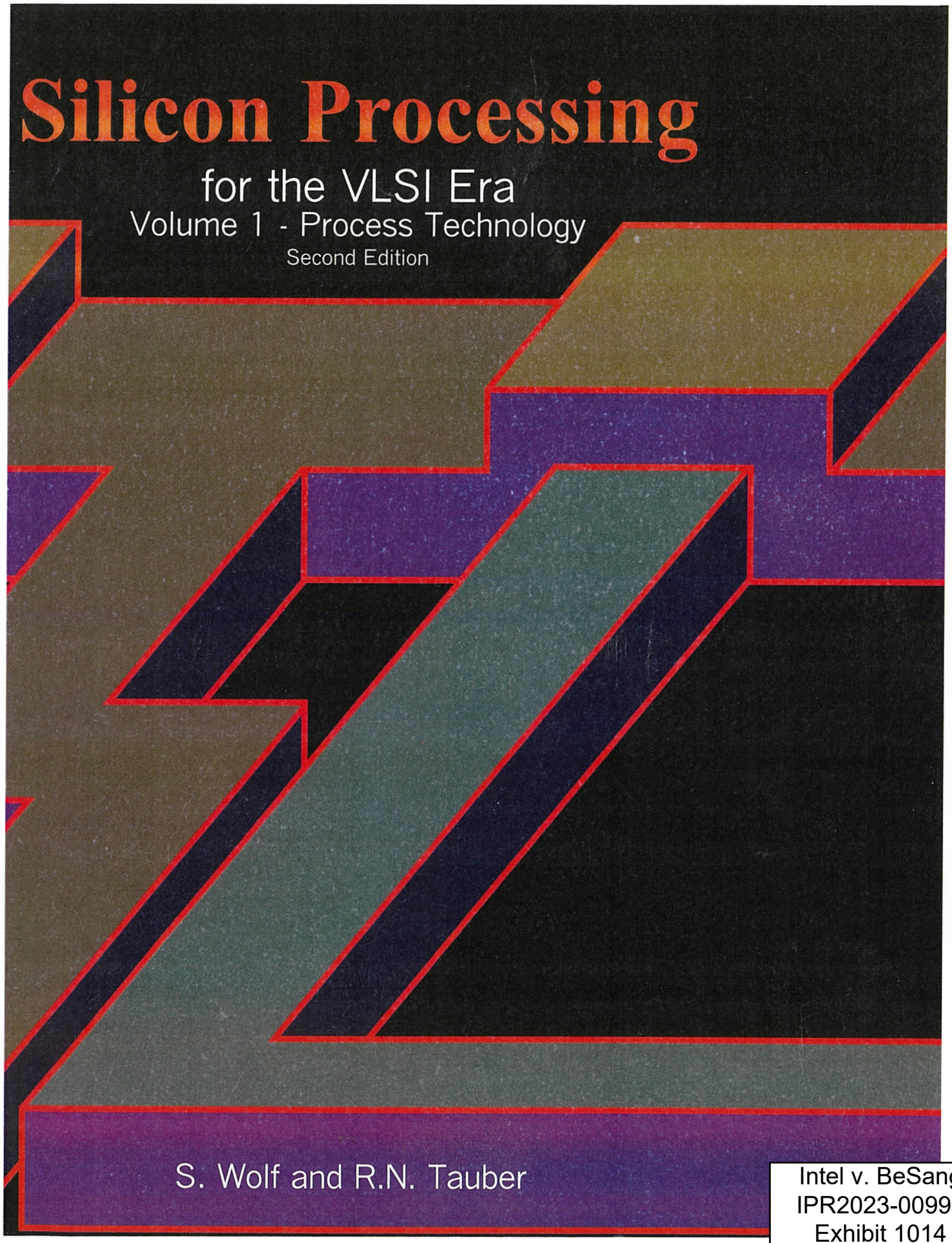
Silicon Processing

for the VLSI Era

Volume 1 - Process Technology

Second Edition



S. Wolf and R.N. Tauber

Intel v. BeSang
IPR2023-00991
Exhibit 1014

SILICON PROCESSING
FOR
THE VLSI ERA

VOLUME 1:

PROCESS TECHNOLOGY

Second Edition

STANLEY WOLF Ph.D.

RICHARD N. TAUBER Ph.D.

LATTICE PRESS

Sunset Beach, California

DISCLAIMER

This publication is based on sources and information believed to be reliable, but the authors and Lattice Press disclaim any warranty or liability based on or relating to the contents of this publication.

Published by:

LATTICE PRESS
Post Office Box 340
Sunset Beach, California 90742, U.S.A.

Cover design by Roy Montibon, New Archetype Publishing, Los Angeles, CA.

Copyright © 2000 by Lattice Press.
All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage and retrieval system without written permission from the publisher, except for the inclusion of brief quotations in a review.

Library of Congress Cataloging in Publication Data
Wolf, Stanley and Tauber, Richard N.

Silicon Processing for the VLSI Era
Volume 1: Process Technology

Includes Index
1. Integrated circuits-Very large scale integration. 2. Silicon. I. Title

ISBN 0-9616721-6-1

9 8 7 6 5 4 3 2 1

PRINTED IN THE UNITED STATES OF AMERICA

To my wife and life editor, Carrol Ann,
and to my children, Jennifer Laura and Stanley Charles Ross

Stanley Wolf

To my partner in life, Barbara

Richard N. Tauber

MASTER TABLE OF CONTENTS

CHAPTER

1. SILICON: SINGLE-CRYSTAL GROWTH & WAFER PREPARATION	1
2. CRYSTALLINE DEFECTS AND GETTERING	35
3. VACUUM TECHNOLOGY FOR ULSI APPLICATIONS	70
4. BASICS OF THIN FILMS	104
5. CLEANING TECHNOLOGY FOR ULSI	119
6. CHEMICAL VAPOR DEPOSITION OF AMORPHOUS AND POLYCRYSTALLINE FILMS	149
7. SILICON EPITAXIAL FILM GROWTH and SILICON ON INSULATOR	225
8. THERMAL OXIDATION OF SILICON	265
9. DIFFUSION IN SILICON	324
10. ION IMPLANTATION FOR ULSI	371
11. ALUMINUM THIN FILMS AND PHYSICAL VAPOR DEPOSITION IN ULSI	434
12. LITHOGRAPHY I: OPTICAL PHOTORESISTS - MATERIAL PROPERTIES AND PROCESS TECHNOLOGY	488
13. LITHOGRAPHY II: OPTICAL ALIGNERS AND PHOTOMASKS	545
14. DRY ETCHING FOR ULSI	655
15. MULTILEVEL METALLIZATION FOR ULSI	719
16. CMOS PROCESS INTEGRATION	807
17. ASSEMBLY AND PACKAGING FOR ULSI	841
APPENDICES	875
INDEX	879

Exhibit 1014 - Page 5 of 218

DETAILED TABLE OF CONTENTS

PREFACE

PROLOGUE

1. SILICON: SINGLE-CRYSTAL GROWTH & WAFER PREPARATION 1

1.1 TERMINOLOGY OF CRYSTAL STRUCTURE, 1
1.2 THE MANUFACTURE OF SINGLE-CRYSTAL SILICON, 5
1.2.1 From Raw Material to Electronic Grade Polysilicon
1.3 CZOCHRALSKI (CZ) CRYSTAL GROWTH, 8
1.3.1 Czochralski Crystal Growth Sequence
1.3.2 Incorporation of Impurities into the Crystal (Normal Freezing)
1.3.3 Modifications Encountered to Normal Freezing in CZ Growth
1.3.4 Czochralski Silicon Growing Equipment
1.3.4.1 Furnace
1.3.4.2 Crystal Pulling Mechanism
1.3.4.3 Ambient Control
1.3.4.4 Control System
1.3.5 Analysis of Czochralski Silicon in Ingot Form
1.3.5.1 Oxygen & Carbon Measurements in Si Using IR Absorbance Spectroscopy
1.4 FLOAT-ZONE SINGLE-CRYSTAL SILICON, 20
1.5 FROM INGOT TO FINISHED WAFER: SLICING; ETCHING; POLISHING, 22
1.6 SPECIFICATIONS OF SILICON WAFERS FOR ULSI, 25
1.6.1 Electrical Specifications
1.6.2 Mechanical/Dimensional Specifications
1.6.3 Chemical/Structural Specifications
1.6.4 Surface/Near Surface Specifications
1.7 THE ECONOMICS OF SILICON WAFERS, 29
1.8 TRENDS IN SILICON CRYSTAL GROWTH AND ULSI WAFERS, 31
REFERENCES, 32
PROBLEMS, 34

2. CRYSTALLINE DEFECTS AND GETTERING 35

2.1 CRYSTALLINE DEFECTS IN SILICON, 36
2.1.1 Point Defects
2.1.2 One-Dimensional Defects (Dislocations)
2.1.3 Area Defects (Stacking Faults)
2.1.4 Volume Defects
2.1.4.1 Classical Homogeneous 3-D Nucleation Theory
2.1.4.2 Nucleation of Volume Defects in Silicon Ingots
2.2 INFLUENCE OF DEFECTS ON DEVICE PROPERTIES, 53
2.2.1 Leakage Currents in *pn* Junctions
2.2.2 Gate Oxide Quality

Exhibit 1014 - Page 6 of 218

2.2.3 Wafer Resistance to Warpage
2.3 CHARACTERIZATION OF CRYSTAL DEFECTS, 55
2.4 OXYGEN IN SILICON, 57
2.5 GETTERING, 58
2.5.1 Basic Gettering Pinciples
2.5.2 Extrinsic Gettering
2.5.3 Intrinsic Gettering
2.5.4 Gettering with Oxygen Precipitates
2.5.5 Summary of Gettering
REFERENCES, 67
PROBLEMS, 69

3. VACUUM TECHNOLOGY FOR ULSI APPLICATIONS 70

3.1 FUNDAMENTAL CONCEPTS OF GASES AND VACUUMS, 70
3.2 PRESSURE UNITS, 71
3.3 VACUUM PRESSURE RANGES, 72
3.3.1 Mean Free Path and Gas Flow Regimes
3.4 THE LANGUAGE OF GAS/SOLID INTERACTIONS, 73
3.5 TERMINOLOGY OF VACUUM PRODUCTION AND PUMPS, 75
3.5.1 Vacuum Pump Types
3.5.2 Pumping Speed and Conductance
3.5.3 Throughput
3.6 ROUGH PUMPS, 81
3.6.1 Oil-Sealed Rotary Mechanical Pumps
3.6.2 Vacuum Pump Oils for Semiconductor Processing
3.6.3 Roots Pumps
3.6.4 Dry Mechanical Pumps
3.7 HIGH VACUUM PUMPS I: CRYOGENIC PUMPS, 87
3.7.1 Cryopump Operation
3.7.2 Cryopump Regeneration
3.8 HIGH VACUUM PUMPS II: TURBOMOLECULAR PUMPS, 92
3.9 TOTAL PRESSURE MEASUREMENT, 95
3.9.1 Capacitance Manometers
3.9.2 Thermocouple Gauges
3.9.3 Pirani Gauges
3.9.4 Ionization (High Vacuum) Gauges
3.10 MEASUREMENTS OF PARTIAL PRESSURE: Residual Gas Analyzers, 98
3.10.1 Operation of Residual Gas Analyzers (RGA)
3.10.2 RGAs and Non-High Vacuum Applications: Differential Pumping
3.10.3 Interpretation of RGA Spectra
3.10.4 RGA Specification List
REFERENCES, 102
PROBLEMS, 103

4. BASICS OF THIN FILMS 104

4.1 THIN FILM GROWTH, 105
4.1.1 Thin Film Nucleation
4.1.2 The Structure of Thin Films
4.2 MECHANICAL PROPERTIES OF THIN FILMS, 108
4.2.1 Adhesion
4.2.2 Stress in Thin Films
4.2.3 Other Mechanical Properties
4.3 ELECTRICAL PROPERTIES OF METALLIC THIN FILMS, 113
4.3.1 Measurement of the Electrical Properties of Thin Films
4.3.2 Electrical Transport in Thin Films
REFERENCES, 118
PROBLEMS, 118

5. CONTAMINATION CONTROL AND CLEANING TECHNOLOGY FOR ULSI 119

5.1 TYPES OF CONTAMINATION IN IC FABRICATION, 119
5.2 SOURCES OF CONTAMINATION IN IC PROCESSING, 120
5.3 THE EFFECTS OF CONTAMINATION ON ULSI DEVICES, 121
5.4 CONTAMINATION PREVENTION MEASURES, 122
5.4.1. Cleanroom Design and Minienvironments
5.4.2 Gowning Procedures
5.4.3 Ultrapure Chemicals
5.4.4 De-Ionized (DI) Water
5.4.5 Machine-Design and Wafer Handling Techniques
5.4.6 Process Modifications
5.5 WAFER CLEANING TECHNIQUES FOR ULSI, 128
5.5.1 Wet-Chemical Removal of Film Contaminants
5.5.1.1 FEOL Wet-Cleaning - RCA Clean
5.5.1.2 Modifications to the RCA Clean
5.5.1.3 Ozone-containing Water
5.5.1.4 Wet Cleaning of Metal-Coated Wafers (BEOL)
5.5.1.5 Spray Processing
5.5.2 Vapor-Phase and Dry Cleaning
5.5.2.1 Vapor-Phase Cleans
5.5.2.2 Ozone/UV Dry Cleans
5.5.3 Photoresist Removal
5.6 PARTICLE REMOVAL, 134
5.6.1 Vibrational Scrubbing (Ultrasonic and Megasonic)
5.6.2 Particle Removal by Brush Scrubbing
5.6.3 Particle Removal by Liquid Jet spraying
5.6.4 Particle Removal by Cryosol Spray Techniques
5.6.5 Particle Removal by DUV-Laser Irradiation
5.7 RINSING AND DRYING WAFERS, 139
5.7.1 Rinsing

5.7.2 Wafer Drying after Rinse
5.7.2.1 Spin-Dryers
5.7.2.2 Isopropyl-Alcohol (IPA) Vapor Dryers
5.7.2.3 Hot DI-Water Drying
5.7.2.4 Marangoni Drying
5.8 PARTICLE DETECTION ON WAFER SURFACES, 143
5.8.1 Automatic Laser Particle Counters for Detecting Particles on Wafers
5.8.2 Automatic Defect Classification (ADC)
REFERENCES, 146
PROBLEMS, 148
6. CHEMICAL VAPOR DEPOSITION of AMORPHOUS & POLYCRYSTALLINE FILMS 149
6.1 BASIC ASPECTS OF CHEMICAL VAPOR DEPOSITION, 150
6.1.1 Grove's Simplified CVD-Film-Growth Model
6.1.2 Gas Flow and Gas-Phase Mass Transfer
6.1.2.1 Gas Flow in CVD Reactors
6.1.2.2 The Stagnant Layer Model
6.1.2.3 Boundary Layer Theory
6.2 CHEMICAL VAPOR DEPOSITION SYSTEMS, 162
6.2.1 Components of CVD Systems
6.2.1.1 Gas Sources and Delivery Systems for CVD
6.2.1.2 Mass-flow Controllers
6.2.1.3 Heating Sources for CVD Reaction Chambers
6.2.2 Terminology of CVD Reactor Design
6.2.3 Atmospheric Pressure CVD Reactors
6.2.4 Low Pressure Chemical Vapor Deposition Reactors
6.2.4.1 Horizontal-Tube LPCVD Batch Reactors (Hot Wall)
6.2.5 Plasma Enhanced CVD: Physics, Chemistry, & Reactor Designs
6.2.5.1 Parallel-Plate Cold-Wall Batch PECVD Reactors
6.2.5.2 Mini-Batch Radial Cold-Wall PECVD Reactors
6.2.5.3 Single-Wafer Cold-Wall PECVD Reactors
6.3 POLYCRYSTALLINE SILICON: PROPERTIES and CVD METHODS, 180
6.3.1 Properties of Polysilicon Thin Films
6.3.1.1 Physical Structure and Mechanical Properties of Poly-Si
6.3.1.2 Electrical Properties of Polysilicon
6.3.2 Chemical Vapor Deposition of Polysilicon
6.3.2.1 Deposition Parameters
6.3.2.2 Structure of Polysilicon - Deposition Condition Dependence
6.3.3 Doping Techniques for Polysilicon
6.3.3.1 Diffusion Doping of Polysilicon
6.3.3.2 Ion Implantation Doping of Polysilicon
6.3.3.3 *In Situ* Doping of Polysilicon
6.4 PROPERTIES and DEPOSITION OF CVD SiO_2, 189
6.4.1 Chemical Reactions for CVD SiO_2 Formation
6.4.1.1 Low-Temperature Silane-Based CVD SiO_2
6.4.1.2 Medium-Temperature LPCVD TEOS SiO_2

6.4.1.3 Low-Temperature PECVD TEOS
6.4.1.4 Ozone TEOS
6.4.2 Step Coverage of As-Deposited CVD SiO_2 Films
6.4.3 CVD & Applications of Undoped and Doped SiO_2 Films
6.4.3.1 Undoped CVD SiO_2
6.4.3.2 Phosphosilicate Glass (PSG)
6.4.3.3 Borophosphosilicate Glass (BPSG)
6.5 PROPERTIES AND CHEMICAL VAPOR DEPOSITION OF SILICON NITRIDE, 202
6.6 OTHER DIELECTRIC FILMS DEPOSITED BY CVD, 206
6.6.1 Silicon Oxynitrides
6.7 CVD OF METALS, SILICIDES, AND NITRIDES FOR ULSI APPLICATIONS, 207
6.7.1 CVD of Tungsten (W)
6.7.1.1 CVD Tungsten Chemistry
6.7.1.2 Blanket CVD W and Etchback
6.7.2 Chemical Vapor Deposition of Tungsten Silicide (WSi_X)
6.7.3 CVD of Titanium Nitride (TiN)
6.7.4 CVD of Aluminum
REFERENCES, 220
PROBLEMS, 224
7 SILICON EPITAXIAL GROWTH and SILICON ON INSULATOR 225
7.1 THE DEVICE APPLICATIONS OF EPITAXY, 226
7.1.1. Why Do We Use CMOS Epitaxial Wafers?
7.2 GROWTH OF EPITAXIAL LAYERS, 228
7.2.1 Atomistic Model of Film Growth
7.3 CHEMICAL REACTIONS USED IN SILICON EPITAXY, 230
7.4 PROCESS CONSIDERATIONS FOR EPITAXIAL DEPOSITION, 233
7.4.1 Silicon Precursors
7.4.2 Doping of the Epitaxial Films
7.4.3 Intentional Doping
7.4.4 Unintentional Doping (Autodoping and Solid State Diffusion)
7.5 DEFECTS IN EPITAXIAL FILMS, 238
7.5.1 Wafer Preparation
7.5.2 Defects Induced During Epitaxial Deposition
7.5.3 Pattern Shift, Distortion, and Washout
7.6 LOW-TEMPERATURE EPITAXY PROCESSES, 243
7.7 SELECTIVE EPITAXIAL GROWTH (SEG), 245
7.8 EPITAXIAL DEPOSITION EQUIPMENT, 247
7.9 CHARACTERIZATION OF EPITAXIAL FILMS, 251
7.9.1 Optical Inspection of Epitaxial Film Surfaces
7.9.2 Electrical Characterization
7.9.3 Epitaxial Film Thickness Measurements
7.9.4 Infrared Reflectance Measurement Techniques
7.9.4.1 ASTM Technique of Epi Layer Thickness Measurement with IR Reflectance

7.9.5 Fourier-Transform Infrared (FTIR) Spectroscopy
7.10 SILICON-ON-INSULATORS (SOI), 256
7.10.1 Silicon-on-Sapphire (SOS)
7.10.2 SIMOX
7.10.3 Wafer Bonding
REFERENCES, 261
PROBLEMS, 264

8. THERMAL OXIDATION OF SINGLE-CRYSTAL SILICON **265**

8.1 THE PROPERTIES OF SILICA GLASSES, 267
8.2 OXIDATION KINETICS: THE DEAL-GROVE MODEL, 268
8.2.1 The Deal-Grove (or Linear-Parabolic) Model
8.3 FACTORS WHICH AFFECT THE ONE-DIMENSIONAL OXIDATION RATE, 276
8.3.1 Crystal Orientation Effects on Oxide Growth Rates
8.3.2 Dopant Effects on Oxidation Growth Rates
8.3.3 Water Effects During Dry Oxidation
8.3.4 The Dependence of Oxidation Rates on Chlorine
8.3.5 Effect of Pressure on Oxidation Rates
8.3.5.1 High Pressure Oxidation (HIPOX):
8.3.5.2 Low Pressure Oxidation
8.4 THE INITIAL OXIDATION STAGE AND THE GROWTH OF THIN OXIDES, 284
8.4.1 Experimental Results for Thin Oxide Growth
8.4.2 Growing Thin Oxides
8.5 THE NATURE OF THE Si-SiO_2 INTERFACE, 288
8.5.1 Fixed Oxide Charge
8.5.2 Mobile Ionic Charge
8.5.3 Interface Trap Charge
8.5.4 Oxide Trapped Charge
8.6 STRESS IN SILICON DIOXIDE, 296
8.7 DOPANT IMPURITY REDISTRIBUTION DURING OXIDATION, 296
8.8 OXIDATION OF POLYSILICON, 298
8.9 THE OXIDATION OF SILICON NITRIDE, 299
8.10 THERMAL NITRIDATION OF SILICON AND SILICON DIOXIDE, 299
8.11 TWO-DIMENSIONAL OXIDE-GROWTH EFFECTS, 300
8.11.1 Birds Beak Encroachment
8.11.2 LOCOS Induced Gate Thinning
8.11.3 Trench Corner Rounding
8.11.4 Gate Birds Beak
8.12 OXIDATION SYSTEMS, 303
8.12.1 Thermal Budget
8.12.2 Thermal Processing Equipment (Tools)
8.12.3 Furnaces
8.12.4 Horizontal Furnaces
8.12.5 Vertical Furnaces
8.12.6 Anatomy of a Vertical Furnace
8.12.7 Fast-Ramp Furnaces

8.12.8 Rapid Thermal Processing
8.12.9 RTP Systems
8.12.10 High Pressure Oxidation (HIPOX) Systems
8.13 THICKNESS MEASUREMENTS OF OXIDE FILMS, 314
8.13.1 Optical Measurements of Oxide Film Thickness
8.13.2 Electrical Measurement of Oxide Film Thickness
8.13.3 Physical Thickness Measurement of Oxide Films
8.14 TRENDS IN THIN OXIDE GROWTH, 317
8.14.1 Limitations of Silicon Dioxide
8.14.2 Reliability Limits for Thin Oxides
8.14.3 Fowler Nordheim (F-N) Tunneling
8.14 4 Alternative Gate Insulator Materials for MOSFETs
REFERENCES, 320
PROBLEMS, 322
9. DIFFUSION in SILICON **324**
9.1 THE MATHEMATICS OF DIFFUSION, 325
9.1.1 Fick's First Law
9.1.2 Fick's Second Law
9.1.3 Solutions to Fick's Second Law
9.1.3.1 Chemical Pre-Deposition
9.1.3.2 Drive-In Diffusion
9.1.3.3 Drive-In From An Ion Implantation Predeposition
9.1.4 Concentration Dependence of the Diffusion Coefficient
9.2 DEFECTS AND DOPANT DIFFUSION, 322
9.2.1 Point Defects in Silicon
9.2.2 Temperature Dependence of the Diffusion Coefficient Under Intrinsic Conditions
9.2.3 Intrinsic Diffusion Coefficients
9.2.4 Fast Diffusers in Silicon
9.3 ATOMISTIC MODELS OF DIFFUSION, 336
9.4 DIFFUSION MODELLING, 339
9.4.1 SUPREM III:
9.4.2 SUPREM III Models for B, As, P, and Sb Diffusion
9.4.3 Modeling Intrinsic Diffusion
9.4.3.1 Boron
9.4.3.2 Arsenic
9.4.3.3 Phosphorus
9.4.3.4 Antimony
9.4.4 Modeling Extrinsic Diffusion
9.4.5 Modeling Diffusion with SUPREM IV
9.5 DIFFUSION IN POLYCRYSTALLINE SILICON, 346
9.6 DIFFUSION IN SILICON DIOXIDE, 347
9.6.1 Boron Penetration of Thin Gate Oxides
9.7 ANOMALOUS DIFFUSION EFFECTS, 348
9.7.1 Electric Field Enhancement

9.7.2 Emitter Push Effect
9.7.3 Lateral Diffusion Under Oxide Windows
9.7.4 Oxidation-Enhanced Diffusion (OED)
9.7.5 Transient Enhanced Diffusion (TED)
9.8 DIFFUSION SYSTEMS AND DIFFUSION SOURCES, 357
9.8.1 Advanced Diffusion Technologies
9.8.1.1 Rapid Vapor Doping (RVD)
9.8.1.2 Gas Immersion Laser Doping (GILD)
9.9 MEASUREMENT TECHNIQUES FOR DIFFUSED LAYERS, 359
9.9.1 Sheet Resistance Measurements
9.9.2 Capacitance-Voltage (C-V) Measurements
9.9.3 Spreading Resistance Profiling (SRP)
9.10 JUNCTION DEPTH MEASUREMENTS: PHYSICAL TECHNIQUES, 363
9.10.1 Angle Lap and Stain
9.10.2 Groove and Stain
9.10.3 Secondary Ion Mass Spectroscopy
9.10.4 Two-Dimensional Depth Profiling
REFERENCES, 367
PROBLEMS, 370

10. ION IMPLANTATION for ULSI **371**

10.1 ADVANTAGES (AND PROBLEMS) OF ION IMPLANTATION, 372
10.1.1 Advantages
10.1.2 Problems/Limitations of Ion Implantation
10.2 IMPURITY PROFILES OF IMPLANTED IONS, 374
10.2.1 Definitions Associated with Ion Implantation Profiles
10.2.2 Theory of Ion Stopping
10.2.3 Models for Implantation Profiles in Amorphous Solids
10.2.3.1 Higher Moment Distributions for Implant Profiles in Amorphous Material
10.2.4 Implanting Into Single Crystal Materials: Channeling
10.2.5 Calculating Implantation Profiles: Boltzmann Transport Equation and Monte Carlo Approaches
10.2.5.1 Monte Carlo Calculations
10.3 ION IMPLANTATION DAMAGE ACCUMULATION & ANNEALING IN SILICON, 386
10.3.1 Implantation Damage in Silicon
10.3.2 Primary Crystalline Defect Damage
10.3.3 Amorphous Layer Damage
10.3.4 Electrical Activation and Implantation Damage Annealing
10.3.4.1 Electrical Activation of Implanted Impurities
10.3.4.2 Annealing of Primary Crystalline Damage
10.3.4.3 Annealing of Amorphous Layers
10.3.4.4 Dynamic Annealing Effects

10.3.4.5 Diffusion of Implanted Impurities
10.4 ION IMPLANTATION EQUIPMENT, 396
10.4.1 Components of an Ion Implantation System
10.4.2 Ion Implanter Types
10.4.2.1 Medium-Current Implanters
10.4.2.2 High-Current Implanters
10.4.2.3 Low-Energy Implanters
10.4.2.4 High-Energy Implanters
10.4.2.5 High-Angle Implanters
10.4.3 Ion Implantation Equipment System Limitations
10.4.3.1 Elemental and Particulate
10.4.3.2 Dose Monitoring Inaccuracies due to Beam Charge-state
10.4.3.3 Implantation Mask Problems
10.4.3.4 Wafer Charging During Implantation
10.4.3.5 Machine-to-Machine Dose Matching
10.4.3.6 "Scan Lock-Up" and Scanned-beam Machines
10.4.3.7 Micro-Uniformity Dose Errors
10.4.4 Ion Implantation Safety Considerations
10. 5 CHARACTERIZATION OF ION IMPLANTATION, 412
10.5.1 Measurement of Implantation Dose and Dose Uniformity
10.5.1.1 Implantation Dose Measurements
10.5.1.2 Implantation Dose Uniformity and Diagnosis of Implanter
10.5.2 Measurement of Implantation Depth Profiles
10.5.3 Measurement of Implantation Damage and Annealing Efficacy
10.6 EXAMPLES OF ION IMPLANTATION PROCESS APPLICATIONS, 417
10.6.1 Selecting Masking Layer Materials and Thickness
10.6.2 Implanting Through Surface Layers
10.6.3 Threshold-Voltage Control in MOS Devices
10.6.4. Shallow Junction Formation by Ion Implantation
10.6.5 High-Energy Implantation
10.6 6 Fabrication of SOI materials
10.6.7 New Techniques For Doping
10.6.7.1 Molecular Ions
10.6.7.2 Plasma Doping or Plasma Immersion Ion Implantation
10.6.7.3 Laser Doping
REFERENCES, 428
PROBLEMS, 432
11. ALUMINUM THIN FILMS AND PHYSICAL VAPOR DEPOSITION IN ULSI 434
11.1 ALUMINUM THIN FILMS IN ULSI, 435
11.2 SPUTTER DEPOSITION FOR ULSI, 438
11.2.1 Introduction to Glow Discharge Physics
11.2.2 The Creation of Glow Discharges
11.2.3 Structure of Self-Sustaining Glow Discharges and Their Dark Spaces

11.2.4 Obstructed Glow Discharges & Dark-Space Shielding
11.3 THE PHYSICS OF SPUTTERING, 443
11.3.1 The Billiard Ball Model of Sputtering
11.3 2 Sputter Yield
11.3.3 Selection Criteria for Process Conditions and Sputter Gas
11.3.4 Secondary Electron Production for Sustaining the Discharge
11.3.5 Sputter Deposited Film Growth
11.3.6 Species that Strike the Wafer During Film Deposition
11.4 RADIO-FREQUENCY (RF) GLOW DISCHARGES, 450
11.5 MAGNETRON SPUTTERING, 456
11.5.1 Magnetron Sputter Sources for ULSI
11.5.1.1 Evolution of Planar Circular Sputtering Sources
11.5.1.2 Deposition Rate and Thickness Uniformity with Circular Planar Magnetrons:
11.6 VLSI AND ULSI SPUTTER DEPOSITION EQUIPMENT, 461
11.6.1 The Components of a Generic Sputtering System
11.6.1.1 Sputtering Targets
11.6.1.2 Vacuum Pumps for Sputtering Systems
11.6.1.3 Power Supplies for Sputtering Systems
11.6.1.4 The Gas Supply for Sputtering Systems
11.6.2 Commercial Sputtering Systems for 125-mm and 150-mm Wafers
11.6.3 Commercial Sputtering Systems for 200-mm Wafers
11.7 PROCESS CONSIDERATIONS IN SPUTTER DEPOSITION, 468
11.7.1 Sputter Deposition of Alloy Films
11.7.2 The Effects on the Sputter Process of the Transport of the Vaporized Atoms Between the Target and the Substrate
11.7.3 Wafer Heating During Sputter Deposition
11.7.4 Faceting and Trenching
11.7.5 Particle Generation in Sputtering Processes
11.7.6 Reactive Sputtering
11.8 STEP COVERAGE & VIA/CONTACT HOLE FILLING BY SPUTTERING, 475
11.8.1 Sputter Deposition of Barrier Layer Films into Contact Holes and Vias
11.8.1.1 Sputter Deposition with Collimators
11.8.1.2 Long-Throw Collimated sputtering
11.8.1.3 Ionized Magnetron Sputter Deposition
11.9 FUTURE TRENDS IN SPUTTER DEPOSITION PROCESSES, 483
11.10 METAL FILM THICKNESS MEASUREMENT, 483
REFERENCES, 485
PROBLEMS, 487
12. LITHOGRAPHY I: OPTICAL PHOTORESIST and PROCESS TECHNOLOGY 488
12.1 BASIC PHOTORESIST TERMINOLOGY, 488
12.2 PHOTORESIST MATERIAL PARAMETERS, 490
12.2.1 Resolution

12.2.1.1 Resolution - Contrast
12.2.1.2 Resolution - Swelling, Proximity Effects, and Resist Thickness
12.2.2 Sensitivity
12.2.3 Etch Resistance and Thermal Stability
12.2.4 Adhesion
12.2.5 Solids Content and Viscosity
12.2.6 Particulates and Metals Content
12.2.7 Flash Point and TLV Rating
12.2.8 Process Latitude and Consistency
12.2.9 Shelf-Life
12.3 OPTICAL PHOTORESIST MATERIAL TYPES, 500
12.3.1 Positive Optical Photoresists
12.3.2 Negative Optical Photoresists
12.3.3 Chemically-Amplified Deep-UV Resists
12.3.4 Multilayer Resist Processes
12.3.4.1 Si-CARL Process
12.3.5 Contrast Enhancement Layers
12.3.6 Silylation-Based Processes for Surface Imaging
12.3.6.1 DESIRE
12.3.6.2 PRIME
12.3.7 The Predicted Role of Multilayer and Surface Imaging Technologies
12.4 PHOTORESIST PROCESSING, 510
12.4.1 Resist Processing: Dehydration Baking and Priming
12.4.2 Resist Processing: Spin Coating
12.4.3 Resist Processing: Soft-Bake
12.4.4 Resist Processing: Exposure
12.4.4.1 Standing Waves
12.4.4.2 Linewidth Variation as Resist Crosses Steps
12.4.4.3 Swing Curves and CD Variation with Resist Thickness
12.4.2.4 Reflective Notching
12.4.2.4 Dyed Photoresists
12.4.2.6 Anti-Reflective Coatings (ARCs)
12.4.2.7 Bottom Anti-Reflective Coatings (BARCs)
12.4.2.8 Top Anti-Reflective Coatings (TARs)
12.4.5 Resist Processing: Post-Exposure Bake
12.4.6 Resist Processing: Development
12.4.7 Resist Processing: After-Develop Inspection
12.4.7.1 Linewidth Variation and Control
12.4.7.2 Linewidth Measurements
12.4.8 Resist Processing: Post-Development-Bake
12.4.9 Resist Processing: Photostabilization of Resists
12.5 PHOTORESIST PROCESSING SYSTEMS, 538
REFERENCES, 541
PROBLEMS, 544

13. LITHOGRAPHY II: OPTICAL ALIGNERS and PHOTOMASKS 545

13.1 THE HISTORY (AND FUTURE) OF MICROLITHOGRAPHY, 546
13.2 BASICS OF OPTICAL SCIENCE FOR MICROLITHOGRAPHY, 548
13.2.1 Basic Terminology of Plane Waves of Light
13.2.2 Diffraction, Numerical Aperture, and Resolution
13.2.2.1 Resolution of the Optical System
13.2.2.2 Resolution - The Rayleigh Criterion
13.2.2.3 Resolution - The Optical Grating
13.2.2.4 Resolution - Fourier Optics Perspective
13.2.2.5 Coherence in Optical Systems
13.2.2.6 Resolution - Modulation Transfer Function
13.2.2.7 Resolution – Impact of the Depth of Focus:
13.2.2.8 A General Resolution Criterion - The Focus-Exposure Process Window:
13.2.3 Resolution Enhancement Techniques Involving the Stepper Optical System
13.2.3.1 Off-Axis Illumination:
13.2.3.2 Multiple Exposures Through Focus (FLEX)
13.3 PATTERN REGISTRATION, 582
13.3.1 Definition of Alignment and Overlay
13.3.2 Interfield and Intrafield Overlay Errors
13.3.3 Interfield Errors
13.3.4 Intrafield Errors
13.3.5 Overlay Metrology
13.4 OPTICAL LITHOGRAPHY EXPOSURE SYSTEMS, 588
13.4.1 Light Sources and Illumination Systems for Optical Lithography
13.4.1.1 Mercury Arc Lamps
13.4.1.2 The Arc-Lamp Illumination System
13.4.1.3 Excimer Laser DUV light Sources
13.5 OPTICAL PROJECTION SYSTEMS, 595
13.5.1 1:1 Scanning Projection Aligners
13.5.2 Reduction Step-and-Repeat Projection Aligners (Reduction Steppers)
13.5.3 Non-Reduction Step-and-Repeat Projection Aligners (1X Steppers
13.5.4 Step-and-Scan Projection Systems
13.5.5 Stepper Wafer Handling System
13.5.6 Temperature, Vibration, and Environmental Control of Steppers
13.6 ALIGNMENT SYSTEMS IN STEPPERS, 605
13.6.1 Off-Axis Alignment Systems
13.6.2 Through-the-Lens Alignment Systems
13.6.3 Alignment Marks and Their Detection
13.6.4 Alignment Strategies
13.7 MECHANICAL ASPECTS OF STEPPER WAFER STAGES, 610
13.7.1 Wafer Stage Positioning and Wafer Chuck Design
13.7.2 Automatic Focussing Systems in Steppers
13.7.3 Automatic Leveling Systems

13.8 MASK AND RETICLE FABRICATION, 615
13.8.1 Terminology and History of Photomasks
13.8.2 Fabrication of Photomasks and Reticles
13.8.2.1 Glass Quality and Preparation
13.8.2.2 Glass Coating (Chrome)
13.8.2.3 Mask Imaging (Resist Application and Processing)
13.8.2.4 Pattern Generation
13.8.3 Mask and Reticle Defects and Their Detection and Repair
13.8.3.1 Repairing Defects in Masks and Reticles
13.8.4 Pellicles
13.8.4.1 Inspecting Masks and Reticles with Pellicles Attached
13.8.5 Critical Dimension and Registration Inspection of Masks and Reticles
13.8.6 Storage, Transport, and Loading of Reticles into the Stepper
13.8.7 Optical Proximity Correction (OPC)
13.8.8 Phase Shift Masks (PSM)
13.9 MICROLITHOGRAPHY TRENDS, 635
13.9.1 The Limits of Optical Lithography
13.9.2 Non-Optical Microlithographic Technologies
13.9.2.1 Electron Beam Direct-Write Lithography
13.9.2.2 Electron Beam Projection Lithography (SCALPEL)
13.9.2.3 Extreme Ultra-Violet Reflective Projection Lithography (EUV)
13.9.2.4 Proximity X-Ray Lithography
13.9.2.5 Ion-Beam Projection Lithography
REFERENCES, 650
PROBLEMS, 654

14. DRY ETCHING FOR VLSI 655

14.1 THE TERMINOLOGY OF ETCHING, 656
14.1.1 Bias, Tolerance, Etch Rate, and Anisotropy
14.1.2 Selectivity, Over-Etch, and Feature Size Control
14.1.3 Determining the Required Selectivity with Respect to Mask Materials, S_{fm}
14.1.4 Determining Required Selectivity With Respect to Substrate, S_{fs}
14.1.5 Combined Impact of the Requirements of Anisotropy and Selectivity
14.1.6 Loading Effects and Microloading
14.2 TYPES OF DRY-ETCHING PROCESSES, 666
14.3 BASIC PHYSICS AND CHEMISTRY OF PLASMA ETCHING, 667
14.3.1 The Reactive-Gas Glow Discharge
14.3.2 Electrical Aspects of Glow Discharges
14.3.3 Heterogeneous (Surface) Reaction Considerations
14.3.4 Parameter Control in Plasma Processes
14.4 ETCHING SILICON & SILICON DIOXIDE IN FLUOROCARBON PLASMAS, 673
14.4.1 The Fluorine-to-Carbon-Ratio Model
14.5 ANISOTROPIC ETCHING AND CONTROL OF EDGE PROFILE, 678

14.6 DRY-ETCHING VARIOUS TYPES OF MATERIALS IN ULSI APPLICATIONS, 681
14.6.1 Silicon Dioxide (SiO_2)
14.6.1.1 Shaping the Sidewlls of Contact Holes and Vias by Dry-Etching
14.6.1.2 Etching 0.25 μm (and Smaller) Contact Holes and Aspect-Ratio Dependent Etching Effects (ARDE)
14.6.1.3 Via Veil Removal After Via Etching
14.6.2 Silicon Nitride
14.6.3 Polysilicon
14.6.4 Refractory Metal Silicides and Polycides
14.6.5 Aluminum and Aluminum Alloys
14.6.6 Organic Films
14.7 PROCESS MONITORING AND END POINT DETECTION, 696
14.7.1 Laser Interferometry and Laser Reflectance
14.7.2 Optical Emission Spectroscopy
14.8 DRY-ETCH EQUIPMENT CONFIGURATIONS, 698
14.8.1 Batch Commercial Dry-Etch System Configurations
14.8.1.1 Barrel Etchers
14.8.1.2 Parallel Electrode (Planar) Reactors
14.8.1.3 Cylindrical Batch Etch Reactors (Hexode Etchers)
14.8.2 Single-Wafer Etchers
14.8.2.1 Single-Wafer Parallel Plate Reactors
14.8.2.2 Magnetic-Enhanced Reactive Ion Etchers (MREIE)
14.8.2.3 Downstream Etchers
14.8.3 High-Density Plasma Sources
14.8.3.1 ECR Plasma Sources
14.8.3.2 Helicon Plasma Sources
14.8.3.3 Inductive (or Transformer) Coupled Plasma Sources
14.8.3.4 Helical-Resonator Plasma Sources
14.8.3.5 High Density Plasma Sources in ULSI Fabrication
14.8.3.6 Electrostatic Chucks
14.9 MISCELLANEOUS PROCESSING ISSUES RELATED TO DRY-ETCHING, 711
14.9.1 Plasma Etching Safety Considerations
14.9.2 Contamination Arising from Dry-Etch Processes
14.9.3 Damage Arising from Dry-Etch Processes
14.9.3.1 Oxide Damage During Polysilicon or Metal Etch Processes
REFERENCES, 715
PROBLEMS, 718
15. MULTILEVEL INTERCONNECTS for ULSI 719
15.1 THE NEED FOR MULTILEVEL-INTERCONNECT TECHNOLOGY, 719
15.1.1 Interconnect Limitations of ULSI
15.1.1.1 Functional Density
15.1.1.2 Propagation Delay

15.1.2 Problems Associated with Multilevel-Interconnect Processes
15.1.3 Terminology of Multilevel-Interconnect Structures
15.2 MATERIALS FOR MULTILEVEL INTERCONNECT TECHNOLOGIES, 724
15.2.1 Conductor Materials for Multilevel Interconnects
15.2.1.1 Requirements of VLSI Conductor Materials
15.2.2 Dielectric Materials for Multilevel Interconnects
15.2.2.1 Requirements of Dielectric Layers in Multilevel Interconnects
15.3 PLANARIZATION OF INTERLEVEL DIELECTRIC LAYERS, 727
15.3.1 Terminology of Planarization in Multilevel Interconnects
15.3.1.1 Degree of Planarization
15.3.1.2 The Need for Dielectric Planarization
15.3.1.3 The Price of Increasing the Degree of Dielectric Planarization
15.3.1.4 Design Rules Related to Intermetal Dielectric-Formation and Planarization
15.3.2 Step-Height Reduction of Underlying Topography
15.3.2.1 Provide a Semi-Planar Surface over Local-Interconnect Levels
15.3.2.2 CVD SiO_2 and Bias-Sputter Etchback
15.3.3 Planarization through Sacrificial-Layer Etchback
15.3.3.1 Sacrificial-Etchback Process Problems
15.4 DOUBLE-LEVEL-METAL (DLM) INTERCONNECTS FOR 1-μm CMOS, 739
15.5 CHEMICAL MECHANICAL POLISHING (CMP), 741
15.5.1 The History of CMP
15.5.2 Modeling the Mechanisms of CMP
15.5.2.1 Metal CMP Mechanisms
15.5.2.2 Silicon Dioxide CMP Mechanisms
15.5.2.3 CMP of Low-*k* Dielectrics
15.5.3 CMP Equipment
15.5.3.1 CMP Polishing Tools
15.5.3.2 CMP Consumables (Polishing Pads)
15.5.3.3 CMP Consumables (Slurries)
15.5.3.4 Slurry Distribution Systems
15.5.3.5 Endpoint Detection
15.5.3.6 Cleaning Issues in CMP
15.5.3.7 CMP Metrology
15.5.3.8 CMP Polisher Tool Reliability
15.5.3.9 CMP Systems and Process Integration
15.5.4 Miscellaneous Problems of CMP
15.5.4.1 Dishing
15.5.4.2 Thickness Non-Uniformity Within a Wafer After CMP
15.5.4.3 Economic Considerations: Throughput and Cost of Ownership
15.6 METAL DEPOSITION AND VIA FILLING, 770
15.6.1 Conventional Approach to Contact and Via Fabrication
15.6.2 Advanced Via Processing (Vertical Vias and Complete Filling of Vias by Metal)
15.6.3 Processing Techniques that Allow for Vertical Vias
15.6.3.1 CVD-W Plugs
15.6.3.2 Filling Vias with Al Deposited by Sputtering
15.6.3.3 Filling Vias by Electroplating of Cu
15.7 TRIPLE-LEVEL-METAL (TLM) INTERCONNECTS FOR 0.5-μm CMOS, 779

15.8 COPPER FOR ULSI INTERCONNECTS, 779
15.8.1 Process Integration Issues of Cu
15.8.2 CVD of Copper (Cu)
15.8.3 Electroplating and Electroless-Plating of Copper
15.9 SPIN-ON GLASS (SOG), 785
15.9.1 SOG Process Integration:
15.9.2 The Etchback SOG Process
15.9.3 The Non-Etchback SOG Process
15.10 LOW-k DIELECTRICS, 791
15.10.1 Process Integration Issues of Low-k Dielectrics
15.10.2 First Generation Low-k Dielectrics ($2.8 < k < 3.5$)
15.10.3 Second-Generation Low-k Dielectrics ($2.5 < k < 2.8$)
15.10.3.1 2nd-Gen Spin-On Dielectrics with $2.5 < k < 2.8$
15.10.3.2 2nd Gen-CVD Dielectrics with $2.5 < k < 2.8$
15.10.4 Ultra-Low-k Dielectrics ($k < 2.0$)
15.11 HIGH-DENSITY-PLASMA CVD (HDP-CVD) OF DIELECTRIC FILMS, 795
15.12 DAMASCENE AND DUAL-DAMASCENE INTERCONNECT STRUCTURES, 797
15.13 DAMASCENE INTERCONNECT STRUCTURE FOR 0.25 μm CMOS, 799
REFERENCES 801
PROBLEMS 806

16. CMOS PROCESS INTEGRATION **807**

16.1 INTRODUCTION TO CMOS TECHNOLOGY, 807
16.1.1 Historical Evolution of CMOS
16.1.2 The Operation of CMOS Inverters
16.2 PROCESS SEQUENCES FOR CMOS:1.2 μm–0.5 μm GENERATIONS, 816
16.2.1 Starting Material
16.2.2 Formation of Wells and Channel Stops:
16.2.3 Definition of Active and Field Regions
16.2.4 Threshold-Adjust Implantation Step
16.2.5 Gate Oxide Growth
16.2.6 Polysilicon Deposition and Patterning
16.2.7 Formation of Source/Drain Regions
16.2.8 Pre-metal Oxide Deposition and Contact Formation
16.2.9 Metal Deposition and Patterning
16.2.10 Intermetal Dielectric Deposition and Via Patterning
16.2.11 Metal 2 Deposition and Patterning
16.2.12 Passivation Layer and Pad Mask
16.3 PROCESS FLOW FOR 0.25 μm CMOS, 828
16.3.1 Starting Material
16.3.2 Formation of the Shallow-Trench-Isolation Structure
16.3.3 Formation of Retrograde Wells and Carrying Out of V_T-Adjust Implants
16.3.4 Formation of the Gate Oxide
16.3.5 Formation of the Gate-Stack Structures

16.3.6 Formation of the Source/Drain Junctions
16.3.7 Salicide Formation
16.3.8 Pre-metal Dielectric Deposition and Contact Formation
16.3.9 Interconnect Structure Formation
REFERENCES 838
PROBLEMS 839

17. ASSEMBLY AND PACKAGING FOR ULSI **841**

17.1 WAFER SORT, 841
17.2 WAFER BACKSIDE PREPARATION, 842
17.3 DIE SEPARATION, 842
17.4 DIE ATTACH, 845
17.4.1 Die-Attach-Related Reliability Issues
17.4.2 Die-Attach Materials
17.4.2.1 Hard Solders That Form Eutectic Die Bonds
17.4.2.2 Silver-Filled Specialty Glasses for Die Attach
17.4.2.3 Epoxy Die-Attach Adhesives
17.4.2.4 Polyimide Adhesives for Die Attach
17.4.3 Die-Attach Equipment
17.5 BOND-PAD TO PACKAGE CONNECTIONS, 851
17.5.1 Wire Bonding
17.5.2 Wire Bond Spacing
17.5.3 Tape Automated Bonding
17.5.4 Flip-Chip Bonding
17.6 INTRODUCTION TO CHIP PACKAGES, 860
17.7 PACKAGING TECHNOLOGY I: HERMETIC PACKAGES (CERAMIC & CERDIP), 861
17.7.1 Ceramic Packages
17.7.2 Glass-Sealed Refractory Packages
17.8 PACKAGING TECHNOLOGY II: PLASTIC PACKAGES, 863
17.8.1 Plastic Package Reliability Problems
17.9 ULSI PACKAGE TYPES, 867
17.9.1 Through Hole (TH) IC Packages
17.9.2 Surface Mount (SM) IC Packages
17.10 TRENDS IN ULSI IC PACKAGES, 869
17.10.1 Packages for Memories and Microprocessors
17.10.2 Packageless Technologies
REFERENCES 872
PROBLEMS 873

APPENDICES **875**

A. PHYSICAL CONSTANTS, 875
B. MATERIAL PROPERTIES OF SILICON, 876
C. ARRHENIUS RELATIONSHIPS, 877

INDEX **879**

Chapter 1

SILICON: SINGLE CRYSTAL GROWTH and WAFER PREPARATION

Silicon (Si) is presently the most important semiconductor for the electronics industry, with VLSI and ULSI technology being based almost entirely on silicon. The name *silicon* comes from the Latin *silex* or *silicis* meaning flint. Second in abundance only to oxygen, silicon comprises 25.7% of the earth's crust. The parameter values of some of the most useful properties of silicon[1] are given in Appendix 2.

Bardeen, Brattain, and Shockley launched solid-state electronics with the invention of the bipolar transistor in 1947.[2] Germanium (Ge) was the original semiconductor material used to fabricate diodes and transistors. The narrow bandgap of Ge (0.66 eV), however, causes reverse-biased *pn* junctions in Ge to exhibit relatively large leakage currents. This limits Ge device operation to temperatures below about 100°C. In addition, integrated circuit planar processing requires the capability of fabricating a passivation layer on the semiconductor surface. Germanium oxide (GeO_2) could act as such a layer but it is difficult to form, is water soluble, and dissociates at 800°C. These limitations make Ge an inferior material for the fabrication of integrated circuits, when compared to silicon. The larger bandgap of silicon (1.1 eV) results in smaller leakage currents and thereby allows silicon devices to be built with maximum operating temperature of about 150°C. The oxide of silicon, SiO_2, is also easy to form and is chemically very stable. Finally electronic grade silicon is about one tenth as costly as germanium. As a result, Si has almost completely replaced Ge for fabricating microelectronic components.

In this chapter we consider the subject of *single-crystal silicon starting material.* We first show how raw material is processed to obtain electronic grade polysilicon. Then we examine the Czochralski and float-zone methods of growing single-crystal silicon from such polysilicon. Next, the technology for forming *silicon wafers* from single-crystal ingots is covered. Finally, a discussion of the material, electrical, and structural characteristics that such silicon wafers must possess in order to be suitable for VLSI (and ULSI) fabrication is presented. In Chap. 2, we cover the crystalline defects that occur in silicon, and techniques for suppressing their formation.

1.1 TERMINOLOGY OF CRYSTAL STRUCTURE

Solid matter exists in crystalline and amorphous forms. In crystalline solids, the atoms which make up the solid are spatially arranged in a periodic fashion. If the periodic arrangement exists throughout the entire solid, the substance is defined as being formed of a *single crystal.* If the solid is composed of a myriad of small single-crystal regions the solid is referred to as *polycrystalline* material. Nevertheless, within any single-crystal region a crystal appears exactly the same at one point as it does at a series of equivalent points. *Amorphous* materials are devoid of long-range periodic structure. The starting silicon material (wafers) on which VLSI and ULSI circuits are fabricated, have single-crystal form.

Exhibit 1014 - Page 23 of 218

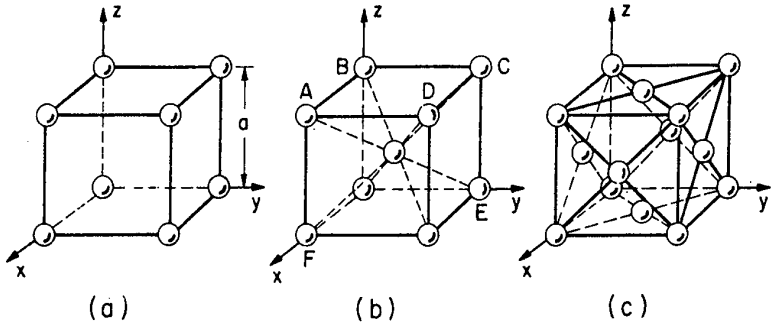


Fig. 1-1 Cubic-crystal unit cells. (a) Simple cubic. (b) Body-centered cubic. (c) Face-centered cubic.

The periodic arrangement of atoms in a crystal is called the *lattice*. The crystal lattice always contains a volume which is representative of the entire lattice (referred to as a *unit cell*), and it is regularly repeated throughout the crystal. As shown in Fig. 1-1, for each unit cell, vectors can be defined (*basis vectors*) such that if the unit cell is translated by integral multiples of these vectors, a new unit cell identical to the original is found. The importance of the unit cell is that the crystal as a whole can be studied by analyzing a representative volume. For example, distances of nearest neighbor atoms can be calculated, as can the periodic properties of the crystal, which define the allowed energy levels.

The simplest lattices are the three cubic lattices seen in Fig. 1-1 (simple cubic [sc], body-centered cubic [bcc], and face-centered cubic [fcc]). The dimension **a** for a cubic cell is called the *lattice constant,* but the distance of nearest neighbor atoms may be smaller than **a**. For example, in the fcc lattice the nearest neighbor distance is one half the diagonal of a cube face, or $(\mathbf{a}\sqrt{2})/2$.

The directions in a lattice are expressed as a set of three integers with the same relationship as the components of a vector in that direction. The three vector components are given in multiples of the basis vectors. For example in cubic lattices, the body diagonal has the components of 1a, 1b, and 1c. Therefore, this diagonal exists along the [111] direction. (The [ ] brackets are used to denote a specific direction.) From a crystallographic point of view, however, many directions in a crystal are equivalent, depending only on the arbitrary choice of orientation for the axes. Such equivalent directions are expressed with angular brackets < >. For example the crystal directions in the cubic lattice [100], [010], and [001] are all crystallographically equivalent (Fig. 1-1), and are referred to as <100> directions.

It is also useful to describe *planes* in a crystal. A set of three integers *h*, *k*, and *l*, called the Miller *indices*,[3] are used to define a set of parallel planes. They are derived as follows (Fig. 1-2):

1. The intercepts of the plane with the crystal axes are found, and these intercepts are expressed as integral multiples of the basis vectors.

2. The reciprocal of the three integers found in step a) are taken.

3. The smallest set of integers *h*, *k*, and *l* are found that have the same relationship to each other as the three reciprocals (i.e., if 1/4, 1/3, and 1/2 are the three reciprocals, then 3, 4, and 6 are the three smallest integers whose relative values are the same as the relative values of the three reciprocals). These integers are the *Miller indices* of the plane, and the plane is labeled (*h k l*). Note that ( ) parentheses are used to denote a specific plane.

One advantage of taking reciprocals of intercepts is that infinities are avoided in the notation that identifies a plane. The reciprocal of the intercept of a plane that is parallel to an axis (and

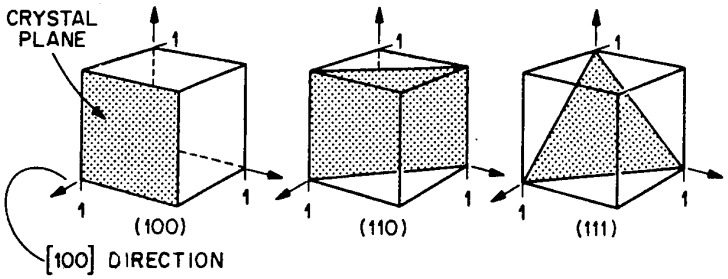


Fig. 1-2 Miller indices of some important planes in a cubic crystal.

thus would intercept the axis at an infinite distance) is zero. From a crystallographic point of view, as in the case of directions, many planes in a lattice are equivalent. The indices of such equivalent planes are collectively indicated by one set of integers, placed within { } parentheses.

It should also be noted that for *cubic lattices*, direction [*h k l*] is perpendicular to a plane with the identical three integers (*h k l*). This fact makes it convenient to analyze cubic cells. That is, if either a direction or a plane is known, its perpendicular counterpart can be quickly determined without further calculation.

Silicon has a *diamond cubic lattice*, a structure which can be represented as two interpenetrating face-centered cubic lattices (Fig. 1-3a). Thus, the simplicity of analyzing and visualizing

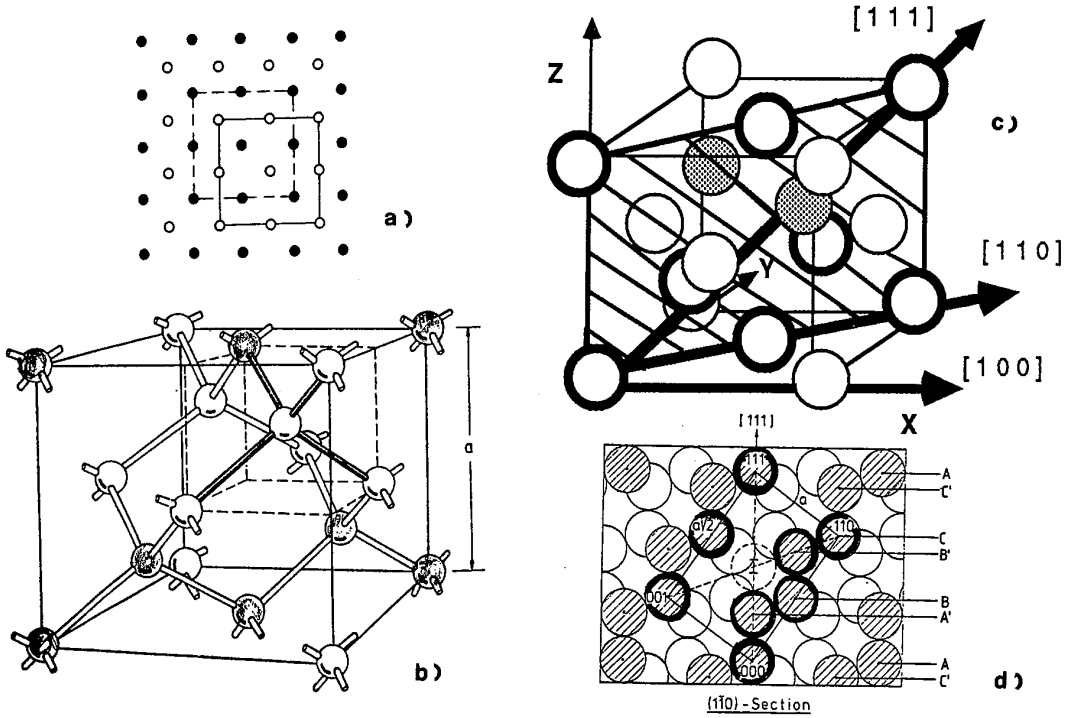


Fig. 1-3 (a) Top view (along any <100> direction) of an extended diamond lattice cell. The open circles indicate one fcc sub-lattice and the solid circles indicate the interpenetrating FCC. (b) Two schematic representations of the diamond lattice unit cell. (c) A (110) section through the diamond lattice unit cell.[8] (d) (110) section showing atoms in the unit cell of (c) and neighboring atoms in the same (110) plane. Reprinted with permission of Springer-Verlag.

cubic lattices can be extended to characterization of silicon crystals. The three dimensional representation of a diamond-lattice unit cell is shown in Fig. 1-3b. In studying the diamond-structure unit cell, it is evident that each atom has four nearest neighbors. The importance of this fact becomes evident when the electronic structure of silicon is considered. That is, since Si is a column IV element and has four valence electrons, each of these electrons is shared with one of its four nearest neighbor Si atoms. (This nearest neighbor atom likewise contributes one of its electrons to the bond.) The two valence electrons shared between nearest neighbor's form a covalent bond and such interactions of electrons in neighboring atoms of a solid serve the function of holding the crystal together. Since all the valence electrons in a perfect Si crystal are involved in shared bonding pairs, none are available as free electrons for conducting electrical current (as long as they remain localized in the bonding region). Such covalently bonded electrons can, however, be thermally or optically excited, and thereby become free to participate in conduction. (Note that in Si, $\mathbf{a} = 5.43$ Å, and the nearest neighbor distance is 2.43 Å).

