters per minute, 316
slurry, 166
SM, stress migration, 251
Smart-cut, 181
SMIF (Standard Mechanical InterFace), 338
smoothing, 166
SOD, spin-on dielectric, 169, 283
soda lime glass, 242
soft bake, 107
soft lithography, 183
SOG, spin-on-glass, 59–60
SOI, silicon on insulator, 4, 43
  applications, CMOS, 266, 370
  applications, MEMS, 297, 375, 378
  bonded, 180
  SIMOX, 164
  Smart-cut, 181
  wafers, 241
solar cells, 9, 237
sol–gel, 56
solubility, 153
SOS, silicon on sapphire, 43
source/drain, 11, 194, 258, 263–265
spacer, 129, 201, 230, 264, 274
SPC, statistical process control, 242
spiking, 135
spin coating, 55, 107
spin-on glass, SOG, 59–60
SPM, sulphuric acid peroxide mixture, 135
SPV, surface photo voltage, 140
spray coating, 107
spray tool, 120
spreading resistance profiling, SRP, 69
sputter etching, 325
sputtering
  bias sputtering, 325
  collimated sputtering, 77
  equipment, 50, 325–326
  etching, 325
  reactive, 325
  yield, 51
SRAM, static random access memory, 349
SRP, spreading resistance profiling, 69
SSP, single-side-polished wafer, 42, 288
stacking fault, 44,
stamping, 183
standard buried collector bipolar transistor, SBC, 270–274

**Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030**

standing waves, 110
statistical process control, SPC, 242
stencil mask, 229, same as shadow mask
step-and-scan, 101
step-and-repeat, 100–101
step-and-stamp, 183
step coverage, 86, 332
stepper (step-and-repeat lithography tool), 100–101
stereomicrolithography, 231
sticking probability, 73, 321
stiction, 219
stoichiometry, 47
Stoney formula, 86
STI, shallow trench isolation, 262
STO, strontium titanate $SrTiO_3$, 48
STP, standard temperature and pressure (273K, 1 atm), 323
straggle, 160
Stranski–Krastanov growth mode, 74
Stress, 43, 83–86, 149
stress migration, 251
Stribeck diagram, 167
stripping, 116, same as resist ashing
structural layer, 217
SU-8 epoxy resist, 18, 218, see thick resist, 108
submicron, dimensions $<1\,\mu$m, 258–261
substitutional diffusion, 155
substrate, 4, 301, 304, see also wafer
sulphuric acid, 135
superlattice, 83
SUPREM simulator, 28
surface analysis, 21–22
surface devices, 9
surface energy, 175
surface micromechanics, 218
surface preparation, 107, 133–140, see also cleaning
surface reaction limited, 52, 120, 330, 331, 335
surface roughness, 42, 78–79, 175, 240
surface stamping, 186
swing curve, 111