On the other hand, when dopant atoms (mostly from Group III or V) are substituted for Si atoms in the crystal lattice, they are said to occupy *substitutional* lattice sites. Such dopants can also become the source of additional conduction electrons or holes. For example, phosphorus is a substitutional *donor*, because four of its valence electrons become covalently bonded with other Si atoms. However, the fifth valence electron is easily separated, and once free, becomes available to conduct current. Group III elements (e.g., boron, aluminum) are substitutional *acceptors*, and become sites leading to *hole* conduction. Atoms that do not occupy lattice sites are described as occupying *interstitial* (e.g. between lattice sites) sites.

It is useful to consider the diamond lattice in further detail. For example, instead of describing it as two interpenetrating face-centered cubic lattices, it is crystallographically more precise to consider it as a *face-centered cubic lattice with two basis atoms at each lattice point.* Thus, to construct a diamond lattice unit cell, one would erect a face-centered cube and place two atoms at each lattice site (one directly on the site, and the other offset from the first at a distance $\mathbf{a}/4$, along all axes). The result would be the unit cell shown in Fig. 1-3b. It should be noted that the specific structure of crystal silicon arises from having to simultaneous-ly satisfy two different conditions: 1) that each atom in the crystal has 4 nearest neighbors; and 2) that the atoms also be as closely packed in space as condition 1) will allow. The packing density of the *diamond lattice* is much lower than that of the *face-centered cubic lattice* (34% versus 74%).

In the study of several processing technologies, reference to various planes and directions in silicon crystals will be made, especially to the (100), (110), and (111) planes and directions. Direct examination of the diamond lattice unit cell does not allow easy conceptualization of the {110} and {111} planes and directions. Thus, we shall pursue a simple procedure to assist the reader in visualizing them. In order to determine the {111} planes most easily, we will make use of the concept that in *cubic lattices* a crystal direction [*h k l*] is perpendicular to the plane (*h k l*).

To begin, we draw a section of a {110} plane of the diamond lattice unit cell as in Fig. 1-3c. Note the $\mathbf{a}\sqrt{2}$ dimension is the length of a cube face diagonal [e.g., along the (110) direction]. All atoms shown in Fig. 1-3c exist along the {110}-plane thus selected. Next, we observe that a [111] direction exists along the diagonal of the {110} plane (Fig. 1-3c). Finally planes perpendicular to <111> directions can be identified as {111} planes.

In Fig. 1-3d, we add the additional neighboring atoms in the crystal that are in the {110} plane, but are outside the unit cell (shaded circles). The other atoms in this figure (unshaded circles) lie in {110} planes below the {110 surface plane which we have cut as our section. From this diagram it can be seen that there are actually three double {111} planes in the

diamond lattice (A, A'; B, B'; C, C') arranged in a regularly repeating *stacking order*. (If one were to examine a set of hard spheres arranged in a close packed fcc structure, the third plane of three stacked {111}-planes would lie such that the sphere centers would not be directly above the centers of the spheres in either of the two planes below, and thus this structure has an ...ABCABCABC... stacking order. On the other hand, it is also possible to place the third layer in a close packed structure such that the centers of the third layer spheres *do lie* directly over the A (or first) level spheres. This structure has an ..ABABAB...stacking order, but this is not an FCC structure.) In addition to illustrating the stacking order of the {111} planes, the diagrams of Fig. 1-3 graphically depict the open structure of the silicon lattice. They also show the unobstructed channels that exist when traversing the Si crystal along <110> directions (perpendicular to the {110} plane shown in Fig. 1-3d).

Various structural properties of silicon depend on crystal orientation, as do many ULSI fabrication steps. For example the tensile strength of silicon is highest for <111> directions. In addition, {111} planes have the highest density of atoms. Therefore {111} planes oxidize more rapidly than {100} planes, as more atoms per unit surface area are available for the oxidation reaction. MOS devices are fabricated on {100} wafers since the smallest surface state densities are observed on such orientations.

Actual solid materials in the form of single crystals differ to some extent from the ideal crystal structure discussed in this section. To begin with, even ideal crystals possess *surfaces* at their boundaries, and at a surface, the atoms are incompletely bonded. In addition, real crystals have a variety of imperfections (referred to as *crystalline defects*) which can significantly alter the properties of the crystal. A detailed discussion of such defects in silicon and their impact on crystal and device properties is given in Chap. 2.

It is also useful to note the units used to specify dopant and contaminant concentrations in silicon. Most often they are given in atoms/cm^3, but sometimes in *parts per million atoms* (ppma). Since there are about $5x10^{22}$ Si atoms/cm^3 in single-crystal Si, an impurity concentration of $5x10^{16}$/cm^3 equals 1 ppma, and other concentration levels can likewise be converted.

1.2 THE MANUFACTURE OF SINGLE CRYSTAL SILICON

The fabrication of VLSI and ULSI takes place on silicon substrates possessing very high crystalline perfection. The critical importance of utilizing single crystal material for the transistor regions of microelectronic circuits was first recognized by Teal in 1949.[4] He reasoned that polycrystalline material would exhibit inadequately short minority carrier lifetimes, due to the defects occurring at the grain boundaries of the polycrystalline grains. Teal also pointed out the need to define the degree of chemical purity, degree of crystal perfection, uniformity of structure, and chemical composition of single crystal material when specifying starting material on which to fabricate semiconductor products. The method of obtaining such single crystal Si for IC fabrication, involves several steps (Fig. 1-4):

1. Raw material (e.g., quartzite, a type of sand), is refined by a complex, multi-stage process which produces *electronic grade polysilicon* EGS,

2. This polysilicon is used to grow single crystal silicon by *Czochralski (CZ) crystal growth* or *float zone (FZ)* growth. Single crystal silicon is commercially available in either {100} or {111}-orientations.

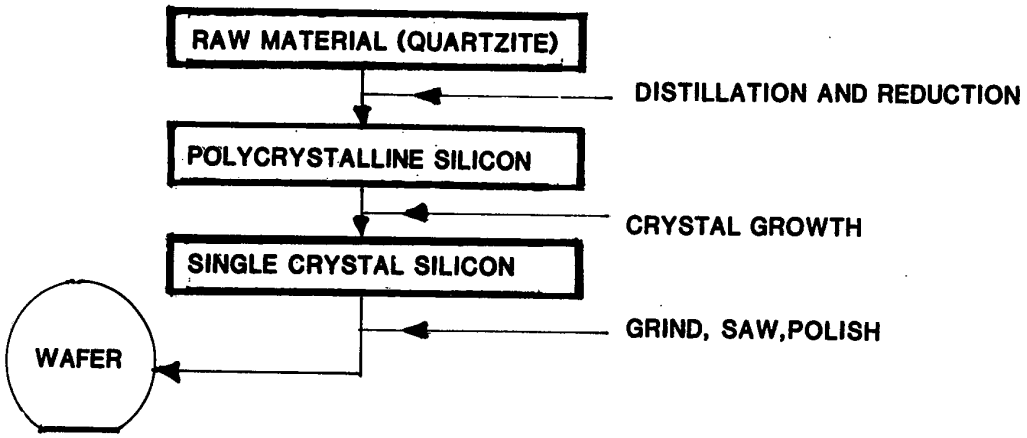


Fig. 1-4 Process sequence from starting material to polished wafer.

In CZ growth, single crystal ingots are pulled from molten silicon contained in a crucible. Czochralski silicon is preferred for ULSI applications since it can withstand thermal stresses better than FZ material,[40] and it is able to offer an internal gettering mechanism that can remove unwanted impurities from the device structures on wafer surfaces. Float zone crystals, because they are grown without making contact to any container or crucible, can attain higher purity (and thereby higher resistivity) than CZ silicon. Devices and circuits calling for high purity starting material (e.g., high voltage, high power devices) are sometimes fabricated from FZ silicon.

1.2.1 From Raw Material to Electronic Grade Polysilicon

Single crystal silicon is grown from melts of *electronic grade polycrystalline silicon* (EGS). Since the CZ process (which is used to grow most single crystal silicon) adds impurities to the resultant ingot, the EGS must be extraordinarily pure. In fact, in order to achieve controlled doping during subsequent single crystal growth, EGS must have impurity levels in the parts per billion atoms (ppba) range (or $10^{13}/cm^3$), making it the purest material routinely available on earth. The raw material from which EGS is refined (quarzite), contains high levels of impurities (e.g., aluminum levels of ~$3x10^{20}/cm^3$). Thus, the refining process must reduce these impurities by approximately *eight orders of magnitude!* Such a radical refinement procedure involves four major stages:[1,5]

1. Reduction of quartzite to metallurgical grade silicon (MGS) with a purity of ~98%;
2. Conversion of MGS to trichlorosilane ($SiHCl_3$);
3. Purification of $SiHCl_3$ by distillation;
4. Chemical vapor deposition (CVD) of Si from the purified $SiHCl_3$, as EGS.

In *Step 1*, SiO_2 in the form of quartzite is reacted with carbon to yield silicon and carbon monoxide:

$$SiO_2\,(solid) + 2\,C\,(solid) \rightarrow Si\,(solid) + 2\,CO\,(gas) \qquad \textbf{(1.1)}$$

Quartzite is a relatively pure form of SiO_2 that occurs in nature. In the presence of carbon (in the form of coal, coke, or wood chips), it is reduced to MGS. The name MGS is derived from the fact that silicon with this purity (~98%) is sufficiently refined to be used as an alloy material

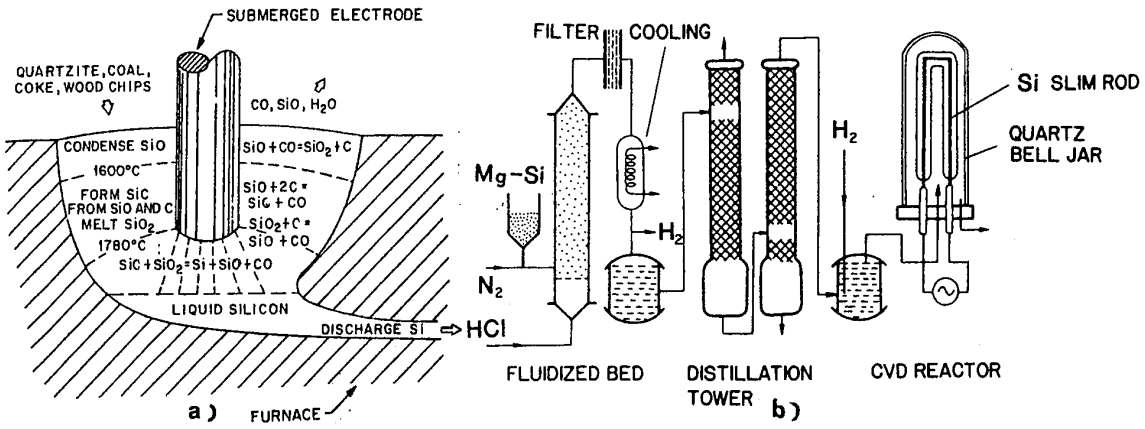


Fig. 1-5 (a) Schematic of submerged-electrode arc furnace for production of MGS. Reprinted with permission of the publisher, the Electrochemical Society. (b) Schematic of fluidized bed, distillation tower, and CVD reactor developed by Siemens. Reprinted with permission of Academic Press.

in the manufacture of aluminum or for producing silicone polymers. Most silicon manufactured every year is consumed by such applications, and only a small fraction is further refined into EGS for electronic applications. The quartzite, coal (or coke), and wood chips (for porosity) are loaded into a submerged electrode arc furnace (Fig. 1-5a). The temperatures within the electric arcs that are generated between the furnace electrodes can exceed 2000°C, and in those regions SiC is formed. The SiC reacts with SiO_2 to form Si, SiO, and CO. The silicon is drawn off as shown in Fig. 1-5a, while the SiO and CO gases escape through the spaces created by the presence of the wood chips. Large quantities of electrical power are consumed in producing the reaction (12–14 kWh/kg). The MGS from this process has Al and Fe as the main impurities.

In *Step 2*, trichlorosilane ($SiHCl_3$) is formed by the reaction of anhydrous hydrogen chloride and MGS. In this step, the MGS is first ground to a fine powder. The powdered Si is then treated with HCl to form $SiHCl_3$. The reaction of the solid Si and the gaseous HCl occurs at 300°C in the presence of a catalyst. Both the formation of the $SiHCl_3$ and the chlorides of the impurities (e.g. $AlCl_3$, BCl_3) takes place in this step. The resultant $SiHCl_3$ is a liquid at room temperatures (boiling point 31.8°C), and hence can be purified by distillation.

In *Step 3*, the $SiHCl_3$ is separated from its impurities by fractional distillation. Upon the conclusion of this step, the $SiHCl_3$ is so highly refined that it is no longer possible to directly determine its impurity levels. That is, the $SiHCl_3$ must be deposited as semiconductor grade polysilicon, then converted to single crystal silicon, and finally measured electrically, in order to determine the impurity levels.

In *Step 4*, the highly purified $SiHCl_3$ is converted back into polycrystalline silicon (EGS) by CVD in the presence of hydrogen. The process takes place in a reactor (Fig. 1-5b) first proposed by Siemens GmbH in the late 1950s, and it is still referred to as the *Siemens process.*

$$2\,SiHCl_3\,(gas) + 2\,H_2\,(gas) \rightarrow 2\,Si\,(solid) + 6\,HCl\,(gas) \qquad \textbf{(1.2)}$$

The starting surface is a thin silicon rod (called a *slim rod*) which serves as a nucleation surface for the depositing silicon. For large rods (200 mm in diameter and several meters long), the deposition process takes several hundred hours.

After deposition, the EGS is processed into three products: 1) one-piece crucible charges; 2) nuggets (random sized pieces); and 3) poly-rods (Fig. 1-6). The first two are used as charges in the CZ growth of single-crystal silicon, while the rods are used for float-zoning single-crystal

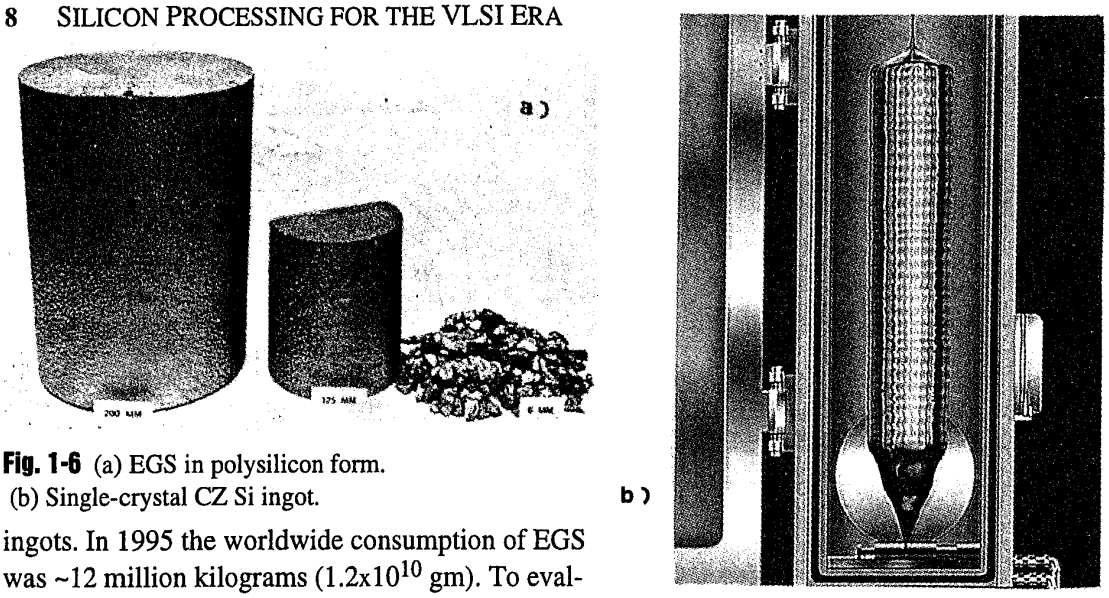


Fig. 1-6 (a) EGS in polysilicon form. (b) Single-crystal CZ Si ingot.

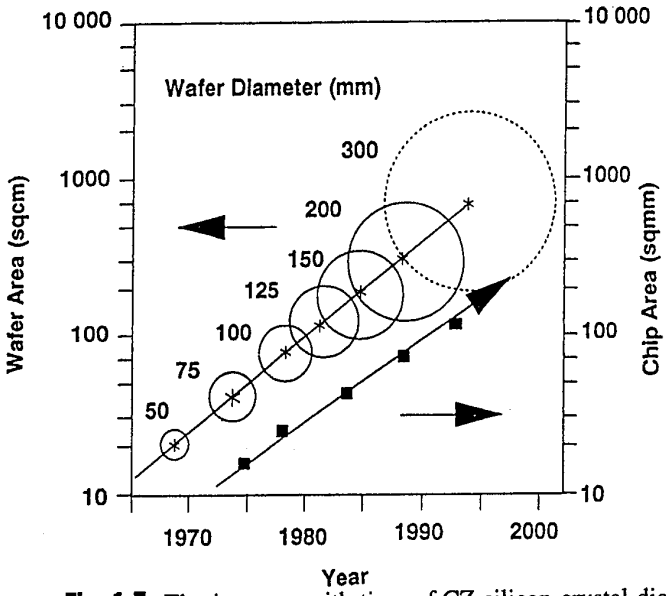
ingots. In 1995 the worldwide consumption of EGS was ~12 million kilograms (1.2×10^{10} gm). To evaluate the purity of EGS (as well as the $SiHCl_3$ as mentioned earlier) conversion to single crystal Si must first be accomplished. This conversion is done by the FZ technique, since FZ contributes negligible levels of impurities to the sample (see Sect. 1.4). The measurement of *n*-type impurities is done after a one-pass float zone. The measurement of residual boron requires multiple passes of the float zone (which sweeps out other impurities but leaves behind the boron, because of its 0.8 segregation coefficient). Thus 90% of the boron is still left in the silicon after several passes. The purity of EGS is on the order of 0.3 ppba, boron. This represents approximately 1.5×10^{13} boron atoms/cm^3, and a resistivity of ~1000 Ω-cm. ASTM Std F-754 describes the evaluation of EGS by float zoning.[6] EGS rods can also be machined to form diffusion tubes and wafer boats. They can withstand several years of temperature cycling between 900 and 1250°C without exhibiting the sagging or deformation of fused silica tubes.

1.3 CZOCHRALSKI (CZ) CRYSTAL GROWTH

Czochralski (CZ) growth, named for its inventor, involves the crystalline solidification of atoms from a liquid phase at an interface. The basic production process for producing CZ Si has undergone remarkably little change since it was pioneered by Teal[4] and others in the early 1950s. (For a more detailed description of CZ growth than is presented here, see Refs. 1, 3, 8, and 9.)

The simplicity of the process and the relatively high degree of crystal purity helped establish the CZ process as the dominant Si crystal growing technology in the early years of the semiconductor industry. The development of dislocation-free ingot growth and automatic diameter control in the late 1960s lead to a rapid growth in the sizes of ingot diameters and crystal melts. Further increases in charge size and diameters were driven by the economic advantages of even larger ingots. As shown in Fig. 1-7, by 1998 300-mm diameter crystals had become available, grown with 150 kg charges. Typical commercial ingots are grown 0.5–2.5 m in length, and a 100 kg charge can grow a 200 mm diameter ingot roughly 1 m long.)

In this section we introduce the most important aspects of CZ growth, focusing on the details that impact the material properties of the silicon wafers for ULSI. These include: a) the sequence of steps in the growth of the ingot; b) the first-order theory of how impurities are incorporated into the ingot (Normal freezing); c) modifications to the normal freezing model that must be added as a result of non-ideal effects encountered in real CZ growth environments (and the impact that this causes on the material properties of CZ wafers); d) the components of CZ crystal growing equipment; and e) the evaluation of CZ silicon crystals in ingot form.



Diameter	Area (Sq. Inches)
2 Inch	3.1
3 Inch	7.1
100mm	12.6
125mm	19.6
150mm	28.3
200mm	50.2
300mm	109.5

Diameter Increase	Percentage Area Increase
2" → 3"	125%
3" → 100 mm	78%
100 mm → 125 mm	56%
125 mm → 150 mm	44%
150 mm → 200 mm	78%
200 mm → 300 mm	125%

Fig. 1-7 The increase with time of CZ silicon crystal diameter and the percentage increase in wafer area with increase in wafer diameter.

1.3.1 Czochralski Crystal Growth Sequence

The steps used in growing a CZ Si crystal are shown in the sequence of photographs of Fig. 1-8a and a schematic diagram of a CZ crystal grower in which this process is performed is given in Fig. 1-12b. A fused silica crucible is first loaded with a charge of undoped EGS together with a precise amount of diluted silicon alloy. The growth chamber is then pumped out, backfilled with an inert gas (to limit incorporation of atmospheric gases into the melt during growth), and the charge is melted (melting point of silicon is 1421°C). Next, a slim seed crystal of silicon (≈5 mm in diameter and 100–300 mm long) with precise orientation is lowered into the molten silicon. The seed crystal is then withdrawn at a controlled rate. Both the seed crystal and the crucible are rotated during the pulling process, but in opposite directions.

The initial pull rate is relatively rapid so that a thin neck (~3 mm in diameter) is produced. As was first discovered by Dash,[10] by forming this thin neck, the growing ingot can achieve a macroscopically dislocation-free crystalline state. At that point the melt temperature is reduced and stabilized so that the desired ingot diameter can be formed (shouldering out step). This diameter is subsequently maintained by precise monitoring of the pull rate. The pulling continues until the charge is nearly exhausted, at which point the ingot is withdrawn to form the *tang* (tail off). A shutdown procedure of the furnace completes the process.

The initial immersion of the cool seed crystal into the melt sets up very high stresses between the heated crystal exterior and the still cool interior. These stresses are high enough to cause plastic deformation and create macroscopic dislocations in the crystal. By rapidly pulling a thin necked region from the melt, these dislocations are driven to the cylindrical surface of the neck, and at that point they cease to propagate into the subsequently solidifying crystal regions (Fig. 1-8b). Thus the growing crystal becomes free of macroscopic dislocations and continues to grow in this way unless there is a significant perturbation to disturb the solidification process (e.g., a particulate in the melt contacts the growing crystal or the crystal is jolted during growth). The tendency for dislocation-free crystals to continue to grow without forming new dislocations is due to the very high stresses that must be imparted to a perfect crystal to create dislocations. Such high stresses are not encountered under well-controlled growth conditions.

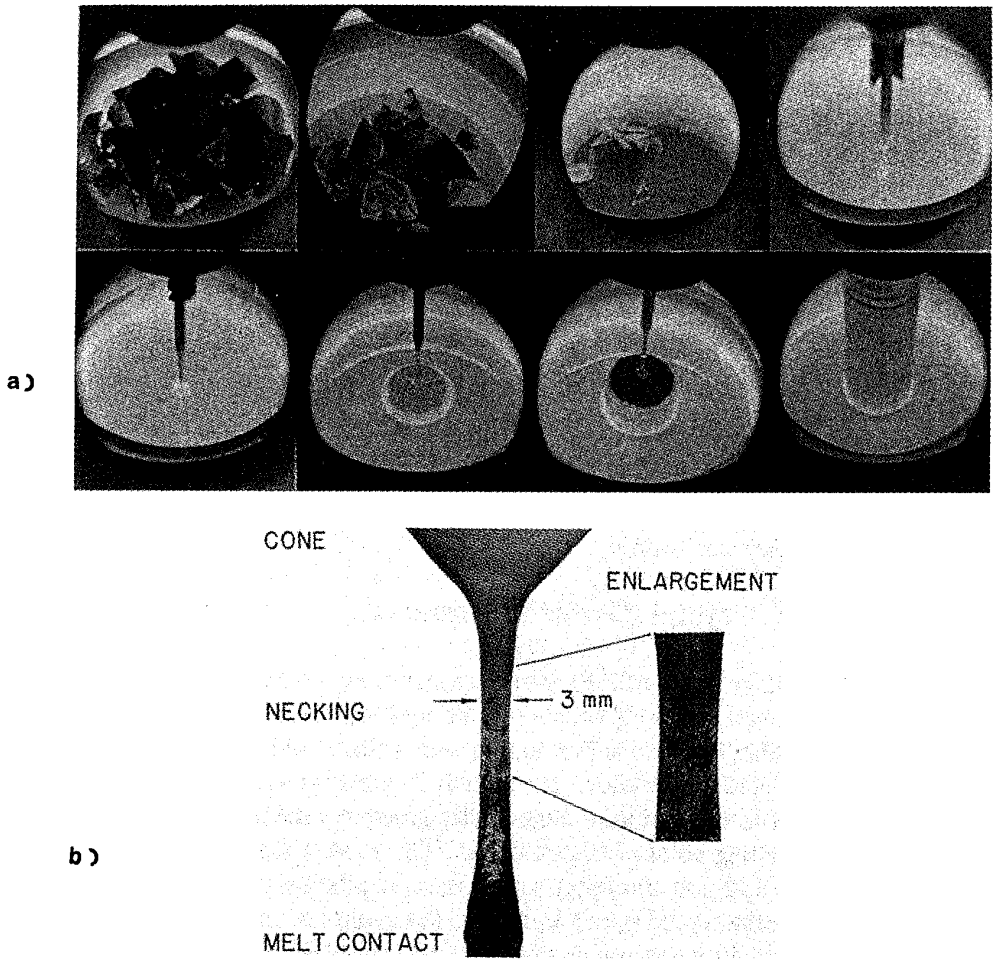


Fig. 1-8 (a) Illustration of several process steps during CZ crystal growth. Courtesy of Dynamit-Noble-Grace. (b) X-ray topograph of seed necking and conical part of crystal. Dislocations generated at the end of the seed crystal that contacted the molten zone grow out to the side surface of the neck and do not propagate into the main crystal.[9] Reprinted with permission of Academic Press.

The references give more details on the Dash process for growing dislocation-free crystals.[8,10]

As the crystal solidifies and is pulled from the melt, the latent heat of fusion L, is transferred to the crystal. This heat is transported from the solid-liquid interface along the growing crystal. Heat is lost from the crystal surface due to radiation and convection. The upper-limit to the pull rate V_{max}, is reached when the maximum heat that can be lost by the ingot is being transferred to the ingot by the solidifying silicon. The value of V_{max} can be calculated from the following relation under the condition of zero thermal gradient in the melt:[11]

$$V_{max} = \frac{k_s}{Ld}\left(\frac{dT}{dx}\right) \tag{1.3}$$

where: d is the density of silicon; k_s is the thermal conductivity of silicon; and (dT/dx) is the temperature gradient along the axis of the ingot.

As the diameter of the ingot is increased, the maximum pull-rate must decrease, because the heat loss is proportional to the surface area of the ingot (which increases only linearly with the

diameter), but the heat gain is proportional to the volume being solidified (which increases as the square of the ingot radius).

The actual pull rate is typically 30% to 50% less than the maximum rate[11] calculated from Eq. 1-3. Such slower rates are required since temperature fluctuations in the melt occur near the interface. (The origin of these fluctuations is discussed in a subsequent section). When such a fluctuation causes a transient increase in the melt temperature, *remelt* of the crystal starts to occur and the nominal pull-rate must be reduced until solidification restarts. When a subsequent temperature reduction takes place, the solidification rate increases, leading to an increase in the ingot diameter. To maintain the desired diameter, the pull-rate at that instant must be increased. Such temperature oscillations thus lead to pull-rate variations and an overall lower average pull-rate. Automatic diameter control (ADC) mechanisms slave the pull-rate changes to optical diameter measurements.

Several ingot properties are impacted by such temperature fluctuations. First, dimensional nonuniformities occur in the diameter of the ingot. Second, these temperature instabilities cause microscopic resistivity variations in both radial and axial directions (e.g., local resistivity variations of ±8% for boron-doped and ±10–20% for phosphorus-doped wafers from this effect are typically observed). This problem and techniques proposed to suppress it are discussed in subsequent sections.

The dislocation-free attribute of CZ crystals refers to macroscopic (edge and screw) dislocations. Small dislocation loops, however, can be formed during the growth of the main crystal by the condensation of excess point defects (Chap. 2). This phenomenon occurs as the crystal cools from the solidification temperature (1421°C) down to about 950°C. During this cooldown, the solubility of the point defects created at temperatures near meltdown is reduced, until the crystal is supersaturated with point defects. If these point defects can diffuse sufficiently, they will tend to arrive at sites at which they can agglomerate and thereby be removed from solution. If the pull-rate is high enough, the solidified region cools with sufficient rapidity that the excess point defects are immobilized (or *quenched*) in the lattice. They thereby have insufficient opportunity to diffuse and agglomerate into clusters large enough to cause dislocation loops. Both interstitial silicon and oxygen have been suggested as the point defect species

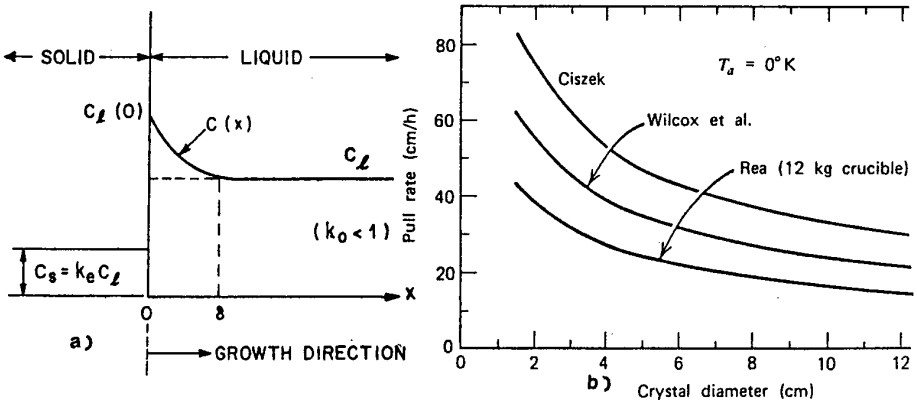


Fig. 1-9 (a) Doping distribution near the melt-solid interface. (b) Crystal pull rate as a function of diameter.[21] From K.V. Ravi, *Impurities in Semiconductor Silicon*, 1981, Copyright © John Wiley and Sons. Reprinted with permission of John Wiley and Sons.

responsible for forming such agglomerates. For crystals in which the pull-rate has not been sufficiently high, the presence of the resultant dislocation loops will cause the crystal to exhibit *swirl* patterns when wafers are subjected to defect-delineation chemical etching. Swirl has also been linked to the growth-induced defects stemming from fluctuations occurring at the liquid/solid interface, but there is no well-accepted theory on their nucleation and growth.[12,13]

When CZ silicon crystals were first grown, the process required highly skilled operators to control all the important growth parameters including melt temperature, ingot diameter, pull-rate, and rotation. Today, such parameters are automatically monitored, resulting in more reproducible control of the physical and chemical aspects of the pulling process. A variety of growth process sequences can be carried out by modern crystal growth systems, with "one button" operations that automatically take the system from meltdown to shutdown.

1.3.2 Incorporation of Impurities into the Crystal (Normal Freezing)

Impurities are incorporated into the CZ crystal as it solidifies, together with the silicon. As pointed out earlier, the impurities are included in the melt by adding precise amounts of silicon alloy (heavily doped with the impurity of choice) into the crucible with the EGS. In the model used to predict how the impurity is incorporated into the crystal (Normal freezing model),[14] three assumptions are made: 1) the impurity is uniformly distributed throughout the melt; 2) once solidification occurs, the impurity does not diffuse in the solidified crystal during the remainder of the ingot growth process; and 3) the impurity arrives at the solid-liquid interface in the same concentration as exists in the melt.

In the course of solidification, however, the following phenomenon is encountered: the solubilities of the impurity in the melt C_l and in the solid C_s at a given temperature in equilibrium are generally not equal. For the dilute solutions that typically occur in silicon, the ratio of these solubilities may be defined as the equilibrium segregation coefficient k_0:

$$k_0 = C_s/C_l \tag{1.4}$$

Let us consider how this phenomenon impacts the impurity concentration in the ingot when k_0 is <1 (since this is the case for most impurities in Si, see Table 1-1). As the silicon solidifies, only a fraction of the impurity concentration that was present in the molten volume that solidified (as specified by k_0), is incorporated into the crystal. The remaining fraction remains in the melt (Fig. 1-9). As the ingot grows the melt volume is consumed, and therefore decreases. However, a fraction of the impurity arriving at the solidifying interface between the melt and solid is continuously rejected by the crystal and stays in the melt. Thereby, the melt becomes progressively more concentrated with the impurity. As a result, as the ingot grows, the molten material arriving at melt-crystal interface contains an ever-larger proportion of impurity atoms. Since the same *fraction* of total impurities present in the arriving materials gets incorporated into the crystal, the ingot also becomes more heavily doped along its length.

Table 1-1 SEGREGATION COEFFICIENTS for COMMON IMPURITIES in SILICON

IMPURITY	Al	As	B	C	Cu	Fe	O	P	Sb
k_0	0.002	0.3	0.8	0.007	4×10^{-4}	8×10^{-6}	1.25	0.35	0.023

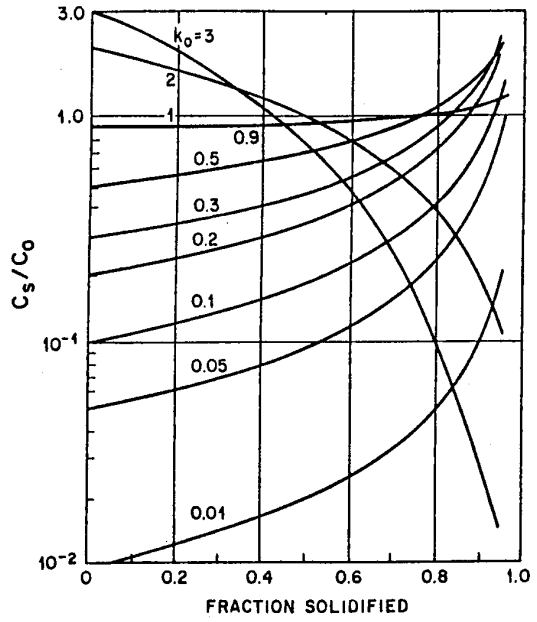


Fig. 1-10 Curves from growth from the melt, showing the doping concentration in a solid as a function of the fraction solidified, with examples of some impurities in silicon.

If a crystal solidifies and incorporates the impurity from the melt in the manner described, a mathematical relation, known as the *Normal freezing relation* can be used to predict the impurity concentration along the length of the pulled ingot:

$$C_s = k_o C_o (1 - X)^{k_o - 1} \qquad \textbf{(1.5)}$$

where: X is the melt fraction solidified; and C_0 is the initial impurity concentration in the melt. Figure 1-10 shows how the predicted impurity concentration in an ingot will vary for various k_0 values (including those of B and P) if Eq. 1-5 is obeyed. Although impurity incorporation does not exactly follow Eq. 1-5, the information in Fig. 1-10 does support the following observations in CZ ingots: 1) when k_0 is close to 1 (e.g., for B, k_0 = 0.8), it is less difficult to achieve relatively level *axial* resistivity values (and, as we shall see, adequate *radial* resistivity uniformity as well); 2) when $k_0 << 1$ (e.g., for P, k_0 = 0.3, or for Sb, k_0 = 0.023), these desired uniformities are more difficult to obtain and control.

Note that starting wafers are available with specific concentrations over the range of 10^{14} to 10^{19} atoms/cm^3 to meet the needs of various device applications. A pure dopant element can be directly added into the Si melt for producing a heavily-doped crystal. However, an impurity-silicon alloy is preferred when a lightly doped crystal is needed. The alloy for such latter applications is added in the form of fine grains, ground from the solidified silicon melt that contains a known concentration of the dopant. The amount of an impurity that can be added into the melt is constrained by its solubility limit. As shown in Fig. 2-2, two *n*-type dopants (arsenic and phosphorus) have solubility limits in Si that exceed 10^{20}/cm^3 at the Si melting temperature, but boron is the only *p*-type dopant whose solubility limit in Si is >1x10^{20} cm^3.

EXAMPLE 1-1: It is desired to grow a CZ silicon ingot containing $1x10^{15}$ boron (B) atoms/cm^3. What concentration of B atoms must be present in the melt to yield this result? For a 60-kg silicon charge, how many grams of B (atomic weight 10.8) are needed?

SOLUTION: Since for B, k_0 is 0:8, the initial concentration required in the melt should be: $1x10^{15}/0.8 = 1.25x10^{15}$ B atoms/cm^3.

Next calculate the volume of the 60-kg silicon charge. The density of the molten charge is 2.53 g/cm^3, so the volume of 60 kg of Si is:

$$\frac{60 \times 10^3 \text{ grams}}{2.53 \text{ grams/cm}^3} = 2.37x10^4 \text{ cm}^3$$

The number of B atoms in the melt is

$$1.25x10^{15} \text{ atoms/cm}^3 \times 2.37x10^4 \text{ cm}^3 = 2.96x10^{19} \text{ B atoms, and}$$

$$\frac{2.96 \times 10^{19} \text{ atoms} \times 10.8 \text{ g/mole}}{6.023 \times 10^{23} \text{ atoms/mole}} = 5.31x10^{-4} \text{ grams of B.}$$

Figure 1-11 shows how impurity segregation during ingot growth impacts the fraction of the ingot length that can meet a given resistivity specification. In this example, wafers must meet a 8–12 Ω-cm resistivity range specification (~0.8–1.2x10^{15}/cm^3 impurity concentration). We see that about 80% of the ingot length can provide such wafers if it is B-doped, but only 50% of the length can be used if the ingot is P or As doped. If we also include the loss of ingot material due to removal of grinding damage the crystal material yields drop to 66% and 40% respectively.

1.3.3 Modifications to Normal Freezing in CZ Growth

In actual CZ growth of silicon, some of the assumptions on which Normal freezing behavior is based are not entirely correct. The most important deviation is that the impurities are not homogeneously distributed throughout the melt at all times. In fact, at the solid-melt interface, a *boundary* layer exists. In this layer, the material is largely unmoved by fluid flow (i.e., it does

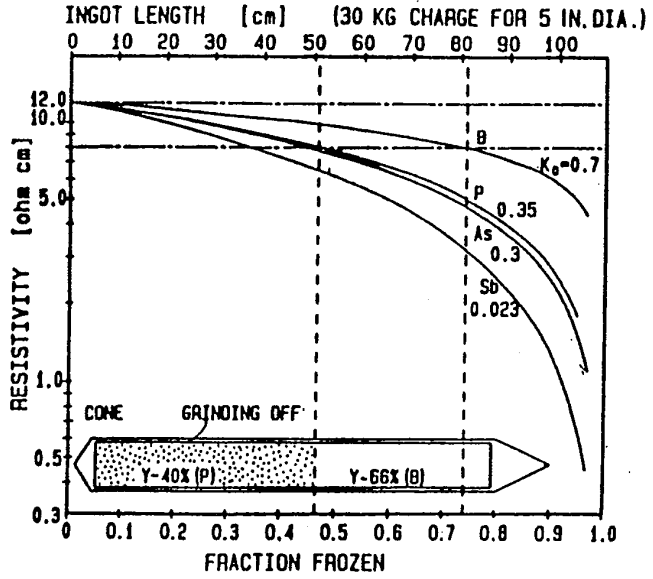


Fig. 1-11 Resistivity distributions for several dopants along the ingot length.

not move by macroscopic flow, but only by diffusion of individual melt atoms), and the impurities rejected by the growing ingot cause the boundary layer impurity concentration to rise above that of the melt. If such a boundary layer did not exist, and if macroscopic motion of the melt caused mixing to establish complete homogeneity of impurities, such local concentration increases of impurities near the interface would not arise.

Convective flow of the melt is always present. In general, *thermal convection* occurs as a result of hot silicon rising along the crucible walls and descending towards the center of the crucible.[45] Temperature gradients and solute concentrations are the driving forces of thermal convection. When such flows become irregular and unstable, the melt exhibits intermittent release of "thermals." (This phenomenon is also known as *pluming*.) Such pluming causes variations in the temperature and boundary layer thickness near the melt-crystal interface. These temperature and boundary layer fluctuations degrade the uniformity of several electrical properties of the ingot (e.g., macroscopic and microscopic radial resistivity).

It is common practice in CZ growth to use *forced convection* in an attempt to suppress thermal convection (and thereby reduce the unwanted effects arising from the flow instabilities). Forced convection is introduced by rotating the crystal during growth. For small melts this technique is relatively effective, but for larger melts only a very limited reduction of thermal convection is achieved through crystal rotation. The crucible is also rotated (in the opposite direction to crystal rotation), to homogenize the melt and achieve thermal symmetry.

It has been determined experimentally that the Normal freezing model must be modified such that the effects of boundary layers are taken into account to give more accurate predictions of the impurity incorporation. The modifications lead to the definition of an *effective segregation coefficient* k_e, which produces closer correlations between observed and calculated axial dopant incorporation in an ingot:[15,16]

$$k_e = \frac{k_0}{k_0 + (1 - k_0)e^{\left(\frac{-VB}{D}\right)}} \qquad \textbf{(1.6)}$$

where: B is the boundary layer thickness; V is the growth velocity (pull-rate); and D is the diffusion coefficient of the impurity in the melt. Equation 1-6 predicts that as the boundary layer thickness increases, k_e increases (thus more impurity is incorporated into the growing crystal).

The concepts used to arrive at the effective segregation coefficient and the thermal convection behavior of the melt described above can help to explain the radial resistivity variation observed in CZ growth wafers (and the cause of microscopic resistivity variations in the ingot). The *macroscopic radial resistivity variation* (Fig. 1-12) arises because the motion of the molten silicon (due to the dominant thermal convection flow in large melts),[17] causes a thinner boundary layer to occur near the periphery of the growing ingot (and a thicker boundary layer at the ingot center). Thus, if $k_0 < 1$, more impurities will be incorporated near the center of the ingot, and progressively fewer towards the edge. The *microscopic resistivity variations* are caused by transient temperature and boundary layer fluctuations brought about by thermal convection flow instabilities.

The values of k_e are larger than those of k_0, and can approach 1 if the parameter (VB/D) is large. This phenomenon (when $k_e \rightarrow 1$) may be used to improve *axial* doping uniformity. By increasing the pull-rate and decreasing the seed rotation, B is increased (since B is inversely proportional to rotation speed). On the other hand, Abe argues that to increase *radial* uniformity B should be thin (and therefore more uniform) across the liquid/solid interface.[9] This calls for a

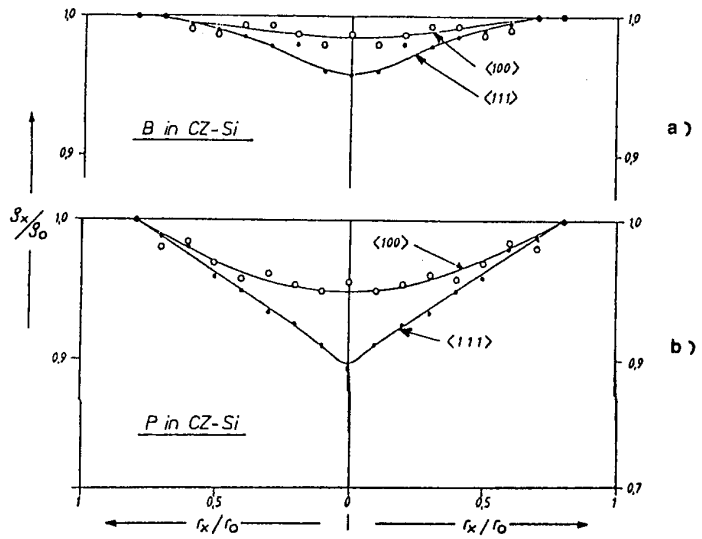


Fig. 1-12 Radial resistivity variation in (a) boron doped and (b) phosphorous doped CZ single crystals showing the influence of the crystal orientation on the dopant concentration.[8] Reprinted with permission of Springer-Verlag.

large rotation rate. As discussed in the chapters' closing section, a variety of other techniques have been investigated to suppress thermal convection in large melts. If such thermal convection can be reduced, more uniform impurity incorporation during CZ growth should be possible.

The incorporation of precisely controlled amounts of oxygen into the crystal[8,18,19,25] is also an important requirement for creating wafers suitable for fabricating VLSI with high yields (see Chap. 2). Ordinarily, incorporation of oxygen is unintentional, and arises from crucible (SiO_2) dissolution during growth, resulting in oxygen concentrations from $5x10^{17}$–$1x10^{18}$ atoms/cm^3. Variables that affect the distribution and concentration of oxygen in the growth crystal include thermal convection currents rising along the crucible walls, oxygen evaporation from the melt surface, incorporation of oxygen into the crystal, and dependence of crucible dissolution on melt level. The number of variables and their potential interactions with one another make it difficult to develop models which accurately predict oxygen incorporation in CZ crystals.

The techniques proposed to suppress thermal convection, however, are also predicted to reduce the impact of crucible dissolution effects on oxygen incorporation. In fact, significantly lower ($2x10^{17}$/cm^3) concentrations and more homogeneous incorporation of oxygen have been observed when thermal convection currents are eliminated and the solid-liquid interface is maintained at a nearly constant temperature. With oxygen concentrations as low as $2x10^{17}$ atoms/cm^3, oxygen-related defects such as swirl, stacking faults, thermal donor generation, and oxide precipitation (see Chap. 2) are reported to be eliminated, but such concentrations are also too low to allow intrinsic gettering techniques to be implemented (also discussed in Chap. 2).

Commercially available, state-of-the-art wafers in 1998 offer 10% axial and 5% radial uniformities of interstitial oxygen (O_i) concentrations (in the range of 25–35 ppma [new ASTM] see Sect 1.3.5). As noted by Zuhlehner,[8] however, controlled oxygen incorporation into CZ Si is still being successfully accomplished only on the basis of semi-empirical relationships.

1.3.4 Czochralski Silicon Growing Equipment

The systems used to grow Czochralski Si crystals[47] (also known as *pullers*), can be considered to consist of four subsystems as shown in Fig. 1-13:

1. *Furnace*: crucible, susceptor, rotation mechanism, heater and power supply, and growth chamber.
2. *Crystal-pulling mechanism*: seed cable or chain, rotation mechanism, seed chuck, and crystal handling device.
3. *Ambient control:* chamber gas source, flow controller, and vacuum/exhaust system.
4. *Control system:* computer-based controller, and sensors.

1.3.4.1 Furnace: The material chosen for CZ silicon crucibles is fused silica. It has a high melting point, good thermal stability, and is relatively non-reactive with molten silicon. The molten silicon to some extent does corrode the fused silica crucible, especially during the higher temperature melt-in period. Thus, substantial quantities of oxygen (and silicon) do become dissolved into the melt. Some of this oxygen becomes incorporated into the crystal, initially taking the form of an interstitial impurity. As discussed earlier and also in more detail in Chap. 2, the presence of this oxygen can have potential beneficial as well as detrimental effects on the crystal properties. The purity of the crucible silica is also important, since the silica impurities can limit the upper values of resistivity possessed by ingots grown in the crucible (the levels of B and P in silicon are typically 8–10 ppma). In addition, silica has no distinct melting point, but gradually softens with increasing temperature.

Near the melting point of silicon, the silica material of the crucible becomes so soft that it requires the support of a heat resistant and rigid outer crucible (susceptor). Such susceptors are fabricated from high purity graphite, which exhibits adequate temperature and cleanliness properties. The other components in the chamber that are subjected to high temperatures are also

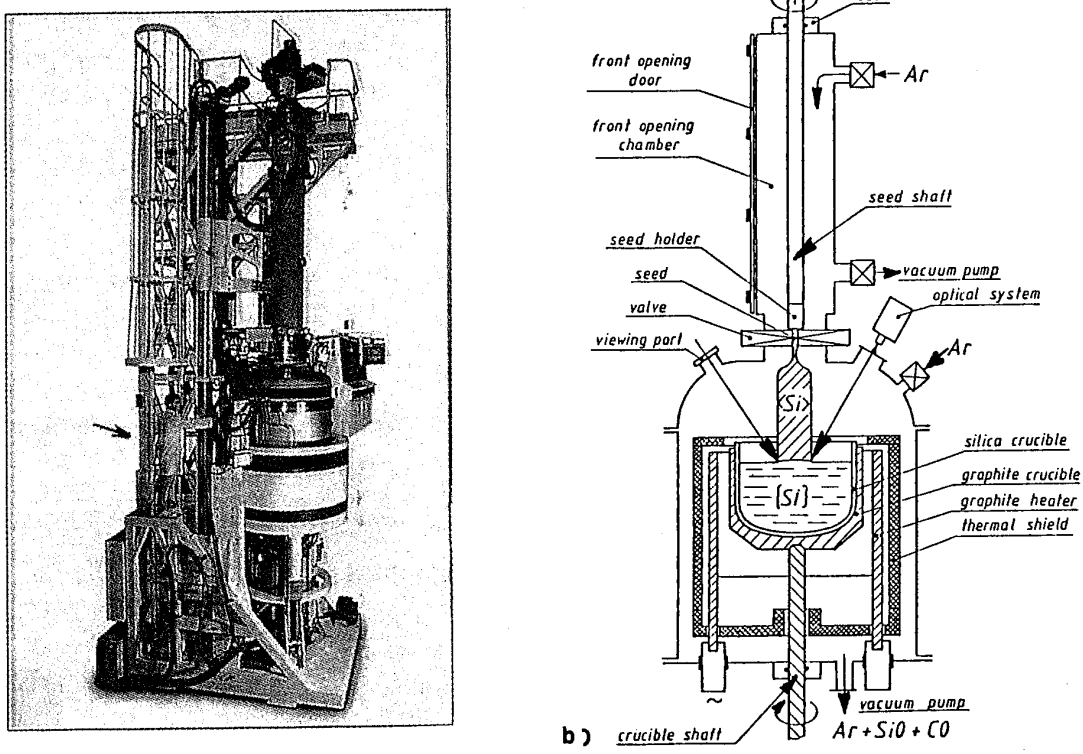


Fig. 1-13 (a) Modern computer controlled CZ puller for 300-mm wafers. Courtesy of Cybeq Systems. (b) Schematic drawing of CZ puller.[8] Reprinted permission of Springer-Verlag.

fashioned from graphite, including the *pedestal* (on which the susceptor sits, and which can rotate as well as raise or lower the susceptor [and thereby also the crucible and melt]), *heater elements*, *insulation package*, and *heat shield*. Carbon contamination of the melt occurs primarily from graphite dust particles, reaction of the hot graphite parts with oxygen introduced from oxygen or water leaking into the chamber, and reaction of the susceptor/pedestal with the silica crucible to form CO. Keeping carbon pieces free of carbon dust, maintaining leak-tight growth systems, and using inert purge gases to sweep out any CO gas that may form from reactions, are effective techniques used to keep carbon incorporation in modern CZ wafers to acceptably low levels.[23] The silicon charge is melted by graphite resistance heaters connected to a dc power supply. Large pullers require tens of kilowatts of power for operation.

The use of CVD silicon nitride as an alternative crucible material has also been demonstrated. Nitride crucibles also exhibit some dissolution. This causes the crystal to be doped with nitrogen, which behaves as a weak donor in silicon.

1.3.4.2 Crystal Pulling Mechanism: The mechanism which pulls the crystal from the melt (and simultaneously rotates it), must perform this function with minimal vibration and sufficient precision. Most systems use a cable or chain to pull the crystal, but their major problem is a tendency for the crystal to swing under certain operating conditions. The process engineer must design the pulling process to avoid these critical conditions. The hot crystals have become too large to be manually handled. Thus, automated unloading systems are used to transport the ingots from the furnace to the cropping and evaluation areas.

1.3.4.3 Ambient Control: As mentioned, the environment in the chamber must be kept free of reactive gases (O, CO, etc.). This is accomplished by filling the chamber with an inert gas (e.g., Ar). During the process, this gas is swept out of the chamber by the vacuum system (together with any CO, SiO, etc. that may be evolving in the chamber), and replaced with fresh inert gas through the ambient gas inlet. Ingot growth at a reduced pressure in the chamber can be used to alter evaporation from the melt. This has been suggested as a method for controlling the concentration of Sb-doped melts and to reduce the oxygen content incorporated into the ingot.[39] A typical consumption rate of inert gas for this purpose is ~1500 liters/kg of silicon grown.

1.3.4.4 Control System: The process parameters, such as pull-rate, crucible rotation, seed rotation, melt temperature, gas flow, and crystal diameter, are controlled in modern pullers with digital computer systems. As mentioned earlier, completely automated cycles (one-button operation) require manual intervention only for loading the crucible with EGS. Recipes that result in various crystal properties are stored in the computer memory, allowing the system to produce nearly identical crystals from run-to-run. Other control systems features may include: central computer interfacing capability, data logging and printing, CRT terminals, and tape or magnetic disk program storage.

Ingot diameters during growth are controlled by focusing an infrared temperature sensor on the melt-crystal interface, and monitoring changes of the meniscus temperature. The pull-rate mechanism and chamber heater is slaved to the output of the sensor, and the diameter is adjusted by changes in the pull-rate. The level of the melt is detected by laser beam reflection. Some aspects of implementing digital control of ingot diameter during growth are discussed in Ref. 46.

1.3.5 Analysis of Czochralski Silicon in Ingot Form

Upon the completed growth of a silicon ingot, a variety of analysis steps are conducted before the ingot is sawn into wafer form. The ingot is weighed and the diameter is measured along its

length. The ingot is visually inspected for evidence of any twinning or other gross crystalline imperfections which may have occurred during growth. The entire crystal is next etched in an acid bath designed to delineate any dislocations that might have been triggered and then propagated through to the tang end of the crystal. Any such dislocated segments must be removed by sawing (see ASTM Std F-47[6] for a discussion of this procedure).

After grinding the ingot to a precise diameter, the resistivity is measured at various points along its growth axis, using either a two-point or four-point probe method (see Chap. 4). This identifies the various resistivity ranges along the ingot length. Radial resistivity measurements are also commonly performed on polished ingot slices at this time for quality or process control.

1.3.5.1 Oxygen and Carbon Measurements in Silicon Using Infrared Absorbance Spectroscopy: The levels of interstitial oxygen and substitutional carbon present in single crystal Si are most commonly determined by using infrared absorbance spectroscopy (but the technique is not valid for polycrystalline silicon). A spectrum of percent transmittance is obtained from a double-sided polished sample of known thickness (typically 0.2–1.0 cm thick, where the technique exhibits its maximum detection capability). The concentration of oxygen and carbon can then be calculated from the IR absorbance due to their presence. Either a dual-beam IR spectrometer or a Fourier transform infrared (FTIR) spectrometer may be employed for this purpose, but FTIR has become the method of choice (due to its increased sensitivity and ease of use). The principles of FTIR analysis are described in more detail in Chap. 7. FTIR analysis can also be used to infer oxygen precipitation rates (Chap. 2). Since FTIR detects interstitial oxygen, a reduction in the oxygen concentration after thermal cycling can be interpreted as a measure of precipitation. Therefore, FTIR measurements can be included among the procedures used to characterize incoming silicon wafers and for process control.

The presence of interstitial oxygen in the Si lattice causes absorption of IR radiation at 1106 cm^{-1} (wavenumber) as a result of the antisymmetric vibration of the Si_2O complex in the host Si crystal (and substitutional carbon causes absorption at 607 cm^{-1} due to the local vibration mode of the carbon). These bands are superimposed on multiple phonon excitations of the Si, and must therefore always be subtracted from the spectrum of an oxygen- and carbon-free sample (Fig. 1-14). Quantitative analysis is based on measuring the percent transmission of wavelengths characteristic of an element, and comparing this to standards. The detection limits at room temperature are $2.5x10^{15}/cm^3$ for oxygen, $5x10^{15}/cm^3$ for carbon, and $2x10^{14}/cm^3$ for nitrogen with FTIR. Samples may be prepared by cutting and polishing segments of a single crystal ingot, or whole wafers may be analyzed.

Two methods can be utilized to measure the absorbance due to interstitial oxygen: a) the *air reference method*; and b) the *difference method*. In the *air reference method*, the percent transmittance relative to air is measured and an appropriate baseline (I_0) drawn at the peak location. The absorbance α, is then calculated according to the relationship:

$$\alpha = (1/x)\cdot(\ln [I_0/I]) \qquad (1.7)$$

where x is the sample thickness (in cm), I is the percent transmitted at peak absorption, and I_0 is the baseline intensity at peak absorption. The value is then adjusted by 0.4 cm^{-1} for 300 K measurements (and 0.2 cm^{-1} for 77 K measurements) to account for the shape of the background (baseline). In the *difference method*, a reference sample is placed in the reference beam of a dual-beam spectrometer. The reference sample is a polished slug of float-zone Si with negligible oxygen or carbon content. Following standard procedures, the absorbance of the test sample rel-

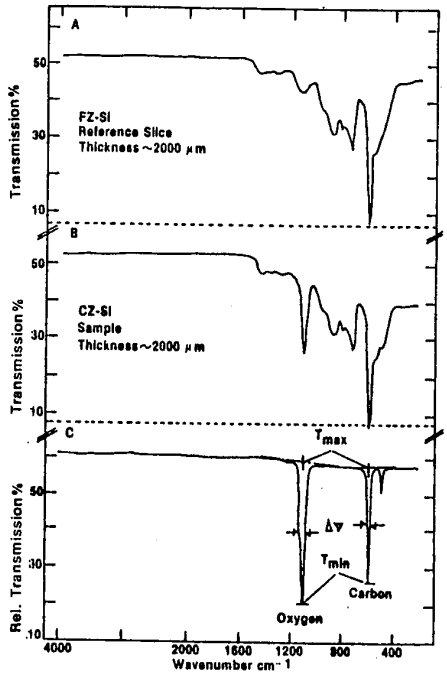


Fig. 1-14 Transmission spectra: a) Silicon slice free of oxygen and carbon; b) Slice containing oxygen and carbon; c) the difference between spectra shown in a) and b).[20] Reprinted by permission of Solid State Technology, published by PennWell.

ative to that of the reference is measured, and the value of α is calculated as in the air reference method (except that no adjustment is made to α, as the baseline is flat and well-characterized).

The value of the interstitial oxygen content in silicon $[O_i]$, is obtained by multiplying α by a *calibration constant*. The literature contains reports of several values for this constant (ranging from 4.9 to 9.63). The discrepancy is due to the difficulty in conducting the experiments from which this value is extracted. The ASTM Std F-121 for measuring $[O_i]$ specified the value of 9.63 for the calibration constant up until the F 121-77 (1977) version. In 1980, ASTM approved a new standard, ASTM F-121-80, in which the accepted value of the constant was changed to 4.9. Thus, it is now customary in literature that discusses $[O_i]$ values to report the calibration constant value that is being used. If 9.63 is utilized, the $[O_i]$ values are referred to as "old ASTM", while use of 4.9 implies "new ASTM" values. In pre-1980 literature the range of $[O_i]$ in CZ wafers was reported as 26–38 ppma, while $[O_i]$ in CZ wafers using "new ASTM" is given as 13–19 ppma. In this text we use "new ASTM" values for $[O_i]$ values. Even more recently (1983), a cooperative effort sponsored by JEIDA in Japan yielded a value of the calibration constant of 6.06.[46]

Substitutional carbon present in single crystal Si can be determined in the same manner as oxygen. The difference method must be employed when analyzing for the presence of carbon, since strong interference by phonon bands makes the air reference method ineffective. The calibration constant for carbon is 2.0 at 300 K.

The analysis of silicon crystals for oxygen and carbon can be carried out using a modified form of the ASTM Stds[6] F-120, F-121, and F-123 and FTIR. Reference 20 discusses an extension of these standard measurements in that it covers techniques for oxygen and carbon

measurements with IR on samples with either one or both sides polished (whereas ASTM F-121 and F-123 only discuss the analysis of double-side-polished slabs).

1.4 FLOAT-ZONE SINGLE-CRYSTAL SILICON

The *float zone* (FZ) process[21,22] is the second major method used to grow single-crystal Si. In FZ, a molten zone is passed through a poly-Si rod of approximately the same dimensions as the final ingot. The Si does not come into contact with any substance other than ambient gas in the growth chamber. Thus, FZ crystals have higher inherent purity than CZ ingots (which incorporate significant amounts of oxygen, carbon, boron, and other metallic impurities from the dissolution of the crucible walls). For applications requiring high-purity starting Si (e.g., high power thyristors and rectifiers, and infrared radiation detectors), FZ material becomes the material of choice. The largest quantities of FZ silicon are supplied in the resistivity range of 10-200 Ω-cm, although material up to 30,000 Ω-cm has been produced. Some economical and technical disadvantages, however, make FZ silicon less suitable for VLSI applications than CZ silicon.

From an economic perspective, FZ growth is more costly than CZ growth. In addition, CZ wafers are available in larger diameter form than FZ wafers. From a technical viewpoint, as-grown FZ crystals generally also exhibit more severe microscopic resistivity variations than do CZ crystals, and the oxygen impurity content of CZ wafers offers added strength and intrinsic gettering potential. The microscopic resistivity variations of FZ silicon can be essentially eliminated with neutron transmutation doping, but this technique is not as attractive for lower resistivity materials. Techniques for incorporating oxygen into FZ silicon have also proved to be difficult to implement. Typical oxygen values for FZ silicon are 0.1-0.3 ppma, and carbon is found at 0.1-1.0 ppma levels. Thus, in the foreseeable future, CZ and FZ methods are likely to continue serving their respective application areas.

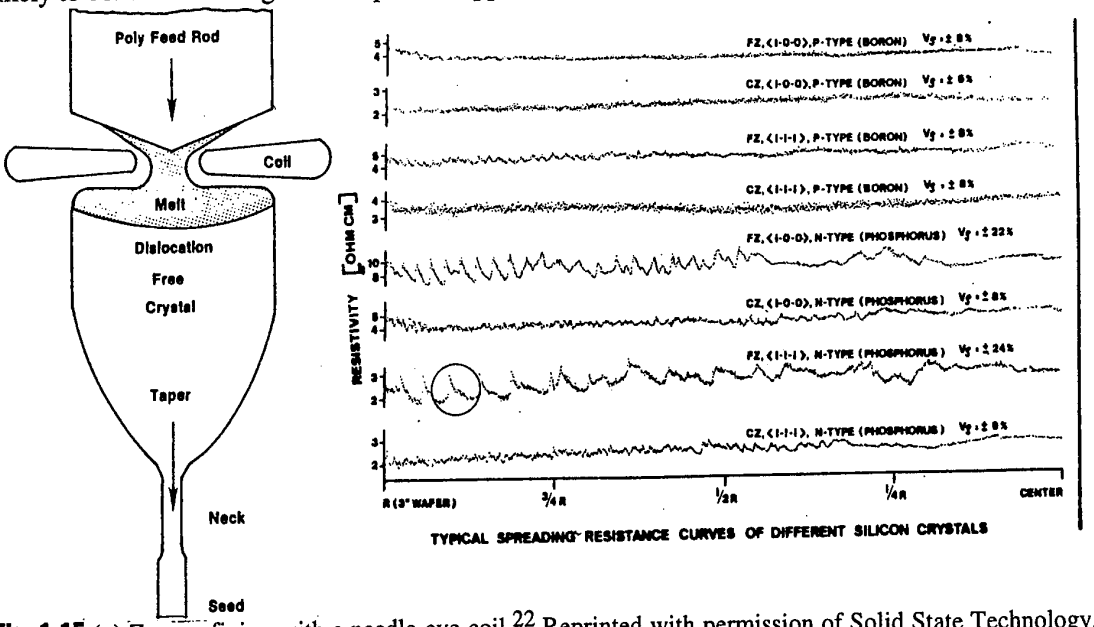


Fig. 1-15 (a) Zone-refining with a needle-eye coil.[22] Reprinted with permission of Solid State Technology, published by PennWell Publishing. (b) Comparison of spreading-resistance measurements of CZ and float-zone (FZ) crystals having the same resistivity specification. The dopant striations in FZ crystals are greater due to the pronounced remelt phenomena at each rotation of the crystal.[8] Reprinted with permission of Springer-Verlag.

Figure 1-15 shows a schematic of the FZ process. A polysilicon rod is mounted vertically inside a chamber under vacuum or an inert atmosphere (e.g., argon). A needle-eye coil provides rf power to the rod, causing it to melt and establish a 2 cm long molten zone. The rf field maintains a stable liquid zone. The molten zone is moved along the rod to effect purification and to establish the crystal. Modern zone-refiners use stationary coils and provide for vertical motion of feed rod and crystal.

1.5 FROM INGOT TO FINISHED WAFER: SLICING, ETCHING, and POLISHING

After a single crystal ingot has been grown, a complex sequence of shaping and polishing steps are performed to produce starting material suitable for fabricating semiconductor devices (i.e., *wafers*).[26,51] The procedure (known as *wafering*) includes the following steps (Fig. 1-16):

1. The grown single crystal ingot undergoes routine evaluation of resistivity, impurity content, crystal perfection, size, and weight. Sections of the ingot that are crystallographically defective, irregularly shaped, or undersized, are cut off and discarded (as are the seed and butt [or *tang*] ends). Loss of the ingot material may approach 50% at this point.

2. Since ingots do not grow perfectly round, nor with sufficiently uniform diameters, they must shaped to the desired form and dimension. Thus, ingots are deliberately grown larger than the final desired wafer diameter and a grinding operation removes the excess material to reduce as-grown crystals to a cylindrical shape of precise diameter. (Exact diameter dimensions are required to make wafers compatible with automated handling procedures.) Since silicon is a hard, brittle material (registering 72.6 on the Rockwell "A" hardness scale), grinding machines with diamond wheels are used to perform this cylindrical shaping process. The grinding step is followed by an etch step that removes the grinding work-damage. The ingot diameter is reduced 0.25–1.0 cm by the grinding and etch steps.

3. For wafers up to 150 mm, *flats* are used for aligning and crystal orientation. For larger wafers ≥ 200 mm, *notches* are predominantly used (see Fig. 1-17). These features are next

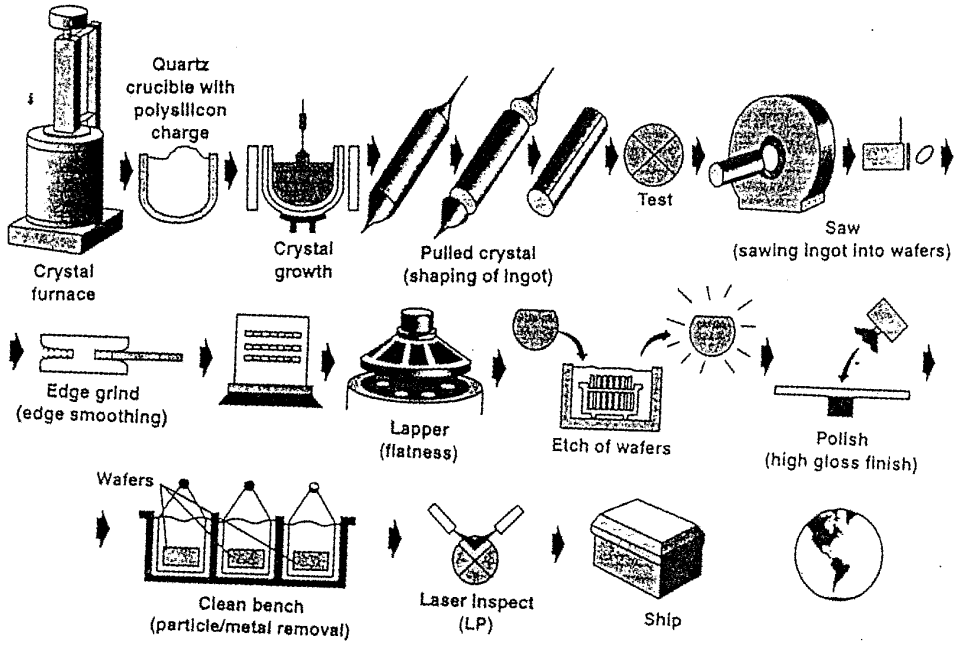


Fig. 1-16 Steps in the wafering process.

ground along the length of the ingot. The largest flat, called the *primary flat*, is usually positioned relative to a specific crystal direction. (The primary flat orientation is found by using x-rays to produce a Laue photograph.) The primary flat is used for several purposes. Automated wafer handling equipment utilizes the primary flat to obtain correct alignment, and in addition, devices on the wafer can be oriented to specific crystal directions with this flat as a reference. Smaller flats are called *secondary flats*, and they are utilized to identify the orientation and conductivity type of the wafer. Since automated equipment relies on the flats for correct operation, the flat dimensions must be precisely machined.

4. The sawing operation[26,27] that produces silicon slices from the shaped ingot also defines the surface orientation, thickness, taper, and bow of the slice. The ingot is rigidly mounted to maintain exact crystallographic orientation during the sawing process. Wafers of <100> orientation are normally cut "on orientation" (e.g., within ± 0.5°).

Wafer thickness is primarily set by the sawing operation, although some material is also removed by subsequent operations (e.g., for a 200-mm wafer of final 725 μm thickness, a total of ~1250 μm of ingot silicon is needed: 725 μm final wafer thickness; 400 μm for kerf loss during sawing; 50 μm lapping loss; 50 μm etching loss; 25 μm polishing loss. Thus, about 8, 200-mm wafers per linear cm of ingot can be produced. As wafers get larger they must be thicker (Table 1-3) to allow them to withstand thermal processes (e.g., epitaxy and oxidation) and handling during fabrication. Thus, fewer wafers per unit length of ingot are possible than for smaller wafers.

The most common mode of slicing has been *inner diameter* (ID) *slicing* (Fig. 1-18) using stainless steel blades with diamond particles bonded to their edges.[26] Continuous monitoring with blade deflection sensors is required to assure that slices are sawn adequately free of bow, taper, work damage, and saw marks. Cutting speeds are in the neighborhood of 0.05 cm/sec. Since one slice is cut at a time by the saw, this is a relatively slow process. Wafers

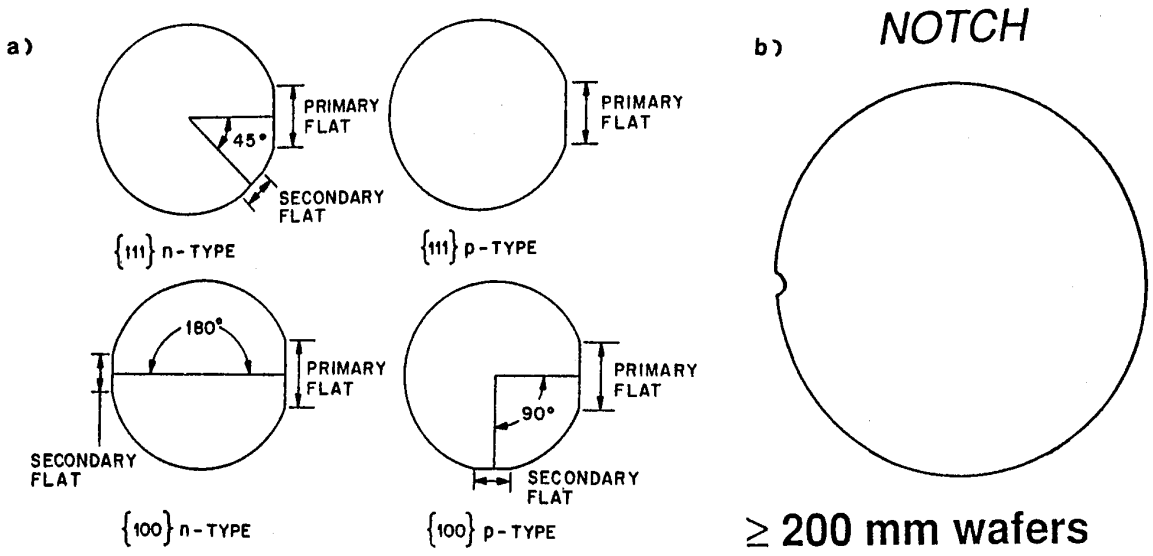


Fig. 1-17 (a) Indentifying flats on a silicon wafer. Reprinted with the permission of the Semiconductor Equipment and Materials Institute. (b) Notched wafers.

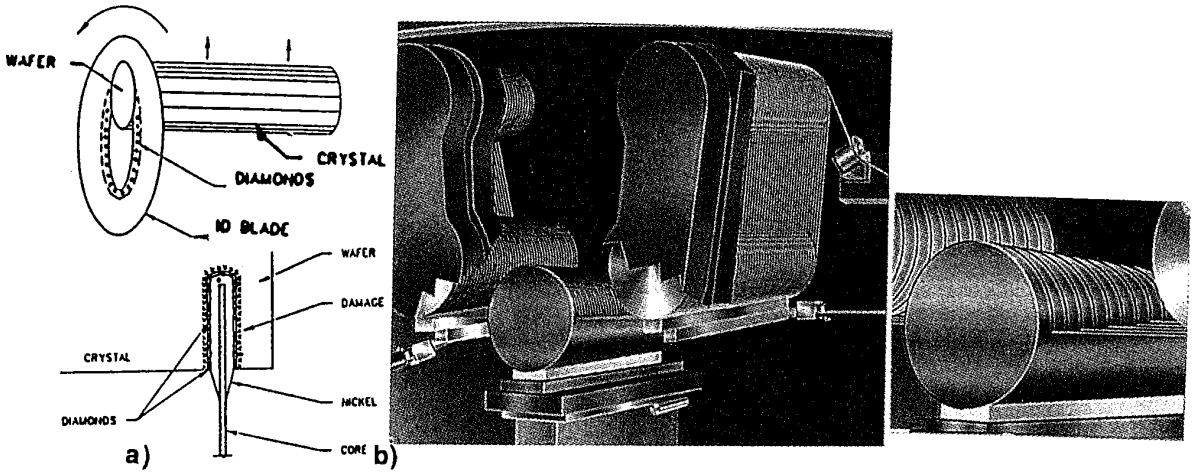


Fig. 1-18 (a) ID saw geometry.[27] Reprinted with permission of Solid State Technology, published by PennWell. (b) Wire sawing an ingot to produce wafers. Reprinted with the permission of the Semiconductor Equipment and Materials Institute.

with 300-mm diameters are sawn with slurry-coated wires instead of inner-diameter saws. More slices per inch of ingot are possible due to smaller kerf loss.[56]

5. A technique using a *laser* to create an alphanumeric, or bar code, identification mark on the front of the wafer near the primary flat or notch (Fig. 1-20) is then employed. *Laser marking* according to SEMI[29] Std. M 1.8, specifies an 18 character field to identify the wafer manufacturer, conductivity type, resistivity, flatness, wafer number, and device type. This allows each wafer to be identified. Wafer traceability, correlation of device characteristics produced on a wafer with the wafer properties, and the ability to prevent in-process rejected wafers from being re-inserted into the fab line are all benefits of utilizing wafer marking. Automatic code readers at various stages of processing can also track wafers and provide data on wafer movement in the fab.[30,31]

6. The thickness of the as-sawn slices is sufficiently variable that a *lapping* and *grinding* step is next employed to bring all slices to within the specified thickness tolerance. This step also serves to reduce bow and warp, and to increase slice flatness. Both sides of the slice are lapped using a mixture of Al_2O_3 and glycerine, and a succession of increasingly finer polishing grits is used in multiple lapping steps. This process can achieve flatness uniformities of 2 μm.

7. The edges of the silicon slices are then shaped (rounded) as shown in Fig. 1-19. Edge rounding substantially reduces the incidence of chipping during normal wafer handling, and helps minimize film buildup on the edge of the wafer during photoresist and epitaxial processes. (Edge-chip defect sites are known to nucleate plastic deformation [*slip* - see Chap. 2] during subsequent thermal processing steps).[32]

8. The work damage and contamination caused by the previous shaping steps is next removed. This damaged and contaminated layer is chemically etched away.[43] A relatively non-porous and clean wafer backside is also produced by this step. The wet etch procedure for this step typically utilizes an etchant solution of hydrofluoric, nitric, and acetic acids to attack the silicon.

9. A chemical-mechanical polishing[43] step is used to produce the highly reflective, scratch, and damage free surface on one or both sides of the wafer. This is accomplished by mounting unpolished slices onto a carrier, and then putting them on a polishing machine.

There, a powered platen drives an appropriate polishing pad material across the wafer surface. A colloidal silica slurry of sodium hydroxide and fine (~10 nm) SiO_2 particles is dripped onto the table. The frictional heat of the sliding mounted wafers causes the sodium hydroxide to oxidize the Si (i.e., the chemical part of the process). The oxide is then abraded away by the silica particles (i.e., the mechanical part). Following the polishing cycle (in which a layer of about 25 μm is removed) the wafers are subjected to series of chemical dips and rinses to remove the polish slurry. A cleaning process concludes this step. An inspection and packing in a particle-free package finish the sequence.

10. Epitaxial layers are required for some ULSI process technologies. Wafer manufacturers are now supplying silicon wafers with a variety of epitaxial layers for CMOS ULSI technologies. Another option being offered is a layer of polysilicon or mechanical damage on the backside of the wafers for extrinsic gettering purposes (see Chap. 2).

1.6 SPECIFICATIONS OF SILICON WAFERS FOR VLSI

There are many properties that a wafer must exhibit (Table 1-2) in order to be an appropriate starting substrate for the fabrication of VLSI (and ULSI). In this section the terms used to specify these properties are defined and measured. Readers are also directed to recommended test procedures that have been developed to verify that received silicon wafers conform to purchase specifications. Many of these tests have been standardized by the American Society for Testing and Materials (ASTM)[6] and the Semiconductor Equipment and Materials Institute (SEMI).[29] The standard test procedures are available (updated when appropriate) in publications annually issued by these organizations. Table 1-3 provides some typical specification ranges for 125, 150, and 200 mm diameter CZ wafers. Note that specifications have become tighter with increasing wafer size due to automatic wafer handling and photolithographic requirements. ADE Corporation is the major supplier of tools used to measure many of these specifications.

1.6.1 Electrical Specifications

Conductivity Type - Specifies whether wafers are *n*- or *p*-type. Information about which element was used to dope the wafer should also be provided.

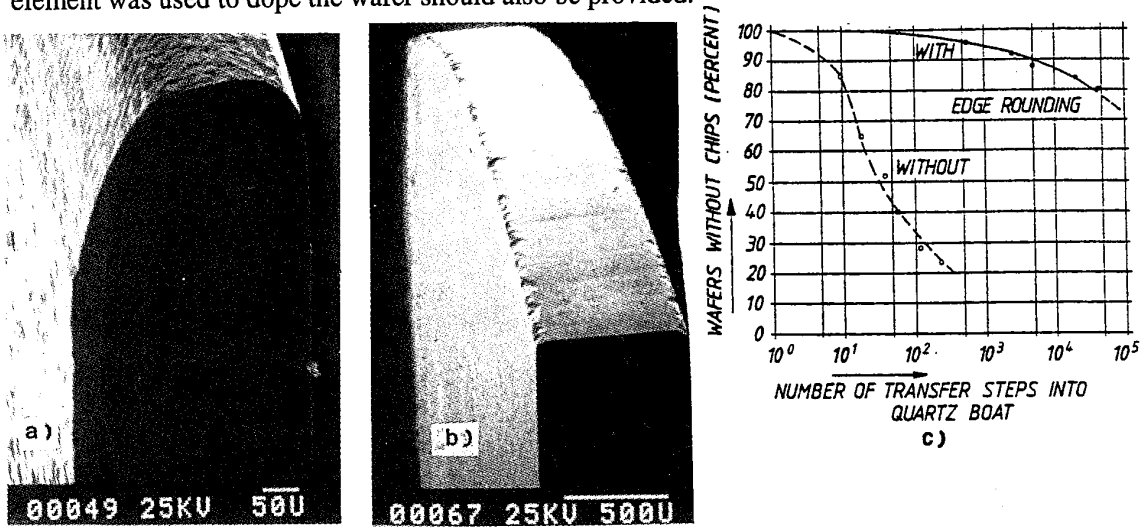


Fig. 1-19 a) Edge-rounded wafers. b) Non-edge rounded wafers.[51] Reprinted with permission of Academic Press. (c) Mechanical strength of edge-rounded vs. non-edge-rounded wafers.[8] Reprinted with permission of Springer-Verlag.

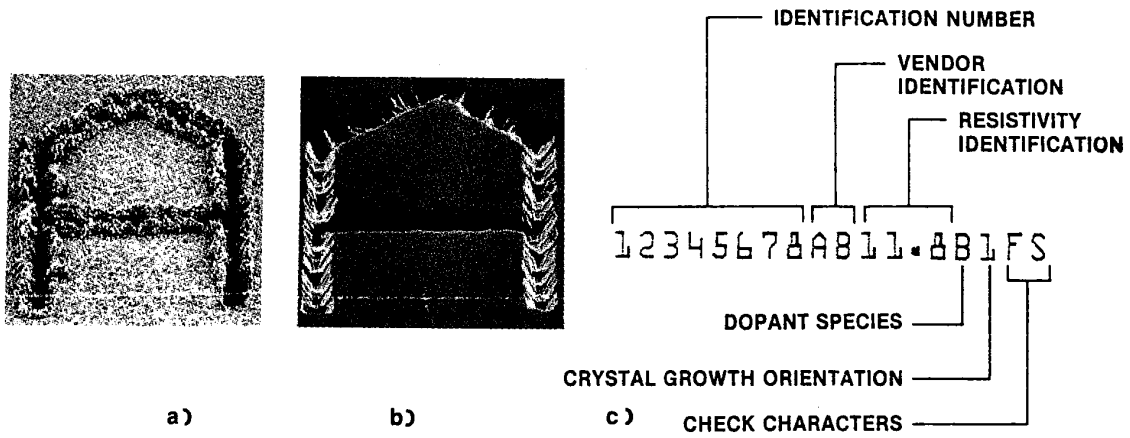


Fig. 1-20 (a) SEM of letter "A" marked on wafer after it appears as marked on lapped wafer (left), after wafer has been etched (center), and after polishing. Courtesy of Wacker Chemitronic. (b) SEMI specification for marking vendor codes on wafers. Dimensions are specified along flat on wafer. (c) SEMI specifications for 18-character vendor code. SEMI standard, used with permission.

Resistivity or Resistivity Ranges (Ω-cm) - specifies the average (or range of) resistivities of the wafers. The resistivity is related to the doping density[49] as shown in Fig. 1-21. The wafer resistivity can be verified using ASTM Std F-43 with a four-point probe.

Radial Resistivity Gradient (% Variation) - gives a measure of the variation of the resistivity between the center and the selected outer regions of a wafer. The radial resistivity variations of wafers are a function of the crystal growth process and dopant type. The ability to predictably control a variety of device characteristics depends on having uniformly doped substrate material (an example is the threshold voltage of MOS devices). Radial resistivity gradient is measured using ASTM Std F-81 with a four-point probe.

Resistivity Variations - are local variations of the resistivity that occur primarily due to remelt phenomena. The effect is more pronounced in FZ than CZ material, although the effect

Table 1-2. COMPARISON OF MATERIAL PROPERTIES AND REQUIREMENTS FOR VLSI AND ULSI

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

Table 1-3 SPECIFICATIONS FOR SINGLE-CRYSTAL SILICON WAFERS

Parameter	150 mm	200 mm	300 mm
Diameter (mm)	150 ± 0.2	200 ± 0.2	300 ± 0.1
Thickness (μm)	675 ± 25	725 ± 20	775 ± 25
Primary flat length (mm)	55–60	notch	notch
Oxygen content (ASTM '79) (ppma)		≥ 23 ± 2	≥ 22 ± 1.5
Flatness (μm) (GTIR)	1.5	1.5	1.5
Site Flatness (μm) (SFQD)		± 0.17 (1998)	± 0.12 (2001)
Surface orientation	(100)	(100)	(100)

can be substantially eliminated in FZ ingots by subjecting them to neutron-transmutation doping. Resistivity variations are measured with a four-point probe (ASTM Stds F-81 and F-525).

1.6.2 Mechanical/Dimensional Specifications

Diameter - specifies the linear dimension across the surface of the wafer, passing through the center of the wafer, excluding flats. This specification has been continuously tightened to improve pre-alignment placement of wafers onto stepper equipment, and to facilitate automated wafer handling. For 200-mm wafers the diameter specification is 200 ± 0.2 mm. It is measured according to ASTM Std F-613 using a ring template.

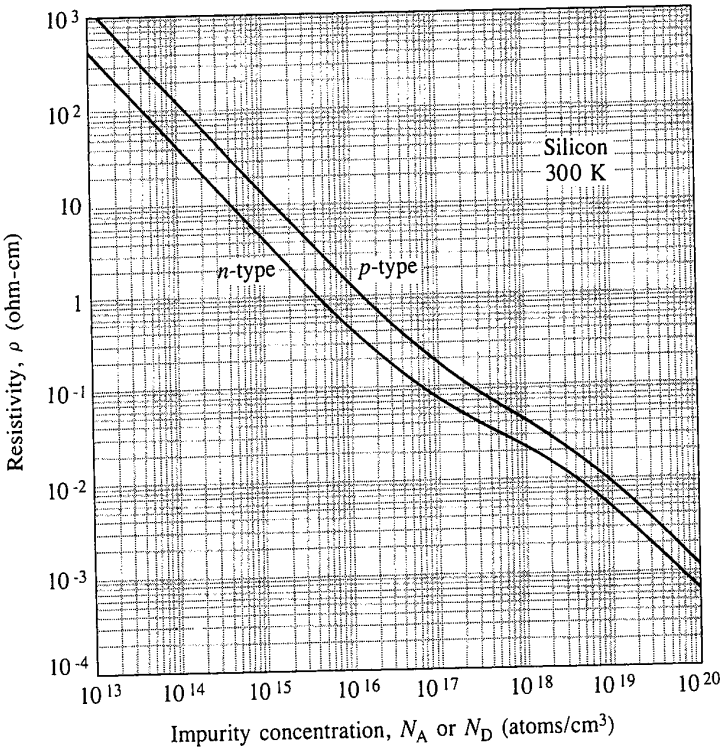


Fig. 1-21 Resistivity versus impurity concentration for Si. From W.F. Beadle, R.D. Plummer, and J.C. Tsai, *Quick Reference Manual for Silicon Integrated Circuit Technology,* 1985, Copyright © John Wiley & Sons. Reprinted with permission of John Wiley & Sons.

Thickness - must be controlled because if wafers are too thin they may break or warp during normal wafer processing, or if too thick, they may not work in all processing equipment and fixtures. Thickness is measured at the wafer center according to ASTM Std. F-533 with a thickness gauge.[51] 200-mm and 300-mm wafers are 725 ± 20 μm and 775 μm ± 25 μm thick, respectively.

Total Thickness Variation (TTV) - Excessive thickness variations may also cause problems in mechanical handling and photolithographic processes especially if stepper processing employs back-surface referencing. TTV is the difference between the maximum and minimum values of thickness on a wafer (Fig. 1-22), and is primarily a function of the sawing process. It is measured according to ASTM Standard F-533 using a non-contact type thickness gauge.[43,50,51]

Bow is the concavity due to sawing, or the deformation from thermal processing of the wafer centerline (Fig. 1-22). Bow is a *bulk* property of the wafer, not of the wafer surface, and units of bow are μm. Bow is measured according to ASTM Std F-534-84. Parameters such as bow and warp are measured on free wafers (rather than on wafers flattened by a vacuum chuck).[43,50,51]

Warp - is the deviation (difference between maximum and minimum distance) exhibited by the centerline of a wafer from a planar condition, when such a deviation includes both concave and convex regions (Fig. 1-22). Warp is also a *bulk* property (expressed in μm) and should not be confused with flatness. Warp can occur during wafer processing, as discussed in Chap. 2. It can be measured according to ASTM Std F-657-80.[43,50,51]

Flat dimensions - must be machined to a precise length and orientation. Edge rounding of the wafer can degrade the ends of the flatted area, and can thus change the effective flat length.[44] Flat length and any such flat edge ambiguity is measured using an optical comparator according to ASTM F-671.

Notches - Notches are features machined on the edge of large (≥ 200 mm) wafers for alignment and crystal orientation purposes (Fig. 1-17b). Notches are 1 mm deep, and the notch

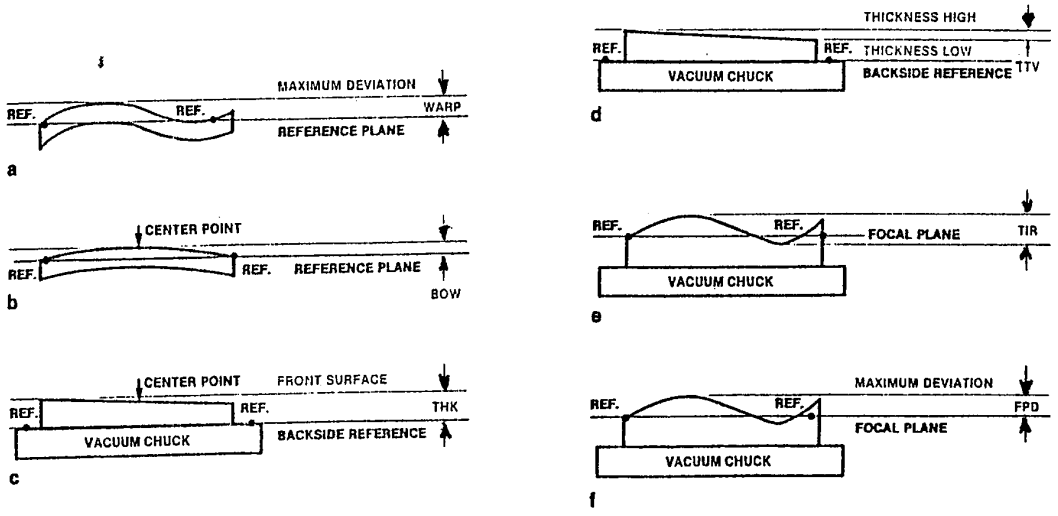


Fig. 1-22 Typical wafer flatness parameters (a) warp (b) bow (c) thickness (d) total thickness variation, TTV (e) total indicator reading, TIR (f) focal plane deviation, FPD.

axis (a line through the pin center to the center of the wafer) is a <110> direction. A pin is used to find the notch, and thus align the wafer in a fixture.

Edge contour - the wafer edges are contoured (rounded) to minimize occurrence of chipping and to help reduce film buildup at the wafer edge during photoresist and epitaxial deposition. The contour shape should meet the SEMI specification M1.5. The contour quality is measured by comparison to a template.

Laser Marking - as described earlier, a laser can be used to create a dot-matrix alphanumeric code on the wafer surface.[30,31] See SEMI specification M1.8 for character generation information and interpretation.

1.6.3 Chemical/Structural Specifications

Bulk Structural Defects - Rapid inspection of large diameter wafers by x-ray topography should reveal no structural defects in starting wafers.

Surface Orientation - specifies the orientation of the surface of the semiconductor wafer and the allowed degree of misorientation. Normally <100> material is cut *on orientation* ±0.5°. Surface orientation of wafers is verified according to ASTM Std F-26.

Oxygen Content - is important because it affects wafer strength and impurity gettering processes as discussed in Chap. 2. Oxygen content is measured using ASTM F-121 and FTIR. The interstitial oxygen content will be between 13 and 19 ppma, and should be tailored to the customers specification and a tolerance of 2 ppma may be acceptable.

1.6.4 Surface/Near Surface Specifications

Flatness - is the maximum peak-to-valley deviation of a wafer surface as measured from a reference plane (Fig. 1-21). Flatness is a *surface* property, expressed in μm and can be measured according to ASTM Std F-775-83. Flatness parameters, wafer thickness, and TTV are usually measured with the wafer pulled flat on a vacuum chuck. Global flatness (GTIR - global total indicator reading) of approximately 1.5 μm will be required for small feature size ($\leq$ 0.35 μm) applications. A local site flatness (SFQD) of ±0.12 μm over a 26x32 mm field of view will ensure that each step and repeat field is sufficiently flat, and thus remains within the depth of focus of lenses designed to resolve 0.18 μm dimensions.[33,43,50,51]

Particles - are detected visually by the unaided eye and a collimated light beam, optical microscopy, or automatic surface scanners (ASTM Std F-154 and F-24). Their presence is quantified as the number of particles/unit area.

1.7 THE ECONOMICS OF SILICON WAFERS

As discussed earlier, the size of silicon wafers has been continuously increasing. The motivation for using larger wafers is economic, and here we discuss this issue. The economic benefits of larger wafers are gained from the additional silicon area they provide (Fig. 1-7). If the manufacturing cost per wafer remains constant (or if its percentage increase is smaller than the percentage increase in silicon area) a reduction in the manufacturing cost per chip may be possible.

Let us address the issue of the silicon wafer costs. We see from Fig. 1-23 that the cost per square inch of silicon on a given wafer size is highest when the new wafer size is first introduced, but then drops with time.[52] For 125-mm prime device wafers the minimum silicon cost reached ~\$1/in^2. For each succeeding wafer size the minimum cost/in^2 is expected to be 1.4x the cost of the previous generation. Thus, for 200-mm wafers the silicon cost is \$2/in^2,

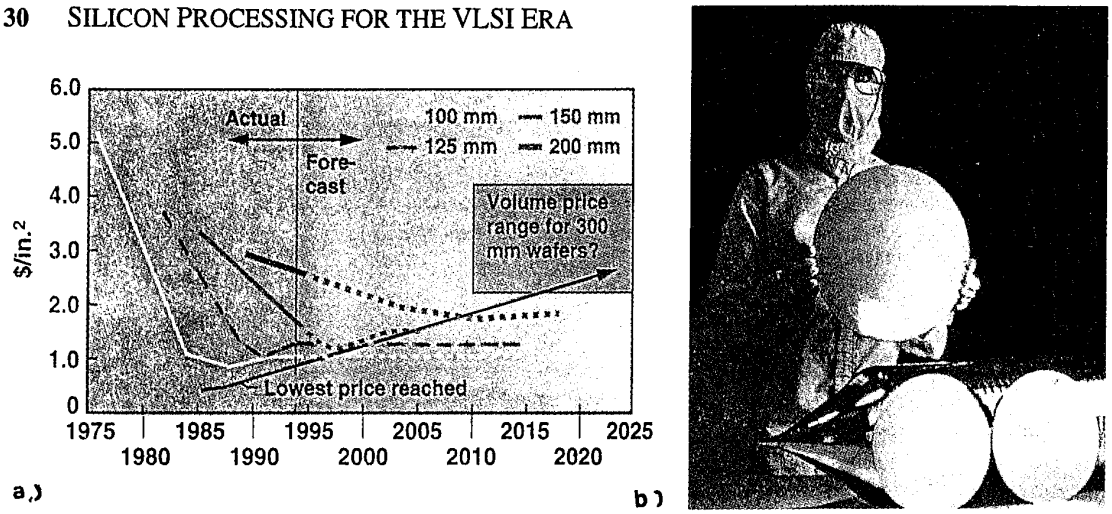


Fig. 1-23 a) Actual wafer pricing cycle by wafer diameter. Reprinted with permission of Semiconductor International.[52] b) Photograph of a 300-mm wafer. Courtesy of MEMC.

resulting in the minimum cost of prime bulk CZ wafers of this size being \$80–\$100. The cost of 300-mm wafers is expected to eventually decline to somewhere above \$3/in^2, with a single bulk CZ wafer thus eventually costing \$350–\$400. In 1998, the cost of prime 300- mm wafers was ~\$1000. Note that adding an epitaxial layer to the wafer approximately doubles the wafer cost, and the cost of 200-mm epi wafers in 1998 was \$150–\$180.

The cost/in^2 increases with wafer size for a number reasons. First, larger wafers are thicker, meaning fewer wafers per cm of ingot length are produced. Second, a smaller percentage of the electronic-grade polysilicon (EGS) ends up in prime wafers. (For 150-mm wafers this percentage was 30%, for 200-mm wafers it dropped to 17%, and for 300 mm wafers it is expected to be only ~10%.[53]) Third, wafer specs are tighter for each generation, reducing the "sweet spot" in the middle of the ingot that yields prime wafers. Finally, epi may be required for DRAM production on 300-mm wafers, which will create a steep rise in average price of prime wafers.

Note that the consumption of wafers of a given size also varies with time (see Fig. 1-24), and each new wafer size remains in production for about 24 years.[54] For example, production of 150-mm wafers began in the late1970s, and their consumption rate peaked in 1997. Their use in

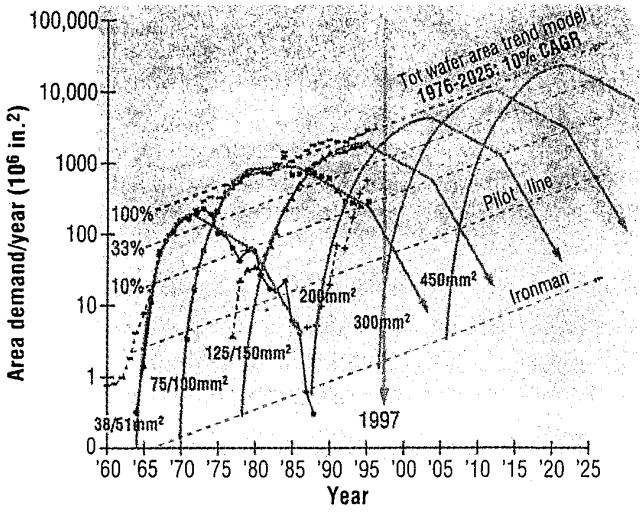


Fig. 1-24 The 24-year single-diameter cycle of silicon wafers.[54]

in production is expected to be phased out by about 2010. Next, 200-mm wafers were first introduced in 1988, and they are expected to reach their peak consumption rate sometime between 2005 and 2010 (depending on the rate of ramp up of the new 300-mm wafer generation).[54] By 1998, however, full-scale-production IC factories using 300-mm wafers were not yet running. As a result, well-designed 200-mm equipment is predicted to have a long working life before it is made obsolete by future wafer-fabs running larger wafer sizes.

The worldwide consumption of single-crystal silicon has been growing at 15–20% per year. In 1992, about 1.9 billion in^2 of silicon were sold. By 1994 this number was 2.7 billion in^2, and by 1997 it had increased to ~4 billion in^2. In 1994 this translated to a total production of about 100 million wafers. The sale of this Si generated an income of $7 billion US in 1997.[55] There are a relatively large number of silicon wafer suppliers. In the late 1990's the largest were Shin-Etsu Handotai (SEH: Japanese), MEMC (German); Wacker-Siltronic (German); Sumitomo-Sitix (Japanese); Mitsubishi Materials Silicon (Japanese); and Komatsu Electric (Japanese).

1.8 TRENDS IN SILICON CRYSTAL GROWTH and ULSI WAFERS

The increasing size of silicon wafers is the most obvious trend in silicon material technology. In 1996 300-mm wafers were introduced, and 400-mm wafers are predicted to be the next size (probably to be introduced in about 2007). The use of larger diameter wafers for maintaining productivity as die sizes increase presents several major challenges to semiconductor manufacturers. Foremost is the enormous cost of building new fabs with equipment capable of handling the larger wafers. Next, patterning techniques must be developed to print smaller feature sizes over larger areas. This also involves the need to maintain ever higher degrees of wafer flatness, which in turn requires the use of low-temperature or RTP processing and such planarizing techniques as CMP. Larger wafers must also be thicker to increase their resistance to warpage. Takasu wrote that the projected thickness may make the wafers too stiff to be adequately held down by vacuum chucking, mandating the use of new wafer hold-down techniques.[32] Larger wafers (and their cassettes) are heavier, and this will require automated wafer transport.

Problems associated with another aspect of larger wafers, that of incorporating dopants and oxygen uniformly throughout an ingot are being addressed in several ways. To increase axial uniformity various procedures have been considered, including the use of double crucibles,[34] continuous liquid feed (CLF) systems,[35] and magnetic Czochralski growth (MCZ).[35,36]

The *double crucible* (Fig. 1-25) uses a large outer crucible and a smaller, low aspect-ratio (wide and shallow) inner crucible. A connection between them is maintained by a capillary tube. By feeding the inner crucible from a larger reservoir in which a uniform doping level is maintained, the doping level in the inner crucible can be held more constant. In this manner, an improved uniformity in the axial doping level can be achieved. In addition, thermal convection in the melt of the inner crucible is reduced as a result of its low-aspect ratio dimensions. Thermal fluctuation effects are thus also less pronounced. The *CLF method* is similar in that additional molten silicon is supplied to the growth crucible from a second crucible (Fig. 1-25b). The feed mechanism utilizes a siphon technique, and the doping uniformity improvements stem from the same phenomena as in the double-crucible method. Since the second crucible is out of the growth chamber, larger total charges may be possible than with the double-crucible.

Magnetic fields surrounding the crucible have been found to suppress the thermal convection in the melt. Lenz's law states that when a conductor crosses magnetic field lines, a current is induced. The magnetic field generated by the induced current opposes the conductor motion. This principle also applies to the silicon melt in a magnetic field. The resultant

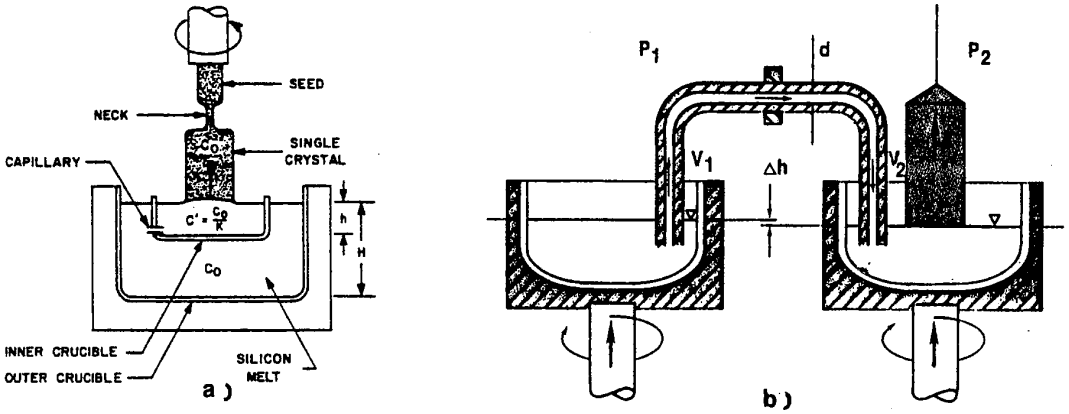


Fig. 1-24 (a) Schematic representation of a double-crucible arrangement.[17] Reprinted with permission of the Electrochemical Society; (b) Concept of continuous liquid feed furnace.[35] Reprinted with permission of Solid State Technology, published by PennWell.

suppression of the thermal convection brings about several benefits. First, smaller temperature fluctuations occur at the crystal-melt interface, and thus ingot diameter and microscopic resistivity variations due to remelt are smaller. Second, the transport of oxygen from the crucible walls to the solidification interface is reduced, and less oxygen is incorporated into the ingot (7 ppma for MCZ, versus 13–19 ppma for conventional CZ silicon). With lower oxygen content oxygen-associated defects (precipitates and stacking faults) are reportedly suppressed. Various MCZ have been developed and are being used effectively to control the oxygen in 200-mm wafers.[35] The oxygen level in 200-mm wafers is 25-35 ppma (new ASTM), and can be controlled to ±2 ppma with a radial uniformity better than ±5%.[40] It has been predicted that MCZ will be a major technology in growing 300-mm wafers in order to achieve the required oxygen control.[42]

Wafer specifications will continue to become more stringent. Increased flatness will be needed to successfully perform submicron lithography. Controlled wafer curvature[32] (concave or convex) may need to be specified depending on the specific types of chucks used to hold down wafers during lithography steps. A technique for determining backside particle existence will be needed, since particles on the backside of a wafer can impact wafer flatness. The concentration of C and O of every wafer will be specified more precisely and a maximum 2 ppma variation in O may be needed to control intrinsic gettering procedures. Wafer cleanliness and front-surface particle control will become even more important. A larger percentage of 300-mm production wafers may be epitaxial wafers, as they offer improved gate-oxide integrity yields.

REFERENCES

1. W.C O'Mara, R.B. Herring, and L.P. Hunt, Eds., *Handbook of Semiconductor Silicon Technology*, Noyes Publications, Park Ridge, NJ, 1990.
2. W. Shockley, "The Theory of *pn* Junctions and Transistors," *Bell Syst. Tech. J.*, **28** 435 (1949).
3. R.A. Laudise, *The Growth of Single Crystals*, Prentice-Hall, Englewood Cliffs, N.J., 1970.
4. G.K. Teal, *IEEE Trans. Electron Dev.*, **ED-23**, 621 (1976).
5. L.D. Crossman and J.A. Baker, "Polysilicon Technology," *Semiconductor Silicon 1977*, Electrochemical Society, Pennington, New Jersey, 1977, p. 18.
6. The American Society for Testing and Materials (ASTM), Committee F-1 on Electronics, Philadelphia, Pennsylvania.
7. J.R. Carruthers, *et al.*, "Czochralski Growth of Large Diameter Silicon Crystal-Convection and Segregation," *Semiconductor Silicon 1977*, Electrochem. So., Pennington, N.J., 1977, p. 61.
8. W.Zuhlehner & D.Huber, "Czochralski Grown Silicon" in *Crystals 8*, Springer-Verlag Berlin 1982.

9. T. Abe, "Crystal Fabrication," in [N.G. Einspruch and H. Huff, Eds.], *VLSI Electronics-Microstructure Science*, Vol. 12, Academic Press, Orlando, Fla. (1985), Chap. 1, p. 3.
10. W.C. Dash, *J. Appl. Physics*, **29**, 705, 736 (1958), *ibid.*, **30**,459 (1959).
11. S.N. Rea, "Czochralski Silicon Pull Rates," *J. Cryst. Growth*, **54**, 267 (1981).
12. H. Kolker, "Behavior of Nonrotational Striations in Si," *J. Cryst. Growth*, **50**, 852 (1980).
13. I. Chikawa & S. Yoshikawa, "Swirl Defects in Silicon Crystals," *Solid State Tech.* **23** 65 (1980).
14. W.R. Runyan, *Silicon Semiconductor Technology*, McGraw-Hill, New York, 1965.
15. J.A. Burton, R.C. Prim, and P. Slichter, *J. Chem. Phys.*, **21**, 1977 (1953).
16. M.H. Liepold, T.P. O'Donnell, and M.A. Hagan, "Materials of Construction for Silicon Crystal Growth," *J. Cryst. Growth*, **40**, 366 (1980).
17. K.E. Benson, W. Lin, and E.P. Martin, "Fundamental Aspects of Czochralski Silicon-Crystal Growth for VLSI," *Semiconductor Silicon 1981*, Electrochem. Soc., Pennington, NJ, 1981.
18. J.R. Patel, "Current Problems on Oxygen in Silicon," *Semiconductor Silicon 1977*, Electrochemical Society, 1977, p. 189.
19. L. Jastzerbski, "Origin and Control of Material Defects in Silicon VLSI Technologies: An Overview," *IEEE Trans. on Electron Devices*, **ED-29**, No. 4, April 1982, p. 475.
20. P. Stallhofer and D. Huber, "Oxygen and Carbon Measurements on Silicon Slices by the IR Method," *Solid State Technology*, August 1983, p. 233.
21. K.V. Ravi, *Imperfections and Impurities in Semiconductor Silicon*, Wiley, New York, 1981.
22. H.G. Kramer, "Float-Zoning of Semiconductor Silicon," *Solid State Technol.*, Jan. '83, p. 137.
23. J.W. Moody and R.A. Frederick, "Developments in Czochralski Silicon Crystal Growth," *Solid State Technology*, August 1983, p. 221.
24. M. Watabe, *et al*, "Oxygen-Free Silicon Single Crystal from Silicon-Nitride Crucible," *Semiconductor Silicon*, Electrochem. Soc., Pennington, N.J., (1981), p. 126.
25. R.S. Swaroop, "Advances in Silicon Technology for the Semiconductor Industry, Part II," *Solid State Technology*, July 1983, p. 97.
26. A. Bonora, "Flex-Mount Polishing of Si Wafers," *Solid State Technology*, **20**, 55 (1977).
27. N. Jackson, "Materials and Technology of Wafering," *Solid State Technol.*, July 1985, p. 107.
28. R.L. Lane, "ID Slicing Technol. for Large Diameters," *Solid State Technol.*, July 1985, p. 119.
29. Semiconductor Equipment and Materials Institute (SEMI), Mountain View, California.
30. I.A. Morris, "Wafer Identification Increases Yields," *Micro. Manuf. and Test*, Sept. 1985, p. 50.
31. J. Scaroni, "A Close Look at Laser Marking of Silicon Wafers," *Solid State Tech.*, July '97 p. 245
32. S. Takasu, "Silicon Wafers for VLSI Technology," *VLSI Technology 1983*, Electrochem. Soc., Pennington, N.J. (1983), p. 490.
33. J.E. Lawrence and H.R. Huff, "Silicon Material Properties for VLSI Circuitry," in *VLSI Electronics*, Vol. 5 (N.G. Einspruch, Ed.) Academic Press, Orlando, Fla. (1982), p. 51.
34. *ibid*, Ref. 17.
35. G. Feigl, "Recent Advances in CZ-Si Crystal Growth, "*Solid State Technology*, Aug. 83, p. 121.
36. T. Suzuki, N. Isawa, Y. Okubo, and K. Hoshi, "CZ Silicon Growth in a Transverse Magnetic Field," *Semiconductor Silicon 1981*, Electrochem. Soc., Pennington, N.J., 1981, p. 90.
37. *ibid*, Ref. 39, p. 404.
38. T. Abe, *et al*, "The Characteristics of Nitrogen in Si Crystals," *VLSI Science and Technology*, Electrochem. Soc., Pennington, N.J, (1985), p. 543.
39. S. Kishino, "Material Characterization for VLSI Applications," *VLSI Science and Technology*, Spring, 1985, Electrochemical Society, 1985, p. 399.
40. R.N. Thomas *et al.*, "Melt Growth in Large Diameter Wafers: Part II," *Solid State Technology*, April 1990, p. 163.
41. H. Huff, "Silicon Wafers Engineered for VLSI Circuits," *Semiconductor Intntl.*, July 1985, p. 82.
42. H. Komiya, "300 mm Technology: Current Status and Future Prospect," *Dig. of Tech Papers*, 1997, Symp on VLSI Tech., p.2.

43. D. Biddle, "Characterizing Semiconductor Wafer Topography," *Microelectroncs Manuf. and Testing*, March 1985, p. 15.
44. H. Huff and R.F. Holt, "Computer Automated IC Manufacturing Demands on VLSI Wafers," *Solid State Technology*, September 1985, p. 193.
45. *ibid.*, Ref. 7.
46. K.M. Kim "Digital Control of CZ Silicon Crystal Growth" *Solid State Technol.* Jan. 1985 p. 165.
47. R. Iscoff, "Crystal Growing & Fabrication Equipment," *Semiconductor Intl.*, Nov. 1983, p. 68.
48. P. Burggraaf, "Silicon Crystal Growth Trends," *Semiconductor International*, Oct. 1984, p. 56.
49. W.R. Thurber, R.L. Mattis, and Y.M. Liu, "Resistivity Dopant Density Relationships for Si," *Semicon. Characterization Techniques*, Electrochemical Society, Pennington, N.J., 1978, p. 81.
50. W.A. Baylies, "Review of Flatness Effects in Microlithography," *Solid State Technol.* **24** 1981.
51. W. Helgeland, *et al.*, "CZ Crystal Growth into the 90's," *Microel. Mfg. & Test.*, May 1986. p.1.
52. W.H. Reed, "Large Diameter Wafer Update," *Semiconductor Internatl,* November 1994, p. 140.
53. D. Rose, "Wafer Makers Challenged by Many Aspects of 300 mm," *Solid State Technology,* February 1996, p. 79.
54. D. Anderson, "Stoking the Productivity Engine with New Materials and Larger Wafers," *Solid State Technology*, March 1997, p. 57.
55. Rose's Annual Forecast for Fab Materials, *Semiconductor International*, April 1995, p.17.
56. G.R. Fisher, "Challenges for 300-mm Polished Wafer Manufacturers," *Semiconductor International,* September 1998, p. 97.

PROBLEMS

1. (a) The lattice constant of Si is 5.43Å. Find the nearest neighbor distance in a diamond lattice which has this lattice constant. (b) Calculate the density of Si from the lattice constant, the atomic weight, and Avogadro's number. Compare the results with those given in Table 1-1.
2. Sketch a view down a <110> directon of a diamond lattice.
3. Show that in fcc lattice, each atoms has 12 nearest neighbor atoms.
4. Show that the maximum fraction of the unit cell which can be filled by hard spheres in the simple cubic, fcc, and diamond lattices are 0.52, 0.74, and 0.34, respectively.
5. Label the planes illustrated in Fig. P1.
6. Determine the number of atoms in the (100), (110), and (111) planes of a Si crystal.
7. A Si crystal doped with boron is to have a resistivity of 4.0 Ω-cm when one-half of the ingot is grown. If the Si charge weighs 60 kg, how many grams of boron-doped Si with a resistivity of 0.02 Ω-cm should be introduced into the melt to achieve the desired ingot resistivity.
8. Plot the doping distribution of (a) arsenic, and (b) phosphorus at distances of 20, 40, 60, 80, and 90 cm from the seed in 100 cm long ingot that has been pulled from a melt with an initial doping concentration of 10^{16} cm^{-3}.
9. The seed crystal used in CZ growth is "necked down" to ~3 mm (Fig. 1-8b), to help produce a dislocation-free state in the pulled ingot. Using the yield strength of silicon ($2x10^6$ g/cm^2), calculate the maximum weight of an ingot that can be supported by such a neck. For ingots of 150, 200, and 300 mm in diameter, calculate their length for this weight.
10. Using the maximum growth rates shown in Fig. 1-9b, calculate the time required to pull a 100 mm diameter, and a125 mm diameter ingot of 1 m in length.
11. Explain the difference between *bow* and *TLV*.

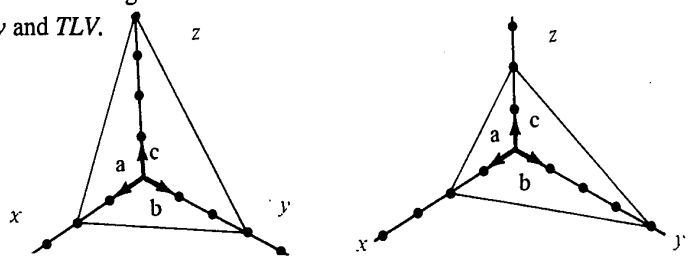


Fig. P1

Chapter 7

SILICON EPITAXIAL GROWTH and SILICON ON INSULATOR

The *epitaxial* growth process provides an excellent means of growing a thin film of single crystal material upon the surface of a single crystal substrate. The term *epitaxy* is derived from two Greek words meaning "arranged upon." If the grown film is the same material as the substrate, the process is termed *homoepitaxy*. Homoepitaxy is generally referred to as *epitaxy* or simply *epi* and these terms will be used interchangeably throughout this text. Homoepitaxy of single crystalline silicon is the most important technological use of epitaxy, and is the primary subject of this chapter.

On the other hand, if the epitaxial film is grown on a substrate that is chemically different, the process is then termed *heteroepitaxy*. The principal applications of heteroepitaxy are in the deposition of silicon on sapphire (SOS) and in the fabrication of heterojunction semiconductor devices. When a heteroepitaxial deposition is made on a substrate that has a lattice constant very different than that of the film, a strained layer can form near the film/substrate interface. If the deposited film is thin enough, the strain can be accommodated elastically, but if the film is too thick the strain becomes relaxed through the generation of defects (e.g., misfit dislocations). These defects are known to propagate from the interface into the device region where they can adversely affect device properties. Conversely, if the lattice mismatch is small (as in the case of the growth of silicon-germanium alloys on silicon substrates), the strain is accommodated elastically. Under these conditions the epitaxial growth is termed *pseudomorphic* growth. Pseudomorphic growth is the method used in the formation of silicon-heterojunction-bipolar transistors (HBTs). Figures 7-1a, 7-1b, and 7-1c show a schematic representation of homoepitaxy, heteroepitaxy (strain relieved), and pseudomorphic epitaxial growth, respectively.

Epitaxial growth can be achieved by depositing the film directly from either the vapor phase,

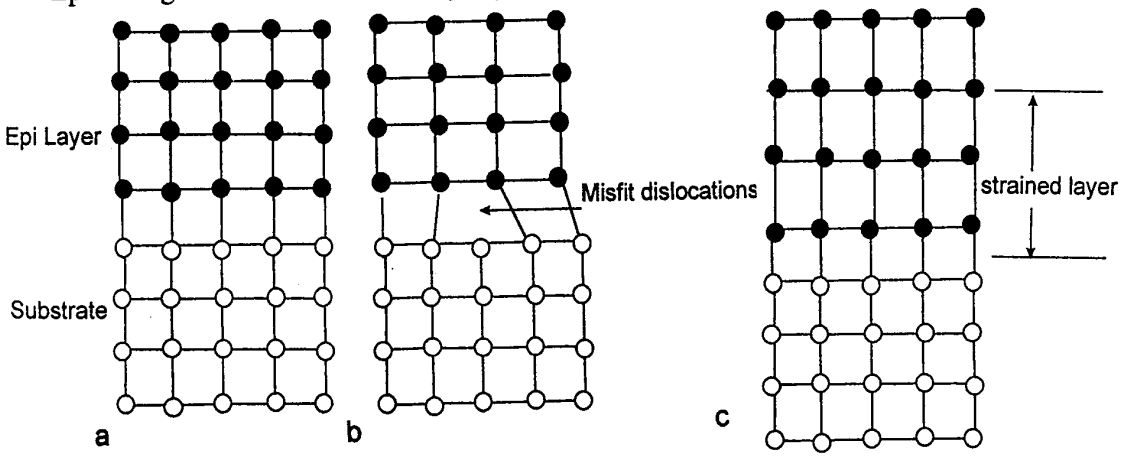


Fig. 7-1 Schematic of epitaxial structures: a) homoepitaxy; b) strain relieved heteroepitaxy (e.g., silicon-on-sapphire [SOS]); and c) psuedomorphic epitaxy.