tantalum, 21, 76, 129, 282
tantalum nitride, 21, 81
TAR, top antireflection (coating), 111
target, 50, see sputtering
TCAD, technology CAD, 27
TCO, transparent conducting oxide, 303
TCR, temperature coefficient of resistivity, 76
TDS, thermal desorption spectroscopy, 24
TED, transient enhanced diffusion, 265
Teflon®, trade name of PTFE, 61
TEM, transmission electron microscope, 17
temperature coefficient of resistivity, TCR, 76
TEOS, tetraethoxysilane, $Si(OC_2H_5)_4$, 53
tert-butanol, 220
test structures, 14, 95–96, 242
texture, 76
TFH, thin film head, 108, 131, 302, 356
TFT, thin film transistor, 302–304
thermal budget, 249
thermal conductivity, 37, 58, 60
thermal desorption spectroscopy, TDS, 24
thermal expansion, 58–60, 373, see coefficient of thermal expansion
thermal isolation, 291
thermal oxidation, 143
thermal stability, 79–81, 248–249
thermal waves, 25, 161, 283, same as modulated photoreflectance
thermocompression bonding, TCB, 177
thermocouple, 317
thin films
  deposition, 49–56, 73–80
  dielectrics, 58–62
  metallic, 56–58
  polymeric, 62–63
  stresses, 83–86
  structure, 73–79
thin film devices, 10
thin film head, TFH, 108, 131, 302, 356
thin film optics in resist, 110
thin film transistor, TFT, 302–304
thinning, 165, 174
threshold limit value, TLV, 347
throughput, 100, 310
TiN, titanium nitride, 77, 278–279
tip, 150, 217, 233, 375
TIR, total indicator reading, 239
titanium, 58, 81–82, 120, 129, 337–338
titanium silicide, $TiSi_2$, 63, 195
TLV, threshold limit value, 347
TMAH, tetramethyl ammonium hydroxide, 205–207
tool, 237, 309, 355, same as equipment
top antireflection (coating) TAR, 111
top gate (TFT), 302
top surface imaging, TSI, 112
total indicator reading, TIR, 239
total thickness variation, TTV, 239, 289
transfer bonding, 376
transient enhanced diffusion, TED, 265
transition width, 68–69
transmission electron microscope, TEM, 17
transparency, 49, 368
transparent conducting oxides, TCO, 303
transport-limited reaction, 52, 120, 330, 331, 335
trench isolation, 262, 274
TSI, top surface imaging, 112
TTV, total thickness variation, 239, 289
tub, same as CMOS well, 255–257, 261
tungsten, 52, 129, 168, 231, 278–280

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030

tungsten lamp, 316
twin-well, 194, 261
TXRF, total reflection X-ray fluorescence, 21, 140

UHV, ultrahigh vacuum, 323
ULK, ultra-low k dielectric material, 283
ULPA, Ultra Low Penetration Air filter, 345
ULSI, ultra large scale integration, 14
ultrahigh vacuum, UHV, 323
ultrasonic cleaning, 141
undercutting, 121–122, 217–219
uniformity, 25
unidirectional flow, 13, 344, same as laminar flow
unintentional processes, 287
unlimited source diffusion, 156
up-time, 310
UPW, ultrapure water, same as DI-water, 12, 116, 140, 346
USG, undoped silica glass, 52
utilization, 310, 327

vacancy, 44, 154
vertical furnace, 315
via, 277, 280
viscosity (of resist), 107–108
VLSI, very large scale integration, 14
void, 58, 251
volatility, 119, 126, 133
volume change, 63, 146
volume devices, 8
volume stamping, 184

wafers, 4, 42, 238–241, 288, 298, 301, 377
  epitaxial, 65, 240, 291
  silicon, 40–43, 239–241, 261, 288
  SOI, 43, 241
  specifications, 42, 238–241, 288–289
wafering, 40
wafer starts per month, WPM, 355
warp, 239
waveguide, 83
well, 194, 256, 261, same as tub
wet bench, 135
wet cleaning, 135–136
wet etching, 120
wet oxidation, 143
WIWNU, within-wafer non-uniformity, 25
WPH, wafers per hour, 309
WPM, wafer starts per month, 355
Wright defect etch, 121
WTWNU, wafer-to-wafer non-uniformity, 25

xerogel, 56
XPS, X-ray photoelectron spectroscopy, 22, same as ESCA
XRD, X-ray diffraction, 20, 48
XRL, X-ray lithography, 102, 184, 368
XRR, X-ray reflectivity, 19
XRT, X-ray tomography, 24

yield, 12, 349, 371, see also sputtering yield
  cost, 358–360
  fab, 349
  models, 349–353
  vs. reliability, 250
yield strength, 36–37, 44
Young's modulus, 35, 57, 60, 61, 63, 83–86, 168, 178, 283

zero anneal, 318
zeta potential, 136
ZnO, 83, 374
zone melting, 39
zone model, 75

**Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1030**