Exhibit 1014 - Page 57 of 218

the liquid phase, or the solid phase. These techniques are termed *vapor-phase epitaxy* (VPE), *liquid-phase epitaxy* (LPE), and *solid-phase epitaxy* (SPE), respectively. VPE is the predominant method used in silicon processing, since it provides a means of achieving excellent control of the film thickness, impurity concentration, and crystalline perfection. A disadvantage of VPE is that it is requires relatively high temperatures (800–1150°C) in order to realize the necessary crystalline perfection. Such high-temperature processing can lead to the occurrence of *out-diffusion* and *autodoping*. These phenomena place a limitation on the ability to control the doping level in the epi layer and are discussed in detail in Section 7.4.4 of this chapter.

LPE has found its widest use producing epitaxial layers of III-V compounds (e.g., GaAs and InP). LPE, however, occurs during the re-freezing of molten layers of silicon on a silicon surface during the laser melting GILD process (see Chap. 9). SPE occurs during the crystalline re-growth that takes place during the low-temperature annealing of amorphous layers formed as a result of damage during high-dose ion implantation (see Chap. 10).

This chapter covers: 1) device applications of epitaxy; 2) fundamentals of epitaxial film growth; 3) epitaxial process considerations, including silicon precursors and doping effects; 4) defects in epitaxial layers; 5) pattern effects; 6) low-temperature epitaxy; 7) selective epitaxy; 8) epitaxial film characterization; and 9) silicon on insulators (SOI). The fundamentals of the kinetics and gas dynamics of CVD (epitaxial growth is a CVD process) are covered in detail in Chapter 6. In this chapter only the relevant results are repeated.

7.1 THE DEVICE APPLICATIONS OF EPITAXY

In the early days of semiconductor fabrication, epitaxial deposition was used to improve the performance of bipolar transistors that were then commonly used. By growing a lightly doped n-type epitaxial layer over a heavily doped n^+ substrate the bipolar device could be optimized for high breakdown voltage of the collector-substrate junction, while preserving a low collector resistance. The low collector resistance was required to improve the device operating speeds at moderate currents. Figure 7-2 depicts a cross-section of a high-speed oxide-isolated bipolar transistor showing the presence of the epitaxial layer. A key feature of bipolar epitaxy is that the film is deposited over patterned, n^+-doped diffused layers called *buried layers* or *sub-collectors*. BiCMOS devices are also be fabricated in epitaxial films deposited over buried layers.

More recently, epitaxial processes have been used to fabricate CMOS integrated circuits. In CMOS circuits the complete device is built in a thin (2–4 μm), lightly-doped p-type (or in some cases, intrinsic) epitaxial layer that is deposited over a uniformly heavily p^+-doped substrate. Figure 7-3 shows a cross section of a twin-well CMOS device depicting the epi layer.

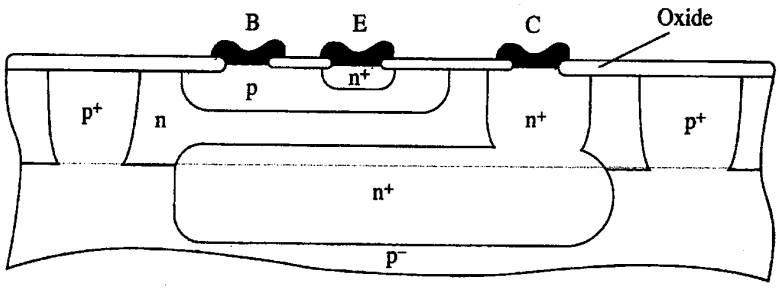


Fig. 7-2 Cross-section of a bipolar transistor built in an epitaxial layer. The light line indicates the boundary of the deposited n-type epi layer. Note that the n^+-buried layer extends into the epi layer as a result of autodoping and out-diffusion.

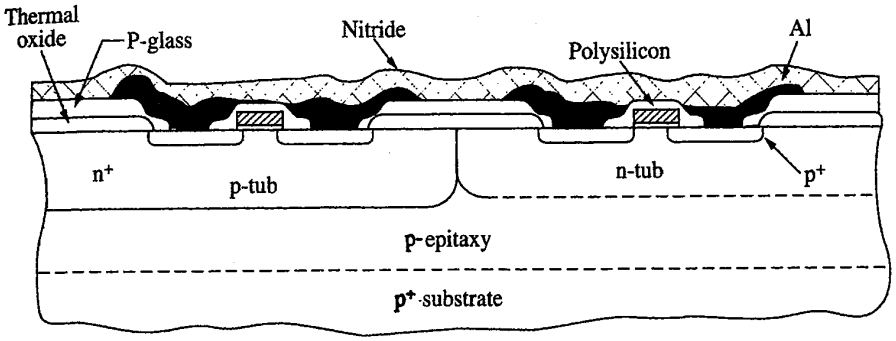


Fig. 7-3 Cross-section of a twin-well CMOS structure built in a p^- epitaxial layer and a p^+ substrate. Note that the p^+-region extends into the epi layer as a result of autodoping and out-diffusion.

In the next section we discuss the reasons for using epitaxial silicon rather than *polished bulk wafers (PBW)* for constructing advanced CMOS devices. It is interesting to note that IC manufacturers usually purchase CMOS epitaxial wafers directly from the wafer suppliers, while for bipolar ICs they process the epitaxy themselves. The reason is that a masking layer is required to produce the buried layer prior to the epi in the bipolar process.

7.1.1. Why Are CMOS Epitaxial Wafers Used?

CMOS epitaxial wafers, when compared to polished bulk wafers provide: a) improved latch-up immunity; b) layers of silicon free of SiO_x precipitates, COPs, or vacancy condensates (all of which are formed during crystal growth); and c) a silicon surface that is smoother than that found on PWBs, with minimal surface damage.

Latch-up in CMOS circuits is a phenomenon that causes the formation of a low resistance path between the supply voltage and ground. As a result, it is possible for the circuit to draw an unusually high amount of current. Such current draw can cause the circuit to cease functioning or even to "burn-up." The device physics underlying latch-up is beyond the scope of this text, but is covered in detail in Vol. 2 of this series. Although there are process and design techniques that can minimize latch-up, the use of epitaxial silicon has become the "standard" approach to fabricating ULSI CMOS microprocessor circuits because it provides latch-up protection.

Polished bulk wafers are known to include SiO_x precipitates, COPs, and vacancy condensate defects at or near the wafer surface. Moreover, despite the best polishing techniques, the wafer surface can retain polishing defects such as micro-damage and surface roughness. Surface defects are known to diminish the gate-oxide integrity and increase junction leakage. Epitaxial layers, on the other hand, do not contain oxygen (hence no SiO_x precipitates), COPs, or vacancy condensates, since these defects only form during crystal growth (during which oxygen is dissolved in the as-grown Si crystal, see Chaps 1 and 2). Since epi is deposited on a polished wafer, the new surface that forms is generally free of polishing micro-damage and surface roughness. As a result, CMOS devices produced in epitaxial material have superior dielectric integrity and junction leakage currents, compared to similar devices made on polished wafers.

The use of epitaxial wafers in the fabrication of microprocessors is now standard, while its use in DRAM devices is rather limited. There are two reasons for this. First, the gate oxide thickness required for DRAM devices is generally thicker than that required for microprocessor devices, even for the same "generation" of technology. These thicker gate oxides are less susceptible to the presence of surface defects and do not benefit as much from the advantages of the epi layer. Second, the cost to purchase a 200-mm epitaxial wafer is about 1.5–1.8 times that of a polished wafer. Since low manufacturing costs for DRAMs are vital for profitability, the

Table 7-1 APPLICATIONS OF EPITAXY

Application	Thickness (μm)	Resistivity Range (ohm-cm)
Bipolar devices	0.8–1.6	0.5–4.0
CMOS devices	2–4	2–10
BiCMOS devices	3–5	2–20
High voltage devices	5–50	50–80

added cost per wafer is often too high to justify its use. However, the use of epi wafers for DRAM manufacturing is under serious consideration as thinner gate oxides become necessary. When the yields on devices fabricated in bulk wafers no longer meets the requirements for DRAM manufacturing, the industry may be compelled to change to epitaxial wafers. On the other hand, the DRAM manufacturers will try to find alternative processes to extend the life of bulk wafers, rather than pay the added cost associated with the use of epitaxial wafers.

Table 7-1 shows some of the applications of silicon epitaxy along with the epi thickness and resistivity ranges that are commonly used. Table 7-2 (which is extracted from the SIA roadmap) shows the epi layer requirements for the most advanced CMOS devices.[1] The two key points illustrated in that table are: 1) thinner epi layers are required (i.e., down to 1 μm); and 2) the structural defect density must be reduced by more than a factor of 2 from its current level. These two discordant requirements represent a significant challenge, since thinner films require lower growth temperatures, and lower temperatures can lead to increased defect density.

7.2 GROWTH OF EPITAXIAL LAYERS

In this section an atomic mechanism that has been proposed to explain epi film growth is first discussed. Then the various chemicals precursors used in conventional epitaxial growth are described. The kinetics of CVD were covered in detail in Chap. 6, and only the results pertinent to epitaxial growth will be repeated here.

7.2.1 Atomistic Model of Film Growth

Whatever *precursor* (source gas) is used and whatever chain of reactions occur (see next section) the final reactant species must be adsorbed on the growing surface where the reaction takes place. The reaction produces silicon and some by-product. In order for film growth to occur, the silicon atoms must remain adsorbed on the surface and the by-products desorbed. The adsorbed silicon atoms are referred to as *adatoms*. If an adatom is in a position such as A in Fig. 7-4, one of several events can take place. If the adatom A remains in its position other silicon atoms can become attached to it and a small silicon cluster or island can form. If numerous clusters form and begin to grow, it is likely that as they merge, a highly defected or even polycrystalline film will form. However, as it turns out, adatoms in position A are quite mobile at the deposition

Table 7-2 EPITAXIAL LAYER REQUIREMENTS vs. TECHNOLOGY NODE
Extracted from NTRS, SIA (1997)[1]

First year product shipment	1997	1999	2001	2003
Feature size (nm)	250	180	150	130
Layer thickness (μm)	2–5	2–5	2–4	2–4
Thickness tolerance (±%)	5	4	4	4
Layer structural defect density (m^2)	≤3.3	≤2.9	≤2.6	≤2.3

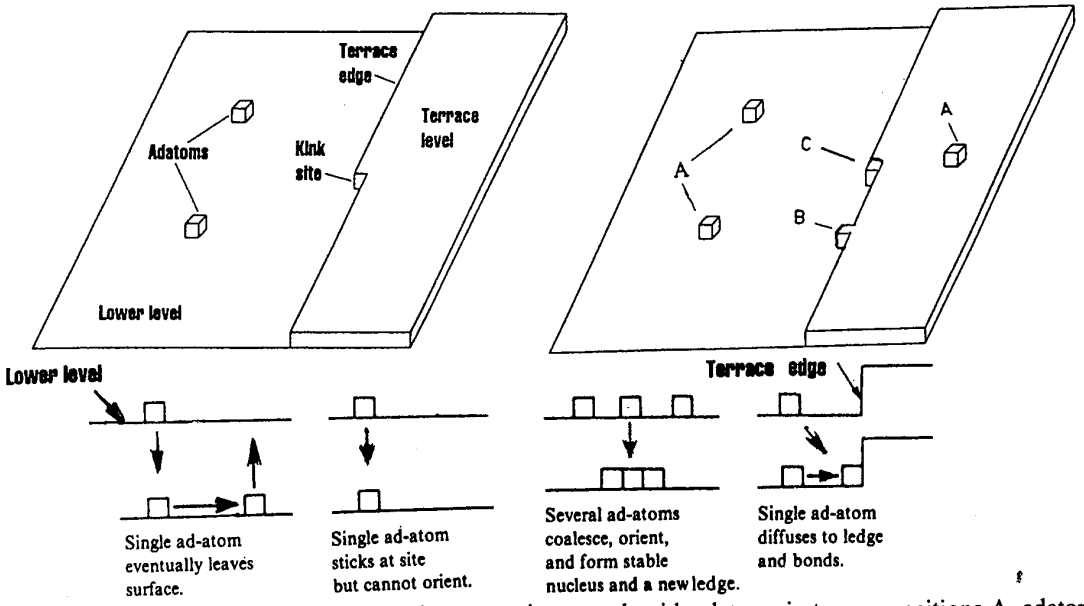


Fig. 7-4 Epitaxial growth model showing step-wise growth with adatoms in terrace positions A, adatoms at steps (or terrace edges) B, and adatoms at stable kink positions C.

temperatures and are likely to migrate along the surface. If a migrating adatom reaches a position along a growing *step* (or *terrace edge*), such as position B in Fig. 7-4, the adatom is more likely to remain in place (since atoms at position B are more energetically stable than at positions A). The stability arises from the fact that more Si-Si bonds have already formed when the adatom is at position B. The most energetically favored site for the adatom, however, is the so-called *kink position* (position-C in Fig. 7-4). When the adatom arrives at such a kink position, one-half of the Si-Si bonds can be established and further migration is thereafter unlikely to occur. For growth to continue, additional adatoms must migrate to kink positions and become incorporated in the growing film. In this respect, film growth proceeds by lateral movement (2-dimensional) of the steps on the lattice surface. After one layer is completed a second layer grows. Since epi proceeds by this lateral growth, an impurity sitting on the wafer surface can impede the motion and produce a stacking fault or dislocation defect in the film. For this reason, extremely clean silicon surfaces are required to achieve high-quality epi growth.

If adatom migration is suppressed during growth, a polycrystalline film is more likely to be deposited. Atom mobility is restricted by high deposition rates and low temperatures. At any particular deposition temperature there exists a maximum deposition rate above which polycrystalline films are deposited, but below which single crystal epitaxial films are formed.[2] Figure 7-5, shows a plot of the effect of growth rate and temperature on the formation of single (monocrystalline) or polycrystalline films. From that curve we observe that monocrystalline growth is favored at high temperatures and low growth rates, and polycrystalline deposition is favored under conditions of high growth rates and low temperatures. The maximum deposition rate at which single-crystal growth can be maintained is found to increase exponentially with temperature. When these results are plotted on an Arrhenius plot (Fig. 7-5) a straight-line dependence is obtained (indicative of a thermally activated process). The value of the activation energy extracted from this curve is ~5 eV. These results may be explained as follows. At high growth rates there is not sufficient time for the adatoms to migrate to the kink sites, thereby polycrystalline growth is favored. As the temperatures increase, however, the surface migration rate also increases, allowing time for silicon to reach the kink sites before additional adatoms form other clusters. Once at the kink site, the film continues to grow by lateral motion. The

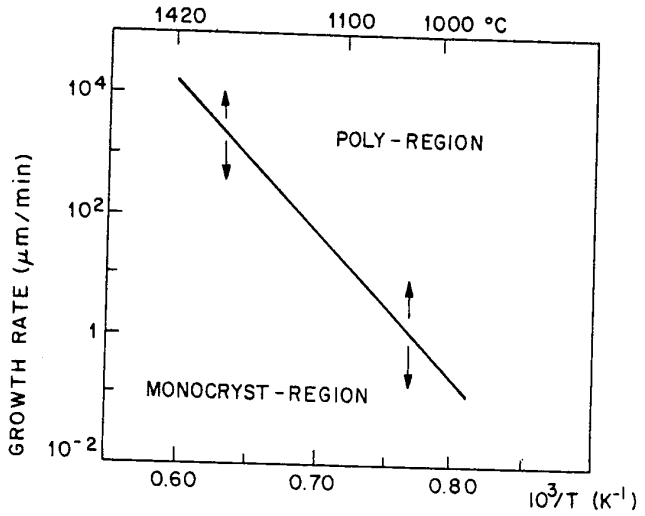


Fig. 7-5 Maximum growth rate for formation of single crystalline silicon as a function of temperature.[7] Reprinted with permission of North Holland Publishing Co.

activation energy of ~5 eV is comparable to that of silicon self-diffusion. Consequently, it has been proposed that the mechanisms for epitaxial single-crystal formation and silicon self-diffusion are the same.

7.3 CHEMICAL REACTIONS USED IN SILICON EPITAXY

There are four major chemical sources (called *precursors*) that are used commercially for epitaxial silicon deposition. They are: 1) silicon tetrachloride ($SiCl_4$) termed *sil tet*; 2) trichlorosilane ($SiHCl_3$) called *TCS*; 3) dichlorosilane (SiH_2Cl_2) called *DCS*; and 4) monosilane (SiH_4). Each precursor has properties that makes it attractive for use under particular deposition conditions and film requirements. $SiCl_4$ has historically been the most widely-used precursor, as well as the most widely studied. However this material finds little application for modern epitaxial deposition. Requirements for thinner epitaxial layers and lower deposition temperatures have instead brought about the increased use of SiH_2Cl_2, $SiHCl_3$, and SiH_4.

The growth rate of silicon epitaxial films depends on several parameters including: a) the precursor used; b) the deposition temperature; c) the mole fraction of reactants present in the vapor phase; and d) the reactor pressure. During epi deposition the precursors are usually introduced to the reactor along with a large excess of hydrogen. The hydrogen serves both as a reducing agent and as a diluent or carrier gas. Figure 7-6 shows the silicon growth rate as a function of temperature for deposition from $SiCl_4$, $SiHCl_3$, SiH_2Cl_2, and SiH_4.[3] There are several notable findings portrayed in this figure. First, the deposition rate depends on which silicon precursor is being used. SiH_4 has the highest deposition rate at all temperatures, followed by SiH_2Cl_2, $SiHCl_3$, and finally $SiCl_4$. Second, two growth regimes are observed. In the high temperature regime (where the deposition rate is relatively insensitive to temperature, namely region B of Fig. 7-6), the controlling mechanism is *gas-phase mass transport* (see Chap. 6) and is strongly dependent on the reactor geometry and gas flow. At lower temperatures, the growth rate is strongly dependent on temperature, indicative of a *thermally-activated process* (region A of Fig. 7-6). Finally, the activation energy (corresponding to the slope of the Arrhenius plot in region A) is essentially independent of the precursor employed. This suggests that the reaction-rate-limiting mechanism may be the same in all cases. One rate limiting mechanism that has been postulated is the desorption of hydrogen from the reacting surface. Hydrogen can occupy surface sites on the silicon, and if not removed can prevent new silicon

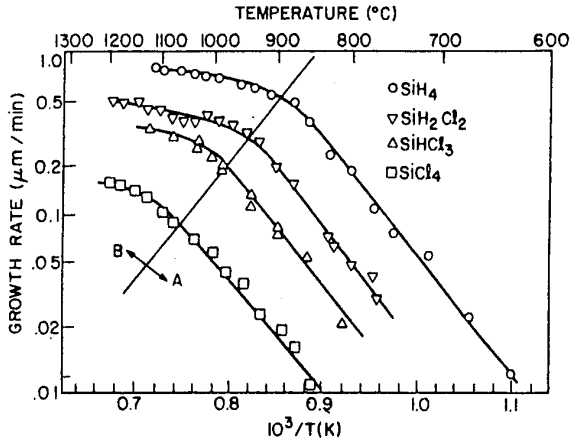


Fig. 7-6 Growth rate of silicon films as a function of temperature for various silicon sources.[13] Reprinted with permission of Philips Research Reports.

atoms from being incorporated into the growing film. This model is substantiated by a study where the growth rate was found to saturate as the DCS concentration increased when a hydrogen carrier was used, but did not saturate when a nitrogen carrier was used.[4] This would confirm that the hydrogen concentration limits the reaction. In another investigation, however, where a direct measurement of surface desorption was made, the results indicated that HCl desorption (instead of hydrogen desorption), may be the rate limiting step.[5]

It should also be noted that the transition temperature where the growth switches from being reaction-rate-controlled to being mass-transfer-controlled depends on: a) the mole fraction of precursor; b) the type of deposition system being used; c) the gas flow rate; and d) the precursor used. Precursors with more Cl require higher temperature to reach the mass-transport-controlled growth regime. Since production depositions are usually performed in region B (where the effect of temperature variations across the wafer may be minimized), it is desirable to use a precursor with less Cl, unless extremely high temperatures can be tolerated.

Overall reactions can be written for each of the precursors, but they do not represent what is happening on the atomic level. For example, the overall reaction for the hydrogen reduction of $SiCl_4$ can be written as:

$$SiCl_4\,(vapor) + 2H_2\,(vapor) \leftrightarrow Si\,(solid) + 4HCl\,(vapor) \qquad \textbf{(7.1)}$$

As will be discussed, the sub-reactions making up this reaction are reversible, and under appropriate thermodynamic conditions the reverse or etching reaction can dominate.

Overall reactions, by their nature, do not provide a complete picture of the reaction sequence, the species present in the gas phase, or the species adsorbed on the substrate. Ban and coworkers studied the thermodynamics of the Cl-H-Si system.[6,7,8,9] In their studies they used a starting mixture of H_2 with either SiH_2Cl_2, $SiHCl_3$, or $SiCl_4$. Using a mass spectrometer to determine the active species present in the reaction atmosphere, they found constituents in the chamber other than those initially introduced. The new species included HCl, and $SiCl_2$. The presence of the compounds implies that a large number of chemical reactions can be taking place simultaneously in the systems.

Figure 7-7 shows the results of a study where the concentration of the chemical species found in the vapor phase during epitaxial deposition from the H_2 reduction of $SiCl_4$ is plotted as a function of distance along the length of a horizontal reactor (using IR spectroscopy to measure these concentrations).[12] Although, only $SiCl_4$ and H_2 were introduced, four constituents were

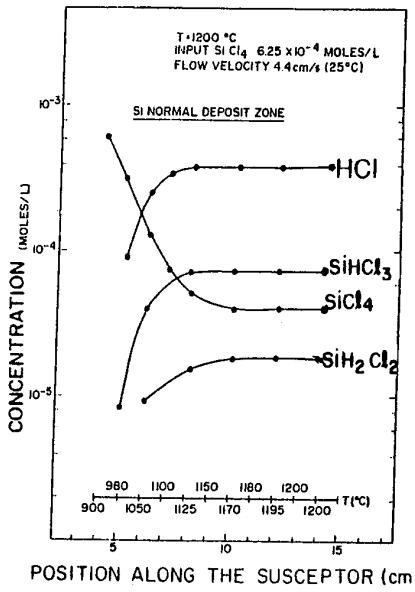


Fig. 7-7 Species present (detected by IR spectroscopy) along the length of a horizontal reactor with $SiCl_4 + H_2$ at 1200°C.[12] Reprinted by permission of publisher, the Electrochemical Society.

found to be present: $SiCl_4$; $SiHCl_3$; SiH_2Cl_2; and HCl. The $SiCl_4$ concentration was found to decrease along the length of the reactor while the concentration of the other three constituents increased. Based on these results the investigators proposed that the overall reaction could proceed as the sequence of steps shown in Eqs. 7-2 to 7-6:

$$SiCl_4\,(vapor) + H_2\,(vapor) \leftrightarrow SiHCl_3\,(vapor) + HCl\,(vapor) \quad \textbf{(7.2)}$$

$$SiHCl_3\,(vapor) + H_2\,(vapor) \leftrightarrow SiH_2Cl_2\,(vapor) + HCl\,(vapor) \quad \textbf{(7.3)}$$

$$SiH_2Cl_2\,(vapor) \leftrightarrow SiCl_2\,(vapor) + H_2\,(vapor) \quad \textbf{(7.4)}$$

$$SiHCl_3\,(vapor) \leftrightarrow SiCl_2\,(vapor) + HCl\,(vapor) \quad \textbf{(7.5)}$$

$$SiCl_2\,(vapor) + H_2\,(vapor) \leftrightarrow Si\,(solid) + 2HCl\,(vapor) \quad \textbf{(7.6)}$$

The sequence shows the presence of $SiHCl_3$ and SiH_2Cl_2 as intermediate species with the final step being the reduction of $SiCl_2$. These equations are valid for the other chlorosilane precursors, since growth could start with $SiHCl_3$ and SiH_2Cl_2 as well, and proceed to the formation of $SiCl_2$, and ultimately Si. It should be noted that all of these reactions are reversible, meaning that under the appropriate thermodynamic conditions the reactions can go to the left, which implies that silicon can also be *etched* in chlorosilane ambients. Etching is favored at both low and high temperatures, while deposition is favored at intermediate temperatures. Figure 7-8, shows the epitaxial growth rate as a function of temperature for $SiCl_4$. It should be noted that the etching reaction (corresponding to a negative growth rate) occurs at temperatures below 900°C and above 1400°C, while deposition occurs in the intermediate temperature range.

As predicted by Eq. 6-7 of Chap. 6, the growth rate of the film during epitaxial deposition should be proportional to the mole fraction Y of the reactant species in the gas phase. Figure 7- 9 shows the growth-rate dependence for silicon deposited by the hydrogen reduction of $SiCl_4$, on the mole fraction of $SiCl_4$. For this reaction, the growth rate departs from linearity at relatively low values of Y, and then starts to decrease when the Y value reaches about 0.1. As the concentration of $SiCl_4$ further increases to 0.27, the reaction reverses and silicon etching occurs.

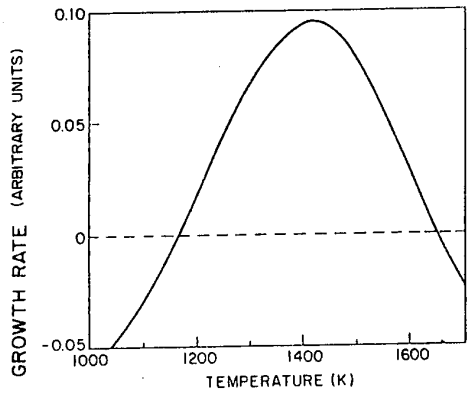


Fig. 7-8 The growth rate of variation of silicon films as a function of deposition temperatures for $SiCl_4$ + H_2.[38] Reprinted with permission of the publisher, the Electrochemical Society.

The silane (SiH_4) reaction is fundamentally different than the reaction of the chlorosilane precursors. The silane reaction proceeds by the overall reaction:

$$SiH_4\ (vapor) \rightarrow Si\ (solid) + 2H_2\ (vapor) \qquad (7.7)$$

which is not reversible at normal processing temperatures.

7.4 PROCESS CONSIDERATIONS FOR EPITAXIAL DEPOSITION

7.4.1 Silicon Precursors

Table 7-3 lists the common silicon precursors and the temperature range and growth rates where they are commonly used. The growth rates can be modified up or down by changing the precursor concentration, deposition temperature, or pressure. In addition to the four precursors discussed previously, we also consider a new precursor, disilane (Si_2H_6), which has found application in low-temperature epi growth.

Deposition from $SiCl_4$ requires high temperatures, and such temperatures can lead to excess

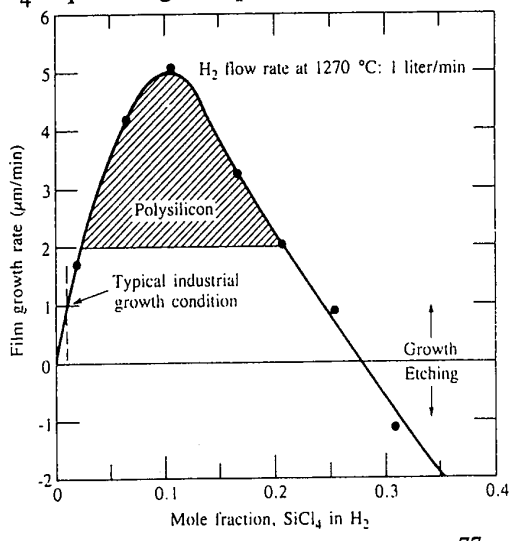


Fig. 7-9 Silicon growth rate as a function of $SiCl_4$ concentration.[77] Reprinted with permission of the publisher, the Electrochemical Society.

Table 7-3 SILICON PRECURSOR DEPOSITION CHARACTERISTICS

Silicon Source	Deposition Temperature (°C)	Growth Rate (μm/min)	Remarks
Silicon tetrachloride ($SiCl_4$) Sil Tet	1150–1250	0.4–1.5	Used in old processes. Very high temperature required for deposition.
Trichlorosilane ($SiHCl_3$) TCS	1110-1150	0.4–2.0	Easily reduced, relatively high deposition temperature required. A popular source.
Dichlorosilane (SiH_2Cl_2) DCS	1020–1120	0.4–3.0	Reasonably low temperatures. Used for selective epi growth (SEG).
Monosilane (SiH_4) - Referred to as "silane"	650-900	0.2–0.3	Used in SOS. Deposits at low temperature. Used in Low temperature Si-Ge epi. Gas phase nucleation is an issue.
Disilane (Si_2H_6)	400–600	<0.1	Used for very low temperature epi. Pyrogenic and toxic.

autodoping and outdiffusion and result in a large transition region. As a result $SiCl_4$ is primarily used in "legacy" processes that have not changed since the more suitable precursors have become available in high-purity form. $SiHCl_3$ has similar properties to those of $SiCl_4$, except deposition can be done at lower temperatures. This source has become very popular for routine epitaxial deposition. It can be used to deposit thick epitaxial layers, since it is easy to achieve deposition rates in excess of 1 μm/min. SiH_2Cl_2 is also widely used to deposit high-quality thin epi layers at even lower temperatures. It has been reported that films prepared from SiH_2Cl_2 have lower defect densities than those prepared either from SiH_4 or $SiCl_4$.[11] Device yields have also been reported to increase when using SiH_2Cl_2 (compared to films deposited at high temperatures from $SiCl_4$).[12] DCS is also the precursor of choice for selective epitaxial growth (SEG). SiH_4 is used to deposit very thin epitaxial layers at temperatures less than 900°C. The major advantage of using SiH_4 is that high deposition rates can be achieved at relatively low temperatures. Also, since no HCl is present in the reaction chamber, gettering of metallic impurities is unlikely. It is possible that the silane can react homogeneously in the gas phase at low temperatures and cause powdery deposits to appear on the reactor wall (or on the wafer). A significant advantage of SiH_4 is that it does not produce pattern shift (discussed in Sect. 7.5.3). When using SiH_4, the H_2 carrier should have very low residual oxidant concentration, since SiH_4 will react with the oxidant to form SiO_2.

7.4.2 Doping of the Epitaxial Films

A key advantage of epitaxial deposition is that the doping level in device regions can be precisely controlled. Dopants that are added to give this precise control are called *intentional dopants*,

since they are desired in the film. Unwanted dopants, however, can also become incorporated in the film during the growth process. These are called *unintentional dopants.*

7.4.3 Intentional Doping

In order to control the conductivity type and carrier concentration (electrical resistivity) of epitaxial layers, dopants are intentionally added to the reaction gases during the growth process. These dopants are introduced through their hydrides: diborane [B_2H_6] for boron; phosphine [PH_3] for phosphorus; and arsine [ASH_3] for arsenic. Ultrapure hydrides are introduced into the reactor from gas storage cylinders, where the hydrides are diluted in hydrogen to concentrations of 20–100 ppm. When the dopant gases are mixed with the main hydrogen flow in the reactor their concentration is further diluted to the parts per billion range. The dopant concentration that is ultimately incorporated in the epi film depends on the relative amount of dopant in the gas stream. Although hydrides are relatively unstable above room temperature, they are prevented from dissociating in the gas phase by the presence of large quantities of H_2 in the reactor. Caution must be exercised when handling these dopant gases since they are extremely toxic.

Unfortunately, there is no simple rule to relate the concentration of dopant atoms in the gas phase with the number of atoms incorporated into the solid Si film. As it turns out, the dopant incorporation rate depends upon substrate temperature, deposition rate, dopant mole volume relative to the precursor mole volume in the gas stream, and reactor geometry. As a result, the dopant incorporation rate must be empirically determined for each reactor and set of growth conditions. For low concentrations of dopants, the amount of dopant incorporated is generally linearly proportional to the gaseous dopant concentration. For high mole volumes of dopant, the dopant incorporation becomes sub-linear with respect to the gas phase composition. Nevertheless, epitaxial films with well-controlled doping concentrations in the range of 10^{14}–10^{20} atoms/cm^3 can be routinely achieved once the process is characterized for a particular tool. A key factor adding to the control of the dopant concentration is the use of mass flow controllers (MFCs) to accurately meter the gas flows (see Chap. 6).

The silicon deposition rate during the deposition of doped epi layers is found to depend upon the dopant type and concentration. For example, PH_3 is known to reduce the deposition rate, while B_2H_6 increases the rate.[13] PH_3, which is readily adsorbed on the surface of the growing film, dissociates to form free hydrogen, which then bonds with silicon to form stable Si-H bonds on the surface. The hydrogen can be desorbed above 400°C leaving behind phosphorus on the surface. This phosphorus passivates the surface and suppresses the growth rate of the epi layer by hindering the 2-D growth of the film. AsH_3 exhibits similar behavior, and at high concentrations can substantially reduce the epi growth rate. B_2H_6, conversely, decomposes readily on the growing surface and deposits boron. This enhances the heterogeneous decomposition of the reactive silicon species and therefore enhances the growth rate. The amount of doping incorporated in the growing film also depends on the silicon deposition rate. At low rates more doping is incorporated for the same gaseous concentration, while at high rates less is incorporated. Arsenic incorporation (from AsH_3) decreases by a factor of 4, as the growth rate is increased 0.2 to 1.0 μm/min.[14]

7.4.4 Unintentional Doping (Autodoping and Solid State Diffusion)

Most epitaxial usage requires an epitaxial deposition of a lightly-doped layer (10^{14}–10^{17} atoms/cm^3) upon a heavily-doped substrate (10^{19}–10^{21} atoms/cm^3). The substrate may be uniformly doped or consist of buried layers. During the epitaxial deposition unintentional

dopants can enter the growing film from several sources. One such source is the solid-state diffusion of dopants from the heavily-doped substrate into the growing film. This diffusion is termed *outdiffusion.* Since diffusion can be readily modeled, the concentration of dopant transported to the film can be determined from the complementary error function (which assumes that the dopant concentration at the epi interface is constant at a value of C_s). The equation used to calculate the doping level caused by outdiffusion into the epi film is given by Eq. 7-8.

$$C_e = \frac{C_s}{2} \operatorname{erfc} \frac{x}{2\sqrt{D_s t}} \tag{7.8}$$

where C_e is the concentration of dopant in the epitaxial layer, C_s is the initial concentration of dopant in the substrate, D_s is the diffusion coefficient of the substrate impurity at the deposition temperature, and t is the deposition time. Implicit in using Eq. 7.8 is the assumption that the epi film growth rate far exceeds the velocity of the diffusion front. Fortunately this is the case for practical deposition conditions.

The second type of dopant transfer is termed vapor-phase *autodoping.* Autodoping emanates from evaporation of dopants from the substrate and re-incorporation of these same dopants into the growing film. There are two types of autodoping that we consider, namely *vertical autodoping* and *lateral autodoping*, both of which are depicted schematically in Fig. 7-10.[15] Vertical autodoping is the incorporation of dopant into the growing film above the buried layer regions, while lateral autodoping is incorporation of dopant beyond the buried layer regions. In CMOS, where the substrate is uniformly doped, only vertical autodoping need be considered. Autodoping can also result from dopant evaporation from the edges and backsides of wafers, from the wafer platform (susceptor) if it is coated with heavily doped material, or from other wafers in a multi-wafer reactor. The doping from the susceptor is responsible for the so-called *memory effect* of epitaxial reactors (i.e., the effect that makes it difficult to grow layers of high resistivity immediately after growing a low-resistivity layer). Autodoping requires that evaporated dopants be trapped within the gas-phase boundary layer, otherwise they will be removed in the system exhaust. Reduced pressure epitaxy tends to minimize the boundary layer and thereby reduce autodoping of some dopants (i.e., As and P, but not B). Outdiffusion and

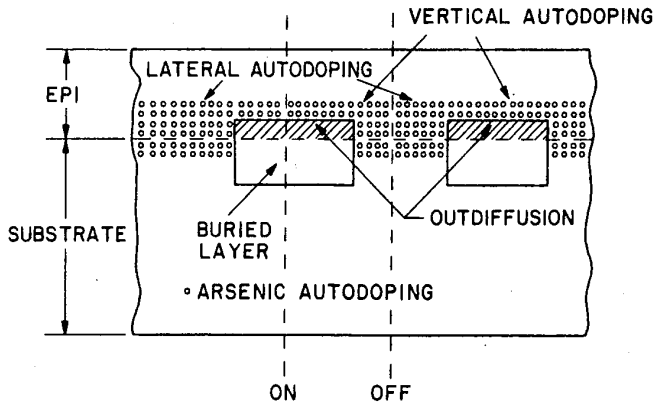


Fig. 7-10 Cross-section of an epitaxial wafer (with an arsenic buried layer) showing the effects of lateral autodoping, vertical autodoping, and outdiffusion.[14] Reprinted with permission of the publisher, the Electrochemical Society.

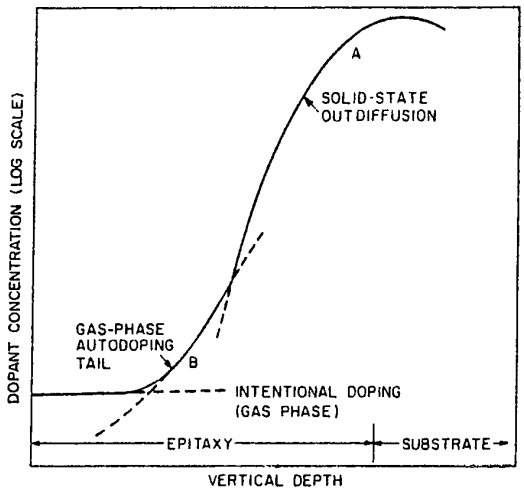


Fig. 7-11 Plot of the expected doping profile of an epitaxial layer grown over a heavily doped substrate showing the several regions of autodoping.[15] Reprinted with permission of the publisher, the Electrochemical Society.

autodoping produce a *transition layer* in which the doping concentration lies between the substrate and epi layer values. Without outdiffusion and autodoping the carrier concentrations would abruptly change from the highly-doped substrate to the lightly-doped epi layer.

The transition layer between the heavily-doped substrate and the lightly-doped epi regions is shown in Fig. 7-11.[16] The region of the epi film close to the substrate shows higher doping levels than intended, since dopant has diffused out of the substrate into the growing film during deposition. (Region A in Fig. 7-11). However, the rapid growth of the film relative to the movement of the diffusion front quickly limits this effect. Thereafter, the transition zone becomes controlled by the incorporation of dopants from the vapor phase. When the dopant atoms in the gas phase evaporated from the substrate exceed those intentionally introduced, the autodoping tail (Region B in Fig. 7-11) develops. Since the high-concentration substrates quickly become covered with more lightly-doped material, the autodoping effect ceases, and the desired concentration is reached. If the whole substrate is heavily doped, however, evaporation from the wafer edges and backside can continue after the front side is sealed. Nevertheless, even this effect decreases with time, since solid-state diffusion from the wafer interior must supply dopant to the surface for evaporation. The extent of the autodoping tail depends on deposition temperature, silicon source, growth rate, reactor geometry, and pressure.

Outdiffusion and autodoping impose a limit on the minimum epitaxial thickness and doping level, and therefore processes that minimize unintentional doping are preferred. The following procedures have been proposed to minimize such effects.

1. To reduce boron autodoping (CMOS applications) use the lowest possible deposition temperature commensurate with good quality films and reasonable deposition rates. By using DCS or TCS rather than $SiCl_4$, deposition can be made at a lower temperature. Low-temperature deposition is not applicable to arsenic autodoping since, in that case, autodoping increases at lower temperatures.

2. For n-type substrates, buried-layer dopants that have low vapor pressure and low diffusivity should be used. For instance, Sb should be selected instead of As or P (because the former has a very high vapor pressure, and the latter has a high diffusivity). It is, however, more difficult to achieve high doping levels in Sb substrates.

3. The backside of heavily-doped substrates (and/or the susceptor) should be sealed with lightly-doped silicon to minimize vaporization.

4. The epitaxial system could be operated at reduced (i.e., sub-atmospheric) pressures to minimize lateral autodoping. At reduced pressure the diffusivity of gas molecules in the gas phase is greater than at atmospheric pressure. Thus, the evaporated atoms can diffuse rapidly to the main gas flow and be swept out of the reactor before becoming incorporated in the film. This method works well for phosphorus and arsenic, but not for boron.

5. A buried layer that is ion-implanted to reduce the substrate surface concentration could be used.

6. A *cap-purge-grow* sequence can be employed (in which a very thin layer of undoped film is first grown over the buried layer or substrate to seal in the dopant). This should be followed by a high flow gas purge to remove dopants in the gas phase, and then by the introduction of the desired dopant during the remainder of the epi growth.

7. The use of a high-temperature HCl etch should be avoided, or if absolutely needed, a low temperature purge after the HCl etch should be utilized to insure that the dopant evolved during the etching is swept out of the system.

7.5 DEFECTS IN EPITAXIAL FILMS

During the epitaxial deposition process it is possible to induce crystalline defects in the growing film. These defects can result from a variety of factors such as: a) reactor contamination; b) inadequate substrate cleaning; and c) crystal imperfections in the substrate. The nature and types of crystalline defects in silicon and their effect on device properties are discussed in more detail in Chap. 2. Here we consider how crystalline defects arise in epitaxial Si films.

7.5.1 Wafer Preparation

A noteworthy concept propounded for many years is that the crystalline quality of the epi layer cannot surpass that of the substrate. This of course relates to the quality of the wafer surface, since defects below the surface should not affect the epi layer. Accordingly, great care must be exercised in substrate preparation. Substrate preparation can include such steps as wafer intrinsic gettering, *ex situ* wafer cleaning, and *in situ* wafer preparation. A typical pre-epitaxial wafer clean may include a double-sided wafer scrub and an RCA-type wet clean. The details of wafer cleaning are presented in Chap. 5.

A thin layer of chemical oxide may remain on the wafer surface after the wet cleaning process, and must be removed prior to epi growth. Several approaches have been implemented to remove this oxide. One is to dip the wafer in a dilute HF-H_2O solution as the last step in the *ex situ* wet clean. HF will passivate the surface of the silicon and prevent oxidation for up to 30 minutes. The time limitation and variability of this approach makes it difficult to successfully perform such a process in a production environment. Therefore, a second approach that etches a thin layer of silicon in the reactor using dilute HCl (1–5%) in H_2 at high temperature ($\geq$ 1140°C) is more widely used. One problem with HCl is that it can transport Fe contamination to the wafer (see Chap. 8). HCl etching can also lead to wafer pitting and poor quality epi.

Hydrogen pre-baking is an additional favored approach to removing the chemical oxide *in-situ*, just prior to the actual epi deposition. This pre-baking step is performed by subjecting the wafer to ultrapure H_2 gas at moderate temperatures (850–900°C) for a short period of time (30–60s). In some cases this pre-bake is performed at the same temperature as the epi deposition to improve the throughput of single-wafer reactors. For the pre-bake to be effective, the oxidant

partial pressure within the reactor (i.e., O_2 or H_2O) must be low enough so that SiO_2 is no longer thermodynamically stable at the process temperature. The pre-bake step works by converting SiO_2 to volatile sub-oxides (e.g., SiO) which sublime from the wafer surface.[17]

Smith and Ghidini studied the stability of SiO_2 in oxygen and water vapor as a function of temperature and partial pressure.[18,19] They developed a set of curves (Fig. 7-12) which depict the partial-pressure/reciprocal-temperature (P-T) range where the oxide is stable versus where the clean silicon surface is stable. The clean silicon surface is stable at high temperatures and low oxidant partial pressure, while the oxide is thermodynamically favored at low temperatures and high partial pressures of oxidant. For oxygen, the demarcation between stable silicon and oxide is distinct, while for water vapor there is a region where both Si and oxide can coexist. Care must be exercised so that the pre-bake is performed in the stable-silicon-surface region and not in the intermediate state. Since oxide removal proceeds by formation and evaporation of SiO, care must also be taken to ensure the process occurs uniformly across the wafer, otherwise the surface can be roughened by the selective removal of oxide. When growing epi at low temperature, very low concentrations of oxidants are required, otherwise the oxide will not be removed. For a 900°C deposition the partial pressure of O_2 should be less than $7x10^{-5}$ torr and the partial pressure of H_2O less than $5x10^{-6}$ torr. Despite the high purity of the hydrogen gas (obtained from boiling a liquid source), it is sometimes necessary to carry out an additional drying step on the hydrogen gas before feeding it into the epi reactor. The drying can be done by passing the H_2 through a heated (400°C) Pd dryer. The rapidly diffusing H_2 passes through the Pd tube while the much slower diffusing H_2O remains behind and is swept out to the exhaust.

7.5.2 Defects Induced During Epitaxial Deposition

The most common defects that occur in an epitaxial film are dislocations and stacking faults. Dislocations can be generated by several mechanisms. First, if the substrate itself is heavily dislocated it is possible for these dislocations to propagate into the growing film. Hence, low-dislocation-density substrates should be used. Second, during heteroepitaxy (or even homoepitaxy where the substrate doping is very different from the film doping), strain relief can manifest misfit dislocations at the epi/substrate interface. Finally, the presence of large thermally-

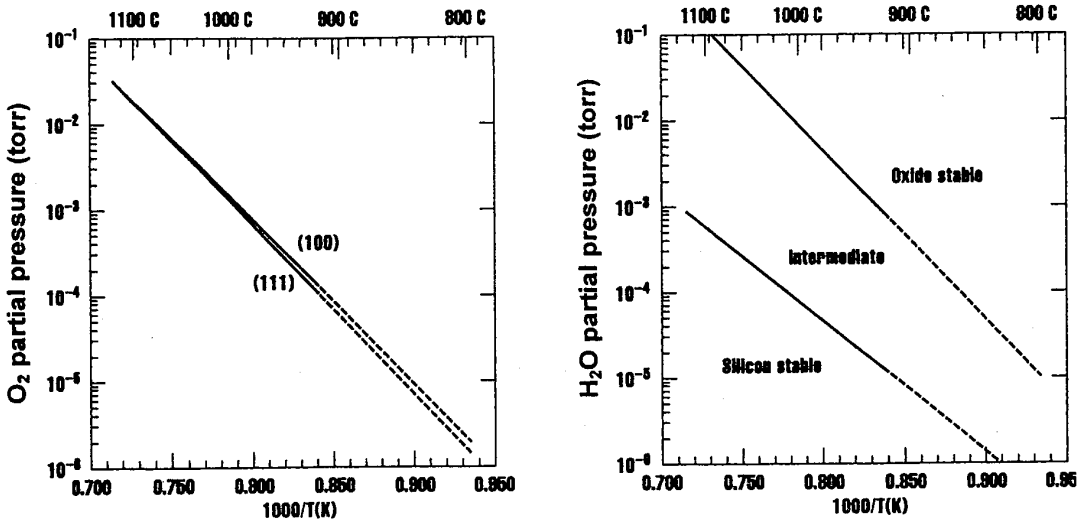


Fig. 7-12 Stability diagram for the surface of silicon and silicon dioxide as a function of temperature and partial pressure of a) O_2 and b) H_2O.[75] Reprinted with permission of Oxford University Press.

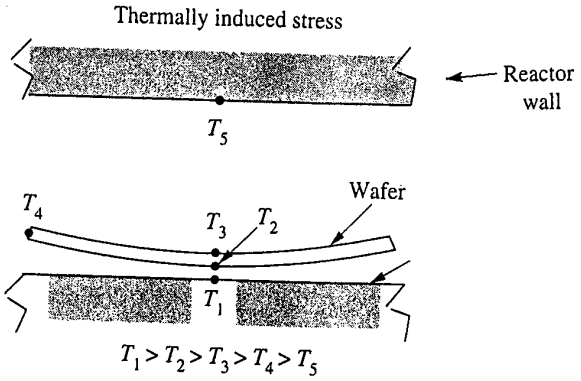


Fig. 7-13 Effect of temperature difference between front and backside of wafer on wafer bow.

generated stresses (that exceed the yield strength of the Si) can cause crystal slip and dislocation propagation. If present, these slip lines can often be visually observed on the periphery of the wafer where the stress is the greatest. Optical scanning systems can measure and classify visual slip damage in epi wafers. More subtle slip damage can be identified using x-ray topography.

Thermal stress in the wafer can result from a temperature gradient, either between the back and front of the wafers, or across the diameter of the wafer. These back-to-front gradients (which are more prevalent in rf heated reactors), can cause a sequence of events that can ultimately lead to slip generation. The rf energy heats the susceptor, which in turn heats the back of the wafers by radiation and conduction. Meanwhile, the front side of the wafer radiates heat to the environment. Thus, the front surface of the wafer becomes several degrees cooler than the back. The hotter backside expands more than the cooler front surface, causing the wafer to assume a bowl-like shape. Such "bowing" causes the wafer edges to lift from the susceptor. The lifted wafer edges receive less heat from the susceptor and cool even further, causing additional wafer bow. Eventually the bowing results in stress levels that exceed the yield strength of the silicon, and crystal slip results. This is schematically depicted in Fig. 7-13. The temperature difference that develops increases for larger diameter wafers. Slip can be substantially reduced in rf-heated reactors by shaping the sample holders into pockets, so the wafers make contact with the susceptor only by their edges (see Fig. 7-14). The temperature difference between the back and front surface can also be minimized by placing heat reflectors above the wafers. Conversely, in IR-heated reactors the front side is already heated by the IR sources, and this also tends

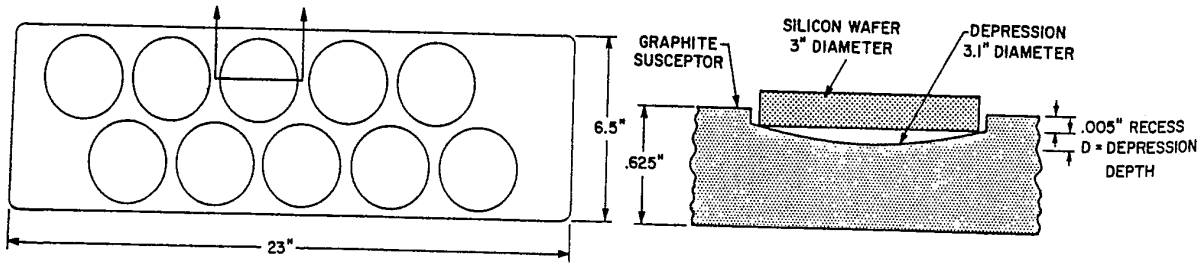


Fig. 7-14 Wafer pockets cut in susceptor to reduce wafer bow effect.[78] Reprinted with the permission of the publisher, the Electrochemical Society.

to minimize the problem. Consequently, rf-heating is now rarely used in epi reactor designs.

In single-wafer reactors, where rapid heating and cooling (RTCVD) is required to achieve good wafer throughput, the temperature gradient across the wafer should never be permitted to exceed the thermal stress level that could generate slip. In other words, the temperature across the wafer must be kept to a low value. Slip formation in RTP systems is discussed in Chap. 8.

Stacking faults (SF) are another common crystalline defect that can form in epi layers. Microscopic surface steps on the substrate, or impurities either on the substrate or within the reactor that fall on the substrate can nucleate such stacking faults. Since the growth of an epitaxial film proceeds by two-dimensional steps moving across the surface along a <110> direction, an encounter with an obstacle on the substrate which upsets this lateral growth could upset the stacking sequence and result in a stacking fault.[20] If the obstacle is on the original substrate, the stacking fault will nucleate immediately and grow through the whole epi layer and emerge at the surface. These obstacles have been attributed to such surface disturbances as slip bands, particles, local regions of silicon dioxide, nitride, or carbide, or metallic impurities. Particulate matter within the reactor that lands on the substrate can nucleate stacking faults anytime in the growth process. Radiantly-heated barrel reactors which have a well formed boundary layer (flow is parallel to the wafer surface) are less prone to this problem since it is difficult for particles to be injected from the gas phase across the boundary layer. Particle transport across such a boundary layer can be further reduced by using low lamp power (as this minimizes electrostatic migration of particles).[21] In addition to causing stacking faults, metallic impurities (Fe, Cu, Ni) that become incorporated and "quenched-in" the epi layer, can (upon further heat treatment) cause impurity precipitation (even at moderate temperatures). Metallic impurity sources include poorly cleaned substrates and metallic impurities within the reactor. *In-situ* HCl etching can serve as an insurance policy against poorly prepared substrates. It should be kept in mind that HCl delivery can also be a source of Fe contamination.

Defects in epitaxial layers can be reduced by the use of denuded zones and oxygen intrinsic gettering (IG) of the starting substrates. (These techniques are introduced in Chap. 2). Borland *et al.*, investigated IG performed on epitaxial substrates prior to epi deposition.[22] These investigators found that devices built into silicon epitaxial layers that were deposited on IG substrate wafers showed improved performance. It was observed that heavily-doped n^+ wafers were resistant to oxygen precipitation, requiring anneals in excess of 40 hours to induce bulk micro-defects. This is one of the reasons that p^+ substrates are primarily used for CMOS epi wafers. Dyson and co-workers used substrates with polysilicon deposited on the backside in addition to employing IG. This combined gettering treatment gave significantly improved epi minority carrier lifetimes.[23]

Another technique for achieving low-defect epitaxial layers is the application of *strain layers*.[24] The strain layer is formed between the substrate and the active part of the epitaxial layer by intentionally alloying the initially deposited silicon with germanium. Since the Si-Ge alloy has a substantially different lattice parameter than the substrate, the mismatch can be accommodated by the formation of misfit dislocations. The epitaxial device layer is then deposited on the Si-Ge strain layer. The misfit dislocations in the strain layer serve as sites for metal precipitation, thereby trapping impurities away from the active device region. Bean reviewed the use of multilayer misfit dislocation extrinsic gettering strain fields.[25] Epitaxial Si-Ge layers were sequentially deposited, with each layer having an increasing concentration of Ge (the density of misfit dislocations increases with Ge concentration). Each layer was then interposed with a strain-free Si layer and a final Si device layer was deposited. The more highly-

doped layers (having greater gettering efficiency) are closer to device regions. The results show that misfit dislocations do not reach the device layer even after high temperature (>1000°C) heat cycling and this method is effective in gettering Au, Cu, Fe, and Ni from epitaxial layers.[26] These techniques, although promising, are typically too costly to implement in a production process.

7.5.3 Pattern Shift, Distortion, and Washout

During the fabrication of bipolar transistors (See Vol. 2 of this series) heavily-doped n^+ (i.e., with Sb or As) buried layers are diffused into a *p*-type substrate prior to epi growth. The drive-in diffusion is usually performed in an oxidizing ambient where an oxide grows on the surface of the wafer. The heavily diffused regions grow a thicker oxide than the surrounding substrate, and consume more silicon. When this oxide is subsequently removed, a step between the buried layers and the substrate remains in the silicon. This step or "feature" is often used for aligning sub-sequent layers to the buried layer. When an epitaxial layer is deposited over the buried layer these steps propagate through the film. Under some conditions the buried layer image at the top of the epi layer is shifted relative to its position in the substrate. This effect is termed *pattern shift*. In other cases the pattern may become distorted or washed out. These effects are termed *pattern distortion,* and *pattern washout*, respectively (see Figure 7-15). The amount of shift or distortion must determined, so that circuit designers can adjust the features on subsequent masks to compensate for these effects. Selecting the correct amount of adjustment, however, is complicated by the fact that these pattern changes are dependent on substrate orientation, deposition rate, reactor operating pressure, reactor type, deposition temperature, and silicon source. If the pattern becomes completely washed out, alter-native alignment strategies must be employed. The substrate crystalline orientation has a significant effect on pattern shift. For the case of (100) oriented wafers (used currently in both bipolar and BiCMOS processing), pattern shift is a minimum when oriented on-axis. On the other hand (111) wafers (which were previously the mainstay of bipolar processing) show a minimum effect when cut off-orientation by 2° to 5° from the <111> direction to the nearest <110> direction. Epi substrates with <111>-orientation are normally supplied 3° off-orientation.

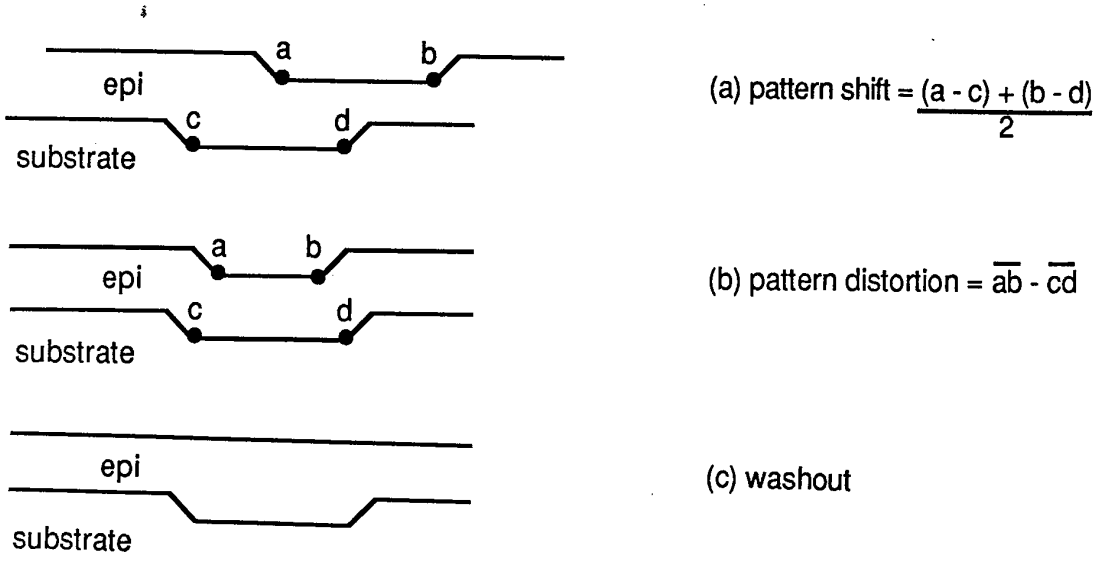


Fig. 7-15 Schematic representation of (a) pattern shift, (b) pattern distortion, (c) pattern washout.

Lee *et al* systematically studied pattern shift as a function of temperature and deposition rate. They found that the shift decreased with increasing temperature, but increased with increasing deposition rate, for both (111) and (100) orientations.[27] Reduced-pressure deposition was found to substantially reduce the pattern shift. The pattern shift can be totally eliminated by using SiH_4 as the source gas. It appears that the presence of a Cl species is required to induce pattern shift. Pattern distortion during epitaxial deposition was studied by Weeks, who found that while distortion occurs in both (100) and (111) oriented substrates, it is exacerbated on (100) substrates.[28] Although reduced-pressure deposition minimizes distortion in (111) material, it is ineffective for reducing the effect on (100) substrates. Unlike the situation with pattern shift, increased temperature and decreased growth rate tend to increase pattern distortion. Furthermore, the use of SiH_4 as the silicon source results in more pattern distortion than when chlorine-based sources are employed. Finally, increased layer thickness minimizes distortion. These opposite dependencies of pattern shift and pattern distortion create a dilemma for the process engineer. Thus, empirical studies must be utilized to fine-tune the process so that both distortion and shift are reduced to a controllable value.

7.6 LOW-TEMPERATURE EPITAXY PROCESSES

The use of raised source/drain selective epitaxy and the development of heterojunction bipolar transistors (HBT) employing strained layers have driven the deposition requirements for epitaxial growth at low temperatures. Numerous *low-temperature epitaxy* (LTE) techniques have been proposed, and three have shown promise. These are *ultra-high-vacuum chemical-vapor-deposition* (UHV/CVD), *low-temperature rapid-thermal CVD* (RTCVD) and *molecular beam epitaxy* (MBE).[29,30] Other LTE techniques have been reported but none have yet met with commercial success. These include: a) ion beam deposition; b) photo- and laser-assisted CVD; c) remote-plasma CVD; d) electron cyclotron resonance (ECR-CVD); and e) plasma-enhanced CVD (PECVD). PECVD epitaxy has been performed at temperatures as low 165°C. But higher temperatures (>300°C) are needed to achieve reasonable (but still not device quality) wafers.[31]

Regardless of the technique used for LTE, it is necessary that the wafer surface be extremely clean prior to the onset of the deposition. The cleaning process may require the use of a dilute HF dip, since oxide removal using hydrogen pre-bake is not effective at very low temperatures. If the application can endure higher temperatures, then an H_2 pre-bake can be used for oxide removal. To maintain the oxide-free surface, a vacuum loadlock must be in place on the system, so that wafers waiting to be processed do not become reoxidized. *In-situ* oxide removal techniques that operate at lower temperatures have been evaluated, but none have met with commercial success. These techniques include: a) silicon-ion-beam cleaning; b) hydrogen-plasma cleaning; c) HF-vapor etching; d) treatment of the surface with GeH_4 to reduce the SiO_2; and e) high-pressure hydrogen cleaning.[32,33,34] Readers are referred to the references for more information on these subjects.

An LTE production technique using ultra-high vacuum chemical vapor deposition (UHV/CVD) in a multi-wafer reactor has been employed successfully to grow epitaxial layers of silicon and silicon-germanium alloys for use in high performance devices.[35] It is reported that device-quality epitaxy can be achieved at a growth temperature of 550°C. The batch of wafers (typically 25) is introduced into the chamber under an H_2 flow. The chamber is pumped down to a base pressure of 10^{-9} torr (UHV), and the precursor gases (silane, germane and diborane) plus hydrogen are introduced. During deposition the pressure is maintained at 10^{-3} torr.

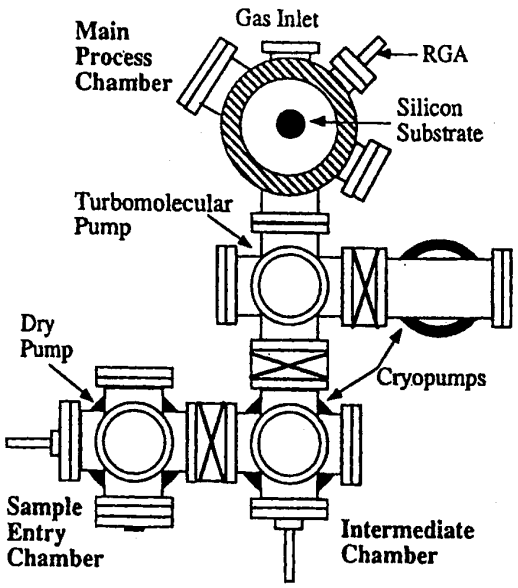


Fig. 7-16 Schematic of a UHV/RTCVD epitaxial system.[36] Reprinted with permission of the American Physical Society.

The growth rate during this process is only about 0.3 nm/min. This very low growth rate provides precise control of the layer thickness, doping concentration, and formation of a pseudomorphic strained layer. Single-wafer, ultra-high-vacuum, rapid-thermal-CVD (UHV/RTCVD) has also been reported as a means of achieving LTE.[36] An advantage of this single wafer technology is that the process can be initiated and terminated by rapid changes in the substrate temperature, rather than changing of gases. Figure 7-16 shows a schematic of such a system.[37] Conventional single-wafer epi reactors such as the Applied Materials Centura® can also be operated at low temperatures (about 600°C) to produce strained layer Si-Ge alloys at a deposition rate of about 10 nm/min.

Another approach to achieving LTE is molecular beam epitaxy (MBE). Although MBE has met with success in growing III-V compound devices, the earliest applications of this technology were for silicon. MBE is now rarely used for growing silicon epitaxial films. The MBE process involves the evaporation of silicon and the desired dopants under extreme ultra-high vacuum (E-UHV) conditions (10^{-11} torr). The evaporant atoms (Si and dopant) are directed at high velocity to the heated substrate (500–700°C) where the atoms then condense and grow epitaxially. Such low-temperature epitaxial deposition is sufficient to eliminate outdiffusion and autodoping and to form controlled strain layers. In fact equally abrupt impurity profiles can be obtained using MBE, whether deposited on heavily-doped arsenic layers or undoped material. There are three techniques that have been used to produce the MBE source gases, namely; Knudson effusion cells, e-beam evaporation, and gaseous sources. E-beam evaporation suffers from the potential of spitting defects, which could result in defects on the wafer.[38] The effusion cell requires such high temperatures that the molten silicon becomes very reactive and is difficult to keep pure. This problem has been ameliorated by using gaseous sources for the silicon and dopants.[39] When using these gaseous sources the distinction between UHV/CVD and MBE begins to blur.[40] Very abrupt changes in doping profile can be achieved by passing or interrupting the dopant fluxes with shutters positioned above the dopant source. The sticking coefficients of the dopants are low, and a much higher dopant flux than the simple ratios of atoms is required to achieve the desired doping level in the film. During film growth atoms from residual gases can impinge on the substrate and may be incorporated as contamination.

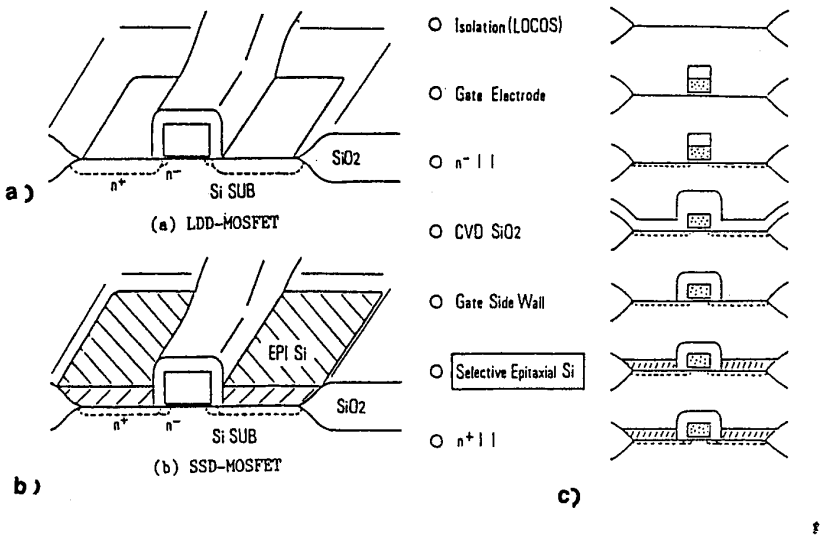


Fig. 7-17 Schematic diagram of: a) an LDD-MOSFET; b) an elevated-source-drain MOSFET fabricated using selective epitaxial growth (SEG); c) fabrication process flow of such a MOSFET.[76] This paper was originally presented at the Spring 1988 Meeting of the Electrochemical Society.

Therefore, background partial pressures of O_2, H_2, and N_2 should be less than 10^{-14} torr. The substrate itself may be a source of contamination (with its chemisorbed oxygen and carbon) and these contaminants must be removed *in-situ* prior to MBE.

7.7 SELECTIVE EPITAXIAL GROWTH (SEG)

During *selective epitaxial growth* (SEG) an epitaxial layer deposits selectively on the single-crystal silicon regions of a microcircuit, while no deposit is formed on the insulator-coated regions of the same IC. The SEG process is favored when silicon nucleation is suppressed on the insulator. Thus, atoms landing on the insulator have time to migrate to single-crystal regions, where nucleation is favored. The nucleation of silicon is least likely on SiO_2, more likely on Si_3N_4, and most likely on silicon.

The earliest application of SEG was to develop an improved device isolation scheme to replace LOCOS.[41,42] (Isolation technology is covered in detail in Vol. 2 of this series.) In SEG isolation, islands of silicon are deposited within a sea of oxide. Devices are then fabricated in the silicon islands and interconnected in the normal fashion. The appearance of shallow trench isolation (STI), however, caused SEG isolation to be virtually abandoned. A more useful application of SEG is to fabricate *raised source/drain* regions. The raised source/drain enables the fabrication of ultra-shallow junction using cobalt or titanium salicide technology. Figure 7-17 shows a cross-section of a CMOS device with SEG raised source/drains.

There are two types of selective epitaxy. *Type-1 SEG* takes place when single-crystal silicon is deposited on bare silicon but no silicon whatsoever is grown on the oxide. It is possible that some small nuclei may form on the oxide, but as long as no continuous silicon film is formed, such nuclei represent only a cosmetic problem. Type-1 selective epitaxy is depicted in Fig. 7-18a. *Type-2 SEG* takes place when epitaxial silicon is deposited on the exposed silicon and polysilicon is simultaneously deposited on the oxide surface. Type-2 selective-epi deposition is depicted in Fig. 7-18b.

Type-1 SEG is favored when high-surface-mobility silicon atoms are deposited from the silicon source. Since nucleation does not readily occur on the oxide, the atoms that land on that

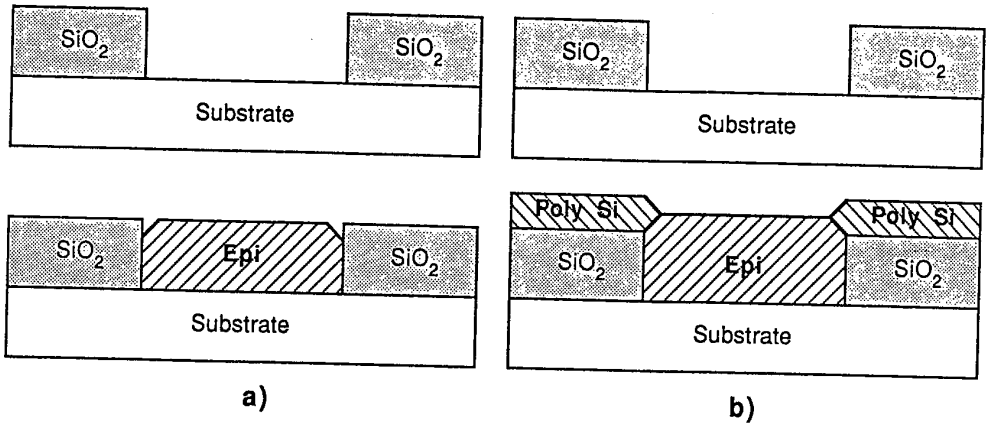


Fig. 7-18 (a) *Type-1* selective epitaxial silicon deposition. (b) *Type-2* selective epitaxial silicon deposition.

surface are able to migrate to sites on the single-crystal material where nucleation is favored. The cleanliness of the oxide surface and the presence of chlorine in the precursor enhances the atom's mobility. The higher the number of chlorine atoms in the silicon source, the better the degree of selectively. Based on Cl-atoms alone the selectivity should obey the following hierarchy: $SiCl_4$, $SiHCl_3$, SiH_2Cl_2, and SiH_4. Indeed, SiH_4 (without added chlorine) shows very little Type-1 selectivity unless deposited under markedly ultra-clean conditions.

Based on the presence of chlorine atoms alone, $SiCl_4$ would be the silicon source of choice for SEG. But the high deposition temperature required for $SiCl_4$ epi can result in the degradation of the SiO_2 mask and loss of selectivity. The resulting defects and contamination on the oxide surface could become a source of silicon nucleation and adversely affect the selectivity. Using Si_3N_4 masks in place of SiO_2 can mitigate this problem, but Si_3N_4 tends to act as a better nucleation surface for silicon deposition. A more suitable silicon source for Type-1 selective epitaxy is SiH_2Cl_2, particularly with the addition of HCl. SiH_4 can also be used if HCl or Cl_2 is added to the reaction. The addition of these components reduces the deposition rate somewhat, but has no adverse effects on the electrical or structural properties of the deposit, and substantially aids in keeping the process selective. Several other factors are known to enhance the selective nature of the silicon deposition. They include: a) performing the process in a reduced pressure reactor; b) increasing the deposition temperature; c) decreasing the mole fraction of silicon source in the gas stream; and d) ensuring the cleanliness of the wafers and the system.[43]

It has been reported that SEG of boron-doped silicon has been achieved using Si_2H_6 (disilane), B_2H_6, H_2, and Cl_2 in a RTCVD reactor operated at 800°C and at a reduced pressure of 24 mtorr.[40] In that case the selective silicon growth rate was 150 nm/min and selectivity to both SiO_2 and Si_3N_4 layers was demonstrated. An advantage of using Si_2H_6 instead of DCS is that the amount of added Cl_2 required for SEG can be substantially reduced. Lower temperature SEG has been reported using SiH_2Cl_2 at 650°C, without the addition of any HCl. To achieve selectivity at such low temperatures an ultra-clean atmospheric reactor was necessary.[44] Finally, despite the absence of any Cl-species in the gas mixture, successful SEG has been demonstrated using only SiH_4 or Si_2H_6 as precursors, but again only under ultra-clean conditions.[45,46] The key point is that wafer and reactor cleanliness are essential conditions for successful SEG.

The presence of crystallographic defects and crystal faceting at the SiO_2/epitaxial interface is common to both Type-1 and Type-2 selective depositions. These defects can affect the performance of devices built into the epitaxial layers. Transmission electron microscopy (TEM)

studies of the interface between silicon and the oxide show lamellar defects resembling microtwins. Typically, the defects are confined to a region close to the interface (i.e., within 0.1–0.3 μm). Epitaxial stacking faults have also been observed in the interface region. Coating the walls of the oxide steps with ~100 nm of polysilicon greatly reduces the crystalline defects in the interfacial region. Fortunately, in the raised source/drain application of SEG only 10–30 nm of silicon is required. As a result, the amount of faceting and defect formation is relatively small, and can be tolerated in the devices.

Type-2 selective deposition is best accomplished by using SiH_4 as the Si source, since selectivity is not critical. In Type-2 SEG an interface is formed between the epitaxial and the polysilicon films. This interface is also a place where defect structures are known to form. The angle of this interface relative to the film growth direction depends upon the crystallographic orientation of the substrate. For (110), (100), and (111) the angles are 90°, 72°, and 70° toward the polysilicon, respectively.

Another type of SEG is termed *epitaxial lateral overgrowth* or *ELO*. ELO occurs when the selective epi film is permitted to continue to grow past the height of the oxide window.[46] At that point the epi continues to grow not only in the vertical direction but also in the lateral direction. Fig. 7-19 shows a schematic of ELO. If the windows are closely spaced it is possible for the lateral regions of the epi to merge and thereby form a continuous silicon film over the substrate. The ELO layer will now be isolated from the substrate either by air or oxide and can serve as a basis for a *silicon-on-insulator* (SOI) substrate. (SOI is discussed in detail in Section 7.10.) The ratio of the lateral to vertical growth rates depends on the window height and the substrate orientation. If the seed window orientation is along the <100> direction on a (100) oriented wafer, the lateral and vertical growth rates will be equal. The main issue with this technique is that the ELO layer can be have a very high defect density at the point where the growth direction changes (i.e., where two laterally growing layers meet).

7.8 EPITAXIAL DEPOSITION EQUIPMENT

Epitaxial reactors are high-temperature, batch- or single-wafer, chemical-vapor-deposition (CVD) systems. Such systems are described in Chapter 6. Epi reactors can be operated at either atmospheric or reduced pressure. The demands of modern epitaxial films place stringent

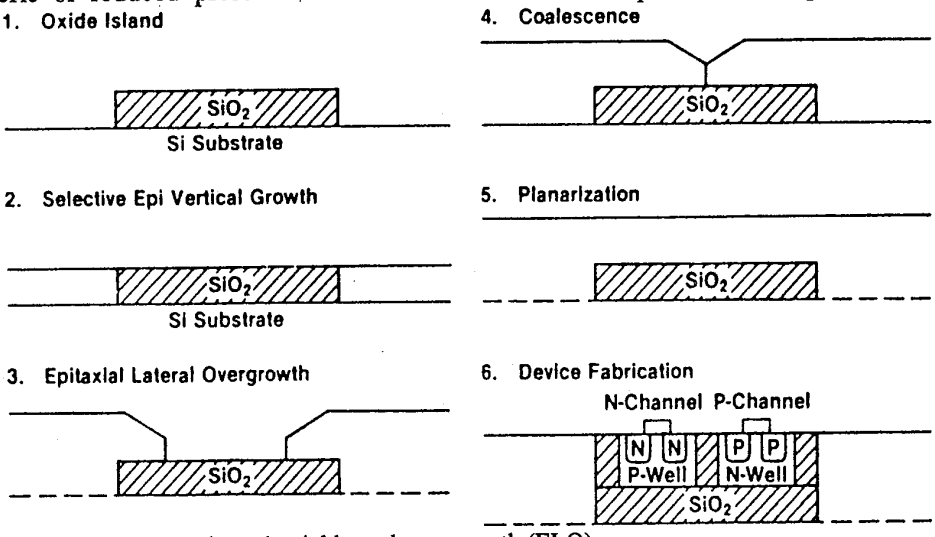


Fig. 7-19 Silicon on insulator by epitaxial lateral overgrowth (ELO).

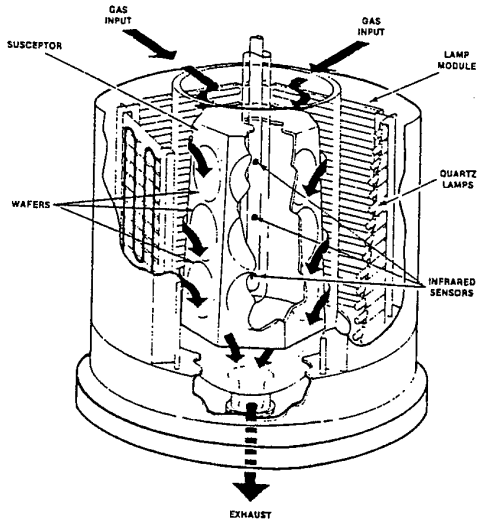


Fig. 7-20 Schematic of a radiantly-heated barrel reactor. Courtesy of Applied Materials, Inc.

requirements on epitaxial process tools. Some of these requirements are: a) tight thickness and resistivity control; b) low number of process-induced defects; c) minimized pattern shift and washout; d) reduced autodoping; e) reduced pressure and temperature operation; f) the capability of handling up to 300-mm wafers (without slip); and g) high productivity. Today, two types of epitaxial process tools are commercially viable: 1) the multi-wafer IR radiantly-heated barrel reactors; and 2) the single-wafer IR lamp-heated reactors. The single-wafer reactors are most popular for large wafer size (≥200 mm). The use of the once-popular rf inductively-heated "pancake" reactor has greatly decreased because of its propensity to cause crystal slip, their low wafer throughput for large wafers, and their inability to achieve low pressures. A discussion of inductively-heated rf pancake reactor can be found in the 1986 edition of this text.

At a minimum, all epitaxial tools consists of: a) a quartz bell-jar or reactor tube (to isolate the epitaxial growth environment); b) a gas distribution system (which provides controlled introduction of the various chemical species); c) an IR (lamp) wafer heating source; d) a susceptor to hold the wafers (usually made of silicon-carbide-coated graphite); and e) an exhaust system to efficiently remove the effluent gases. Temperature monitoring and control is achieved by focusing an IR-sensitive optical pyrometer either on the wafer surface or on the susceptor. The indicated temperature of such a measurement is often 50–100°C below the actual temperature, due to the dependence of the emissivity on temperature.

A radiantly-heated barrel reactor is shown schematically in Fig. 7-20. The susceptor has a hexagonal cross-section and is fabricated from silicon-carbide-coated graphite. Each face of the hexagon can accommodate one or several wafers depending on their diameter. The wafers are held at an angle 2.5° to the vertical, and the gases flow parallel to the wafer surfaces. (This reactor can be considered as a horizontal reactor laying on its side). The reactants are injected at the top of the chamber and products are exhausted from the bottom. The gas flow parallel to the wafer surface favors the formation of a boundary layer above the wafers. The presence of such a layer tends to increase inter-wafer autodoping and source-gas depletion effects. The negative effects of the boundary layer can be minimized by operating the system at reduced pressures. The wafers are heated directly by using water-cooled quartz lamps which line the inside of the reactor. Heating the wafer from the front decreases the occurrence of thermally induced slip. Such systems are capable of operating at atmospheric or reduced pressure (80–100 torr). Another advantage of such configurations is that the wafers are held in near-vertical positions

where they are less susceptible to particulate contamination. The barrel reactor is capable of resistivity and thickness uniformity better than ±5% across the wafer, wafer-to-wafer, and run-to-run, but is limited when high productivity is required for large diameter wafers.

With the increased use of large diameter wafers, the single-wafer reactor (which is more economical than barrel reactors for large wafers), has become predominant. Single-wafer epitaxial reactors are manufactured by Applied Materials Inc. and by ASM International.[79] Figure 7-21 shows the cross section of a single-wafer epitaxial reactor. The chamber is constructed from fused silica (quartz). All stainless-steel components are outside of the chamber, since stainless steel can corrode when it reacts with chlorinated silicon precursors. Because the epi process requires extremely high temperatures (1050–1150°C) there is a tendency for the chamber walls to become hot. To prevent runaway heating of the chamber walls, water and air cooling are utilized, and this allows the wall temperature to be maintained between 400–600°C.

The epitaxial deposition process in a single-wafer process tool is started by robotically transporting the wafer into the reaction chamber, and placing it on a silicon-carbide-coated graphite susceptor. Since the chamber dome and susceptor remain relatively hot from having processed the previous wafer (600–750°C), the new wafer quickly heats to the temperature of its surroundings. The susceptor (along with the wafer) is rotated at 30 rpm to even out temperature and gas-flow fluctuations. This rotation of the wafer gives substantially improved process uniformity.

The wafer is rapidly heated (18°C/sec) to the process temperature by applying the appropriate power to the four independently-controlled banks of quartz IR lamps. The wafer is heated both from the top (by the upper lamp assembly) and the bottom (through the susceptor which is heated by the lower lamp assembly). Such dual-sided heating minimizes front-to-back temperature differences. There are also two banks of lamps which independently heat the center and the edge of the wafer on both the top and the bottom. This feature provides uniform temperature across the wafer particularly during the temperature ramp. Such temperature uniformity is necessary to prevent thermal-stress-induced crystal slip. The temperature is detected from both the wafer (pyrometer #1) and the Susceptor (pyrometer #2). These pyrometers are

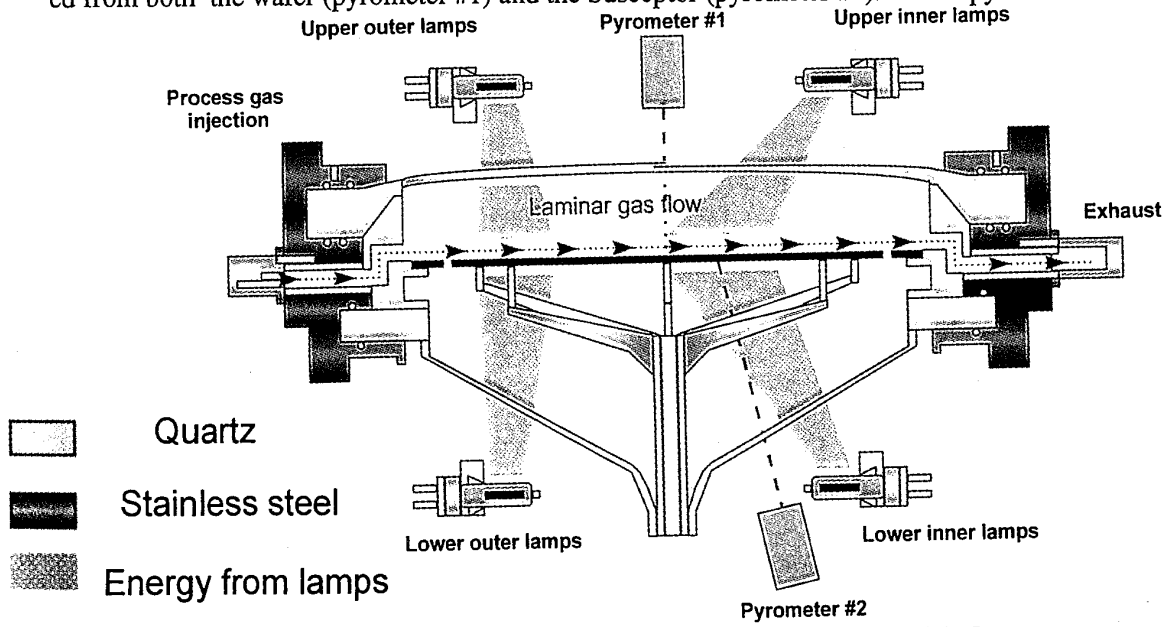


Fig. 7-21 Cross-section of a single-wafer epitaxial reactor. Courtesy of Applied Materials, Inc.

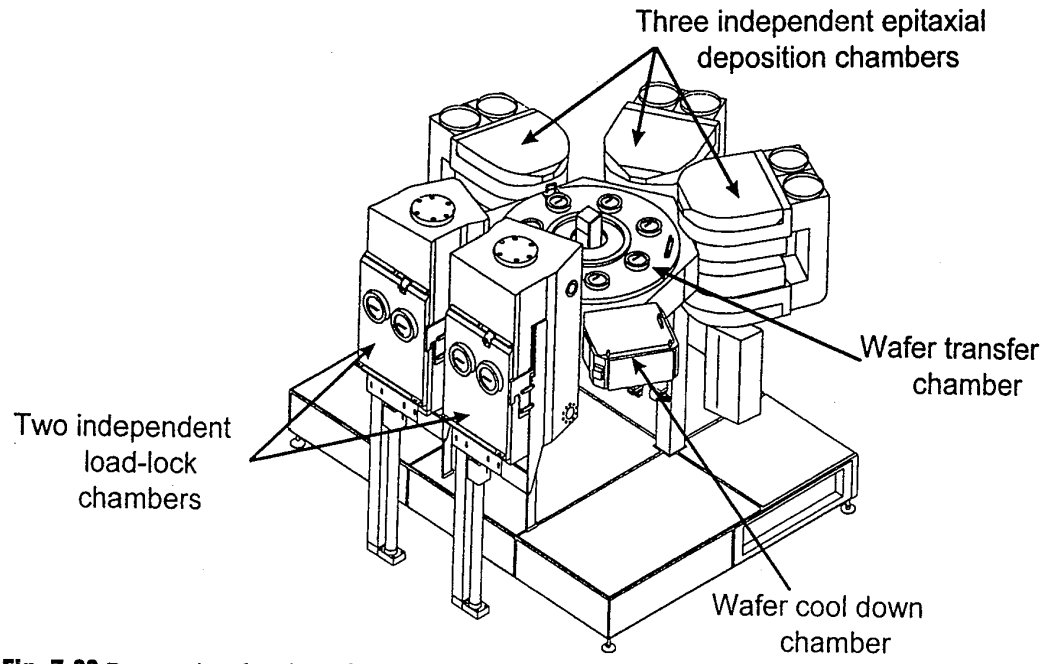


Fig. 7-22 Perspective drawing of a commercial single-wafer epitaxial reactor on a central handling platform showing three reaction chambers and one cool-down chamber. Courtesy of Applied Materials, Inc.

the wafer (pyrometer #1) and the susceptor (pyrometer #2). These pyrometers are calibrated against a thermocouple to compensate for temperature-dependent emissivity variations. Once the wafer reaches temperature, it is stabilized and "pre-baked" in H_2 (for about 1 minute). The H_2 pre-bake removes any chemical or native oxide that may exist on the wafer surface.

The process gases (precursors and dopant) are pre-mixed in a manifold before they enter the chamber. The gases enter from a side port (along with the H_2 carrier) and flow over the wafer in a laminar manner where they then exit the system from a side exhaust. This type of gas flow is characteristic of a horizontal reactor. By attaching a dry vacuum pump to the exhaust the system can also be operated at a reduced pressure of about 80 torr. The operating pressure can be further reduced (<20 torr) if a Roots blower is added to the pumping stack.

After the epitaxial deposition process is completed the flow of the precursor and dopant gases is stopped (H_2 flow continues), and the wafer is allowed to cool to about 750°C. When the wafer reaches 750°C it is removed from the process chamber and placed in a cool-down chamber. Once in the cool-down chamber the wafer rapidly cools to <200°C, whereupon it can be placed back into the wafer cassette. Since one cool-down chamber can service more than one process chamber, the Applied Materials system is configured with three process chambers and one cool- down chamber on a single central robot system or so-called *mainframe*. Such a configuration is depicted in Figure 7-22. Since silicon will deposit on the warm walls of the reactor and on the hot susceptor, it is necessary to clean the chamber on a regular basis. The chamber is cleaned by heating the susceptor to 1200°C and introducing an HCl/H_2 mixture into the chamber to etch the silicon from the walls and the susceptor. Depending on the total amount of silicon deposited, such a cleaning step should be carried out after approximately every five wafers. With a three reactor/one cool-down configuration and a cleaning step after every 5 wafers, a throughput of about 30 wafers-per-hour can be achieved with this system. Table 7-4 shows the standard process conditions used for atmospheric- and reduced-pressure single-wafer epitaxial deposition. Figure 7-23 shows a typical process cycle for an atmospheric epitaxial deposition.

Table 7-4 STANDARD PROCESSING CONDITIONS FOR ATMOSPHERIC- AND REDUCED-PRESSURE OPERATION OF THE APPLIED MATERIALS CENTURA® EPI SYSTEM

Reactor Type	Precursor	Deposition Temperature (°C)	Deposition Rate (μm/min)
Atmospheric	$SiHCl_3$	1125–1150	1.5–4.0
Reduced Pressure	SiH_2Cl_2	1050–1100	0.2–1.5

The film properties measured on a wafer from a single-wafer reactor normally exceed those from a barrel reactor, since each wafer is processed nearly identically. Such effects as the depletion of the precursor concentration and autodoping from other wafers is nonexistent. Typically the thickness uniformity from a single-wafer reactor is measured to better than 3% (3σ) within-the-wafer, and 5% (3σ) wafer-to-wafer. The resistivity can be controlled to better than 5% (3σ) within-the-wafer, and 7% (3σ) wafer-to-wafer. There is no slip detected to within 3 mm of the wafer edge. Metallic contamination can be controlled to $< 1x10^{11}$ atoms/cm^3, and particulate contamination from this reactor is less than 0.02 defects/cm^2 (at ≥ 0.3 μm).

7.9 CHARACTERIZATION OF EPITAXIAL FILMS

Characterization of the grown epitaxial film is performed to assess the epi-dependent properties of product devices early in the fabrication sequence. If the epi quality is poor it is unlikely that high quality devices can be made in the film. Epitaxial layers are characterized for: a) surface quality; b) crystallographic defects; c) electrical properties (principally, resistivity, carrier lifetime, and dopant profile); and d) film thickness. Many of the characterization techniques are common to those used for bulk silicon and diffused silicon layers, and are discussed in more detail in other chapters of this book.

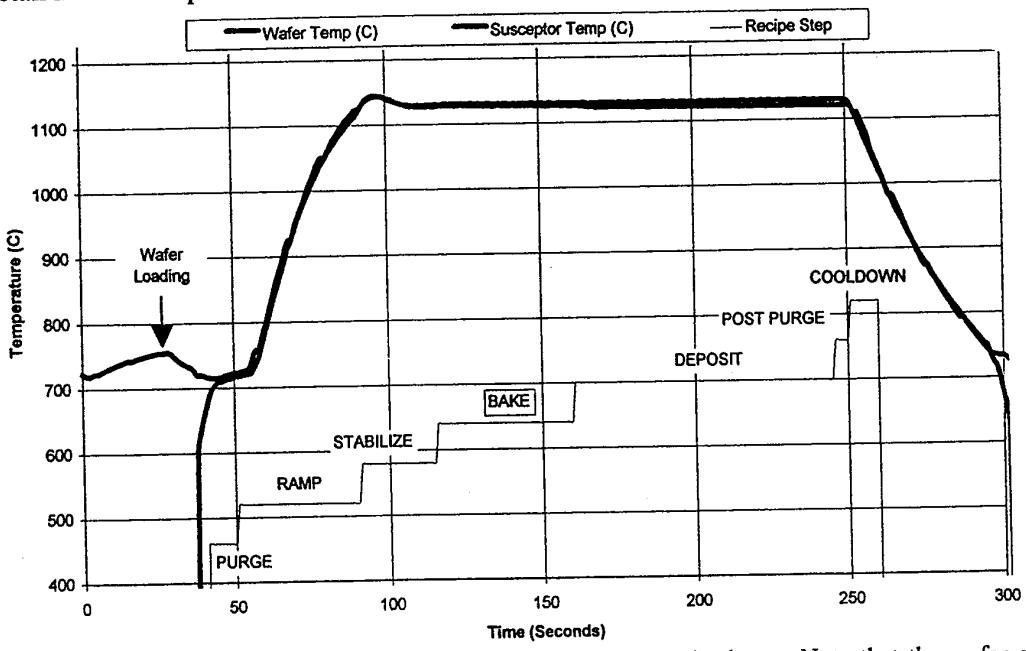


Fig. 7-23 Typical epitaxial deposition process. Each process step is shown. Note that the wafer quickly heats to the susceptor temperature and tracks the susceptor temperature even during the cool-down. Courtesy of Applied Materials, Inc.

7.9.1 Optical Inspection of Epitaxial Film Surfaces

Much can be learned about the quality of the epitaxial layer by simple visual inspection under UV-light. For more detailed inspection an optical microscope or a surface-scanning-laser defect counter (See Sect. 5.8) may be required. These techniques can ordinarily elucidate the presence of slip lines, haze, pits, particles, scratches, and spikes*.[48] Microscopic evaluation at low magnifications (75–200X) can reveal stacking faults and pyramidal defects. The use of Normarski phase-contrast microscopy displays these defects more readily, since Normarski can detect surface features with a height difference of only 3–5 nm. Automated laser scanners (which detect light scattering centers) can be used to identify, map, and quantify defects in epitaxial layers. X-ray topographic methods are used to reveal small amounts of crystal slip in the wafer. Crystal slip becomes more important for large wafer sizes, therefore when evaluating very large wafers (≥300 mm) special attention should be paid to evaluating slip.

In some cases the crystallographic defects must be *decorated* in order to be observed. Such decoration is performed by etching the epi surface with a selective etchant, such as a Dash, Sirtl, Wright, or Secco etching solution. These chemical solutions contain HF, an oxidizing acid (e.g., HNO_3), and acetic acid, and they preferentially etch the silicon around the defect. This enlarges the defect or decorates the defect making it easier to observe by optical microscopy. For extremely small defects (i.e., where very high magnification is required), scanning electron microscopy (SEM) may be employed. For detailed analysis of the crystallographic nature of the defects, transmission electron microscopy (TEM) can be performed. TEM is not routinely used, since it requires time consuming and tedious sample preparation.

7.9.2 Electrical Characterization

The two electrical properties that have the most profound influence on the operation of a device fabricated in epitaxial layers are the epi film resistivity and the minority-carrier lifetime. Control of the film resistivity is usually attained by controlling the ratio of dopant gas to silicon-bearing gas during the deposition. However, variations in dopant incorporation, the effect of autodoping, and the possibility of depletion effects in multi-wafer tools may be sufficient to cause deviations from the desired resistivity. Resistivity is usually monitored on a routine basis. The measurement is performed on a monitor wafer that may be placed in a multi-wafer deposition, or on a sampling basis in a single-wafer reactor. The monitor wafer must be of the opposite conductivity type when using a 4-point probe to determine the sheet resistivity. The shape of the doping profile can be obtained from a spreading resistance probe (SRP), capacitance-voltage (C-V), or secondary ion mass spectroscopy (SIMS) measurement. These techniques are described in Chap. 9 for diffused layers, and only the special requirements of epi will be presented here.

When making a *four-point probe measurement* the resistivity of the film, ρ, is given by:

$$\rho = R_s t_e F_1 \tag{7.9}$$

where R_s is the measured 4-point probe sheet resistivity, t_e is the epitaxial thickness, and F_1 is the geometrical correction factor. The same care that is required when measuring shallow diffused layers should be exercised when measuring thin epitaxial layers, since the probe tips can penetrate the epi layer and reach the substrate.

* Spikes are protrusions on the silicon surface that do not exhibit alignment with the crystalline directions. It is thought that the spikes arise from defects on the substrate that interfere with the 2-D growth of the film and are actually related to the onset of 3-D growth.

C-V measurements are made by first fabricating either a *pn* junction or a Schottky barrier in the epi film. Both techniques are sensitive in the doping range of 10^{14}–10^{16} atoms/cm^3, but suffer from limited depth profiling capabilities and excessive time for sample preparation. Using a liquid-Hg probe to form the Schottky diode does offer the advantage of rapid turnaround for specimen preparation. These techniques may be used in the early stages of development but are not considered routine in-line measurements.

The *spreading resistance probe* (SRP) measurement is applicable over a broad range of dopant concentrations (10^{14}–10^{20} atoms/cm^3). The SRP technique is useful for determining the shape of the transition region, particularly near the high doping side. By observing the shape of the transition region the amount of outdiffusion and autodoping can be ascertained. Secondary-ion mass spectroscopy (SIMS), on the other hand, has a relatively poor sensitivity to the common dopants at the doping levels used in the epi layers. The sensitivities are: 10^{15}, 10^{16}, and $2x10^{17}$ atoms/cm^3 for B, Sb, and P, respectively. As a result, SIMS analysis is limited to the epi layers near the substrate/epi interface, where the dopant concentration is highest.

The *minority-carrier lifetime* in the epitaxial layer provides one method to indirectly determine the presence of metallic impurities (such as Fe, Cu, Ni). These impurities reduce the minority-carrier lifetime in the silicon. A common way to measure lifetime is the *Zerbst method.*[49] To make the Zerbst measurement, the capacitance of an MOS capacitor (which has been fabricated in the epitaxial material) is monitored as a function of time after the device is driven into deep inversion. Details of the measurement are presented in the excellent reference book by Schroder.[50] A deficiency of this technique is that the measurement is confined to a thin layer near the surface, and the lifetime that is measured is not the equilibrium lifetime (since the measurement is confined to the space charge region). It is also possible to use the *surface photovoltage* (SPV) technique to directly measure minority carrier concentration and indirectly Fe contamination.[51] Readers are referred the Schroders' text for details on SPV.[52]

7.9.3 Epitaxial Film Thickness Measurements

Thickness and thickness uniformity are critical parameters of epitaxial layers, and they must be accurately measured and controlled. Several bipolar transistor device parameters (such as the junction breakdown voltage, junction capacitance, transistor gain, and ac performance), depend directly on the epi thickness. Epitaxial thickness measurements fall into two categories, namely destructive and nondestructive. Although a variety of such techniques will be briefly described, it should be noted that the most widely used is the non-destructive method based on the optical properties of the epitaxial structure (infrared reflectance). Junction depth techniques such as: 1) *angle-lap and stain*; 2) *groove and stain*; and 3) *tapered groove and profilometry* are all methods (albeit destructive) for directly measuring the epitaxial thickness. To achieve staining differentiation when using these destructive methods, it is necessary to deposit the epi layer on a monitor wafer of the opposite conductivity type. Details of junction depth measurements are discussed in Chap. 9. These destructive techniques measure the "physical" thickness of the film but have difficulty accounting for the transition region in a device wafer.

The physical layer thickness can also be established by measuring the weight gain of the silicon wafer after deposition. Since 2.0 μm of added silicon on a 200-mm wafer represents the addition of about a 106 mg of weight, this measurement is easily made using a modern analytical balance. It should be understood, however, that only an average thickness can be calculated from this measurement, and any weight gain includes silicon that may have been deposited on the backside. This technique measures the true physical thickness of the epi film.

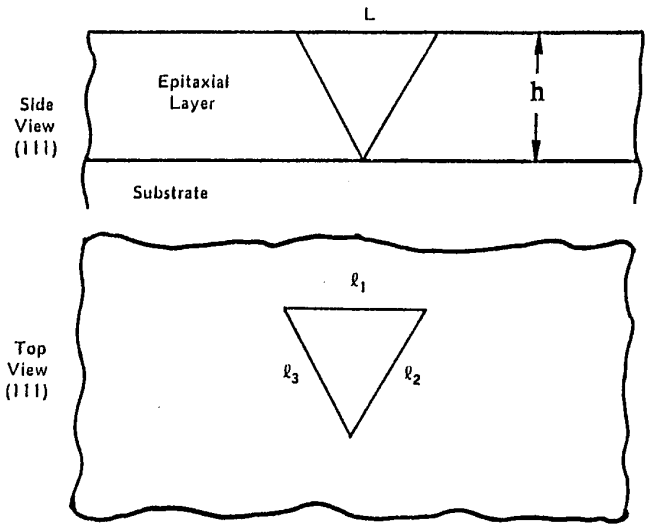


Fig. 7-24 Geometrical basis of the stacking fault method for determining the thickness of an epitaxial layer for the <111> case.

Another nondestructive technique is based on measuring the length of the side of a stacking fault on the film surface. However, this technique is only valid if the stacking fault was nucleated at the film/substrate interface. In some cases, it may be necessary (but undesirable) to delineate the stacking fault more clearly by etching techniques (which thereby make the measurement destructive). This technique is based on the premise that the stacking fault edge length is related to the epitaxial thickness, t_e, by the relationship:

$$t_e = KL \tag{7.10}$$

where L is the length of the fault side [the side of a square for (100) oriented silicon, and the side of an equilateral triangle for (111) silicon]. The K value is a geometric factor that depends on the crystal orientation. The values of K are 0.707 for (100) and 0.616 for (111) orientations. Figure 7-24 shows the geometrical relationship expressed in Eq. 7-10.

7.9.4 Infrared Reflectance Measurement Techniques

When electromagnetic radiation (particularly in the IR range) is incident upon a lightly-doped epitaxial film on a heavily-doped Si substrate the following sequence of events can occur. First, the incoming radiation is partially reflected from the air/film surface and partially transmitted through the film where it strikes the substrate. The radiation reaching the substrate is also reflected (or may be partially transmitted for lightly-doped or sapphire substrates) and this reflected radiation combines with the radiation reflected from the air/film surface. The combined radiation results in an interference pattern, which amounts to intensity variations of the reflected radiation depending on the wavelengths of incident radiation. That is, the interference patterns which are formed depend on the frequency (wavelength) of the incident energy as well as on the optical properties of the film and the substrate and the thickness of the film. Lightly-doped epitaxial layers are relatively transparent in the infrared spectrum in the range of 2.5 to 50 μm. Accordingly, this wavelength range is typically used in the IR-reflectance measurement technique. The heavily-doped substrate behaves as a reflective surface for radiation in this range, and thus the IR reflectance technique is applicable for determining epi layer thickness.

7.9.4.1 ASTM Technique of Epi Layer Thickness Measurement with IR Reflectance: The American Society for Testing Materials (ASTM) has developed a standard technique for determining the epitaxial film thickness based IR reflection.[53] The thickness measured by the IR

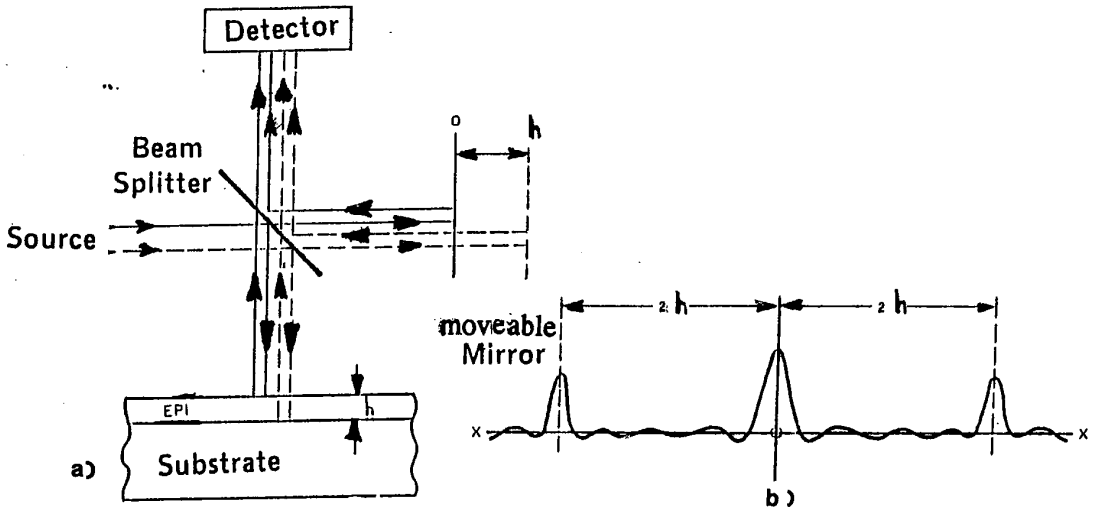


Fig. 7-25 Schematic representation of interferometric measurement of epitaxial layer thickness. (a) interferometer showing path lengths. (b) interferogram showing a coherence peak and side-bursts.

reflectance technique gives a relatively good indication of the added physical thickness if substrate outdiffusion and autodoping can be neglected. For low temperature deposition (≤1050°C), where these effects are minimal, the transition region can introduce an error of 2–3% with respect to the physical thickness (measured by weight gain).

7.9.5 Fourier-Transform Infrared (FTIR) Spectroscopy

The use of Fourier-transform infrared spectroscopy has led to more accurate and rapid measurements of epitaxial layer thickness than the use of standard IR reflectance techniques. A detailed discussion of the technique is beyond the scope of this book, but is well documented in other publications.[54] Briefly, a Michelson interferometer (which consists of a fixed and a movable mirror) is used to collect reflectance data from the epi wafer for all wavelengths (frequencies) used in the IR source. Figure 7-25 shows a schematic of such an interferometer.

Since the substrate and the epi layer normally have substantially different doping levels, they will have different indicies of refraction. As a result the IR will reflect from the epi/substrate interface and the reflection spectrum will exhibit interference patterns depending on the thickness and optical constants of the film. In practice, the incident IR beam is split, and one path travels to the substrate, where it is reflected and is detected by the detector. The second light path is reflected from the movable mirror and is similarly detected by the detector. When the two path lengths of the split beam are equal (solid line in Fig. 7-25), the beam is nearly coherent, and a large signal is detected. When the movable mirror is at a distance equal to the *optical path length* of the IR beam in the epitaxial layer, a second coherence is achieved resulting in what is termed *side-bursts*. The distance between the center and first side-bursts are just equal to twice the product of the epi thickness and index of refraction of the epi film. The resulting interferometer spectrum showing the central and side bursts is called an interferogram, and an example of such an interferogram is shown in Fig. 7-26. By performing a Fourier transform on the interferogram, the normal spectrum of the reflected light can be obtained.

In order to determine the thickness of the layer from the FTIR measurement it is necessary to include correction factors for phase shift and index of refraction variations. Since these parameters vary, depending on the substrate and layer doping, the measurements are empirical in nature. A commercially available system from the ADE Corporation uses so-called "model-based" techniques, and correlates the complete reflection spectrum from the epitaxial specimen

with the Drude general multi-layers reflectance model (using an iterative process).[55] The ADE instrument measures the absolute IR reflectance of the sample over a broad spectral range. A synthetic reflectance spectrum is then calculated from first principles, using three independent parameters: epi layer thickness, transition layer width, and substrate carrier concentration. A least-square fitting parameter is also calculated to serve as a measure of the confidence of the fit between the calculated and measured spectrum. The transition region is modeled by assuming it consists of uniformly-doped very thin layers of epi that decrease (staircase) as the transition region moves away from the substrate. As a result, this technique not only gives the value of the epi thickness, but also the transition-layer width and the substrate doping concentration. This technique is highly repeatable (≤ 0.002 μm) and works well for epi film thickness as low as 0.6 μm. The technique is extremely rapid (one measurement/sec) and has a relatively small spot size (~2-mm). This permits measurements out to near the edge of the wafer, thereby complying with 2–3 mm edge exclusion. Since the measurement is extremely rapid, it is possible to map the thickness across the wafer and thereby investigate the uniformity of the deposition.

7.10 SILICON-ON-INSULATORS (SOI)

Silicon-on-insulator technology has been available since the late 1960's. However, it has taken more than 30 years for the first mainstream SOI commercial devices to be offered. The advantages of SOI technology are well known. First, circuits fabricated in SOI wafers have reduced parasitic capacitance when compared to bulk or epi wafers. Less capacitance translates to lower power or higher speed. Second, SOI devices have improved radiation-induced single-event-upset (SEU) immunity, and thus they are useful for space applications. Even large memory arrays that are used in terrestrial applications are subject to SEU and may require SOI circuits. Third, SOI devices are completely free of latchup. Finally, the fabrication process in SOI wafers can be simplified by reducing the number of masks by as much as 30%. However, SOI wafers are more expensive to fabricate and may contain more defects than bulk or epi-silicon. As the requirements for improved speed, reduced power, and less expensive processing

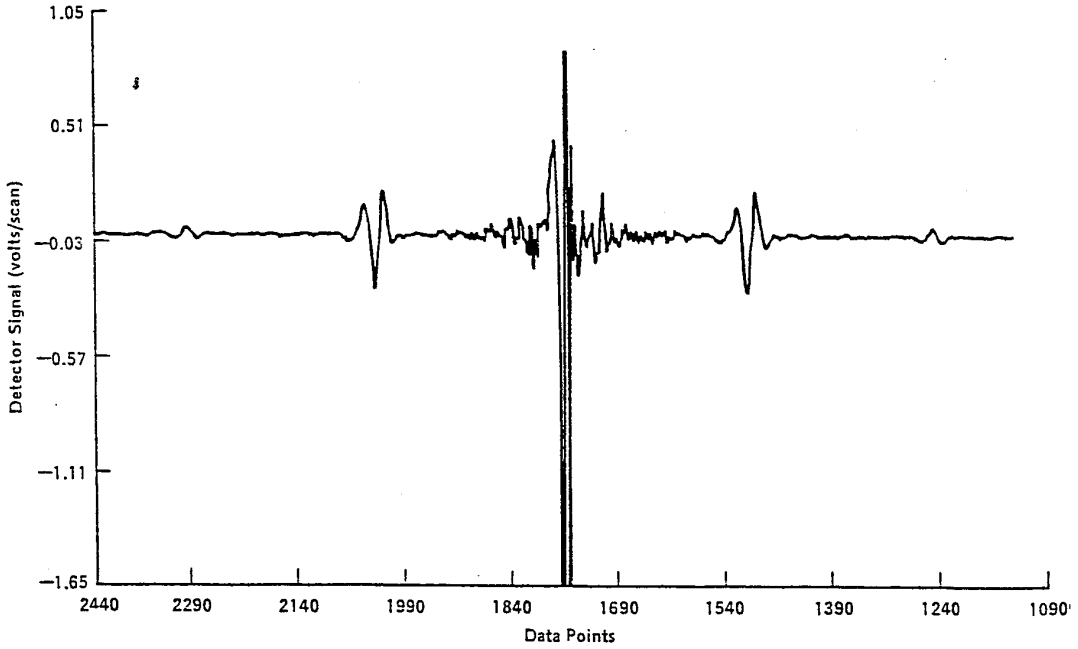


Fig. 7-26 Typical interferogram obtained for an epitaxial thickness measurement.

have become more important, a keen interest in SOI technology has been re-kindled among the mainstream device suppliers.[56]

Currently there are three useful approaches for making SOI wafers. They are: 1) *Silicon-On-Sapphire (SOS)*; 2) *Separation by IMplanted OXygen (SIMOX)*; and 3) *wafer bonding (WB)*. The insulating substrate for SOS wafers is sapphire (i.e., very pure aluminum oxide [Al_2O_3]}. On the other hand, the insulator for the SIMOX and wafer bonding technologies is silicon dioxide. Various (and to date unsuccessful) attempts have been made to grow single-crystal silicon (or large-grain polycrystalline silicon) directly on amorphous insulators.[57] When building devices in SOI wafers, small islands of silicon are formed (typically by dry etch techniques) on top of the insulator. Individual devices are then fashioned in the islands and these devices are then interconnected in the conventional way. SOI technology has been the subject of intense research and development in the late 1990s and a number of symposia are held on a regular basis that address this area.[58,59]

7.10.1 Silicon-on-Sapphire (SOS)

The work on SOS dates back to the 1960s when there was interest in this technology for its well-known immunity to radiation damage. There are several excellent review articles on this subject and the reader is referred to them for additional details.[60,61] Since SOS technology is limited to fabrication of MOS devices, a (100) orientation of the crystalline-silicon film is desirable. This orientation has been heteroepitaxially grown on several different sapphire orientations, specifically (0112), (1012), and (1102). The lattice parameter mismatch between the epi layer and the substrate is sufficiently large to cause the generation of misfit dislocations, stacking faults, and edge dislocations in the epi layer. The defect concentration is heaviest near the substrate/epi interface and decreases as the film grows.[62] Figure 7-27 shows the stacking fault density as a function of epitaxial film thickness in a thick SOS film. However, current requirements for fabrication employ so-called *fully-depleted devices* and these must be built in an epi film only 0.1–0.3 μm thick. These films can be highly faulted, and require additional processing to improve their crystalline quality.

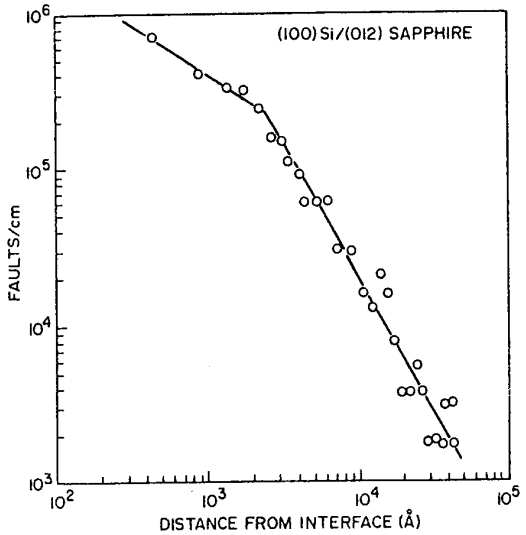


Fig. 7-27 Stacking fault density as a function of the distance from the silicon/sapphire interface in an MOS structure.[36] Reprinted with the permission of the American Physical Society.

The accepted way to improve the film quality is the technique of amorphization (by implanting a high dose of Si) followed by solid-phase-epitaxial regrowth (SPE).[63] In thick SOS films only a single amorphization is required. The ion implantation is first targeted to cause the region near the interface to become amorphized. Then the undamaged single crystal Si near the surface is used as seed material for the epitaxial regrowth of the amorphous layer. Since the regrowth is primarily along the <100> direction, stacking faults and twins do not propagate, leaving a controllable number of dislocations in the film. For thinner SOS films a double SPE (DSPE) process may be required. In this process, the Si region near the sapphire interface is first amorphized by an high-energy ion implantation, and the surface regions are used as a seed for regrowth.[64] This is followed by a low-energy ion implantation that then amorphizes the surface region of the Si. The previously re-crystallized bottom layer is now utilized as the seed to regrow the surface amorphous layer. Devices fabricated in such material show higher surface mobilities and lower junction leakage than either untreated SOS, or material obtained from a single SPE process.[65] Peregrine Semiconductor, which markets high performance SOS-rf devices for the commercial market, builds these devices using standard ULSI processing techniques on 200-mm SOS wafers. SOS films may require additional characterization over that used for conventional epi films. These methods are comprehensively reviewed by Cullen *et al.*[66]

7.10.2 SIMOX

A widely used approach to manufacture SOI wafers is that of SIMOX. Although SIMOX wafers can be obtained from several commercial suppliers, they are still significantly more costly than bulk or even epi wafers.[67] The SIMOX technology is based on the use of very high dose oxygen ion implantation to produce subsurface layers of silicon dioxide.

Figure 7-28 shows a schematic of how SIMOX layers are fabricated. The sequence of steps is as follows:

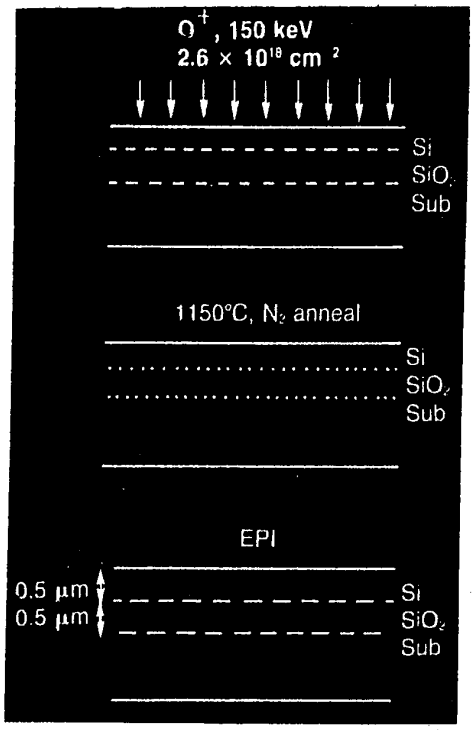


Fig. 7-28 A typical example of the formation of a buried oxide in a silicon substrate.

1. The starting material is typically a (100) device quality wafer. The wafer is first subjected to a high dose (~$2x10^{18}/cm^2$) oxygen (O^+) ion implantation step at high enough energy (150–300 keV) so the peak (projected range) of the implant is deep within the silicon (0.3–0.5 μm). This step is usually carried out with the wafers held at > 400°C to ensure that the silicon maintains its crystallinity during the implantation.
2. The wafers are given a post-implant anneal in N_2 for sufficient time (3–5 hours) at a relatively high temperature (1100–1175°C). This step forms a buried oxide (BOX) layer of silicon dioxide near the peak of the implantation and removes many of the defects (dislocations) formed during the ion implantation step. The BOX depth is defined by the depth of the ion implantation and the BOX thickness is limited to about 0.5 μm.
3. If the device requires thicker silicon than can be achieved with the depth of ion implanted oxygen, an epitaxial Si layer can be deposited to provide this additional thickness.

The are several shortcomings associated with SIMOX technology. First, the formation of SIMOX wafers requires extremely high doses of oxygen, and this necessitates very long implantation times. Second, transition metals may be sputtered from the beam-line or end station components during the implantation, and these metals can become incorporated in the SIMOX wafers and reduce the minority carrier lifetime of the devices. A SIMOX dedicated ion implanter from IBIS Technology is now available. This implanter can furnish adequate throughput and acceptably-low levels of metallic contamination. This production tool has an output of 70 mA of O^+ at an acceleration voltage 210 keV. The total transition metal contamination on the wafers (from Ti to Zn) has been measured to be less than $5x10^{10}$ atoms/cm^2. A third issue is, that despite the high temperature anneal, the surface silicon may retain a dislocation density on the order of $10^3/cm^2$. The presence of such defects can cause reduced oxide integrity and increased junction leakage. However, devices fabricated in the best SIMOX material can be used to build excellent quality CMOS and bipolar ULSI circuits.[68] Finally, SIMOX wafers are more costly than bulk wafers, although the cost is expected to sharply decrease as the popularity of this technique expands.

7.10.3 Wafer Bonding

Another popular technology for producing SOI wafers is *wafer bonding*.[69] Wafer bonding is a simple concept. Basically two oxidized Si wafers (or one oxidized wafer and one bare wafer) are fused together, face-to-face, through a furnace process. Then most of the silicon from one of the wafers is removed and the device is built into the remaining silicon over the now-buried oxide. Unlike SIMOX, the wafer bonding technique can form BOX layers up to 50 μm thick.

There are several approaches for achieving bonding. In one method, two oxidized silicon wafers are held together (face-to-face) by applying external pressure to the wafers while they are exposed to an oxidizing ambient at about 700°C.[70] Bonding is believed to take place by the conversion of oxygen, trapped between the wafers, to SiO_2. The conversion consumes the oxygen and creates a partial vacuum that forces the wafers into intimate contact, allowing the chemical bonding reaction to proceed. The chemical reaction is believed to be the polymerization of silanol bonds to form siloxane networks. A second approach employs a moderate voltage across two oxidized wafers placed face-to-face at temperatures in the range of 1100–1200°C.[71] Since no mechanical force is required when the wafers are first placed in the furnace, this method is less likely to introduce crystalline defects. A third approach is to use one oxidized wafer and one bare wafer. Both are treated with an acid solution that forms O-H

groups on the surface of the wafers. After drying the wafers are placed face-to-face at room temperature in clean air. After this treatment a self-adhesive contact is formed which holds the wafers together during the bonding process. The bonding is completed by annealing the two wafers at 1100°C in a nitrogen ambient for 4 hours.[72] Recently it has been demonstrated that bonding can occur at temperatures as low as 150°C, if the process is performed in a moderate vacuum (e.g., several torr).[73] It is believed that the vacuum significantly reduces the amount of trapped nitrogen at the interface. Trapped nitrogen can prevent an intimate contact of the bonding surfaces during annealing, and hinders formation of the required covalent bonds.

Once the wafers are bonded it is necessary to remove most of the silicon from one of the wafers to produce the SOI structure. There are several proposals being evaluated for this removal. In one method the wafer that will ultimately contain the device consists of a lightly-doped epitaxial layer on a heavily-doped wafer. After oxidation and bonding of the lightly-doped surface of this wafer to another wafer, the heavily doped substrate is removed by either preferential etching or careful lapping (chemical mechanical polish) of the heavily-doped region of the wafer. The epi layer that remains bonded to the second wafer determines the thickness of the active device layer and the sum of the oxides on the two wafers determines the thickness of the buried oxide. One of the issues with this approach is that two silicon wafers are required to make one device SOI wafer.

A recent method, called Smart Cut,® allows the sacrificial wafer to be reused and only one wafer is required to make a bonded SOI substrate. This one-wafer technology is termed Unibond.® Figure 7-29 shows how the process works. Two device-quality wafers (labeled A

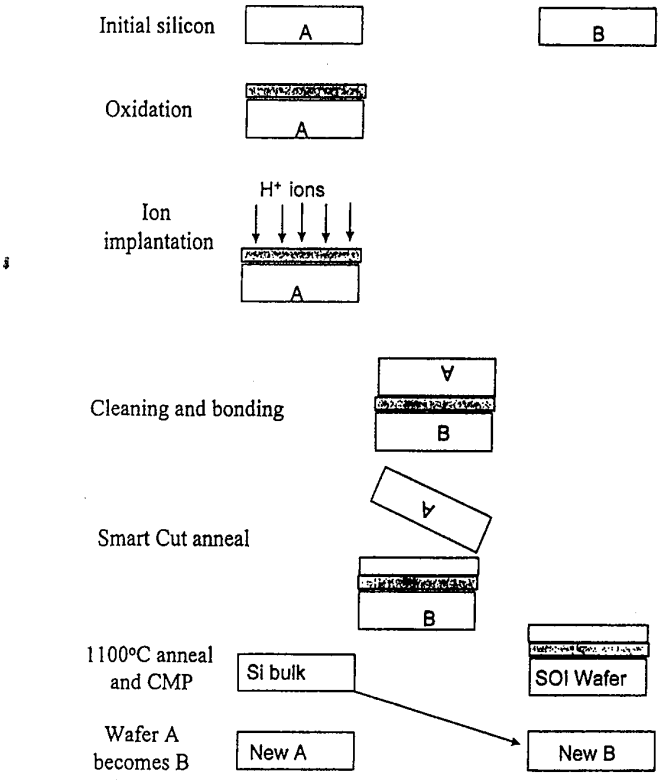


Fig. 7-29 Process flow for fabricating Smart Cut® wafers.

and B) start the process. Wafer A is oxidized (to about 0.5 μm) to form what will become the buried oxide. Then wafer A is ion implanted with H^+ to about 0.2 μm below the oxide surface. The depth of the H ion implantation determines the thickness of silicon above the BOX. Wafers A and B are now cleaned and low temperature bonded face-to-face. The bonded wafers are next annealed (at $\leq$600°C). During this anneal the implanted hydrogen forms a gas and wafer A delaminates at the peak of the implanted hydrogen. Wafer B, which now has a BOX layer becomes the SOI wafer, while wafer A becomes the sacrificial wafer. Both wafers are annealed at 1100°C to remove defects. This is followed by a "light" CMP to refresh the surfaces. Wafer A, which is now a fraction of a micron thinner than it was before the start of the process is used as wafer B when the process starts over. The method has been used to prepare commercially available 200 mm, SOI wafers with a silicon thickness above the BOX of 0.2 μm and a uniformity of better than 5%. The dislocation density is $<10^2/cm^3$, the surface roughness <2Å (rms), and metal contamination $<5x10^{10}/cm^3$.[74]

REFERENCES

1. The National Technology Roadmap for Semiconductors - Technology Needs - 1997 Edition, Semiconductor Industry Association, p. 64.
2. J. Bloem, "Nucleation and Growth of Silicon by CVD," *J. Cryst. Growth,* **50**, 581 (1980).
3. F.C. Eversteyn, "Chemical Reaction Engineering in the Semiconductor Industry," *Philips Research Rept.*, **19**, 45 (1974).
4. W.A.P. Claasen and J. Bloem, "Rate Determining Reactions and Surface Species in CVD of Silicon II: The SiH_2Cl_2-H_2-N_2-HCl System," *J. Crystal Growth*, **50**, 807 (1980).
5. P.A. Coon, M.L. Wise, and S.M. George, "Modeling Silicon Epitaxial Growth with SiH_2Cl_2," *J. Crystal Growth*, **130**, 162 (1993).
6. V.S. Ban and S.L. Gilbert, *J. Crystal Growth.*, **32,** 284 (1975).
7. V.S. Ban and S.L. Gilbert, *J. Electrchem. Soc.*, **122**, 1382 (1975).
8. V.S. Ban, *J. Electrchem. Soc.*, **122**, 1389 (1975).
9. V.S. Ban, *J. Electrchem. Soc.*, **125**, 317 (1978).
10. J. Nishizawa and M. Saito, "Growth Mechanism of Chemical Vapor Deposition of Silicon," *Proc of the VIIIth International Conf. on Chemical Vapor Deposition,* 1981, Electrochemical Society, NJ, 1981, p. 317.
11. H.R. Chang, "Defect Control for Silicon Epitaxial Process Using Silane, Dichlorosilane, and Silicon Tetrachloride," *Proc. of the Symposium on Defects in Silicon*, W.M. Bullis and L.C.Kimerling, Eds., Electrochem. Soc., NJ, pp. 549-557 (1983).
12. S.B. Kulkarni, "Defect Reduction by Dichlorosilane Epitaxial Growth", *Proc. of the Symposium on Defects in Si*, W.M. Bullis, and L.C. Kimerling, Eds, Electrochem. Soc., p. 558-567 (1983).
13. M.L. Yu, D.J. Kitkavage, and B.S. Meyerson, "Doping Reaction of PH_3 and B_2H_6 with Si (100)," *J. Applied Phys.*, **59,** 4032 (1986).
14. R. Reif, T. Kamins, and K. Saraswat, "A Model for Dopant Incorporation into Growing Silicon Epitaxial Films," *J. Electrochem. Soc.*, **126**, 644 (1979).
15. G.R. Srinivasen, "Autodoping Effects in Si Epitaxy," *J. Electrochem. Soc.*, **127**, 1334 (1980).
16. G.R. Srinivasen, "Kinetics of Lateral Autodoping in Silicon Epitaxy," *J. Electrochem. Soc.*, **125**, 146 (1978).
17. G.W. Rubloff, "Defect Microchemistry in SiO_2/Si Structures," *J. Vac. Sci. Technology*, **A8**, 1857 (1990).
18. F.W. Smith and G. Ghidini, "Reaction of Oxygen with Si(111) and (100): Critical conditions for Growth of SiO_2," *J. Electrochem. Soc.*, **129**, 1300 (1982).
19. G. Ghidini and F.W. Smith "Interaction of H_2O with Si (111) and (100)," *J. Electrochem. Soc.*, **131**, 2924 (1984).

20. K.V. Ravi, *Imperfections and Impurities in Semiconductor Silicon*, Wiley, New York, 1981.

21. R.E. Logar and J.O. Borland, "Silicon Epitaxial Processing Techniques for Ultra-Low Defect Densities," *Solid State Technology*, p. 133, June 1985.

22. J.O. Borland *et al.*, "Influence of Epi-Substrate Point Defect Properties on Getter Enhanced Si Epitaxial Processing for Advanced CMOS and Bipolar Technologies," *VLSI Science and Technology - 1984*, K.E. Bean and G. Rozgonyi, Eds., Electrochem. Society, NJ, p. 93.

23. W. Dyson *et al.*, "n^+ and p^+ Substrate Effects on Epitaxial Silicon Properties," *VLSI Science and Technology - 1984*, K.E. Bean and G. Rozgonyi, Eds., Electrochem. Society, NJ, p. 107.

24. S.M. Ali *et al.*, "Extrinsic Gettering with Epitaxial Misfit Dislocations," *Semiconductor Processing*, ASTM, STP 850, Gupta (1984).

25. K.E. Bean, W.R. Runyan, and R.G. Massey, "Silicon Epitaxial Growth: A Critical Technology," *Semiconductor International*, p. 136 (May 1985).

26. S.M. Ali *et al.*, "Extrinsic Gettering Via the Controlled Introduction of Misfit Dislocations," *Appl. Phys. Lett.*, February 1985.

27. P.H. Lee, M.T. Wauk, R. Rosler, and W. Benzing, "Epitaxial Pattern Shift Comparisons in Vertical, Horizontal & Cylindrical Reactor Geometries," *J. Electrochem. Soc.*, **124**, 1824 (1977).

28. S.P. Weeks, "Pattern Shift and Pattern Distortion during CVD Epitaxy on <111> and <100> Silicon," *Solid State Technology*, **24**, 111 (1981).

29. J.C. Bean, "Silicon Molecular Beam Epitaxy: 1984-1986," *J. Cryst. Growth*, **81**, 411 (1987).

30. D.J. Eaglesham, G.S. Higasi, and M. Cerullo, *Appl. Phys. Lett.*, **59**, 685, (1991).

31. C.-H. Chen and T.-R. Yew, "Silicon Epitaxial Growth by Plasma-Enhanced Chemical Vapor Deposition from SiH_4/H_2 at 165-350^{0}C," *J. Cryst. Growth*, **147**, 305, (1995).

32. R. Wise, D. Frystak, J. Barnett, P. Grothe, B. Fowler, and E. Forest, "Investigation of Single-Wafer Low Temperature Epitaxial Silicon Deposition with Clustered HF Pre-Clean," *Electrochem. Soc. Proc;* **Vol. 96-5**, Electrochem. Soc., NJ (1996).

33. C.-L. Wang, S. Unnikrishnan, B.-Y. Kim, D.-L. Kwong, and A.F. Tasch, "The Viability of SeH_4-Based *In Situ* Clean for Low Temperature Silicon Epitaxial Growth," *J. Electrochem. Soc.*, **143**, 2387 (1996).

34. K. Oda and Y. Kiyota, "H_2 Cleaning of Silicon Wafers Before Low-Temperature Epitaxial Growth by Ultrahigh Vacuum/Chemical Vapor Deposition," *J. Electrochem. Soc.*, **143**, 2361 (1996).

35. B.S. Meyerson, "Low Temperature Si and Si:Ge Epitaxy by Ultrahigh Vacuum/Chemical Vapor Deposition: Process Fundamentals," *IBM J. Res. Dev.*, **34**, 806 (1990).

36. K.E. Viplette, P.A. O'Neill, M.C. Ozturk, K. Christensen, and D. Maher, "Low Temperature Selective Silicon Epitaxy by Ultra High Vacuum Rapid Thermal Chemical Vapor Deposition Using Si_2H_6, H_2 and HCl," *Appl. Phys. Lett.*, **68**, 66 (1996).

37. M. K. Sanganeria, M. Ozturk, G. Harris, K. E. Violette, I. Ban, C.A. Lee, and D. Maher, "Ultrahigh Vacuum Rapid Thermal Chemical Vapor Deposition of Epitaxial Silicon onto (100) Silicon," *J. Electrochem. Soc.*, **142**, 3961 (1995).

38. E. Kasper and K. Worner, "Application of Si-MBE for Integrated Circuits," in *VLSI Science and Technology 1984*, K.E. Bean and G.A. Rozgonyi, Eds, Electrochem. Society, NJ, p. 429.

39. H. Hirayama, T. Tatsumi, and N. Aizaki, "Gas Source Molecular Beam Epitaxy Using Disilane," *Appl. Phys. Lett.*, **52**, 1484 (1988).

40. C.Y. Chang and S.M. Sze, *ULSI Technology*, McGraw Hill, New York (1995) p. 130.

41. J. Borland and C. Drowley, *Solid State Technology,* August 1985 p.141.

42. J. Borland *et al.*, "Silicon Epitaxial Growth for Advanced Structures," *Solid State Technology,* January 1988, p. 111.

43. J. Bosch, "Epitaxial Process is Highly Selective in Depositing Silicon," *Electronics International*, pp. 59-60, Jan. 31, 1980.

44. T.O. Sedgwick, P.D. Angnello. M. Berkenblit and T.S. Kuan, *J. Electrochem. Soc.*, **138**, 3961 (1991).

45. M. Murata, N. Nakamura, M. Kato, N. Mikoshiba and T. Ohmi, *Appl. Phys. Lett.*, **54**, 1007 (1989).
46. T. Tasumi, K. Aketagawa, M. Hiroi and J. Sakai, *J. Cryst. Growth*, **120**, 275, (1992).
47. L. Jastrzebski, J.F. Conroy, and R. Soydan, "Issues and Problems Involved in Selective Epitaxial Growth of Silicon in SOI Fabrication," *J. Electrochem. Soc.*, **136**, 3506 (1989).
48. C.H.J. van den Breckel, "Characterization of Chemical Vapor Deposition Processing," *Philips Research Reports*, **32**, 118 (1977).
49. D.K. Schroder and H.C. Nathanson, *Solid State Electronics*, **13**, 577 (1970).
50. D.K. Schroder, *Semiconductor Materials and Characterization*, John Wiley & Sons (1990) p. 417.
51. E.O. Johnson, "Measurement of Minority Carrier Lifetime with Surface Photovoltage," *J. Appl. Phys.*, **28**, 1349 (1957).
52. D.K. Schroder, *Semiconductor Materials and Characterization*, John Wiley & Sons (1990) p. 378.
53. "Thickness of Epitaxial Layers of Silicon on Substrates of Same Type by Infrared Reflectance," 1984 Annual Book of ASTM Standards, Vol. 10.05, p. 213, ASTM, Philadelphia (1984).
54. A.L. Smith, *Applied Infrared Spectroscopy*, John Wiley & Sons, New York, 1979.
55. S. Charpenay, P. Rosenthal, G. Kneisal, C. Gondran, and H. Huff, "Model Based Analysis for Precise and Accurate Epitaxial Silicon Measurements," *Sol. State Technology* , p.161, July, 1998.
56. IBM Microelectronics
57. H.W. Lam, A.F. Tasch, Jr., and R.F. Pinzzotto, "Silicon-on-Insulator for VLSI and VHSIC," in *VLSI Electronics*, Ed. N.G. Einspruch, Vol. 4, p. 1, Academic Press, NY (1982).
58. For example, *Proceedings of the Fourth International Symposium on Semiconductor Wafer Bonding: Science, Technology, and Applications*, Eds. U. Gosele, H. Baumgart, H. Abe, C. Hunt, and S. Iyer, Electrochem. Soc. Pennington, NJ (USA), 1998.
59. IEEE International SOI Conference, Electron Device Society, held in Stuart, FL, Oct. 5-8, 1998.
60. G.W. Cullen and C.C. Wang, Eds., *Heteroepitaxial Semiconductors for Electronic Devices*, Springer-Verlag, 1978.
61. G.W. Cullen *et al.*, *J. Crystal Growth*, **56**, 287 (1982).
62. M.S. Abrahams and C.J. Buiocchi, "Cross Sectional Electron Microscopy of Silicon on Sapphire", *Appl. Phys. Lett.*, **27**, 325 (1975).
63. A. Gupta and P.K. Vasudev, "Recent Advances in Hetero-Epitaxial Silicon-on-Insulator Technology, Part II," *Solid State Technology*, **26**, 129 (1983).
64. R. Reedy, T.W. Sigmon, and L.A. Christer, "Suppressing Al Outdiffusion in Ion Implantation Amorphized and Recrystallized Silicon on Sapphire Films," *Appl. Phys. Lett.*, **22**, 537 (1983).
65. G.A. Garcia, and R.E. Reedy, "Electron Mobility Within 100 nm of the Si/Sapphire Interface in Double-Solid Phase Epitaxially Regrown SOS," *Electron. Lett.*, **22**, 537 (1986).
66. G.W. Cullen, M.T. Duffy, and R.K. Smeltzer, "Recent Advances in Heteroepitaxial Si on Sapphire Technology," *VLSI Science and Technology/1984*, Electrochem. Society, NJ, 1984, p. 230.
67. One supplier of SIMOX wafers is IBIS Technology Corp.
68. C.E. Chen *et al.*, *1984 IEDM Tech. Digest*, 701 (1984).
69. *Wafer Bonding and SOI Technologies*, Abstracts of the Fall, 1997 Meeting of the Electrochemical Society, Paris, France.
70. J.B. Lasky *et al*, "SOI by Bonding and Etchback," *Tech. Dig. IEDM*, 1985 p.684.
71. R.C. Frye *et al*, *J. Electrochem. Soc*, **xx**, 1673 (1988).
72. H. Ohashi, *Tech. Dig. IEDM*, 1986 p. 210.
73. Q.-Y. Tong, W.J. Kim, T.-H. Lee, and U. Gosele, "Low-Vacuum Wafer Bonding," *Electrochemical and Solid State Letter*, **1**, 52 (1998).
74. Unibond Wafer Specification Sheet. SOITEC USA Inc. Peabody, MA, (1998).
75. After S.A. Campbell, *The Science and Engineering of Microelectronic Fabrication*, Oxford University Press, Oxford/New York, 1996, p. 343. Redrawn from data presented in Refs. **19** and **18**.

76. T. Makino *et al., Ext. Abs. Electrochem. Soc. Mtg., Spring 1988,* Abs. No.193, p. 301, 1988.
77. H.C. Theurer, *J. Electrochem. Soc.*, **108**, 649 (1965).
78. McD. Robinson *et al.,* "Low-Dislocation Density RF-Heated Epitaxial Silicon," *J. Electrochem. Soc.*, **129**, December 1982, p. 2858.
79. McD. Robinson and L.H. Lawrence, "Characterization of High Growth Rate Epitaxial Silicon from a New Single Wafer Reactor," in D. Gupta, Ed., *Semiconductor Fabrication: Technology and Metrology,* ASTM STP 990, 1989, p. 30-42.

PROBLEMS

1. Why is epitaxial deposition done in cold-wall reactors (see p. 167 for definition of this term)?
2. What is the minimum percentage of HCl that could be used at 1250°C without causing etching of the silicon substrate?
3. Find the mole fraction of $SiCl_4$ that will result in the maximum film growth rate. Why is epitaxial deposition not carried under these conditions?
4. Determine the time necessary to grow a monolayer of Si epitaxially under the conditions shown in Fig. 6-3b of Chap. 6, at 1200°C.
5. Calculate the activation energies from the Arrhenius plots of Si growth rate versus temperature (Fig. 7-6).
6. Explain why a highly abrupt junction is difficult to achieve using epitaxial deposition.
7. Explain how mask compensation is implemented when laying out a pattern to be transferred to an epitaxial wafer.
8. Sketch a stacking fault in an epitaxial layer on a {100} surface. The fault propagates from layer to layer, retaining its shape and increasing in size with each successive atomic plane. Show the planes along which this fault will propagate. Find the relationship between the height of the stacking fault and its sides.
9. Determine the growth rate of an epitaxial film grown from a $SiCl_4$ source at 1200°C. The gas phase mass transfer coefficient of the reactor is h_g = 5 cm/sec, the surface reaction coefficient is $k_s = 10^7$ exp (-1.9 eV /kT) cm /sec, and $C_g = 5x10^{16}$ cm^{-3}. By how much will the growth rate change if the reaction temperature is decreased by 2°C?
10. Calculate the thickness for <111> silicon of an epitaxial layer in which the etch pits have a dimension of 1.838 μm.

Chapter 11

ALUMINUM THIN FILMS and PHYSICAL VAPOR DEPOSITION in ULSI

This chapter is divided into three major topics involved in the formation of interconnect structures in VLSI. These topics are:

1. *The Properties of Aluminum (and Aluminum Alloy) Thin Films in ULSI.* This section considers the advantages and limitations of Al and Al-alloy thin films for ULSI applications. It should be noted that many of the fabrication processes that involve aluminum are also discussed elsewhere in the chapter (e.g., the deposition of aluminum films by sputtering), or in other chapters of this volume (e.g., dry etching of aluminum in Chap. 14). Several issues associated with the use of thin film aluminum are addressed in other parts of this text as well, including: a) step coverage; b) effects of gas incorporation in the deposited film; c) electromigration (Vol. 2 of this series); d) hillocks in Al thin films and their suppression (Vol. 2); and e) contact formation to silicon, including spiking, contact electromigration failure, and contact structures/barrier layers (here and Vol. 2).

2. *Physical Vapor Deposition (PVD) by Sputtering.* The discussion on sputtering is quite extensive, as this is the primary PVD method utilized in ULSI applications. The sputtering presentation covers the following subjects: a) glow-discharge physics; b) the physics of sputtering; c) deposition kinetics of sputtered films; d) sputter system configurations (dc, rf, and magnetron); e) commercially available sputtering equipment for microelectronic applications; and f) sputter deposition processing issues.

3. *Thickness Evaluation of Opaque (Especially Metallic) Thin Films.*

Since the key topic of this chapter is *physical vapor deposition* (PVD), it is useful to define the concept and introduce the common characteristics shared by PVD processes. All PVD processes proceed according to the following sequence of steps:

1. The material to be deposited (from a solid or liquid source) is physically converted to the vapor phase.
2. The vapor is transported across a region of reduced pressure (from the source to the substrate).
3. If the substrate is located near the surface being vaporized, some of the vapor produced can condense to form a thin solid film on the substrate.

In sputter deposition, the conversion to a gaseous phase (Step 1) is done by the physical dislodgment of surface atoms by momentum transfer. This step will receive the most attention in the chapter, because some aspects of film condensation and formation were discussed in earlier chapters dealing with thin films, epitaxy, and CVD. It should also be noted that many concepts

Exhibit 1014 - Page 97 of 218

Table 11-1 PROPERTIES OF ALUMINUM AND ALUMINUM ALLOY THIN FILMS

Name	Symbol	Melting Point (°C)	Al/Si Eutectic (°C)	Density (g/cm^3)	Resistivity ($\mu\Omega$-cm)
Aluminum	Al	660	577	2.70	2.7
Aluminum/ 4% Copper	Al 4%Cu	650	~577	2.95	3.0
Aluminum/ 2% Silicon	Al 2%Si	640	~577	2.69	2.9
Aluminum/ 4% Copper 2% Silicon	Al 4%Cu 2%Si		~577	2.93	3.2

used to describe PVD processes, including vacuum pressure, mean free path, impingement rate, monolayer time, vapor pressure, and gas flow regimes were presented in Chap. 3.

11.1 ALUMINUM THIN FILMS IN ULSI

Aluminum has historically been the third material in the trinity of matrix substances employed to fabricate Si-based solid-state components (the other two being Si and SiO_2). Aluminum and Al-alloy thin films were selected as the workhorse materials for interconnect structures for the first 30 years of the IC industry. As of 1999 Al-alloys continue to be the most widely used materials for IC metallizations, although copper films are beginning to emerge as probable replacements. The dominance of Al for interconnect applications arose because of its low resistivity (ρ_{Al} = 2.7 $\mu\Omega$-cm), and its compatibility with the other two matrix substances.[1,2] Aluminum thin films adhere well to SiO_2, because (during the thermal step that sinters the Al-Si contacts) some of the Al atop the SiO_2 reduces some of the SiO_2 layer to form a thin layer of Al_2O_3 at the Al/SiO_2 interface. This promotes good adhesion between SiO_2 and Al.

Table 11-1 lists the material properties of Al thin films of most interest for silicon device fabrication. The table also indicates that Al alloys are utilized in microelectronic applications, and in fact, films of such alloys are used instead of pure-Al films. The Al alloys that have found most use include: a) Al:1wt%Si; b) Al:2wt%Cu; c) Al:1wt%Si:2wt%Cu; and d) Al:1.2%Si: 0.15wt%Ti. Each will be discussed below. Note that the relatively low values of the melting point of Al (660°C) and the Al-Si eutectic temperature (577°C) restrict the value of subsequent processing temperatures once Al films have been deposited.

Thin films of sputtered aluminum are typically deposited in the thickness range of 500–1500 nm, and these films have a polycrystalline structure. Early microelectronic devices utilized evaporation to deposit pure Al and Al:Si or Al:Cu alloys. However, the stringent alloy-composition requirements of advanced devices (as well as some other limitations of evaporation) gave sputtering an advantage over evaporation for depositing metal films. The development of magnetron sputtering (which allows aluminum to be deposited at deposition rates of up to ~1000 nm/min), caused sputtering to entirely displace evaporation as the PVD technology for depositing Al-alloy films in VLSI and ULSI fabrication. Aluminum alloy films also replaced pure Al films because the alloys possess enhanced properties for certain interconnect requirements, including superior contact formation characteristics and better resistance to electromigration. A brief introduction to these topics follows (see also Vol. 2 of this series).

The use of *Al:Si* in place of *pure Al* stems from the problem of *junction spiking* that occurs at the interface of pure Al and Si. That is, the solubility of silicon in aluminum (Fig. 11-1) rises as the temperature increases (e.g., to about 0.5% at 400°C). When a Si substrate is put into intimate contact with a pure Al film, the wafer becomes a source of Si that can dissolve in the solid Al solvent. In addition, the diffusivity of silicon along the grain-boundaries of the alum-inum film at 400°C is quite high. Thus, a significant quantity of Si can move from the region beneath the Al/Si-substrate interface into the aluminum film. Simultaneously, Al from the film will move to fill the voids created by the departing Si (Fig. 11-2). If the aluminum penetrates deeper than the *pn* junction depth beneath the contact, the junction will be electrically shorted (junction spiking). One way to suppress this set of events is to use an alloy of Al with a con-centration of Si that exceeds the solubility limit of Si at the maximum process temperature, *in lieu* of pure Al. In that case, when the (Al:Si alloy)/Si interface is heated the Al:Si alloy film cannot dissolve additional silicon originating from the substrate, and junction spiking is avoided.

Although this remedy was once effective, it became inadequate for ULSI applications. The shallower junctions and narrow metal lines are less tolerant of the Si precipitation occurring in these Al alloy films as they cool down from the 400°C steps. The Al-alloy films become super-saturated with Si as they cool down, causingitto precipitate. In addition, electromigration failure of the Al:Si-silicon contacts becomes more severe as the contacts get smaller. Thus, the use of Al-Si alloy films to combat junction spiking in ULSI devices was abandoned in favor of contact structures employing barrier layers. The most widely used barrier layer in ULSI is titanium nitride (TiN). Discussion of the deposition of such TiN layers by PVD is given in a later section, and by CVD in Chap. 8. More information on contacts between interconnect materials and the Si substrate is also found in Chap. 15. (See also Vol. 2 of this series for a detailed discussion of the topic of metal-silicon contacts and the inadequacies of Al:Si for VLSI and ULSI).

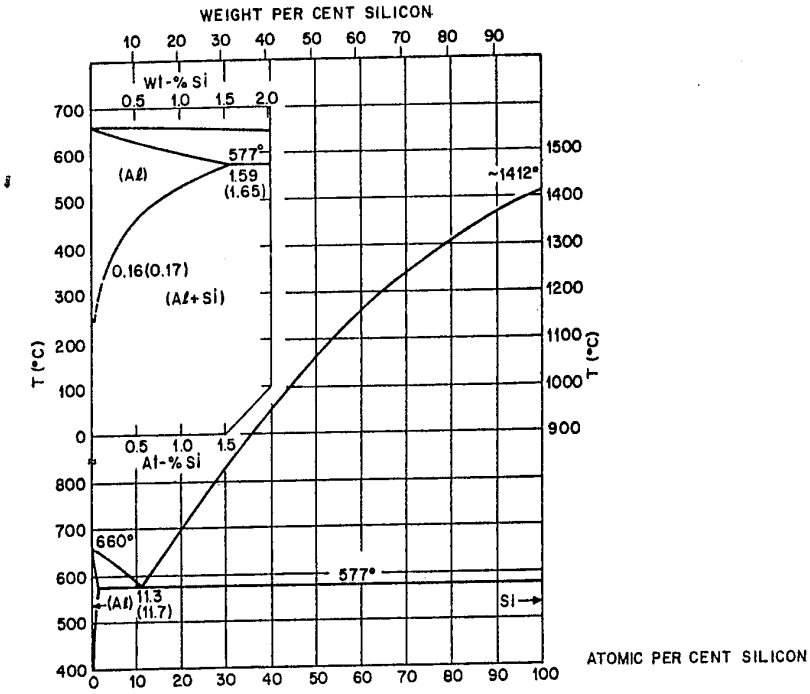


Fig. 11-1 Phase diagram of the aluminum-silicon system. From M. Hansen and A. Anderko, *Constitution of Binary Alloys,* 1958. Reprinted with permission of McGraw-Hill Book Co.[3]

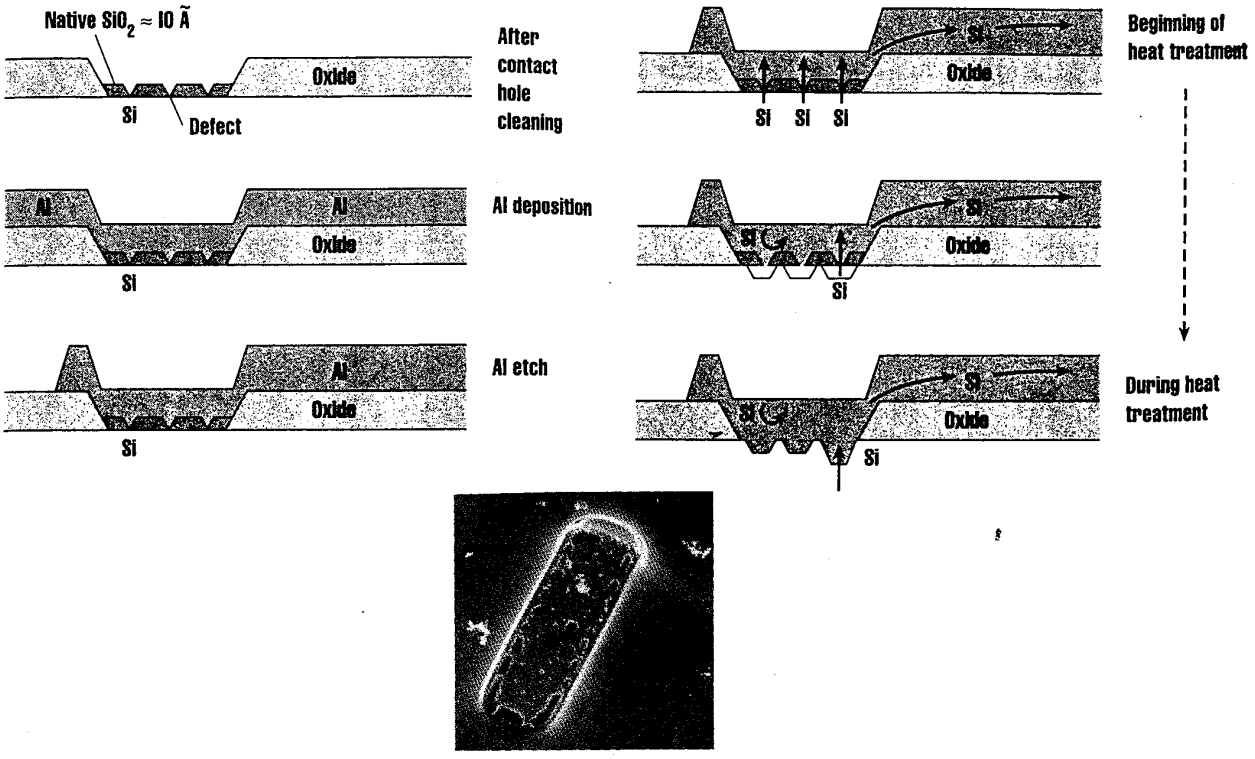


Fig. 11-2 Junction spiking and silicon migration during contact sintering.

It has also been found that the addition of small quantities (0.1–4%) of other materials (e.g., Cu [0.5–4wt%] and Ti [0.1–0.5wt%]) to the aluminum improves the electromigration resistance and reduces the propensity of aluminum thin-films to form hillocks (i.e., protrusions on the Al film surface). One penalty for enhancing these characteristics of Al films is that their resistivity is increased (by 10–30%) over that of pure aluminum films (see Table 11-1). Other drawbacks that must be accepted to gain the improvements offered by Al alloys include added process complexity, and more stringent process controls. An appropriate combination of alloy composition, film deposition, annealing, and etch processes must be selected, developed, evaluated, and maintained. Furthermore, some processes are more difficult to perform on alloy films than on pure Al films, such as dry etching of Al:Cu films (Chap. 14). Details of the issues that must be considered in developing and evaluating processes that utilize Al alloy films are discussed elsewhere, both in this text and in Vol. 2.

Some other characteristics of Al thin films need to be presented at this point. Aluminum readily forms a thin native oxide (Al_2O_3) on its surface upon exposure to oxygen, even at room temperature. The presence of this oxide can affect the contact resistance when another metal layer is deposited onto the Al, and can inhibit both the sputtering of an aluminum target and the etching of aluminum films. Aluminum thin films can also suffer corrosion problems as a result of some fabrication processes. For example, if phosphorus-doped SiO_2 is deposited onto Al films, phosphoric acid (HPO_3) can be formed if moisture is absorbed by the glass. This acid will attack Al and cause corrosion. In addition, dry etching of Al may leave chlorine residues on the Al surfaces. Exposure of these residues to ambient moisture can lead to the formation of HCl. If Cu is present as an alloy in the Al film, severe corrosion can occur (as the $CuAl_2$ compound and the Al form the microelectrodes of a battery, and HCl acts as the electrolyte). Both of these problems are discussed elsewhere in the text in more detail (see Index, under *Al, corrosion*).

11.2 SPUTTER DEPOSITION FOR ULSI

Sputtering is a term used to describe the mechanism in which atoms are ejected from the surface of a material when that surface is struck by sufficiently energetic particles. It has become the dominant technique for depositing a variety of metallic films in VLSI and ULSI fabrication, including aluminum alloys, titanium, titanium:tungsten, titanium nitride, tantalum, and cobalt.

Sputtering displaced the original PVD technique for depositing metal films (evaporation) for the following reasons:

1. Sputtering can be accomplished from large-area targets, which simplifies the problem of depositing films with uniform thickness over large wafers.
2. Film thickness control is relatively easily achieved by selecting a constant set of operating conditions, and then adjusting the deposition time to reach it.
3. The alloy composition of sputter-deposited films can be more tightly (and easily) controlled than that of evaporated films.
4. Many important film properties, such as step coverage and grain structure can be controlled by varying the negative bias and heat applied to the wafers. Other film properties (including stress and adhesion), can be controlled by altering such process parameters as power & pressure.
5. The surface of the substrates can be sputter-cleaned in vacuum prior to initiating film deposition (and the surface is not exposed again to ambient after such cleaning).
6. There is sufficient material in most sputter targets to allow many deposition runs before target replacement is necessary.
7. Device damage from x-rays generated during electron-beam evaporation is eliminated (although some other radiation damage may still occur).

As is true with other processes, however, sputtering also has its drawbacks. They include:

1. Sputtering processes involve high capital equipment costs;
2. Since the process is carried out in low-medium vacuum ranges (compared to the high vacuum conditions under which evaporation is conducted), there is greater possibility of incorporating impurities into the deposited film.
3. Better step coverage can generally be achieved using CVD.

In general, the sputtering process consists of four steps:

1. Ions are generated and directed at a target.
2. The ions sputter target atoms;
3. The ejected (sputtered) atoms are transported to the substrate.
4. Upon reaching the substrate they condense and form a thin film.

Although it is of interest to note that sputtering can be conducted by generating the energetic incident ions by other means (e.g., ion beams), in virtually all VLSI and ULSI sputtering processes their source is a glow-discharge. The discussion of sputtering in this section will be limited to *glow-discharge sputtering*.[4,5,6]

11.2.1 Introduction to Glow Discharge Physics

The energetic particles used to strike target materials to be sputtered in ULSI sputter deposition systems are generated by glow-discharges.[4,5] A *glow-discharge* is a self-sustaining type of plasma (a *plasma* is defined as a partially ionized gas containing an equal number of positive and negative charges as well as some number of neutral gas particles). In Fig. 11-3 a simple *dc-diode type* system that can be employed to study properties of glow discharges used in

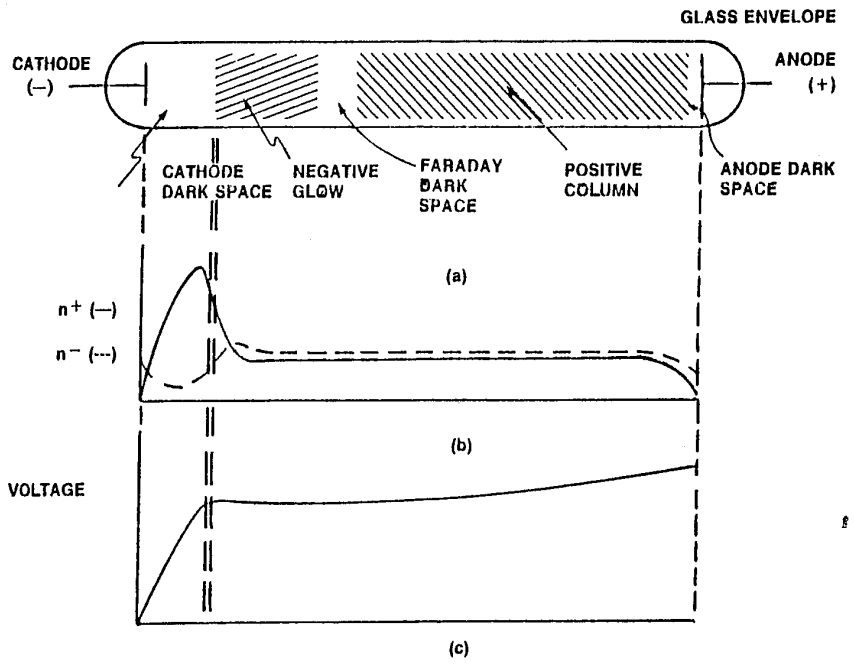


Fig. 11-3 (a) Structure of a glow discharge in a dc diode system. (b) Charged particle concentration in a glow discharge. (c) Voltage variation in a dc diode glow discharge.

sputtering is shown. It consists of a glass tube that is evacuated and then re-filled with a gas at low pressure. Within the tube there are two electrodes (a positively charged *anode* and a negatively charged *cathode*) and a dc potential difference is applied between them.

11.2.2 The Creation of Glow Discharges

Consider the system shown in Fig. 11-3 to examine the case when a tube is filled with Ar at an initial pressure of 1 torr, the distance between the electrodes is 15 cm, and a 1.5 kV potential difference is applied between them. At the outset no current flows in the circuit, as all the Ar gas atoms are neutral and there are no charged particles in the gas. The full 1.5 kV is thus dropped between the two electrodes. If a free electron enters the tube (most likely created from the ionization of an Ar atom by a passing cosmic ray), it will be accelerated by the electric field existing between the electrodes (whose magnitude is: E = V/d = 1.5 kV/15 cm = 100 V/cm).

The average distance that a free electron will travel at $P = 1$ torr before colliding with an Ar atom (i.e., the mean free path λ) is 0.0122 cm (Chap. 3). Most electron-atom collisions are *elastic*, in which virtually no energy is transferred between the electron and gas atom. Such elastic collisions occur because the mass of the electron is much smaller than that of the atom. Thus, the minumum distance an electron must travel before it can undergo an *inelastic* collision (in which significant energy *is* transferred to the atom, either by the excitation of an atomic electron to a higher energy level, or to cause its escape from the atom) is about ten times λ, or 0.122 cm. If this is the minimum distance that must be traveled by electrons between inelastic collisions, there must be a significant number of electron path lengths in the range of 0.5–1.0 cm. If a free electron travels 1 cm in the 100 V/cm electric field, it will have picked up 100 eV of kinetic energy. With this amount of energy, the free electron can transfer enough energy to an Ar electron to cause it to be excited or ionized. If this transferred energy E is less than the ionization potential (e.g., 11.5 eV< E < 15.7 eV for Ar), the orbital electron will be excited to a

higher energy state for about 10^{-8} sec, and then return to the ground state with the emission of visible-light photons. Such excitation is the source of light emission in glow-discharges.

If the energy transferred is greater than the ionization potential (i.e., >15.7 eV for Ar), a second free electron (and a positive ion) will be created. Subsequently, both free electrons will become accelerated again, and the opportunity for cascading the number of free electrons (and creating a condition known as *gas breakdown*) exists. Figure 11-4 shows the breakdown voltage required to initiate discharge, as a function of the product of the pressure, P, and the electrode spacing, d. When the condition of gas breakdown is reached, current flows in the external circuit as the collision-generated free electrons are collected by the anode. Each new ionization event, however, will take place closer to the positively charged anode (as the electrons are accelerated in its direction). Therefore, unless there is a mechanism available for generating additional free electrons for sustaining the current flow, the current will increase to a maximum and quickly decay to zero. When a sufficient number of electrons are available to maintain the discharge, it is said to be *self-sustained.* The source of such electrons is considered later.

11.2.3 Structure of Self-Sustaining Glow Discharges and Their Dark Spaces

A self-sustaining discharge in a system (such as the one shown in Fig. 11-3), exhibits certain characteristics. First, at equilibrium the voltage drop between the electrodes is reduced (e.g., in this example, from 1.5 kV to about 150 V), and the discharge current builds up to where the voltage drop across the current-limiting resistor is equal to the difference between the supply voltage and the electrode potential difference. Second, the discharge has a particular structure, as shown in Fig. 11-3a. The most important region of the discharge is the *Crookes dark space* between the negative glow and the cathode. In Fig. 11-5, one sees that the positive ions of the discharge are present in higher density in front of the negatively charged cathode, producing a localized *space charge* there. Any electrons near the cathode are rapidly accelerated *away* from it, due to their relatively small mass. The much more massive ions are accelerated *toward* the cathode, but much less quickly. Thus, on average, they spend more time than electrons traversing the Crookes dark space, and at any instant their concentration in the dark space is greater than that of electrons.

This has the net effect of greatly increasing the electric field immediately in front of the cathode, and also of screening the remainder of the discharge from the cathode voltage. As a result, the electric field in the rest of the discharge is rather low and uniform, and the glow regions of the discharge are therefore more truly plasma-like (as defined at the outset of this

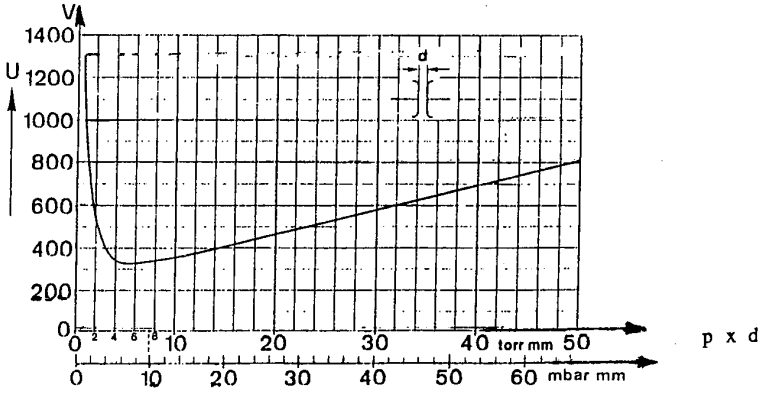


Fig. 11-4 Breakdown voltage, U, between two parallel plane electrodes in a homogeneous electric field as a function of gas pressure, P, and electrode distance, d, for air (Paschens curve).

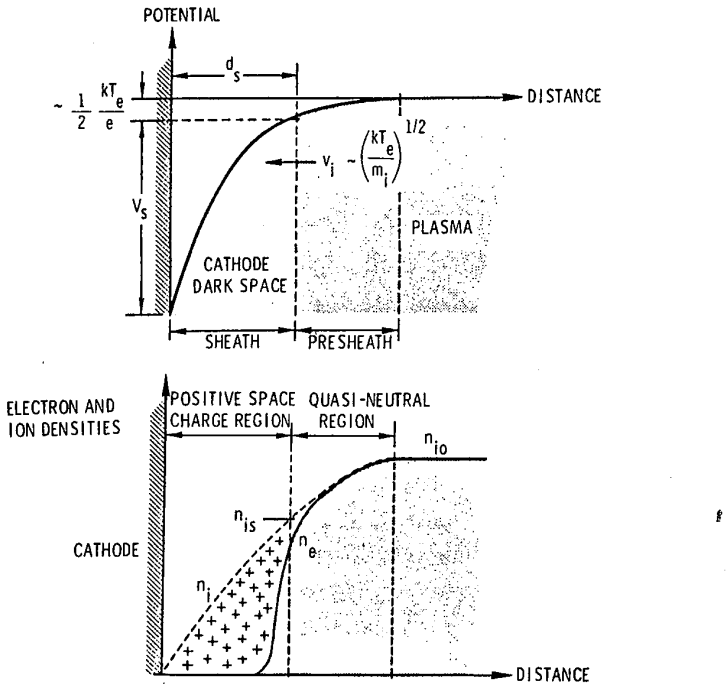


Fig. 11-5 Schematic representation of the positive space charge region that develops over a cathode.

section than the dark-space regions (which do not contain equal concentrations of positive and negatively charged particles). The largest fraction of the voltage dropped between the anode and cathode is thus dropped across the Crookes dark space, and therefore charged particles (ions and electrons) experience their largest acceleration in this region.

The reduced luminosity of the dark space is also due to the electric field present across the dark space. That is, since electrons in the dark space are strongly accelerated and traverse through the space quickly, the electron density at any instant in the dark space is drastically reduced (Fig. 11-5). Although there are fewer electrons, those present can gain high energies from acceleration by the electric field. When they collide with gas atoms in the dark space, they are more likely to cause *ionization*, rather than *excitation* events. As a result fewer light-generating electron-atom collisions occur than in the negative and positive glow regions.

When positive ions from the negative glow region enter the Crookes dark-space, they experience the strong local electric field and become accelerated toward the cathode. They also have a high probability of exchanging their charge with neutral atoms in the dark space. In doing so, they retain their momentum, while losing their charge. The formerly neutral recipient atom becomes an ion, and only at that time can begin accelerating towards the cathode. The effect of this *charge transfer mechanism* was studied by Davis and Vanderslice,[7] and a summary of their results is shown in Fig. 11-6. The data concludes that *virtually no ions reach the target with the full dark space energy.* In addition, it implies that the *cathode is bombarded by energetic neutral atoms as well as energetic ions*, and thus both species apparently produce sputtering events.

The source of electrons that sustains dc discharges is the cathode, which emits secondary electrons when struck by ions. (The mechanisms of such emission are discussed in detail in a subsequent section.) Thus, upon entering the dark space from the cathode, the electrons are accelerated by the dark space field toward the anode. Upon colliding with atoms in the dark

space, they cause ionization, and are also slowed. As slower electrons, they are less energetic and thus the probability of colliding with Ar atoms and causing excitation rather than ionization events is increased. The edge of the negative glow region therefore provides an indication of where such slowed electrons start to be created in large numbers (as a result of high energy, inelastic collisions). If the pressure in the tube is reduced, the distance that the emitted electrons travel before colliding and becoming slowed, becomes longer. Consequently, the dark space lengthens. If the pressure becomes too low, the dark space extends the full length between the electrodes, and the glow-discharge extinguishes. For dc-diode sputtering systems, this effect limits the minimum practical pressure for sputter deposition to the range of 10–40 mtorr (1.3–5.2 Pa). The long dark spaces of dc diodes are also an indication that the electrons are not efficiently utilized to produce ions. How magnetic fields are employed to increase the electron path length in the course of traversing the region between cathode to anode (magnetron sputtering) will be described later. This increases the probability that the electrons undergo an ionizing collision before being collected by the anode. The dark space in such magnetron sputtering systems is typically reduced from 1–2 cm to ~0.5 mm.

11.2.4 Obstructed Glow Discharges & Dark-Space Shielding

In dc diode configurations there must be sufficient electrons emitted at the cathode to keep the discharge self-sustaining, and these electrons must undergo an adequate number of ion-producing collisions with the sputter gas. These electrons are created by ion bombardment of the cathode. Thus, there only needs to be a sufficient number of ions produced in the dark space (and in the glow regions) to produce enough secondary electrons to sustain the glow-discharge. As a result, as long as enough ions are produced in the dark space and glow regions, the exact location of the anode is not important. In fact, if the anode is gradually moved closer to the cathode, the positive column shown in Fig. 11-3 first disappears, then the Faraday dark space vanishes. Finally, a substantial fraction of the negative glow may be extinguished, before any significant effect on the electrical characteristics is observed. If the anode gets too close to the

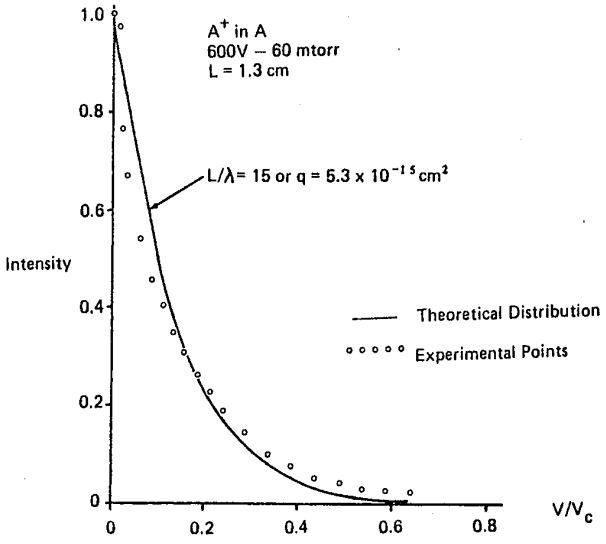


Fig. 11-6 Energy distribution for Ar^+ from an argon discharge.[7] Reprinted with permission of the American Physical Society.

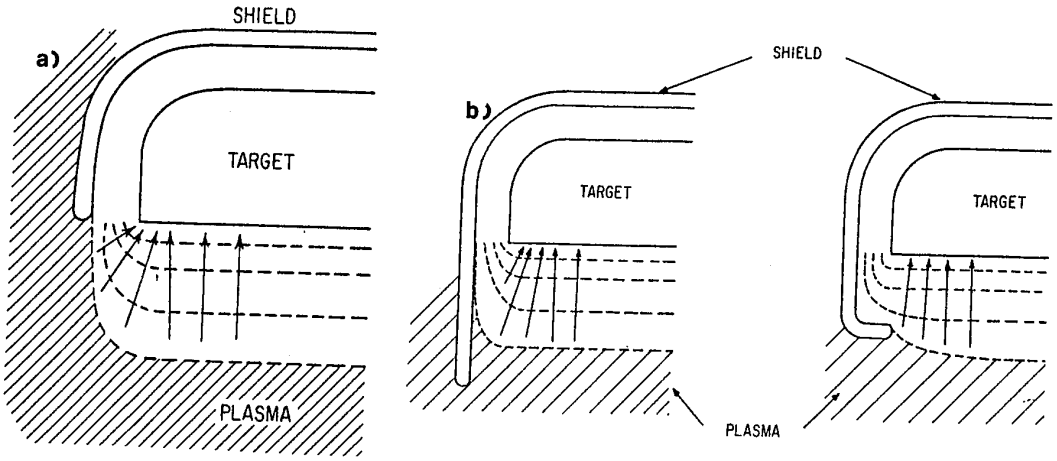


Fig. 11-7 (a) Potential distribution in vicinity of cathode shield (b) Reducing rim effect by extending cathode shield. (c) Reducing rim effect by wrapping shield around the cathode.[8] From L. Maissel and R. Glang, Eds., *Handbook of Thin Film Technology* 1970. Reprinted by permission of McGraw-Hill Book Co.

dark space however, the ion production rate becomes reduced, and the voltage across the electrodes must rise to increase the secondary electron emission. Such a glow is known as an *obstructed glow.* In most practical sputter deposition systems the glow is obstructed. That is, in order to most effectively collect the sputtered material onto the substrate, the anode (on which the wafers are sometimes mounted) is placed as close to the cathode as possible (typically just far enough away to avoid extinguishing the negative glow).

It is typically necessary to insure that sputtering is allowed to occur only at the front side of the target, as the backside contains cooling coils and attachment fixtures which are definitely not to be sputtered. To guarantee that no sputtering takes place except from desired surfaces, a shield of metal (at a potential equal to that of the anode) is placed at a distance less than the Crookes dark space at all other cathode surfaces (Fig. 11-7). Since no discharge will occur between two electrode surfaces separated by less than this distance, such shielding (termed *dark-space shielding*) is effective in preventing sputtering from unwanted cathode surfaces.

11.3 THE PHYSICS OF SPUTTERING

When a solid surface is bombarded by atoms, ions, or molecules, many phenomena occur. The kinetic energy of the impinging particles largely dictates which are the most probable events. For low energy particles (<10 eV), most interactions occur only at the surface of the target material. At very low energies (<5 eV) such events are limited to reflection or physisorption of the bombarding species. For low energies which exceed the binding energy of the target material (5–10 eV), surface migration and surface damage effects can take place. At much higher energies (>10 keV), the impinging particles travel well into the bulk of the sample before slowing down and depositing their energy. Thus, such particles are most likely to be embedded in the target, and this mechanism is the basis of ion-implantation. At energies between the two extremes, two other effects also arise: 1) some fraction of the energy of the impinging ions is transferred to the solid in the form of heat, and lattice damage; and 2) another fraction of such energy causes atoms from the surface to be dislodged and ejected into the gas phase (*sputtering*).

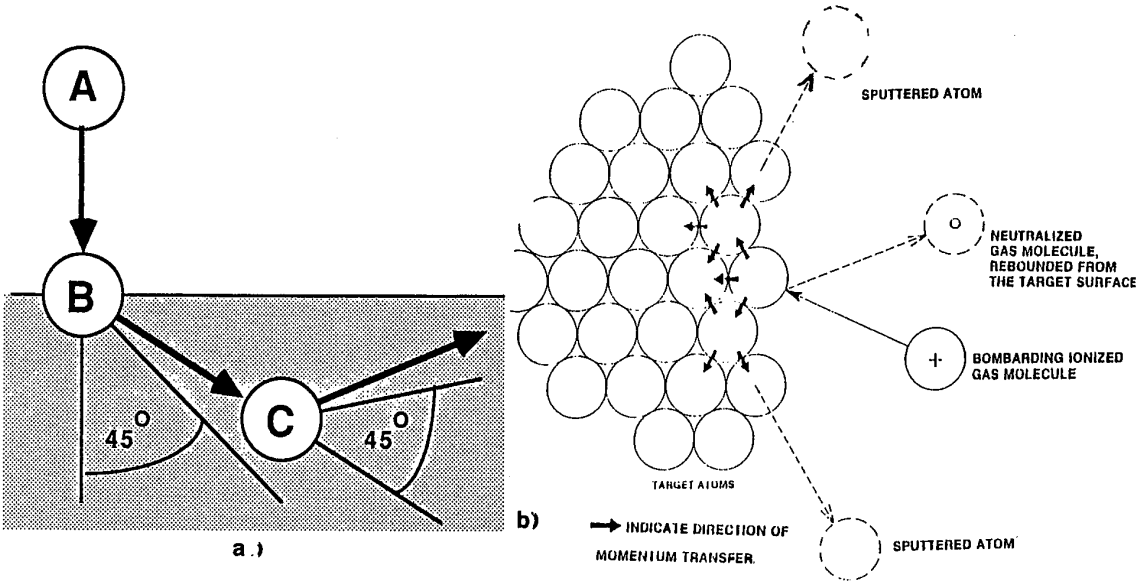


Fig. 11-8 (a) Binary collision between atom A and B, followed by a binary collision between atom B and C. (b) Collision process responsible for sputtering and fast neutral generation.

11.3.1 The Billiard Ball Model of Sputtering

The exact mechanisms which lead to the ejection of atoms under ion bombardment are not known, and a comprehensive theory of sputtering is not likely to be developed in the near future since many parameters are involved. These include the kinetic energy of the ions, lattice structure, and binding energy of lattice atoms. Some of the details, however, are reasonably well understood and can be aptly described with a relatively simple momentum-transfer model. G.K. Wehner, whose theoretical work first established a solid scientific basis for sputtering, often described sputtering as a game of three-dimensional billiards, played with atoms.[9] Using this analogy, it is possible to visualize how atoms may be ejected from a surface as the result of two binary collisions (Figs. 11-8b and 11-8c) when a surface is struck by a particle with a velocity normal to the surface (e.g., atom A in Fig. 11-8b). Note that a *binary collision* is one in which the primary incoming particle (e.g., atom A) strikes a single object (e.g., atom B in Fig. 11-8b), and gives up a significant fraction of its energy to the struck atom, while retaining the remaining fraction. As a consequence of the collision, atom B may leave the point of impact at an angle greater than 45°. If atom B then undergoes a secondary collision with atom C, the angle at which atom C leaves the secondary impact point may again be greater than 45°. Thus, it is possible that atom C can have a velocity component greater than 90° (and thus be directed toward the surface). As a result, there is a finite probability that atom C will be ejected from the surface as a result of the surface being struck by atom A.

When the directions of sputtered atoms from the surface of polycrystalline materials (and most cathode materials in sputter applications are polycrystalline) are measured for the case of *normal incidence*, it is found that the ejected atoms leave the surface in essentially a cosine distribution. A cosine distribution, however, does not describe the sort of small-angle ejections that would be expected from the simple collision processes described above. Evidently in actual sputtering events, more than two collisions are involved, and the energy delivered by impinging ions *during normal incidence* is so randomly distributed that the effect of the incident momentum vector is lost. Note that the energy range of sputtered atoms leaving the target is

typically 3–10 eV, and the bombarding species also recoil from the cathode face with some energy. Thus, the target surface is a source of sputtered atoms and energetic backscattered species.

For the case when the surface is bombarded by ions at an oblique angle (i.e., 45°–90°), there is a higher probability that the primary collision between incident ions and surface atoms will lead to sputtering events. Furthermore, oblique incidence confines the action closer to the surface, and thus sputtering is enhanced. In cases of oblique bombardment, the incident momentum vector becomes important, and sputtered atoms are ejected strongly in the forward direction. In addition, the *sputter yield* (defined as the number of atoms ejected per incident ion), may be as much as an order of magnitude larger than that resulting from normal incidence by bombarding ions (Fig. 11-9). This effect also leads to *faceting*, which is covered in Sect. 11.7.4.

11.3 2 Sputter Yield

The sputter yield is important because it largely (but not completely) determines the rate of sputter deposition. Sputter yield depends on a number of factors besides the direction of incidence of the ions, and these include: a) the target material; b) the mass of bombarding ions; and c) the energy of the bombarding ions. There is a minimum energy threshold for sputtering that is approximately equal to the heat of sublimation (e.g., 13.5 eV for Si). In the energy range of sputtering (10–5000 eV), the yield increases with ion energy and mass. Figure 11-10 shows the sputter yield of copper as a function of energy for various noble gas ions. The sputter yields of various materials in argon at different energies is given in Table 11-2.[9]

Several matters related to sputter yield should be noted. First, although the sputtering yields of various materials are different, as a group they are much closer in value to one another than, for example, the vapor pressure of comparable materials. This makes the deposition of multilayer or multi-component films much more controllable by sputtering. The details of sputtering from multi-component targets are discussed in the section on *Process Considerations in Sputter Deposition.* Second, since the bombarding ions are by no means monoenergetic in glow-discharge sputtering, it is not necessarily valid to use the sputter yield values for pure metals if alloys, compounds, or mixtures are being sputtered. Tabulations of sputtering yields, however, are useful for obtaining rough indications of deposition or etch rates of various materials.

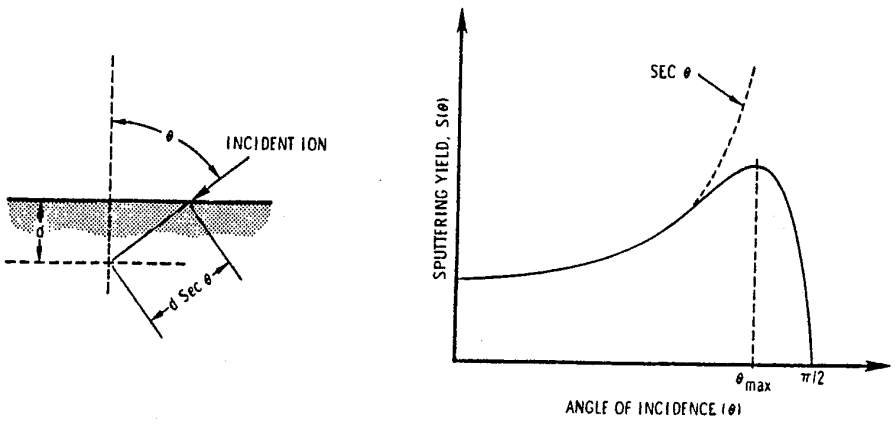


Fig. 11-9 Schematic diagram showing variation of sputtering yield with ion angle of incidence.

Table 11-2 SPUTTER YIELDS FOR METALS IN ARGON (ATOMS/ION)

Target	At.Wt./Dens.	100 eV	300 eV	600 eV	1000 eV	2000 eV
Al	10.0	0.11	0.65	1.2	1.9	2.0
Au	10.2	0.32	1.65	2.8	3.6	5.6
Cu	7.09	0.5	1.6	2.3	3.2	4.3
Ni	6.6	0.28	0.95	1.5	2.1	
Pt	9.12	0.2	0.75	1.6		
Si	12.05	0.07	0.31	0.5	0.6	0.9
Ta	10.9	0.1	0.4	0.6	0.9	
Ti	10.62	0.08	0.33	0.41	0.7	
W	14.06	0.12	0.41	0.75		

11.3.3 Selection Criteria for Process Conditions and Sputter Gas

The information gleaned from sputter yields and the physics of sputtering can also be applied toward an understanding of how process conditions and materials are selected for sputtering including: a) type of sputtering gas; b) pressure range of operation; and c) electrical conditions for the glow discharge.[4,8] That is, in purely physical sputtering (as opposed to reactive sputtering) it is important that the ions or atoms of the sputtering gas not react with the growing film. This limits the selection of sputtering species to the noble gases. Furthermore, argon is generally the gas of choice, since it is easily available (hence low in cost), and its mass is a good match to those of the elements most frequently sputtered (Al, Cu, Si, and Ti), giving adequate sputtering yields for these elements. The *pressure range of operation* is set by the requirements of the glow discharge (lower limit ~2–3 mtorr for magnetron sputtering) and the scattering of sputtered atoms by the sputter gas (upper limit 100 mtorr). In addition, a desired goal of sputter deposition is to obtain maximum deposition rates. As a result, *electrical conditions* are selected

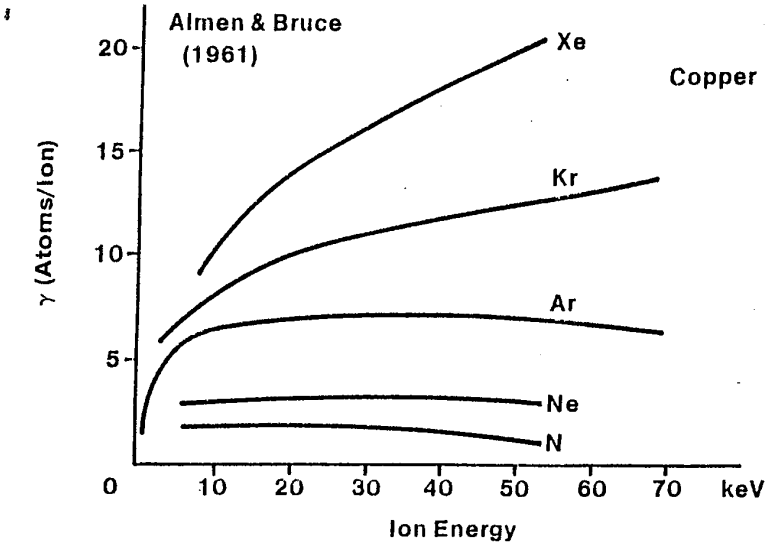


Fig. 11-10 Sputtering yields of the noble gases on copper, as a function of energy.[10]

to give a maximum sputter yield per unit energy. That is, as energy is increased, each energy increment gives a progressively smaller increase in sputter yield. This occurs because higher energy ions implant themselves, and thus end up dissipating a greater proportion of their energy via non-sputtering processes. The most efficient ion energies for sputtering are typically obtained for electrode voltages of several hundred volts.

In general, the higher the current at the cathode, the higher is the film deposition rate, since more ions are striking the cathode (and thus cause more sputtering). The product of the cathode current and the electrode voltage gives the input power of the sputtering process. In magnetron sputtering, cathode current densities of 10–100 mA/cm^2 at a few hundred volts are typical.

From the foregoing discussion on the mechanism of sputtering one can see that sputtering is a highly inefficient process. In fact, ~70% of the energy consumed during the sputtering process is dissipated as heat in the target, and ~25% by emission of secondary electrons and photons from the target. This heating can raise target temperatures to levels capable of damaging the target, associated vacuum components, or the material that bonds the target to the backing electrode. The target must therefore be cooled to avoid these problems. As we will discuss later, this is normally accomplished by bonding the target to a water-cooled Cu backing plate.

11.3.4 Secondary Electron Production for Sustaining the Discharge

Near the outset of the discussion on glow-discharges, it was observed that glow-discharges must be continuously provided with free electrons to keep them sustained. In most dc-sputtering systems, the source of such electrons is secondary-electron emission from the target. An important mechanism that generates such secondary electrons is Auger emission, although other emission mechanisms are also responsible for their production. The *Auger emission* process occurs according to the sequence of events shown in Fig. 11-11. Target electrons require a minimum amount of energy $q\Phi$ to escape into free space (where Φ is the *work function* of the target material, and $q\Phi$ is typically 3–5 eV). However, when a positive ion (such as Ar^+) comes close to the target surface it appears as a potential well of 15.67 eV to target electrons occupying energy states near the Fermi level. Thus, some of the target electrons can escape by tunneling into these potential wells. The energy difference between 15.67 eV and the mini-mum free-space escape energy is released in the form of a photon. If this photon (e.g., of energy 10.67 eV = [15.67 eV - 5 eV]) is absorbed by another electron near the target surface, this electron (i.e., the Auger electron) may posses enough energy to be emitted from the target into free space. Since this sequence of events is rather improbable, several ions must strike the cathode in order for a single Auger electron to be emitted. The secondary electron yield per bombarding ion γ_i has been measured to be 0.05–0.1 for metal targets. Thus, each secondary electron needs to generate as much as 20 ion-electron pairs before it reaches the anode to sustain the discharge.

Besides providing a source of electrons to sustain the discharge, electron emission is also important to the sputtering process in other ways. First, ions bombarding the cathode are neutralized. Thus, it is highly probable that each ion that closely approaches the target will extract an electron from the target, and return to the discharge as a neutral atom. Second, the total target current, I, is the sum of the ion flux striking the target, I_i, and the secondary electron current leaving the electrode, I_e. Since γ_i is larger for dielectric materials, their I_e is also larger. This implies that for the same cathode current, I, dielectrics will sputter more slowly than metals. That is, in dielectric sputtering a larger fraction of I is due to electron emission, and thus for equal values of I, a smaller ion flux is striking the dielectric target.

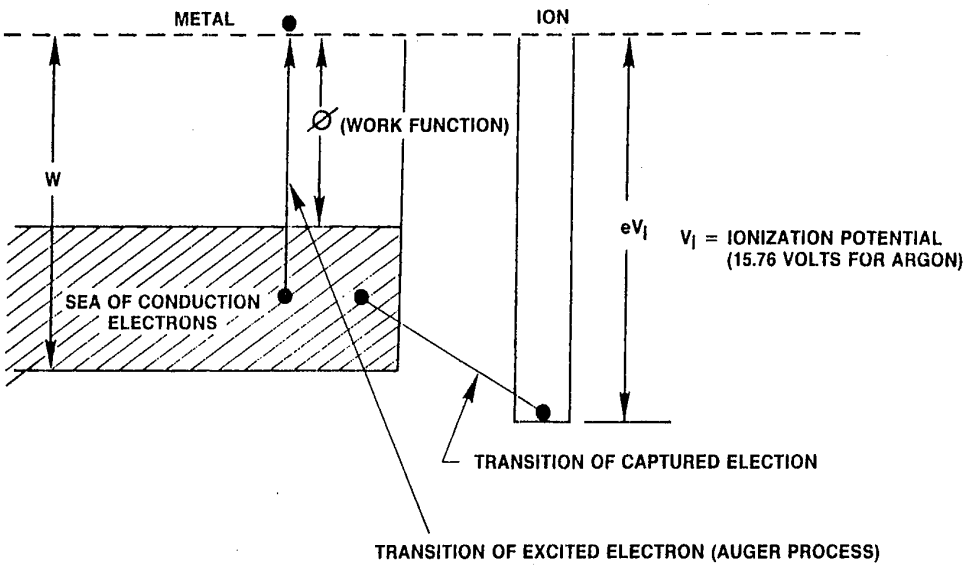


Fig. 11-11 Potential energy diagram for ion approaching a metal target.

11.3.5 Sputter Deposited Film Growth

Upon being ejected from the target surface, sputtered atoms have velocities of $3\text{-}6 \times 10^5$ cm/sec and energies of 10–40 eV. It is desirable that as many of these sputtered atoms as possible be deposited upon the substrates to form the specified thin film. To accomplish this, the target and wafers are closely spaced, with spacings of 5–10 cm being typical. The mean free path, λ, of sputtered atoms at typical sputter pressures is less than 5–10 cm (e.g., at 5 mtorr, $\lambda \cong 1$ cm). Thus, it is likely that sputtered atoms will suffer one or more collisions with the sputter-gas atoms before reaching the substrate (Fig. 11-12). The sputtered atoms may therefore: a) arrive at the substrate with reduced energy (~1–2 eV); b) be backscattered to the target or the chamber walls; or c) lose enough energy so that thereafter, they move by diffusion in the same manner as neutral sputter gas atoms. These events imply that the sputtering gas pressure can impact various film deposition parameters, such as the deposition rate and composition of the film.

The formation and growth of the thin film on the substrate proceeds according to the general discussion on thin film formation given in Chap. 4. Therefore, this discussion is restricted to the events that uniquely impact the formation and growth of glow-discharge sputtered films.

The substrate surface (onto which the desired film of the sputtered target material is deposited) is also subjected to impingement by many species. The sputtered atoms arrive and condense onto the substrate. For a typical deposition rate of 20 nm/min, a monolayer of deposited film will form approximately every second (assuming the size of a typical atom is ~0.3 nm). For this case, and even for much higher deposition rates of Al (e.g., 620 nm/min), the heat of condensation is not necessarily the most important source of substrate heating.

11.3.6 Species that Strike the Wafer During Film Deposition

In addition to the sputtered atoms, the substrate is also struck by many other species (Fig. 11-13), the most important being: 1) *fast neutral sputter gas atoms*, which retain significant energy after striking and recoiling from the cathode. As they impinge upon the substrate, some may embed themselves in the growing film; 2) *negative ions*, formed near the cathode surface by

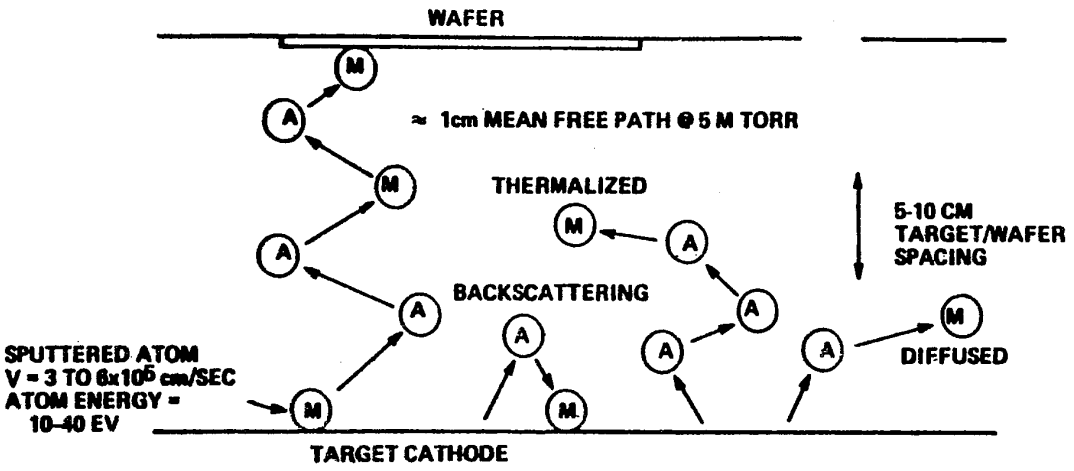


Fig. 11-12 Gas scattering events.

the reaction of secondary electrons and impurity gas atoms, such as O and N. These can acquire substantial energy from the dark space electric field, and thus also strike the substrate with appreciable energy; 3) *high energy electrons*, which are accelerated across the dark space, and then impinge on the wafers. (In dc-diode and rf-diode sputtering they can represent a major source of substrate heating); 4) *low energy neutral sputter gas atoms* (Ar), which strike the substrate at a high flux (e.g., at 20 mtorr the Ar flux is several thousand times greater than the flux of sputtered material). However, the sticking coefficient of low-energy neutral Ar atoms is considered to be negligibly small, and the only incorporation of Ar atoms into the growing film arises from the embedding of fast neutrals; 5) *contaminants present in the form of residual gases* in the chamber can be incorporated into the growing film, particularly if they are chemically active. For example, if a sputtering system contains oxygen in the partial pressure of 10^{-6} torr,

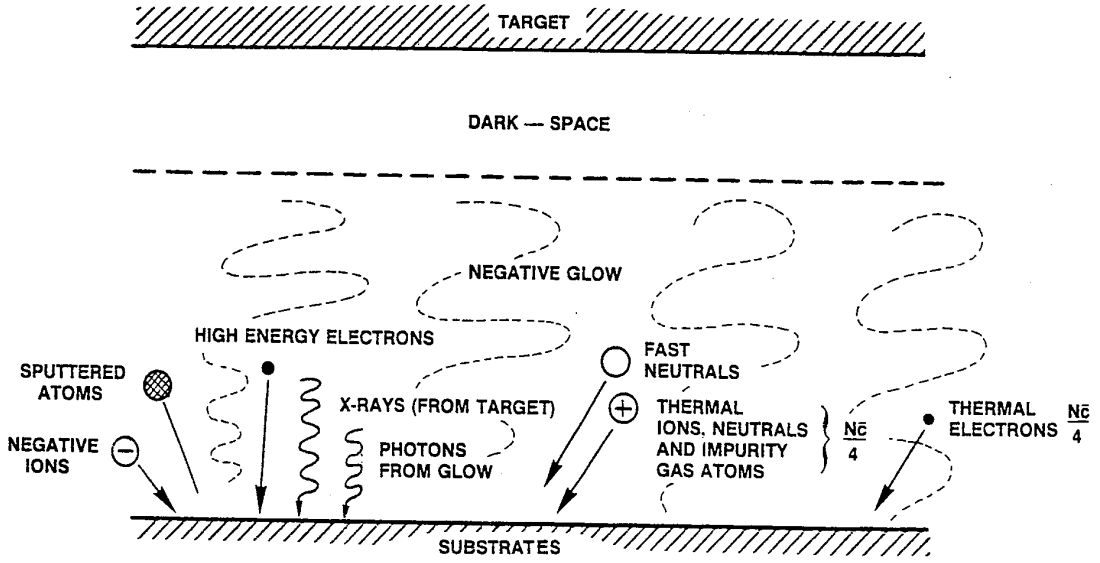


Fig. 11-13 Species arriving at the substrate in a sputtering system.

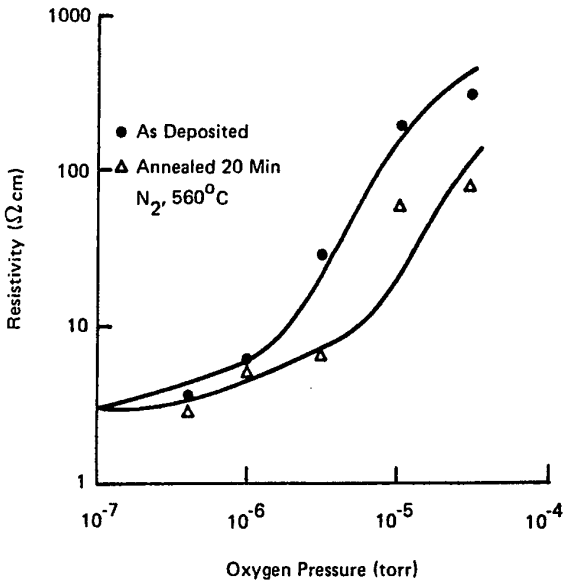


Fig. 11-14 Room temperature resistivity vs oxygen pressure during evaporation of 500 nm Al films at 200°C and 2 nm/s. Reprinted with permission from Transactions of the Metallurgical Society, Vol. 242, pp. 205, a publication of the Metallurgical Society, Warrendale, Penn.

the substrate surface will be struck by ~10^{15} oxygen atoms/sec (or 1 mono-layer/sec).

In Fig. 11-14[11] observe that at O_2 partial pressures of 10^{-6} torr and Al deposition rates of 120 nm/min, the resistivity of the Al film has increased significantly due to the oxidation of some of the Al material. The sources of residual gases can include: **1**) impure sputtering gas; **2**) outgassing from wafers/palette (or wafer holder); **3**) outgassing from chamber walls; and **4**) leaks. To minimize the presence of residual gases in sputtering, it is advisable to use highly pure sputter gases, and to continually evacuate and refill the sputter chamber to flush any residual gases being introduced through outgassing or leaks. Additional details on the effects of residual gases on sputter deposited films is given in the section on *Sputter Processing Issues.*

So far in this discussion, the glow-discharge and sputtering phenomena we have been describing were produced by the application of a *dc voltage* between the electrodes. Sputtering performed with such systems is known as *dc glow-discharge sputtering.* This diode configuration is the simplest glow-discharge-based sputtering system, and consequently it is appropriate to consider its characteristics at the outset of the discussion on sputtering. In general, however, dc-diode sputtering processes possess severe limitations that render them unsuitable for most ULSI sputtering applications. As a consequence, two other glow-discharge-based sputtering modes are used instead to sputter deposit thin films in semiconductor manufacture: **1**) radio-frequency (rf) sputtering; and **2**) magnetron sputtering.

11.4 RADIO-FREQUENCY (RF) GLOW DISCHARGES

One of the limitations of dc diode systems is that they cannot sputter insulators (dielectrics). This is due to the fact that glow-discharges cannot be maintained with a dc voltage if the electrodes are covered with insulating layers. This represents a significant drawback, as there are several important applications which call for maintaining a continuous discharge in the face of dielectric-covered electrodes, including: a) the *in situ* sputter removal of thin *native-oxide* layers from Si and Al surfaces prior to the deposition of overlying films (i.e., *in situ* implies that sputter etch and deposition are performed sequentially without breaking vacuum); and b) the

condition of glow-discharge must be continuously maintained in some dry-etching process chambers that are equipped with electrodes coated with insulating materials.

The claim that a glow-discharge cannot be sustained if an insulator is sputtered using dc voltages is supported by the following argument. When the negatively charged cathode in a dc diode system is bombarded by ions, an electron is stripped from the cathode surface each time an impinging positive ion is neutralized. If the cathode material is a *conductor*, such electrons can be replaced by electrical conduction, and the cathode surface maintains the negative potential required to sustain the discharge. If the material is an *insulator*, however, the electrons lost from the cathode surface are not replaced, since electrical conduction from the insulator interior to the sputtered surface is not possible. Thus, the front surface accumulates a positive charge that increases with time of bombardment. This causes the potential difference between the cathode surface and that of the anode to decrease (Fig. 11-15). As soon as it drops below the value needed to sustain the discharge, the discharge extinguishes. In practical glow-discharge configurations, the time for the insulator surface to acquire this charge is ~1–10 μs (see Prob. 11-4).

A technique was developed to allow replenishment of such lost electrons to the insulator surface. This involves the application of an ac voltage (rather than a dc voltage) to the electrodes, and sputtering processes based on applying this method are known as *rf sputtering*.[4,8] This solution, however, introduces other complications. That is, in order to achieve a practical sputtering process by the application of ac voltages, the following conditions need to be established: a) the electrons lost from the insulator surface by the mechanism observed in dc glow-discharge sputtering must be periodically replenished; b) the glow-discharge must be continuously sustained under ac conditions over the full period of the ac waveform; c) an electric field configuration must be created in the process chamber which allows ions of sufficient energy to bombard and sputter the target insulator; d) sputtering in the process chamber should be suppressed at all surfaces except the target surface (i.e., sputtering should not occur at the chamber or non-target electrode surfaces); and e) the rf power must be efficiently coupled to the glow-discharge to maximize sputter deposition rates. Rf sputtering will be discussed by explaining how each of these requirements is met. The system used to apply an ac signal to the process chamber is shown in Fig. 11-16. The two electrodes are referred to as *A* and *B*, and no longer as cathode and anode, since the signal polarity applied to both electrodes alternates between positive and negative values.

Assume that the rf generator in Fig. 11-16 produces a signal of alternating polarity with an amplitude, V_0, that is large enough to cause breakdown of the process chamber gas. When such a signal is applied to the electrodes, positive ions from the discharge are accelerated toward an

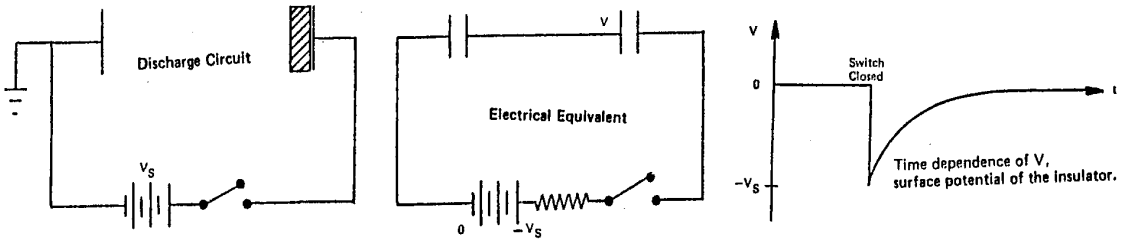


Fig. 11-15 Surface charging of an insulating cathode. From B. Chapman, *Glow Discharge Processes*, 1980, Copyright © John Wiley & Sons. Reprinted with permission of John Wiley & Sons.[4]

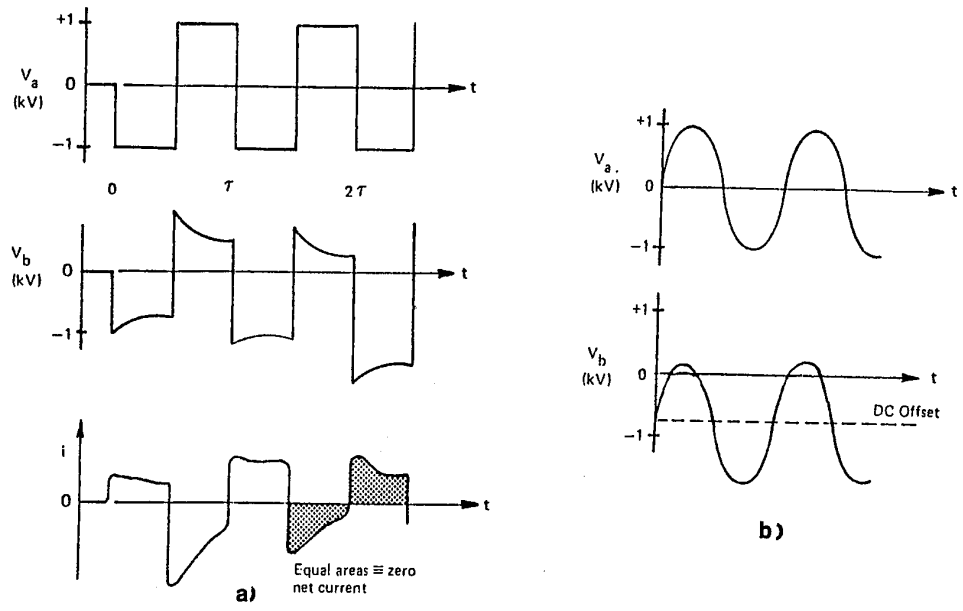


Fig. 11-16 (a) Voltage and target current waveforms when the circuit of Fig. 11-15 is square-wave excited. (b) Voltage waveforms at generator (V_A) and target (V_B) in a conventionally sinusoidally excited rf discharge. From B. Chapman, *Glow Discharge Processes*, 1980, Copyright © John Wiley & Sons. Reprinted with permission of John Wiley & Sons.[4]

electrode when it is subjected to the negative-bias portion of the waveform (and electrons during the positive-bias portion). In *steady-state conditions* (to be defined), the positive charge at the target surface (e.g., electrode *A*) accumulated during the negative part of the cycle, is replaced by impinging electrons during the positive part of the cycle. The net surface charge accumulated by the insulator during each complete cycle of the waveform is zero, and the lost electrons are thus replenished, and condition (a) is satisfied (Fig. 11-17).

As mentioned earlier, the time required to sufficiently charge the insulator surface during the negative part of the cycle to extinguish the glow-discharge is only 1–10 μs. Therefore, low frequency ac signals to the electrodes (e.g., 60 Hz) are ineffective at maintaining a continuous glow-discharge. The discharge remains ignited for only a very small fraction of the input signal period (e.g., 1–10 μs of the 16 ms period of a 60 Hz signal). Thus, in order to satisfy condition (b) above, the charging time of the insulator surfaces implies that an ac signal with a minimum frequency of 50 kHz-1 MHz must be applied to the electrodes. In practice, most ac glow-discharges are operated at 13.56 MHz (i.e., glow-discharge sputtering processes and plasma/RIE dry etch processes). The applied power has a frequency in the radio-frequency range (hence the name *rf sputtering*), and the 13.56 MHz frequency was chosen because it is one of the frequencies designated by international communications authorities at which (a limited amount of) electromagnetic energy can be radiated without interfering with other radio-transmitted signals.

At this point it is also important to mention that rf excited discharges can be sustained without relying on the emission of secondary electrons from the target. The mechanism that allows rf excited glow-discharges to be thus sustained is not well understood, although several theories have been advanced. One theory postulates that ionization is produced by free electrons oscillating in the very weak rf field that penetrates the plasma. Such oscillations, combined with properly timed elastic collisions between the electron and gas atoms, permit the electrons to gain sufficient energy to cause ionization (despite the weak field). Two important effects are assoc-

iated with this phenomenon. First, it is observed that rf-excited discharges remain sustained at lower minimum pressures than dc discharges (e.g., 10 mtorr versus 40 mtorr). Second, is that rf excitation provides a mechanism to create glow-discharges for dry-etch and PECVD processes without needing to be concerned about a source of electrons to sustain the discharge.

To meet condition (c) above, an electric-field in front of the target electrode must exist to accelerate the glow-discharge ions to sufficient energy to produce sputtering. Such an electric-field is produced in rf excited systems by a phenomenon known as *self-bias.* To study this phenomenon, examine the events that occur when the ac signal is initially applied to the insulating electrode (Fig. 11-16). At the time the switch is closed (t = 0), the electrode surfaces contain no charge. The applied negative bias causes positive ions from the discharge to strike electrode *A* (an insulator), and to leave behind a positive charge on the surface as electrons are removed. During the positive portion of the cycle, the electrode attracts and collects electrons. The mass of the electrons is much smaller than that of ions, and the electric field accelerates them more rapidly to the electrode surface. Thus, during the two halves of the first ac cycle many more electrons are initially collected at the electrode than ions. At the end of the first complete period, the electrode surface therefore has a net negative charge. As more cycles transpire, the negative charge continues to build up. Such negative surface charge, however, also has the effect of repelling some electrons during the positive part of the cycle, and attracting positive ions more strongly during the negative parts. In the steady state, an equal number of positive ions and electrons strike the electrode during each complete signal period. The *average* charge build-up on the electrode which occurred during the first few cycles is negative, and remains constant in value under steady-state conditions. This also results in a negative dc offset voltage between the electrode and the discharge. Such a *self-bias voltage* is almost equal to the peak of the applied voltage (Fig. 11-16b) of the rf input signal. If the potential difference between the discharge and the self-biased electrode is sufficiently large, ions will be accelerated strongly enough toward the cathode to cause sputtering there [thus satisfying condition (c)].

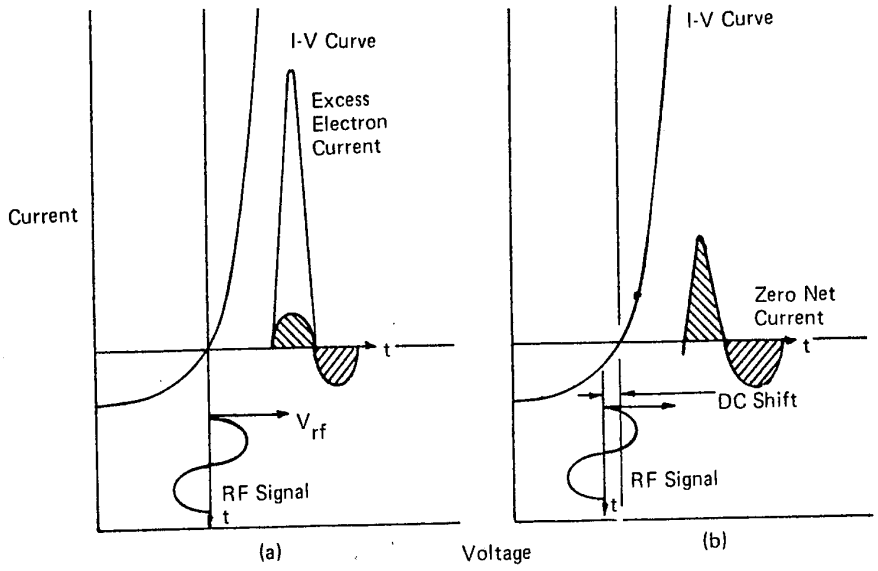


Fig. 11-17 Self-biasing of a dielectric surface.[12] Reprinted with permission of American Phys. Soc.

Unfortunately, this effect can occur at both electrodes, as well as at any surface in the chamber connected electrically to the rf signal. To restrict the sputtering only to the target surface, the electrode configuration in the system must be altered. First, the non-sputtering electrode is grounded, and this may also be a convenient location for placing the substrate wafers onto which the sputtered film is to be deposited. Next note that there are voltages V_A and V_B (called *sheath voltages*) occurring between the discharge and electrodes *A* and *B* respectively. To restrict sputtering only to the target (e.g., electrode *A*), it is necessary for sheath voltage V_A to be large, and V_B to be as small as possible. Assume that the electrodes, *A* and *B*, have areas A_A and A_B, respectively.

To see how to establish such sheath voltages, an expression that describes the relationship between ion flux through the sheath, and the voltage across the sheath (Fig. 11-18) is used. If a space-charge-limited current is assumed, the ion current flux, J, is given by Child-Langmuir equation:[13]

$$J = \frac{KV^{3/2}}{\sqrt{m}D^2} \tag{11.1}$$

where V is the voltage drop across the sheath, D is the dark-space thickness, m is ion mass, and K is a constant. Since the current density of positive ions must be equal at both electrodes, Eq. 11-1 indicates that:

$$\frac{V_1^{3/2}}{D_1^2} = \frac{V_2^{3/2}}{D_2^2} \tag{11.2}$$

Since there is a large voltage drop across the dark space, this implies that the dark space is a region of very limited conductivity. Thus, the electrode dark-space plasma can be modeled as a capacitor, with capacitance proportional to electrode area and inversely proportional to D:

$$C \propto A/D \tag{11.3}$$

Now, an rf voltage will divide between two capacitances in series according to:

$$V_A/V_B = C_B/C_A \tag{11.4}$$

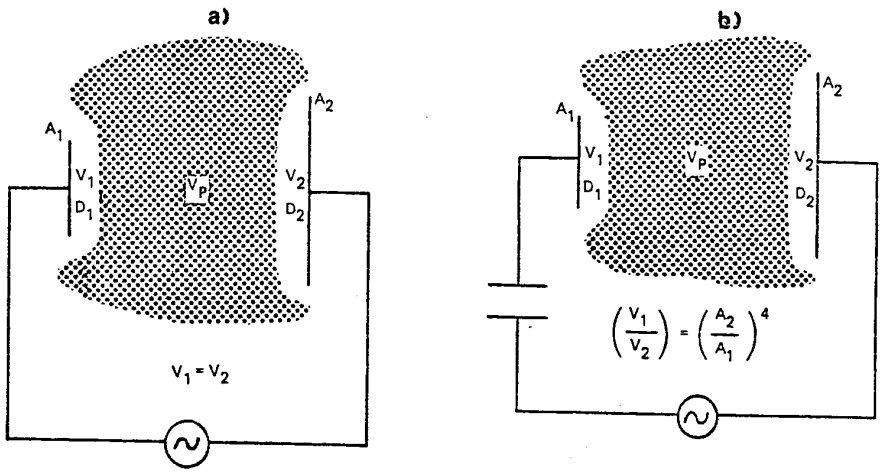


Fig. 11-18 (a) Voltage distribution with equal area electrodes and no blocking capacitor. (b) Voltage distribution with unequal area electrodes and blocking capacitor.[13] Copyright 1970 by International Business Machines Corporation; reprinted with permission.

Using Eqs. 11-3 and 11-4, one gets:

$$V_A/V_B = (A_B/D_B)(D_A/A_A) \tag{11.5}$$

and substituting into Eq. 11-2:

$$V_A/V_B = (A_B/A_A)^4 \tag{11.6}$$

Thus, the larger dark-space voltage will develop at the electrode having the smaller area. The fourth-power dependence is not observed in real systems (the value of the exponent is more typically determined experimentally to be < 2), but the principle of increasing the sheath voltage ratio by controlling the relative electrode areas is valid. In actual sputtering and dry-etch systems, the larger electrode consists of the entire sputtering chamber, including the wafer holder, which has a common electrical ground with the rf power supply. The relatively large area causes a small dark space voltage to exist, and no sputtering takes place from these surfaces. The target material is attached to the smaller electrode, which therefore develops a large dark-space voltage, and can thus experience strong sputtering. In this manner, condition (d) is satisfied.

Since the dark-space voltages V_A and V_B will not be equal if the areas of the electrodes are not equal (and the plasma is a region of equipotential), a dc voltage between the electrodes will exist. If the target is a conductor, a blocking capacitor (Fig. 11-19) is needed to prevent this self-bias voltage from being grounded through the rf generator. If the target were an ideal insulator, it would not be necessary to use an external *blocking capacitor.* It is common practice however, to design systems with external blocking capacitors to remove any possibility of leakage through or around the edges of the target in the presence of the plasma. The size of the capacitor must be large enough that it exhibits low impedance at the frequency of the rf signal.

The final requirement (e) that must be met to realize a useful rf-sputter process is to be able to couple the maximum power from the rf generator to the discharge. The output impedance, Z_o, of rf power supplies used in sputtering (and other IC applications) is designed to be purely resistive (with the value $Z_o = 50\ \Omega$), while the load impedance of the discharge X_L is much larger and the impedance is also largely capacitive. This impedance mismatch is remedied by inserting an *impedance matching network* between the output of the generator and discharge chamber (Fig. 11-19). Most rf systems use feedback control to tune the network by automatically maintaining minimum reflected power. This also protects the generator and coupling lines.

Rf diode sputtering systems behave in other respects much like dc diode systems. That is, secondary electrons are emitted by the target, and subsequently accelerated through the dark space. Most still end up being collected by the grounded electrode, and thus such electrons usually represent a source of substantial substrate heating. In addition, the target is also the source of fast neutral sputter gas atoms, negative ions, and electromagnetic radiation. These

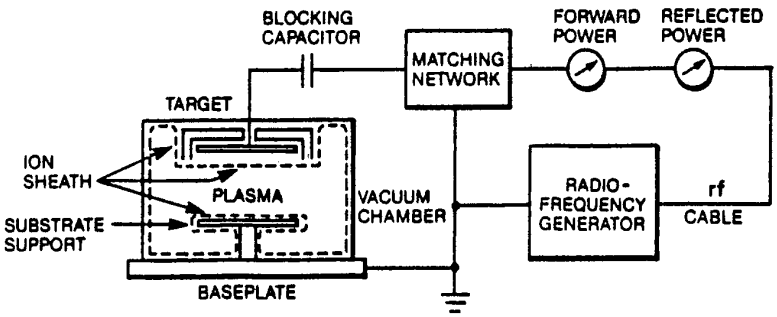


Fig. 11-19 Radio frequency sputtering system.

species are generated in the same manner as in dc diode systems. Furthermore, rf diode configurations suffer the same inefficiency of secondary electron utilization. Since most electrons emitted from the target are collected by the anode, and do not cause ionization, ion production (and consequent ion bombardment of the target) is not significantly greater than in dc diode systems. Thus, for processes that require conductive materials to be deposited at high rates (e.g., aluminum or aluminum alloys) rf diode sputtering is not an optimum process.

11.5 MAGNETRON SPUTTERING

As noted in Sect. 11.3.4, secondary electrons emitted from the target must produce about 20 ion-electron pairs before reaching the anode in order to sustain the glow discharge. Since the cross-section for electron impact ionization of the gas is fairly small, in dc diodes it is usually necessary to operate at fairly high pressures (20–200 mtorr) to have any reasonable ion current. Such high gas pressures, result in poor transport of the sputtered atoms away from the cathode (see Sect 11.7.2), and consequently result in very slow deposition rates. As a result, dc diodes are rarely used in practical applications. To make sputtering a viable production process, alternative sputtering sources which provide significantly higher sputter deposition rates are needed. Higher deposition rates can only be produced by sources that operate at higher current densities, lower voltages, and lower pressures than are possible in dc-diode configurations.

The alternative sputter source that satisfies these requirements relies on a mechanism termed *magnetron sputtering*. Magnetron sources increase the percentage of electrons that cause ionizing collisions by utilizing magnetic fields to help confine the electrons near the target surface. As a result, in magnetron sputtering processes, current densities at the target are increased to 10–100 mA/cm^2, compared to about only 1 mA/cm^2 for dc-diode configurations. In addition, magnetron sources can sustain a glow discharge at much lower pressures than can dc diodes (typically pressures as small as only 1 mtorr are needed, instead of 20 mtorr for dc diodes). Since the mean free path of gas atoms λ at P = 1 mtorr is ~6.6 cm (see Eq. 3-5), this distance is comparable to the distance between the target and the wafer (which is 4–10 cm in modern sputter chambers). Thus, most sputtered atoms in magnetron sources arrive at the wafer having traveled in a straight-line path (i.e., without having undergone any collisions with the sputter-gas atoms). On the other hand, since λ = 0..3 cm at P = 20 mtorr, sputtered atoms in dc diode systems undergo many collisions with gas atoms as they cross the space between the target and the wafer. Magnetron sputtering sources were introduced in 1974, and because of their clear superiority quickly became the dominant source configuration for sputtering applications.

To describe the principle of magnetron sputtering in detail, first consider the motion of a charged particle (e.g., an electron, or an Ar ion) having a charge q, a mass m, and a velocity v. Assume the electron is moving perpendicular to a uniform magnetic field **B** as shown in Fig. 11-20. Such a charged particle experiences a force, $\mathbf{F} = q\,(\mathbf{v} \times \mathbf{B})$ which is perpendicular to both the direction of its motion and the direction of **B**. If no other forces are exerted, the particle will continuously be acted upon by **F**, causing it to move in a circular path with radius, r_g, whose value is calculated from:

$$r_g = mv/q\mathbf{B} \qquad \textbf{(11.7)}$$

For electrons and Ar^+ ions, the equation for r_g can also be expressed in terms of the energy, E, of the particles (in eV), and of the magnetic field **B** (in gauss) in the following way:

$$r_g\,(\text{electron}) = 3.37 \times [E^{1/2}\,(\text{eV})/\mathbf{B}\,(\text{gauss})] \qquad \textbf{(11.8)}$$

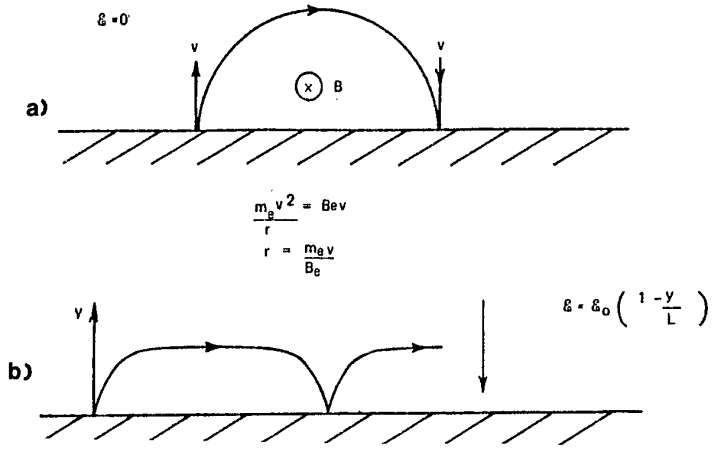


Fig. 11-20 (a) Motion of an electron in a region of magnetic field **B** parallel to the surface. (b) Motion of an electron ejected from a surface with velocity **v** into a region magnetic field **B** parallel to the surface, with no electric field, and (c) with a linearly decreasing field. From B. Chapman, *Glow Discharge Processes*, 1980, Copyright © John Wiley & Sons. Reprinted with permission of John Wiley & Sons.[4]

$$r_g\,(Ar^+) = 9.11 \times 10^2\,[E^{1/2}(eV)/\mathbf{B}\ (gauss)] \qquad \textbf{(11.9)}$$

From these equations, it can be seen that Ar^+ ions have motion radii which are ~300 times larger than the radii of electrons for equivalent conditions of **B** and particle energy. Thus, as an ion moves through the region of the magnetron source containing the magnetic field, the radius of its circular path, r_g, is so large that it essentially moves in a straight line when crossing the dark space of the cathode. This indicates that the direction of electron motion in magnetron discharges is strongly influenced by the magnetic field, while the magnetic field does not significantly change the direction of ions as they cross the dark space.

In the next case, examine the situation of an electron which is emitted from a sputtering target surface. Let there be a magnetic field, **B**, parallel to the surface, and an electric-field perpendicular to the surface (due to the dark space). Assume the electron is initially at rest (e.g., having just been emitted from the target), and the magnitude of the electric-field is given by:

$$E = E_0\,(1 - y/L) \qquad \textbf{(11.10)}$$

where E_0 is the E-field magnitude at the surface of the target, L is the length of the dark space, and y is the vertical distance above the target. The electron is then initially accelerated away from the target in a direction perpendicular to the target surface by the electric field, but simultaneously it experiences a force due to the magnetic field, $\mathbf{F} = q(\mathbf{v} \times \mathbf{B})$.

This force changes the direction of the electron's motion so it is no longer perpendicular to the surface. If the magnetic field is strong enough to deflect the electron so it begins move toward (rather than away from) the target surface before leaving the magnetic field, its overall trajectory will be roughly cycloidal as shown in Fig. 11-20c. That is, as it approaches the target surface after having been deflected, the electron will also be *decelerated* by the electric field. As it slows down r_g is decreased, causing the electron to pivot sharply until it is again moving away from the surface. At that point, the next period of the cycloidal motion is initiated, with the particle again being repelled by the E-field. The maximum excursion of the electron (with charge q and mass m) away from the target during such motion y_{max}, is given by:

$$y_{max} = \frac{1}{B}\left[\frac{2m}{q}(V - V_T)\right]^{1/2} \qquad \textbf{(11.11)}$$

where V_T is the negative target voltage, and V is the potential in the dark space at y_{max}.

The net result is that secondary electrons are trapped near the target surface by the combination of the magnetic and electric fields, and continue to move with a cycloidal (or *hopping*) motion until they collide with an Ar atom. Even though not every such collision produces an ionization event, many more ions are generated than if the secondary electron was to escape from the discharge without undergoing any collisions with Ar. The magnitude of the electron velocity *parallel* to the target surface, v_D (also known as the $\boldsymbol{E}$ x **B**, or *magnetron drift velocity*), is given by:

$$v_D = \frac{10^8 E_\perp (\text{V/cm})}{|\text{B}|(\text{gauss})} \tag{11.12}$$

The most widely used magnetron electrode configuration in ULSI sputtering systems is the *circular magnetron.* Figure 11-21 shows a drawing of such a circular magnetron with permanent magnets behind the target surface. Note that in the axisymetric dc magnetron shown in this figure, the permanent magnet placed behind the target generates magnetic field lines that enter and leave through the cathode plate as shown.

In *circular magnetrons* the target surface is planar, and the **B**-field is created by permanent magnets behind the target.[4,5,6,15] It is common practice to use the ground shield, chamber walls, and other grounded hardware for the anode. During deposition the substrates are stationary in front of the target. In examining the diagram of the circular magnetron, three things should be noted. First, half-way between the magnet poles the magnetic field lies parallel to the target face. This is the region where the magnetron drift velocity, v_D occurs. Second, the $\boldsymbol{E}$ x **B** field formed in this region closes on itself. Thus, a continuous circular path that the "hopping" electrons can traverse is established. Third, the plasma density is greatest where $\boldsymbol{E}$ x **B** is maximized. The region of the target beneath the ring-shaped volume where the plasma density is highest is also sputtered the most rapidly, and this target erosion pattern is called the *race-track.*

Figure 11-22 also shows the hopping motion of an electron as it follows the race track. Near the center of the track the maximum excursion, y_{max}, is ~0.5 cm. The closed path allows the secondary electrons to travel around the circular path several (3–10) times.[16] This makes the coupling of the electron energy to the plasma much more efficient than in dc diodes and results in very low discharge voltages, high currents (in the ampere range), and very little voltage increase as the current is increased. Magnetrons generally follow an empirical current-voltage relationship described as:

$$I = kV^n \tag{11.13}$$

Where I is the discharge current, V is the voltage, k is a system-dependent constant, and n is typically in the range of 5–20. It was first thought that a secondary electron merely drifted around the drift path until it hit a gas atom and ionized it, and then continued around the loop.[17] This is probably not an accurate scenario. Instead, it appears that the electron has many more effective collisions with other electrons, resulting in a general heating of the electron distribution. Therefore, ionization by the high-energy tail of the distribution seems to be the primary means of ionization, not direct collisions between secondary electrons and gas atoms.[18]

11.5.1 Magnetron Sputter Sources for ULSI

The *circular planar magnetron sputter source* configuration described in the previous section is the most widely used magnetron source for ULSI sputter deposition (with some modifications,

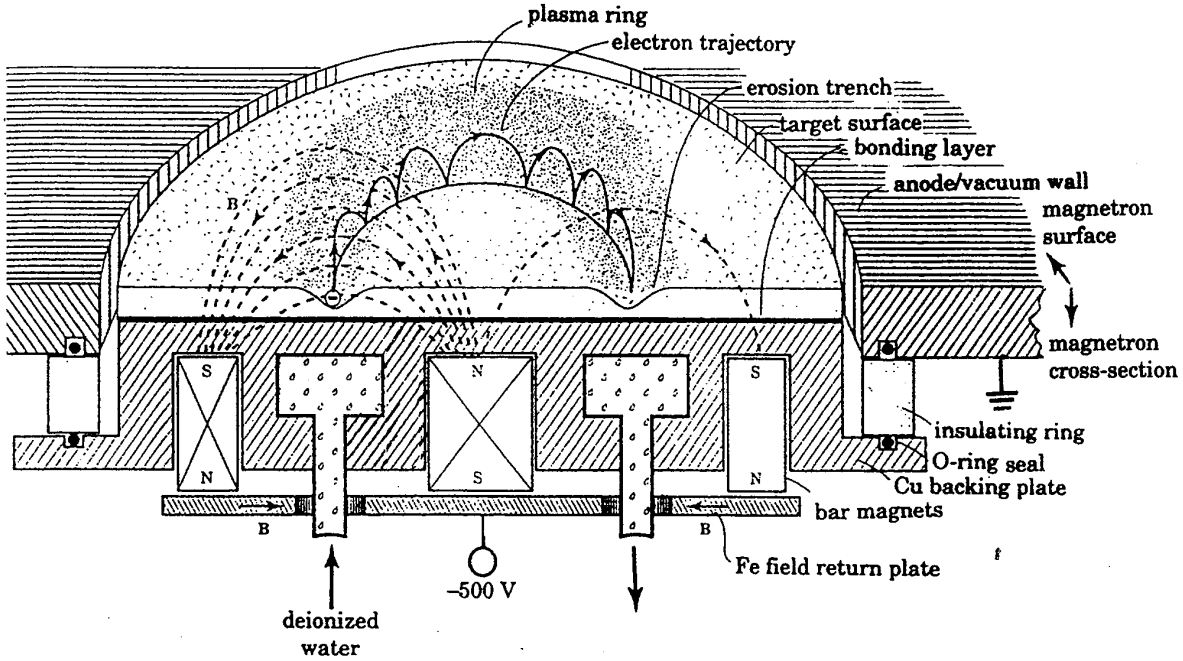


Fig. 11-21 Cross section of circular magnetron sputter source, with permanent bar magnets behind the target, and depicting the resulting magnetic field, race track, and electron orbits.[19]

modifications, as will be described later). In earlier generations of IC technology, however, other magnetron source configurations found much wider use than they do today. These included the *S-Gun* (or *Sputter Gun*) and the *rectangular planar magnetron.* More information about these two magnetron source configurations can be found in the first edition of this text. Here we will confine our discussion to the circular planar magnetron sputter source.

11.5.1.1 Evolution of Planar Circular Sputtering Sources: As noted earlier, the circular planar magnetron sputter sources originally used a ring of permanent bar magnets (plus one central one) located behind the target (see Fig. 11-21). Using the strongest Nd-Fe-B magnets, this magnetic field over the target can approach 1 kilogauss (or 0.1 in SI units). Although this source configuration is a vast improvement over the dc-diode sputtering configuration (principally in terms of much higher sputter deposition rates), it creates several major new drawbacks. Most of them are associated with the fact that for a magnetron source with permanent magnets, the glow discharge region of maximum intensity is an annular ring above the target surface. Hence, most of the sputtering from the target also occurs from the ring-shaped region of the target immediately beneath this intense plasma ring. Several problems arise from this circumstance. First, the deposition rate across the wafer is not as uniform as with a dc-diode source, and steps must taken to mitigate the resultant film thickness non-uniformity (as discussed in the following subsection). Second, the target utilization is poor. That is, the target erodes rapidly in the ring region, but not elsewhere, resulting in a deep groove in the target face along this annular region. The target must be replaced before most of its material is used. This not only requires that the system be shut down for a target change (which reduces the overall uptime of the sputter tool), but also much of the highly-refined (and therefore, expensive) target material must be discarded. Both factors drive up the cost of the sputter process. Finally, on regions of the target surface away from the ring region there can be a net *build-up* of sputter-deposited film (due to redeposition of some sputtered atoms after collision with the sputter gas atoms, see Sect. 11.7.2). Some of the sputtered materials (e.g., Ti:W or TiN) adhere poorly to their own target surface

(which if it is not sputtered, may contain a layer of surface native oxide). Thus, if these materials redeposit onto the target with sufficient thickness, they peel off and become a source of particle contamination in the sputter chamber. (Note that periodic heating of the target during deposition (and cooling between wafers), exacerbates this problem.)

The problem of non-uniformity of the sputtered film thickness as described above has been addressed in a number of ways: 1) in many early magnetron sputter systems the substrates were translated past the sputter target (see Sect. 11.6); 2) the size of the target (relative to the wafer) and the distance between the target and wafer can be optimized to give the best uniformity (see Sect. 11.5.2); and 3) the permanent magnet behind the target can be replaced with a movable magnet array (Fig. 11-22).[20] This array is slowly spiraled behind the target during sputter deposition, and the peak magnetic field is not kept fixed just above the annular ring. More of the target surface is thus exposed to strong ion bombardment, and much more area of the target surface ends up being eroded. The circular planar magnetron source with movable magnets is said to exhibit *full-face erosion.* As one can observe in Fig. 11-23, where a spent circular-magnetron sputtering target is being changed, the entire surface of the target may be eroded to some degree, but erosion is not uniform across the entire area. That is, a spent target has a central mound of thicker material with two or three eroded rings surrounding it. Nevertheless, besides improving the uniformity of sputter deposition rate, full-face erosion consumes more of the target before it has to be replaced (*improved target utilization*), and fewer particles are formed (because re-deposition of material on the target face is suppressed).

11.5.1.2 Deposition Rate and Thickness Uniformity with Circular Planar Magnetrons: Two of the most important characteristics of a thin-film-deposition process are *film deposition rate* and *film thickness uniformity across a wafer.* A sufficiently high deposition rate must be possible to make the process economically viable. The thickness must also be sufficiently uniform to allow control of other processes such as etching, and to guarantee that electrical specifications of features formed from the deposited film are met.

When the circular planar magnetron sputtering source is used in an Applied Materials Endura® single-wafer sputtering chamber (described in more detail in Sect. 11.6.3), it has been

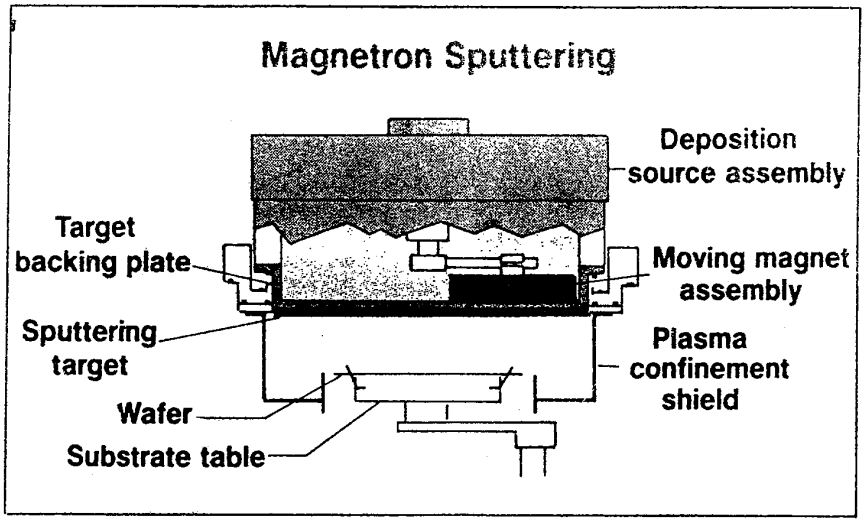


Fig. 11-22 Circular planar magnetron source with rotating movable magnets.[20] Reprinted with permission of Semicondtor International.

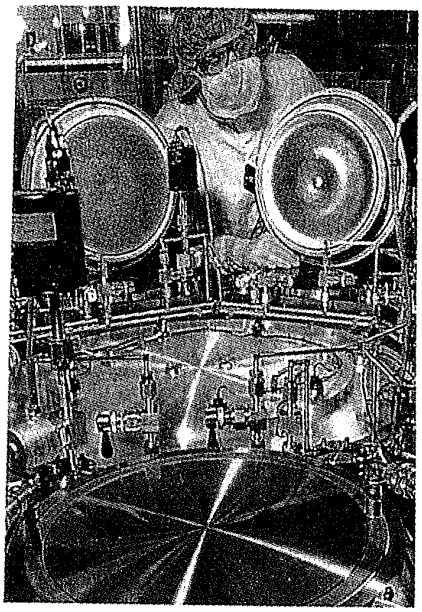


Fig. 11-23 An spent aluminum circular-planar-magnetron target (right) is being changed.

reported that deposition rates of 1200-nm/min can be achieved when depositing Al:(0.5%)Cu films on 200-mm wafers.[30] In fact, the maximum sputter deposition rates with such sources are usually not limited by any aspect of the plasma (as in dc and rf diode cases), but rather by the ability to cool the cathode target to keep it from melting.

The fact that the annular shape of the plasma in a circular planar magnetron results in a non-uniform erosion of the target also causes the deposition onto nearby wafers to be non-uniform. In fact the deposition profile mirrors the race track when the wafer is close to the target, but this deposition profile becomes more smeared as the sample is moved further from the target (see Fig. 11-24a). Thus, a tradeoff is often made between deposition rate and thickness uniformity (i.e., uniformity improves as the target-to-wafer spacing is increased, but the deposition rate decreases). Note also from Fig. 11-24 that the target diameter is made larger than the wafer diameter to improve thickness uniformity. As an example, for 200-mm (8 in) wafers, the target diameter is typically 330 mm (14 in). Figure 11-24b shows such a target sputtering onto a smaller wafer. In practice, thickness uniformities of ±5% (3σ) within a wafer can be achieved.

11.6 VLSI AND ULSI SPUTTER DEPOSITION EQUIPMENT

A number of different sputtering systems have been designed for commercial use in IC fabrication. Only the dominant sputtering systems designs that were used in VLSI and ULSI generations will be described (i.e., those used for IC fabrication on 125-mm, 150-mm, and 200-mm wafers). Sputtering systems used for earlier IC technologies are described in the first edition of Vol. 1. In this section the generic components of a sputtering system are discussed first. Then the sputtering systems used in 125-mm and 150-mm wafer processing are briefly examined. The focus of this discussion is on the sputtering systems used for 200-mm wafers (and larger).

11.6.1 The Components of a Generic Sputtering System

The schematic of a basic sputtering system is shown in Fig. 11-25. It consists of the following subsystems: a) the sputter chamber, in which the substrate holder and sputtering source reside (the source also includes the target); b) vacuum pumps; c) power supplies (dc and/or rf); d) sputtering gas supply and flow controllers; e) monitoring equipment (pressure gauges, volt-

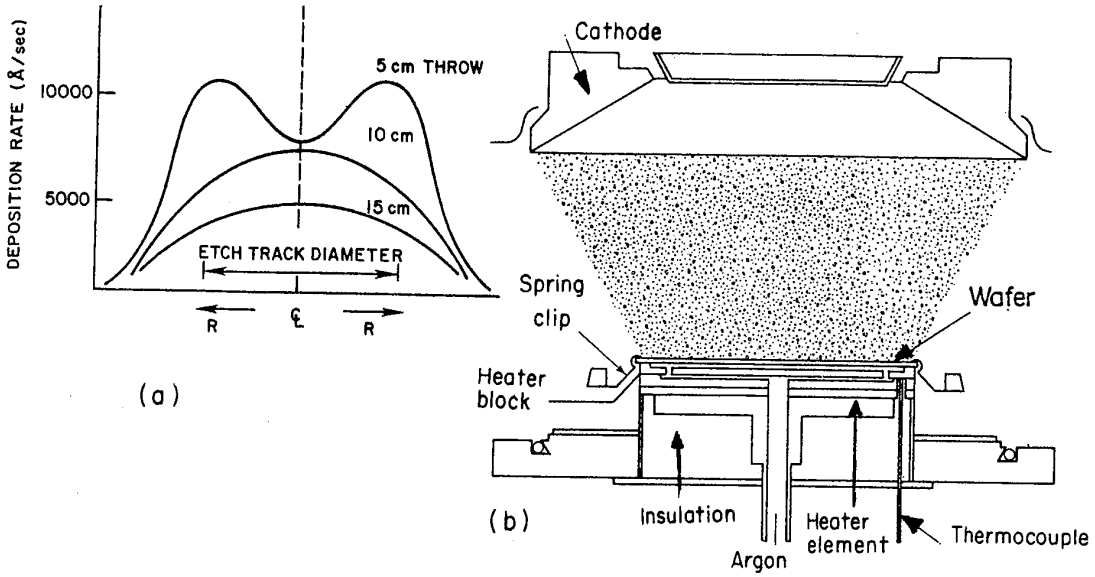


Fig. 11-24 a) The deposition profile and deposition rate on a sample in front of a magnetron cathode as a function of sample distance.[21] Reprinted with permission of Academic Press. b) Sputter deposition from a planar magnetron target whose diameter is larger than that of the wafer.

(pressure gauges, voltmeters, and residual-gas analyzers); f) wafer holders and handling mechanisms; and g) microcomputer controller.

In modern sputter tools the sputter chambers are isolated from the ambient by a load lock. That is, a cassette of wafers to be sputter deposited is placed into the loadlock. Then the valve connecting the loadlock to the outside ambient is closed. The loadlock is pumped down, and upon reaching the base pressure of the transfer chamber the vacuum valve to the transfer chamber is opened. A robotic arm in the transfer chamber removes a wafer from the cassette and brings it into the transfer chamber. The loadlock/transfer chamber valve is then closed. The sputter chamber awaiting the wafer has also been pumped to the appropriate low base pressure (e.g., 10^{-6}–10^{-8} torr). The vacuum valve to the sputter chamber is opened, and the wafer is then

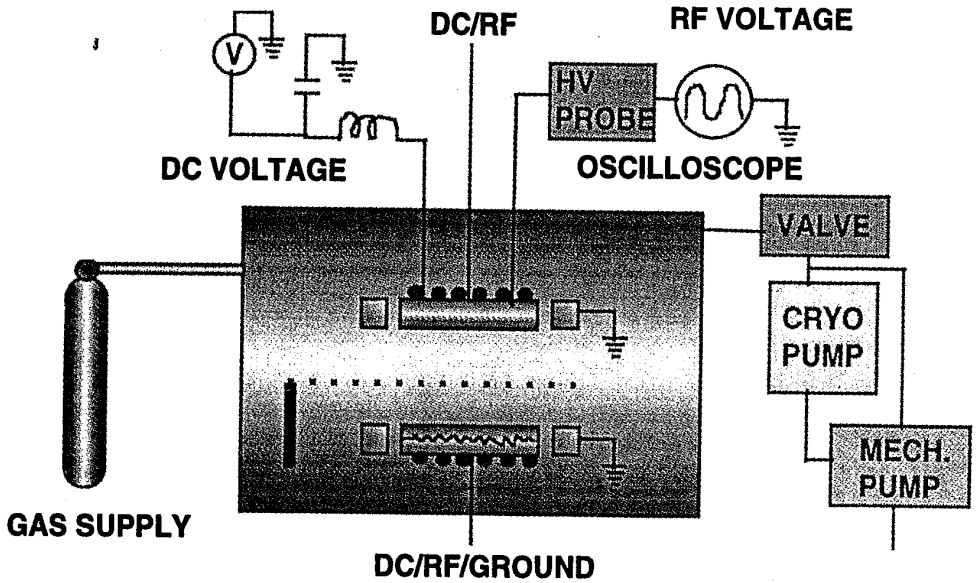


Fig. 11-25 Schematic drawing showing some of the components of a sputtering system.

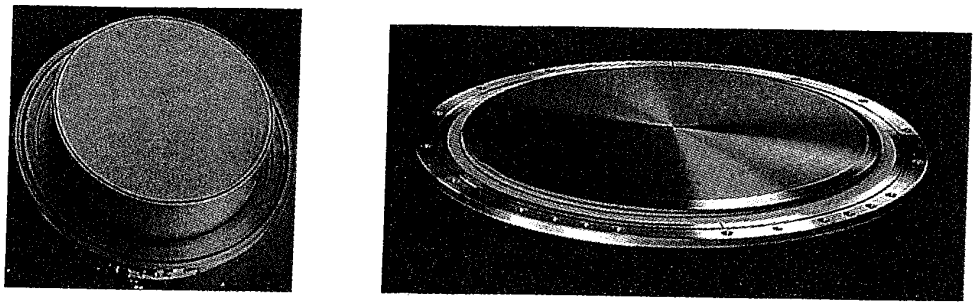


Fig. 11-26 Examples of sputtering targets for planar circular magnetron sources.

inserted into the sputter chamber by the robot arm. After the arm withdraws, the vacuum valve between the sputter chamber and the transfer chamber is closed, and sputter gas is flowed into the chamber to the appropriate pressure (e.g., 5 mtorr) and flow rate. The sputter process can then be started.

11.6.1.1 Sputtering Targets: *Sputtering targets* consist of the material that is to be sputter-deposited onto the wafer. Targets in circular magnetrons are circular discs 3 to 10 mm thick, as shown in Fig. 11-26. These are bonded to a water-cooled Cu backing plate for good thermal contact, usually by soldering with a low-melting-temperature metal (or with a conductive epoxy). Since over 70% of the energy incident onto the target goes into target heating, targets must be adequately cooled to prevent the following problems: warpage; deposition-rate changes due to thermal restructuring of the target material; delamination from the backing plate; or even target melting. De-ionized water is used as the cooling liquid to prevent electrolytic corrosion between the electrically biased backing plate and the grounded water supply. The entire cathode assembly is floated off ground by a ceramic insulator ring.

Targets consisting of metals with relatively low melting points (e.g., Al, Cu, and Ti) are fabricated by melting and casting the metals in either vacuum or protective ambients. The starting materials to fabricate the targets must be very pure (e.g., up to 99.9999 pure [or six nines] for aluminum targets). In the case of cast alloy targets, fine grain size is preferred to minimize segregation of constituent elements into nodules during ion bombardment (since such nodules are a source of large particle ejection from the target during sputtering). Targets of high-melting-point refractory metals (e.g., W:Ti) and compound materials are generally formed from powders by hot-pressing. Although there may be no alternative, this fabrication method is less desirable than the vacuum-melting process. Less-than-theoretical densities are achieved with powder processing. As a result, such targets release trapped gases, particulates, and exhibit occlusions as their surfaces are eroded. The latter can give rise to non-uniform target erosion, arcing, and deposited films of inferior quality. Some device applications (e.g., memory circuits) also demand material very low in alpha-particle emitting elements (i.e., less than 0.01 ppb U and Th content).[37]

Pre-sputtering of targets is done to clean their surfaces prior to film deposition. Once the target surface is clean, sputter deposition onto product wafers can take place. Arcing may occur at the target surface in dc sputtering until the surface is cleaned. Thus, it is common practice to *condition* freshly installed targets by slowly increasing the applied power, until all the material that could give rise to the arcing has been sputtered away.

11.6.1.2 Vacuum Pumps for Sputtering Systems: *Vacuum pumps* are also an important part of the sputtering system (as will be described in more detail in Sect. 11.6.3). Modern sputtering systems use cryopumps to keep the sputter chambers under high (or ultra-high) vacuum between times of sputtering.[19] Cryopumps offer higher water pumping speeds than turbo pumps (see

condition freshly installed targets by slowly increasing the applied power, until all the material that could give rise to the arcing has been sputtered away.

11.6.1.2 Vacuum Pumps for Sputtering Systems: *Vacuum pumps* are also an important part of the sputtering system (as will be described in more detail in Sect. 11.6.3). Modern sputtering systems use cryopumps to keep the sputter chambers under high (or ultra-high) vacuum between times of sputtering.[19] Cryopumps offer higher water pumping speeds than turbo pumps (see Chap. 3) and clean operation. State-of-the art sputtering chambers typically operate with base pressures of $1x10^{-8}$ torr, or less. Prior to beginning the sputter process, the chamber is backfilled with Ar gas. Since an open-flow process configuration is used, once the sputter process is initiated, the cryopump is used to continually pump the sputter gas (Ar) out of the chamber. That is, during sputtering Ar is metered into the chamber at a controlled flow rate, and the pump simultaneously evacuates it while maintaining the pressure in the chamber at the desired process pressure (e.g., 2–5 mtorr). Such continuous renewal of the Ar gas is another measure taken to ensure that the partial pressures of contaminant gases in the chamber are kept to adequately low levels.

11.6.1.3 Power Supplies for Sputtering Systems: Sputtering systems are typically provided with both a *dc* and an *rf power supply.* Dc power supplies can be built to supply up to 20 kW of power, whereas rf power supplies are limited to ~3 kW. As a result, dc magnetron sputtering can provide higher deposition rates than rf magnetron sputtering. Consequently dc-magnetron sputtering is the operational mode of choice for high-deposition-rate processes. Dc power supplies for sputtering must be designed to be capable of withstanding arcs without damage.[12] Such arcs are thought to arise from the local dielectric breakdown of native oxides on freshly installed targets, or from other surface regions on which contaminant insulator layers have formed. Conversely, rf sputtering removes such regions as a matter of course, and rf supplies are not subjected to arcing. The dc supplies, however, must be designed to handle such local arcing without shutting down or damage, as the regions responsible for such arcing are normally cleaned within minutes after the power is supplied to the system, and arcing then ceases. On the other hand, the supply must be able to sense large arcs caused by such events as metal flaking between the target and ground, and to be able to switch off soon enough to prevent damage to the target.

11.6.1.4 The Gas Supply for Sputtering Systems: As noted earlier, Ar is used as the process gas for almost all sputtering applications (with one exception being the reactive sputtering of Ti, where a mixture of Ar and nitrogen is used to deposit TiN films). The gas is metered into the chamber by a mass-flow controller. The purity of the Ar gas is its most important characteristic. As will be discussed in Sect. 11.7, if contaminant gases (especially H_2O, O_2, and N_2) enter the sputter chamber in even low parts-per-million amounts, the characteristics of the films being deposited can be significantly altered. Thus, ultra-high-purity Ar must be used, and the gas lines that deliver the Ar to the chamber must be kept leak tight. Note that at the times when a wafer is not loaded into the process chamber, the Ar flow is typically shut off, allowing the chamber to return to an ultra-high vacuum base pressure.

11.6.2 Commercial Sputtering Systems for 125-mm and 150-mm Wafers

The most widely used sputtering systems for 125-mm and 150-mm wafers were the Varian 3180 models (for wafers up to 125-mm in diameter) and the Varian 3290 models (for wafers up to 150-mm in diameter). They employed a *static* sputtering mode, insofar as the wafer and target

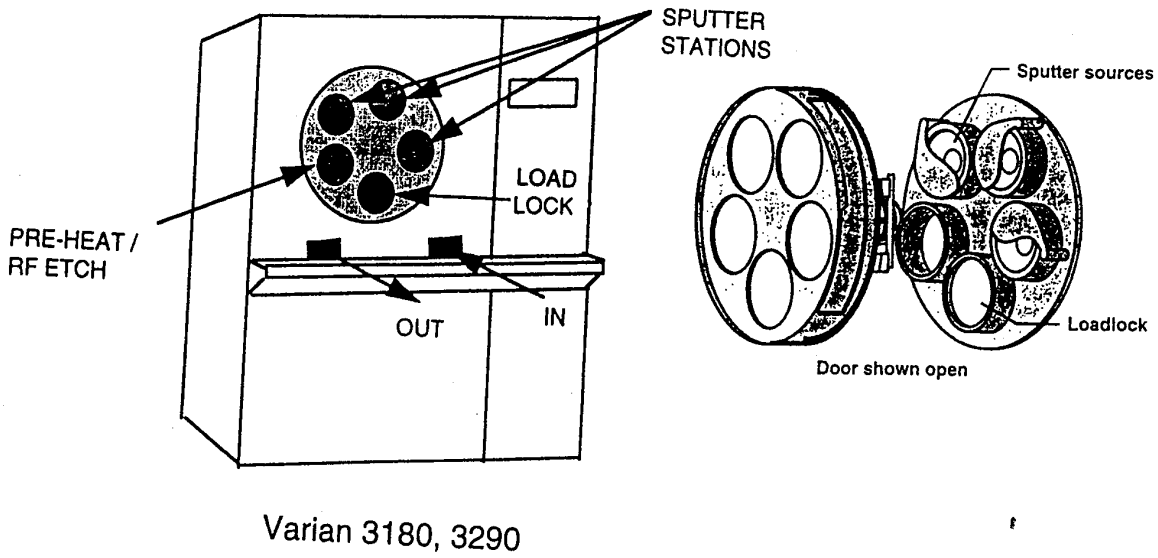


Fig. 11-27 The dominant sputtering system for 125-mm and 150-mm wafer processing, the Varian 3180 (for wafers up to 125-mm in diameter) and the 3290 (for wafers up to 150-mm in diameter).

remained stationary during the sputtering process (Fig. 11-27).[22] This differs from the sputtering systems of earlier generations, in which the wafers were moved in front of the target during sputtering to improve deposition-rate uniformity. Unfortunately, mechanical movement during the sputtering process produces vibrations that significantly increase particle concentrations in the sputter chamber. Thus, moving sputtering systems were largely abandoned once 125-mm wafers were fabricated. The sputter sources in the 3180 were initially S-Guns® (see 1st Ed. for information on S-Guns), but later circular planar magnetron sources were introduced. The wafers and target were both vertically positioned (termed *side sputtering*) and the target-to-wafer distance was 5–10 cm. One wafer at a time was sputter deposited by each target, and there were three sputtering target stations in the 3180 model. Thus, up to three wafers could be sputtered simultaneously by the tool, and if a different target material was placed at each station, multi-layered films could be deposited by moving a wafer between stations. Note that in bipolar processes of that era, the contact/interconnect metallization consisted of three layers, namely PtSi (the contact metal layer), Ti:W (the diffusion barrier layer), and Al:Cu (the main interconnect material layer). Thus, a single tool that could sputter deposit all three layers was useful.) For NMOS processes of that time, only a single layer metal was needed (typically Al:Si), and hence all three stations could be equipped with identical (Al:Si) targets to increase throughput. These sputtering systems also contained an rf sputter-etch/heating station and were designed for cassette-to-cassette operation. The system is operated by placing one or two cassettes on a load station. The wafers are then fed sequentially through the process chamber by means of a transfer plate. The plate rotates clockwise and steps the wafer through the heating station and the three sputtering stations. After the wafers have stepped through all four process stations, they are returned to the wafer cassette. Note that the entire chamber containing the three sputtering stations and the rf etch/heating station is pumped by a single pump and the vacuum level is the same throughout the entire chamber. The base vacuum pressure (when no wafers are being processed) is 5×10^{-7} torr. The tool was not redesigned to accommodate 200-mm wafers.

11.6.3 Commercial Sputtering Systems for 200-mm Wafers

The sputtering systems for 200-mm (and larger) wafers are, by and large, configured as *cluster tools*. In a cluster tool a number of sequential process steps can be carried out without exposing the wafer to the atmosphere between steps. This is accomplished by moving the wafer from one process chamber to another through an evacuated transfer chamber, using a robotic transfer mechanism. The sputtering system that came to dominate 200-mm technology was the Applied Materials Endura.® It was introduced in 1990 and by 1995 had captured almost 80% of the market share for 200-mm sputtering systems.[23] According to an article in *Semiconductor International* in 1991, "Applied was able to develop a machine ... suited to the next few device generations without needing to make any startling technological breakthroughs. There is nothing breath-taking about the Endura 5500, nothing glitzy, nothing here that speaks of genius - though a good deal of intelligence is evident, yes. As it happened that was enough."[24]

The Endura® won this large market share for ULSI sputtering applications for several reasons, but probably the foremost was that it proved to be a highly reliable multichamber process tool. In one reported manufacturing evaluation, the Endura® demonstrated a long mean-time-between-failure (MTBF) record (i.e., > 120 hrs) and had an uptime that exceeded 85%.[25] In addition, it offered some unique technological advantages. Figure 11-28 is a top-down schematic drawing of the Endura.® The core platform of the system consists of two *transfer chambers*. A number of process chambers (including up to four sputter chambers) can be bolted to this core platform. The two transfer chambers are: 1) a *buffer-transfer chamber* that can receive and deliver back wafers to *two cassette loadlocks* (LLA and LLB), and can then sequentially feed the wafers to a *de-gas/wafer orienting chamber* (chamber E) and to the *sputter pre-clean chamber* (chamber A); and 2) *a high-vacuum transfer chamber* that receives wafers from the pre-clean chamber A and transfers them to the sputter chambers (PVD chambers 1, 2, 3, and 4), as needed, and then delivers the sputter-coated wafers back to the cooldown chamber B.

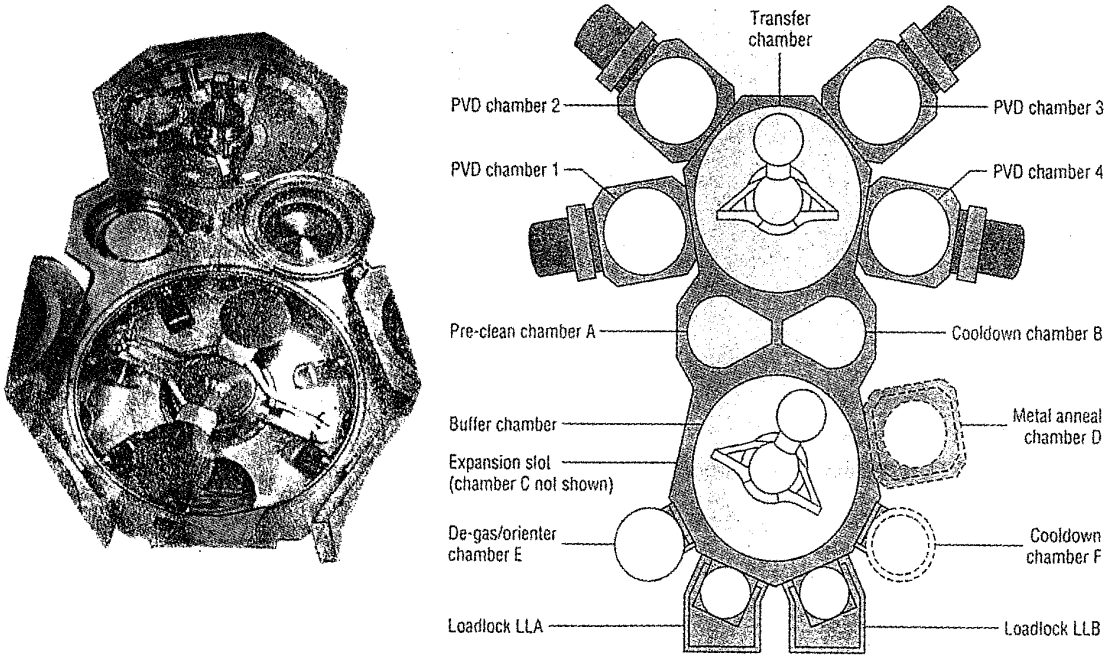


Fig. 11-28 Layout of an Endura 5500® sputtering system. Courtesy of Applied Materials.

The two-transfer-chamber configuration allows for a *staged vacuum architecture*, and that is one of its chief technological benefits.[26] Specifically, there are *five vacuum stages* in the system and two load locks. The vacuum base pressure in the *first stage* of the system (i.e., in loadlocks LLA and LLB) is 10^{-5} torr. In the *second stage* (the buffer-transfer chamber) the vacuum base pressure is 10^{-6} torr. In the *third stage* (the pre-clean chamber A, and the cooldown chamber B) the base pressure is 10^{-7} torr, and these chambers also act as the second loadlock (i.e., between the buffer and high-vacuum transfer chambers). The *fourth stage* is the high-vacuum transfer chamber, and the base pressure there is 10^{-8} torr. Finally, the *fifth stage* is in the sputter chambers themselves, and the base vacuum pressure here is in the *ultra-high-vacuum* range (i.e., 10^{-9} torr). Note that this final vacuum stage is kept at a much higher base vacuum level than in the sputtering systems used for the 125-mm and 150-mm generations, and this represents another of the first applications of ultra-high-vacuum in an IC process tool. It was one of the major advances that the Endura® offered over sputtering systems of previous generations. The staged vacuum architecture also provides for efficient recovery of vacuum levels since the staged vacuum approach is designed for only one decade of change in the vacuum level between any isolated chamber and the associated transfer chamber. In addition, since each process chamber is isolated by a transfer chamber from all the other process chambers, it is possible to break into a particular chamber to change a target without interrupting the rest of the system. This is different from the situation that existed in previous sputterers. That is, in those systems all the sputter stations were interconnected to a common pumping system, and thus it was necessary to shut down the entire tool to change a single target.

Prior to the development of the Endura® it had been learned that electromigration and hillock formation in aluminum films (as well as stress formation in the film)[27] was profoundly affected by contamination (especially the presence of water vapor in even trace amounts in the Ar sputter gas and the presence of native oxide on the Si surface).[28,29] Thus, the Endura® system included two cleaning chambers (instead of just the one found in the 3180 and 3190). In the first one (the de-gas/wafer orienting chamber [chamber E]) water vapor adsorbed on the wafer surface is driven off by heating the wafer. Since this step is carried out in a chamber that is vacuum-connected to the buffer-transfer chamber, much of the oxygen and water on the wafer surface is eliminated prior to its introduction into the high-vacuum section of the platform core. Chamber A is the second pre-clean chamber. Its function is to remove native oxide and other contaminants by use of a low-energy ion bombardment (i.e., using a soft Ar rf-sputter process that is designed to effect such removal with minimum damage to the silicon surface). The ultra-high-vacuum base pressure in the sputter chambers themselves is another important factor that contributes to keeping the level of contaminants in the sputtered film at lower levels than could be achieved in earlier generations of sputterers. A separate cooling chamber between the high-vacuum and buffer-transfer chambers (Chamber B in Fig. 11-28), performs a secondary function of minimizing cross-contamination between these two stages of the platform core. Finally, only one slit valve (that is the valve that isolates a process chamber from a transfer chamber) is open to a transfer chamber at any time so, with the proper pump package, cross contamination among process chambers can be eliminated.

The sputtering chambers of the Endura® use circular planar magnetron sources with movable magnets, as described in Sect. 11.5.1.1. Different magnet assemblies are used for different target types, and the magnet is moved slowly behind the target. The chamber volumes are rectangular in shape, but shields surround the target cylindrically. Wafers are loaded into the chamber and

are then lifted up to be surrounded by the shield. Thus, the electrode-to-wafer distance be adjusted to take into account aging of the target. That is, as the target erodes its thickness decreases. To compensate for this the wafer is moved closer to the target.

The deposition rate of Al:Cu:Si films can be as high as 1.2 μm/min, and for TiN 0.15 μm/min. The basic throughput of the Endura® platform for depositing a three-layer metallization film (Ti, TiN, and Al:Cu: Si) configured with one PVD chamber for each material (and a single de-gas cleaning chamber) was found to be 25 wafers per hour. Film thickness uniformities of better than ±5% and particle densities of < 0.1 cm^{-2} were also obtain ed.[21] Figure 11-29 depicts a perspective schematic drawing of the Endura, ® and Fig. 11-30 shows a photograph of a target being changed in this tool. Other suppliers of ULSI sputtering systems (who in 1999 shared the remainder of the market share in roughly equal amounts) include Novellus (which acquired Varian's PVD business in 1997), Ulvac, Anelva, and Tokyo Electron (TEL).

11.7 PROCESS CONSIDERATIONS IN SPUTTER DEPOSITION

The goal of a deposition process for producing interconnect layers is to create films with all of the properties required for the ULSI circuits being fabricated. In this section the desirable properties of an interconnect film are listed, and various factors that impact the quality of such characteristics are discussed. Note that some of these factors were already considered in previous sections (e.g., deposition rate and thickness uniformity), but others need to be addressed.

A deposited interconnect film must satisfy a large number of requirements in order to effectively perform its role in ULSI integrated circuits. The most important characteristics impacted by the deposition process that a film must possess include: a) correct nominal thickness; b) thickness uniformity of at least ± 5% (within a wafer, and wafer-to-wafer); c) adequately low resistivity; d) resistivity uniform to at least ± 5%; e) good adhesion to underlying and overlying layers; f) step coverage of at least 50%); g) high electromigration resistance and good resistance to hillock formation is demonstrated; i) controllable reflectivity; j) low-resistance contacts to Si and/or other interconnect layers can be made; k) films can be deposited as an alloy with tightly controlled composition; l) films may be need to possess a multilayer structure; m) grain sizes in the film should be controllable; n) deposited films are low in alpha-particle emitting elements (uranium and thorium); o) stresses in the deposited film are low; and p) there is a reasonable latitude in the process conditions that effect the desired film properties (a so-called "robust process window").

In the remaining subsections of Sect 11.7 aspects of the sputtering process that impact a few of the above properties will be the focus, namely: alloy composition, the impact of transporting the vaporized atoms between the target and the wafer, and the deposition of compound materials by reactive sputtering (e.g., TiN). In Sect. 11 the issues involving step coverage and the filling of contact holes and vias using sputter deposition are considered. Additional information on stress, adhesion, and resistivity of interconnect films is found in Chap. 4. A detailed discussion of electromigration, hillocks, and stress-voiding is undertaken in Vol. 2 of this series.

11.7.1 Sputter Deposition of Alloy Films

Many sputtering applications call for the deposition of alloys rather than pure films (e.g., Al:Cu or Ti:W). Multi-component targets are usually used to deposit such alloys. Under steady-state sputtering conditions the elements are sputtered from the target in the same ratio as are present in the alloy. This is one of the primary reasons why sputtering displaced evaporation for PVD depositions. It is more difficult to control the alloy composition of a deposited film within

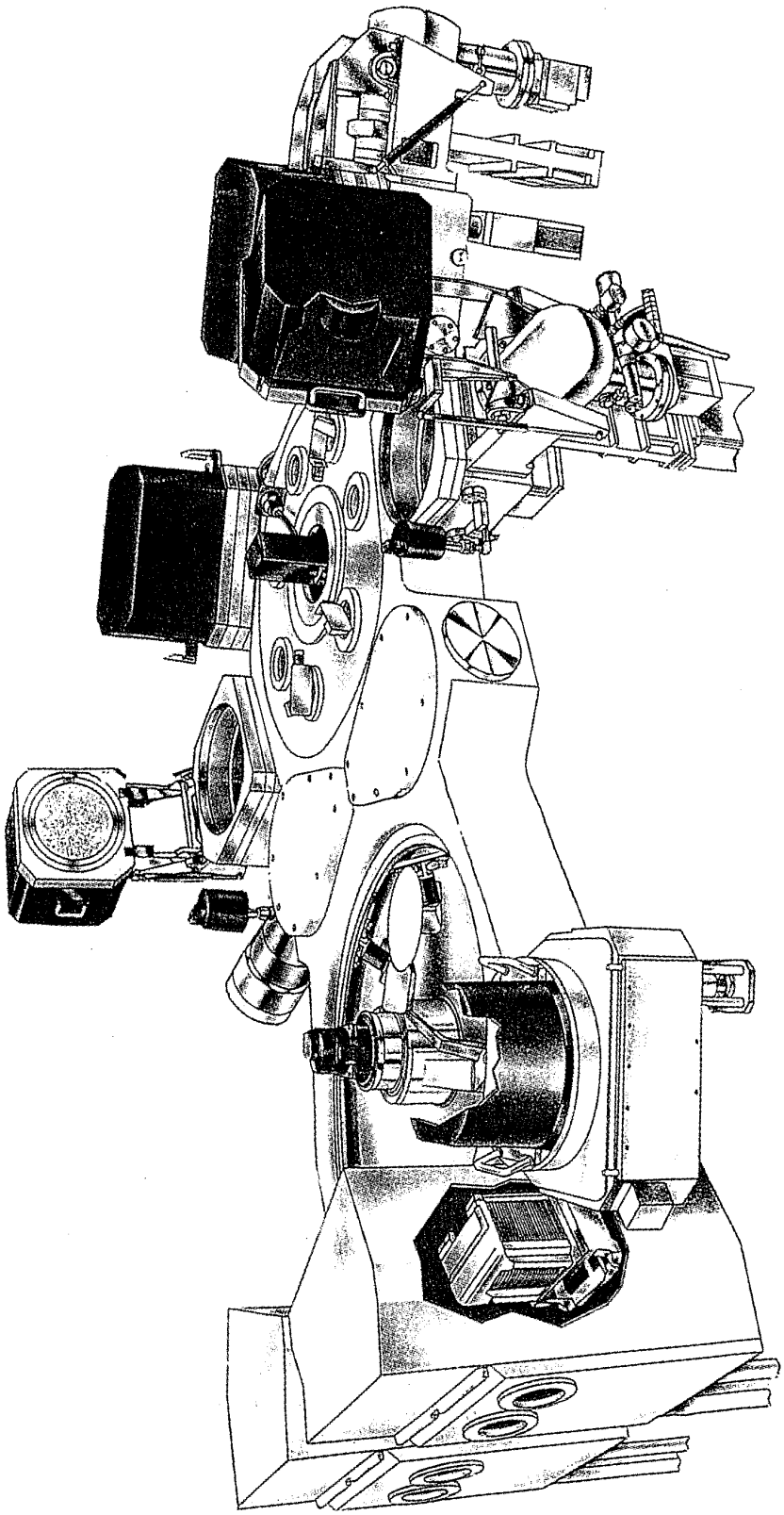


Fig. 11-29 Perspective schematic drawing of an Endura 5500® sputtering system. Courtesy of Applied Materials.

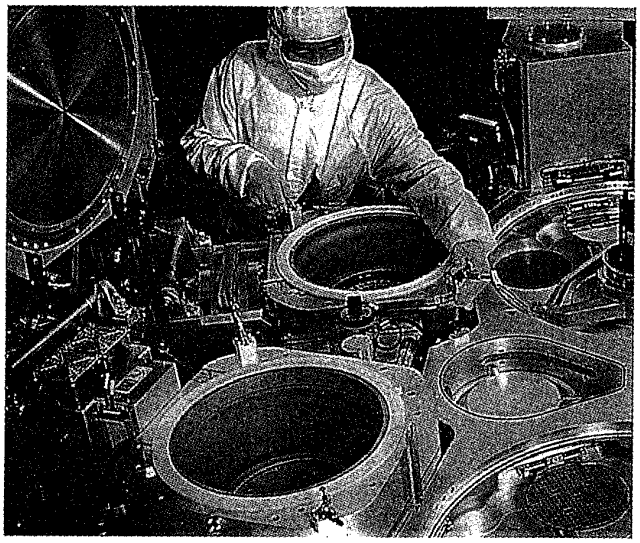


Fig. 11-30 Changing a target in an Endura sputtering system. Courtesy of Applied Materials.

a large wafer (and from wafer-to-wafer) with evaporation. Why the composition of sputtered films can more closely track that of the target alloy composition will now be discussed.[4,31]

Consider the case where a metal film is to be formed from a target consisting of a binary alloy with a number of A atoms (n_A), and a number of B atoms (n_B), such that the total number of atoms is n. Also assume that the B atoms are uniformly distributed throughout the matrix of A atoms. The target concentrations are thus $C_A = n_A/n$, and $C_B = n_B/n$, with sputtering yields S_A and S_B, respectively. Note that a specific target material will be used as an example (that is, one with alloy composition of $A_{0.5}B_{0.5}$ and in which the sputter yields are $S_A = 0.2$ and $S_B = 0.1$), but the result that is obtained holds for any alloy composition and sputter yields. For instance, in this example (shown in Fig. 11-31) the surface will initially have a nominal composition ratio of $C_A/C_B = 1$, and the ratio of the sputtered fluxes, ϕ_A and ϕ_B, in this case is given by:

$$\frac{\phi_A}{\phi_B} = \frac{S_A C_A}{S_B C_B} = 2 \quad (\text{since } C_A = C_B) \tag{11.14}$$

If n_g sputtering gas atoms impinge on the target, the total number of A and B atoms ejected from the surface are $n_g C_A S_A$ and $n_g C_A S_A$, respectively. Thus, the target surface concentration ratio is altered to:

$$\frac{C'_A}{C'_B} = \frac{C_A(1 - n_g S_A / n)}{C_B(1 - n_g S_B / n)} \tag{11.15}$$

instead of C_A/C_B. After the first layer of atoms from the surface is sputtered away, the value of C_A'/C_B' is less than 1 (i.e., 0.9). That is since $S_A > S_B$, the surface will begin to become enriched in B atoms, which now begin to be sputtered in greater quantity, so ϕ_A/ϕ_B decreases. Its value can now be calculated from:

$$\frac{\phi_A}{\phi_B} = \frac{S_A C'_A}{S_B C'_B} = \frac{S_A C_A(1 - n_g S_A / n)}{S_B C_B(1 - n_g S_B / n)} = 1.8 \tag{11.16}$$

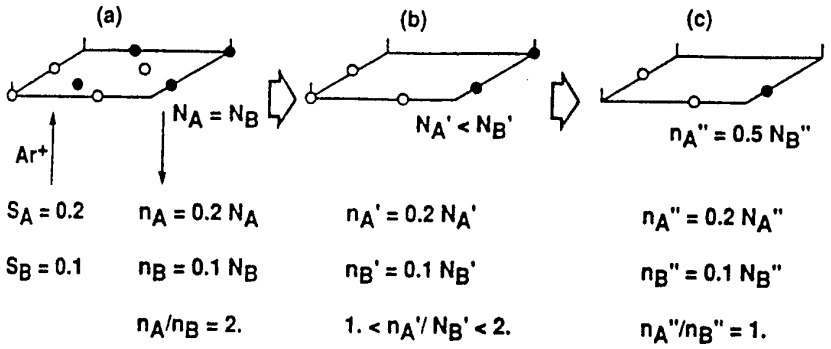


Fig. 11-31 (a) Development of target surface and sputtered flux due to continued sputtering of an alloy = C_B and sputtered flux ratio $\phi_A/\phi_B = 2$ leads to (b) a surface depleted in A so that $C_A'/C_B' \leq 1$, but $\phi_A' \geq \phi_B'$ leading to further depletion until c) $C_A''/C_B'' = 0.5$ and $\phi_A'' = \phi_B''$.

The target surface composition eventually reaches $C''_A/C''_B = 0.5$. At that point the *sputtered flux ratio* reaches its steady-state value $(\phi_A/\phi_B)_{eq} = 1$. The flux ratio of atoms sputtered from the surface when $C''_A/C''_B = 0.5$ represents the original alloy composition ratio, and this is a steady-state condition. Since the atomic species are being removed at the same rate, the surface composition of the target is stable and this process will continue until the target is consumed. A new target, however, must be conditioned to reach this steady-state condition by sputtering a few hundred layers off of its initial surface. Note that the argument given here assumes that minimal solid-state diffusion occurs within the target during sputtering (and that the ratio of the concentrations of the two materials is constant throughout the entire target), enabling the required altered target surface concentration to be maintained.

Consider also a specific example, that of sputtering from an Al:2at%Cu alloy target. Since Cu has a slightly higher sputter yield than Al (i.e., S_{Cu} ~2.7 vs. S_{Al} ~2, for 1000 keV Ar ions), the ratio of Cu to Al atoms sputtered is initially higher than the ratio of Cu to Al in the target. But this situation does not persist for long. As the surface gets relatively depleted of Cu, the relative number of sputtered Al atoms increases (and that of Cu decreases), until at steady-state, the material is leaving the surface in the ratio of 98 Al atoms to 2 Cu atoms. For this to occur, the target surface composition must be altered to have a composition ratio of $C''_{Al}/C''_{Cu} = \phi_{Al}S_{Cu}/\phi_{Cu}S_{Al} = 98(2.7)/2(2) = 66.1$. This is equivalent to about 98.5 Al and 1.5 Cu.

The composition of the deposited film, however, may not exactly mirror the composition of the target because of other factors. For example, if the alloy constituent atoms differ greatly in mass, the lighter atoms will be more strongly scattered by Ar atoms as they move from the cathode to the substrate (see Sect. 11.7.2). In addition, if bias-sputtering is utilized, resputtering of the deposited film may alter the film composition (often in unexpected ways). Thus, the film composition must be measured for each process condition. Again, however, once a set of process conditions is fixed, it is possible to obtain reproducible results from run-to-run.

11.7.2 The Effects on the Sputter Process of the Transport of the Vaporized Atoms Between the Target and the Substrate

As mentioned in Sect. 11.3.5, events that occur as the sputtered atoms are transported in the reduced pressure gas phase between the target and the wafer can have an impact on some of the sputter-deposited film characteristics. That is, the atoms ejected from the target may collide with the sputter gas atoms, and these collisions will change the direction of their trajectories. (see

Fig. 11-12). When the pressure in the sputter chamber is high (as is the case for dc-diode sputtering), the mean free path is short, and many collisions occur. A large fraction of the sputtered atoms thus get scattered to the sidewall of the chamber (or are re-deposited onto the target surface), and this contributes to the low deposition rates observed with dc-diode sputtering. Since lower pressures can be used with magnetron sputtering, this drawback is mitigated. However, another problem arises.

That is, atoms whose masses are smaller than (or are close to) that of Ar, will suffer larger directional changes than those whose mass exceeds that of Ar. This is important when a binary alloy material is being sputtered, and the atoms of one of the alloy materials is much smaller than the other, and those of the other have a mass greater than that of Ar. The smaller mass atoms will be scattered more during the vapor transport, and thus a smaller concentration of them will end up in the deposited film than are present in the target. The collisions will cause more of the smaller-mass atoms to be scattered to the chamber sidewalls, where they will condense to form a solid film (instead of onto the wafer).

This situation applies to sputtering titanium:tungsten (Ti:W) and aluminum:copper (Al:Cu). In the case of Ti:W, the deposited film contains less Ti than the target alloy (~ 50% less). The mass of a Ti atom is much smaller than that of W (48 vs. 184 amu). Since the mass of Ti is only slightly larger than that of Ar (whose mass is 40 amu), Ti atoms are strongly scattered by the Ar gas atoms (Fig. 11-32). But, the trajectories of the heavier W atoms are impacted much less by collisions with Ar. Thus, to deposit a film having 10% Ti and 90% W (by weight) at a sputtering pressure of ~5 mtorr, the target must contain ~19% Ti.[32] This has two important implications:

1. The resistivity of the deposited film depends on the Ti concentration since $\rho_{Ti} = 48$ $\mu\Omega$-cm, and $\rho_W = 5.5$ $\mu\Omega$-cm. In some applications the resistivity of the Ti:W film is a critical parameter (e.g., in the fuse links of some memory devices).
2. Some back-scattered Ti atoms re-deposit on the target surface and form dendrites. These can flake off if they become too large, and hence are a potential particle source.[32]

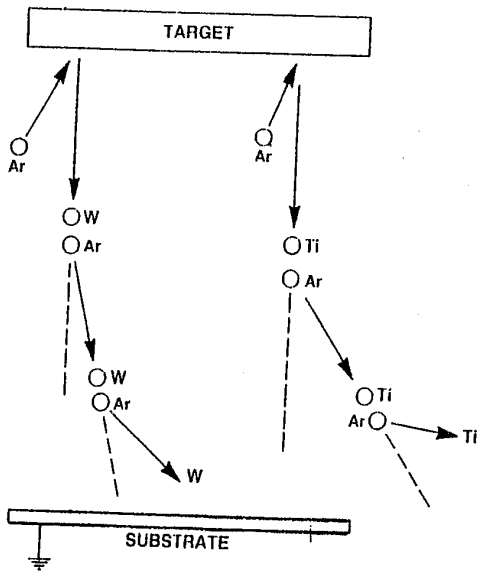


Fig. 11-32 Titanium:tungsten gas scattering model.

For the case of Al:Cu sputtering, since the masses of Al, Ar, and Cu atoms are 27, 40, and 64 amu, respectively, this effect will result in the deposition of films that have a slightly larger Cu concentration than that present in the target.[33] It is also worth mentioning that sputtering with a heavy gas such as Kr (whose mass is 84 amu) is often conterproductive, due to the gas scattering mechanism. That is, the sputter yield may be slightly increased, but the probability of an Al or Ti atom getting to the sample to form a film is reduced significantly.

11.7.3 Wafer Heating During Sputter Deposition

Heating of the wafers can occur during sputter deposition. This heating may be a natural consequence of the process, or may be deliberately produced by auxiliary heating procedures. In some cases heating produces beneficial effects, especially if it can be controlled, while in other instances the heating can degrade the deposited film properties.

Unintentional heating is primarily caused by the following phenomena: a) high-energy secondary electrons striking the substrates; b) the heat-of-condensation of the depositing film (this component is more important in high deposition rate processes); and c) the dissipation of the kinetic energy of the arriving film atoms. In magnetron sputtering, fewer high-energy electrons escape from the target region to strike the substrates, and electron capture shields are utilized to suppress this effect even further. Thus, in modern sputtering systems, this heating effect has been significantly reduced. The heat of condensation and kinetic energy of the incoming film atoms will cause only small temperature increases in low-deposition-rate processes, while in high deposition-rate processes (e.g., the 1 μm/min Al depositions), temperature increases of ~50°C to 100°C have been reported.[34]

Heating by auxiliary means to improve certain film properties (e.g., step coverage) is carried out by lamps, backside heating with hot Ar gas, and other techniques. Although it is difficult to measure the exact temperature on the wafer surface during deposition, in practice it is possible to reproduce the same heating from run-to-run.

11.7.4 Faceting and Trenching

Faceting and trenching are two effects arising from sputtering that can impact the results of sputter etch processes, as well as the reactive-ion etching processes discussed in Chap. 14. The effect of *faceting* is illustrated in Fig. 11-33a-d, in which the sidewall of a feature is seen to develop an increasingly larger facet as the sputtering continues. (The term *facet* originated as a description of the *small faces* that are cut from the surface of a jewel.) The faceting effect arises from the fact that sputtering yield is greater from surfaces which are inclined at a non-90° angle to the incoming ions (see Fig. 11-10). The facet usually starts on corners (which always have some rounding), but is also more pronounced in bias-sputtering processes due to the increased electric field at sharp corners. The facet is inclined in the direction of the incident angle corresponding to the angle of maximum sputtering yield. The faceting effect can *directly* alter substrate step profiles if an unprotected substrate is sputter etched. On the other hand, as shown in Fig. 11-33a-d, if a protective layer (e.g., photoresist) is used during a sputter etch or RIE process, the facet developed in the masking layer can still be *indirectly* transferred into the etched film, if the etching proceeds long enough that the facet intersects the surface.

The phenomenon of *trenching* is illustrated in Fig. 11-33e. It stems from the enhanced ion flux at the base of a step due to ion reflection off the side of the step, which causes etching to occur more rapidly there. It is discussed further in Sect. 14.6.3 in Chap. 14.

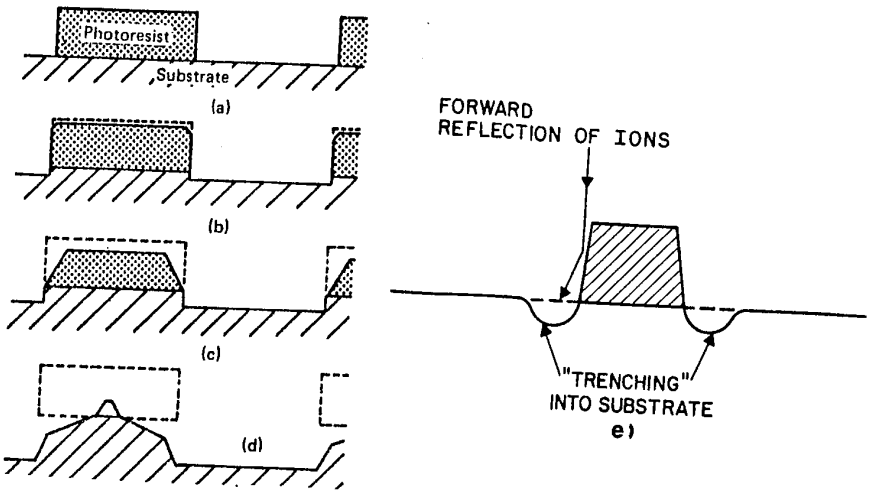


Fig. 11-33 Faceting results from the dependence of sputter etch rate on the angle of incidence of ions striking the surface. (a) prior to etching, (b) initiation of the facet, (c) facet intersects the substrate surface, (d) substrate is exposed and forms its own facet. (e) Trench formation arises from an excess flux of ions resulting from reflection off the sidewall. Copyright, 1983, Bell Telephone Laboratories, Incorporated, reprinted by permission.

11.7.5 Particle Generation in Sputtering Processes

By 1994, for ULSI applications it was required that less than 0.07/cm^2 particles (0.3 μm or smaller) be produced by a sputtering process. Particles arise in sputtering processes primarily from the buildup of barrier-layer material (e.g., TiN or Ti:W) into films on the shields and the target surface itself (i.e., due to re-deposition of sputtered material onto areas of the target surface that subsequently get little erosion). That is, the sputtering shields get coated with highly-stressed, brittle barrier metal films (i.e., TiN). This buildup is continuous and increases as the number of wafers is processed. Ultimately, these films can delaminate, flake off, and shower the substrate with particles. To reduce particulate generation, equipment vendors offer process-specific shields that are designed to improve the adhesion of films that condense and build up on them. The roughening of the shield surface and smoothing of sharp edges on the shields also make films less likely to flake off.[35] The shields must nevertheless be replaced on a regular basis, normally after processing a pre-determined number of wafers.[37] To prevent film buildup on the target surface, magnetron sources that produce full target surface erosion should be used.

11.7.6 Reactive Sputtering

The introduction of reactive gases into the sputtering chamber during the deposition process allows compound films to be deposited by sputtering, in a process termed *reactive sputtering.* The main application of reactive sputtering in ULSI involves the sputter deposition of the barrier film material TiN.[58] Here a target of pure Ti is sputtered in a glow discharge containing a gas mixture of Ar and N_2. By using a pure Ti target (instead of a TiN target formed by hot isostatic pressing of a TiN powder), the target purity is increased and its fabrication cost is lowered. The TiN compounds are formed when sputtered Ti atoms adsorb on the wafer surface and then react with the N gas atoms that impinge on the wafer surface. The plasma provides enough energy to dissociate the nitrogen molecules into atomic nitrogen, which can subsequently interact with Ti to form TiN. The N_2 partial pressure determines the stoichiometry of the deposited film. Stoichiometric TiN is formed when the N_2 partial pressure exceeds a

value that is dependent on sputtering power and total pressure. The typical N_2:Ar ratio of the sputter gas is around 1:1. A typical plot of deposition rate and TiN composition versus N_2:Ar ratio is shown in Fig. 11-34. Films with desired properties may be formed by optimizing the N_2 partial pressure, the total pressure, the sputtering power, and the substrate temperature. The metallic cathode experiences the same environment as the depositing film. If there is enough reactive gas present to react with the metal film on the wafer surface, generally the target surface also reacts. The surface of the target thus becomes a TiN compound, which has a much lower sputter yield (and sputtering in this regime is referred to as the *nitride mode*). As shown in Fig, 11-34, this produces a sharply lower deposition rate (e.g., 150 nm/min at 4 kW in an Endura® on 200-mm wafers).[30] Note that the residual partial pressures of H_2O and O_2 in the sputter chambers should be kept as small as possible to avoid contamination of the sputtered film. Modern sputtering systems whose base pressures approach 10^{-9} torr satisfy this requirement.

11.8 STEP COVERAGE AND VIA/CONTACT HOLE FILLING BY SPUTTERING

In most applications it is desired that thin films maintain a uniform thickness and freedom from cracks or voids. As thin films cross steps that occur on the surface of the underlying substrate, they may suffer unwanted deviations from the ideal, such as thinning or cracking. A measure of how well a film maintains its nominal thickness is expressed by the ratio of the minimum thickness of a film as it crosses a step, t_s, to the nominal thickness of the film on flat regions, t_n (Fig. 11-35). This film property is referred to as the *step coverage* of the film and is expressed as the percentage of the nominal thickness that occurs at the step:

$$\text{Step Coverage (\%)} = (t_s/t_n) \times 100\% \qquad \textbf{(11.17)}$$

Step coverage of 100% is ideal, but each process is normally specified by a lesser minimum value that is acceptable for a given application. It is important, however, that circuit designers understand the actual step coverage value, and derate the maximum allowable current in an interconnect line to reflect the step coverage limitation. The height of the step and the *aspect-ratio* of the feature being covered also determine the expected step coverage. That is, the greater the height of the step, or the aspect ratio (i.e., the height-to-spacing ratio of two adjacent steps), the more difficult it is to cover the step without thinning of the film, and hence the worse

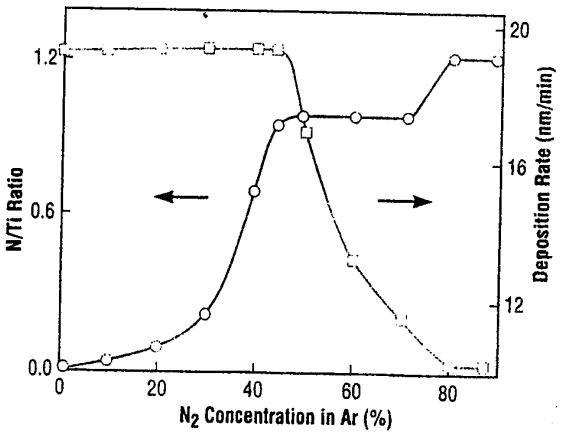


Fig. 11-34 Deposition rate and N:Ti ratio for Ti sputtered in a reactive ambient of N_2:Ar ratios. Reprinted with permission of Solid State Technology, published by PennWell.[38]

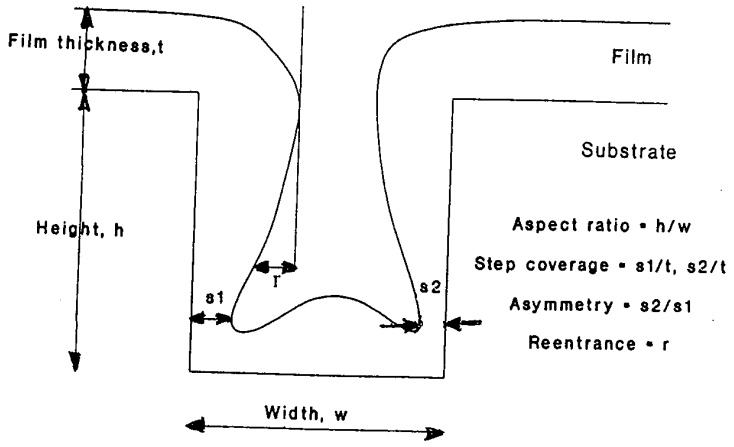


Fig. 11-35 Definition of *step coverage* and related terms.

the expected step coverage. Besides the step height and aspect ratio, the step coverage depends on two other factors which are not controllable by the sputtering process, the *shape* and the *slope* of the step. In general, the smoother the contour and the smaller the slope of the step, the better the expected coverage. Assuming that the four above factors have been fixed as a result of previous process steps, there are then several conditions of the sputtering process that can be adjusted to improve step coverage, and those factors will be discussed.

One technique involves the *sputter etching* of the substrate surface even as sputter deposition (or deposition by CVD), is simultaneously taking place. The sputter etching mechanism exploits the phenomenon of the faceting effect described in Sect. 11.7.4. That is, as a film is deposited into a trench or a contact hole (or via) the material deposited on the top corner is also sputter etched more rapidly than on regions of the substrate that face the ions striking the substrate in a perpendicular direction (i.e., the flat regions shown in Fig. 11-35). Thus, the space at the top of the trench or contact hole is kept from being closed off, and more material can continue to enter the trench or contact hole opening. This assists in the filling of the opening. (More details of how this technique is implemented are given in Chap. 15, Sect. 15.6.)

Heating the substrate has also been generally shown to improve step coverage during sputter deposition. This effect occurs because the migration of the surface atoms of deposited films increases with substrate temperature (see Figs. 11-36a & b). As a result, the atoms tend to move

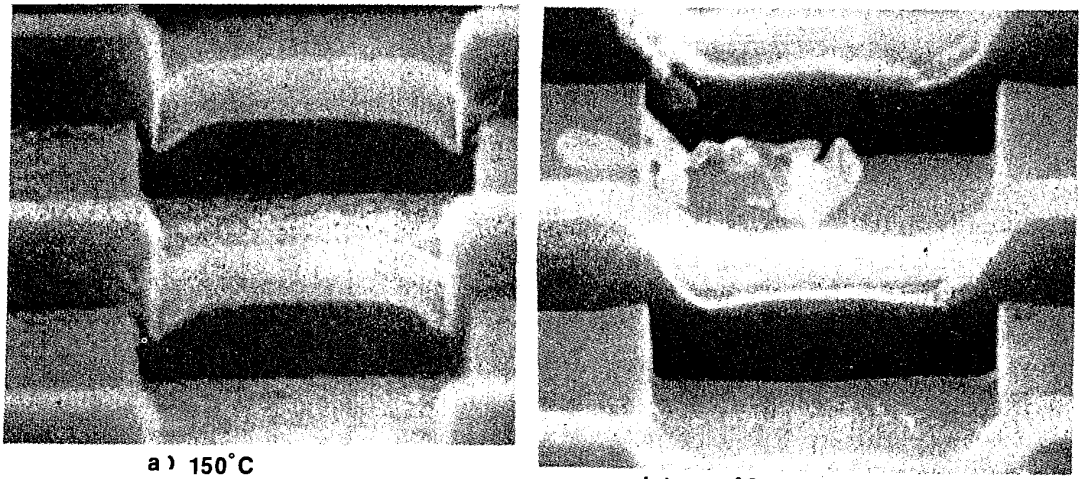


Fig. 11-36 Effect of heating the wafer during sputtering on step coverage of the deposited Al film.

toward regions where the least amount of material has been deposited, thereby averaging out the film thickness (i.e., the concentration gradient of material favors diffusion toward less populated locations). For aluminum films, the substrates must be heated to > 250°C before significant coverage improvements are observed. The optimum temperature and deposition parameters depend on the particular step topography, and must be determined for each process. More information of filling contact holes and vias with Al by sputtering is provided in Sect. 15.6.3.2.

A very important point to note when considering the deposition of a film into a trench or a contact hole is the following. Assume a contact hole has the shape of a cube with side **a**, and the top face open (representing a contact hole with an aspect ratio of 1). In this case there are five inner surfaces of the cube that must be covered by material which would otherwise cover area $\mathbf{a}^2$ on the flat region (see Fig. 11-37). However, the total area to be covered now is $5\mathbf{a}^2$. Thus, if the material is evenly distributed onto all five sides of the contact hole, the film thickness will only be 20% as thick as that on the flat areas outside of the contact hole. Obviously, if the width of the hole remains constant and its depth is increased, the aspect ratio will increase (e.g., if the contact-hole opening-size remained at **a**, but the depth became 2**a**, the aspect ratio will increase to 2, and the area that will have be covered by the material of $\mathbf{a}^2$ will now be $10\mathbf{a}^2$). This is one of the key reasons it becomes increasingly difficult to fill holes as their aspect ratio is increased. Furthermore, if the film on the upper flat regions must be etched away after the hole has been filled, a much thicker film has to be removed (which makes the etch process task significantly more time consuming and difficult).

11.8.1 Sputter Deposition of Barrier Layer Films into Contact Holes and Vias

In the previous discussion it was noted that it becomes progressively harder to fill contact holes as the aspect ratio increases because the material that enters the hole must be shared by the increasingly large total area of the contact sidewalls and bottom. However, another related problem makes the task of filling high aspect ratio holes even more difficult. The discussion of this problem begins by recalling that the sputtered atoms undergo collisions as they are transported in the gas phase in the space between the target and the substrate. In addition, the sputtered atoms leave the target surface from each point at which sputtering occurs in directions that roughly exhibit a cosine distribution (Fig. 11-38).[40] As a result of these two mechanisms, the sputtered atoms will arrive at the upper corners of the trench or contact hole on the wafer from a larger arrival angle than they will at locations within the hole. Thus, the film deposition

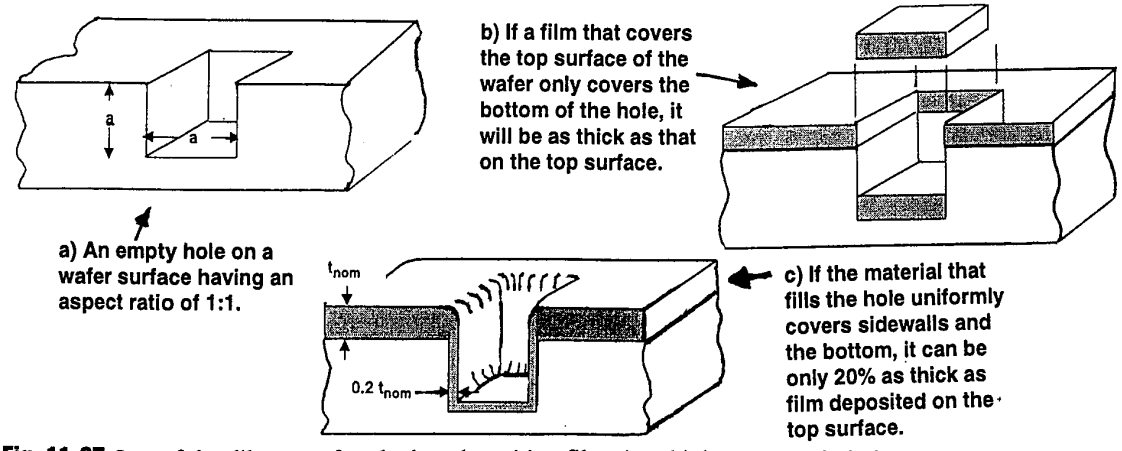


Fig. 11-37 One of the dilemmas faced when depositing films into high-aspect-ratio holes.

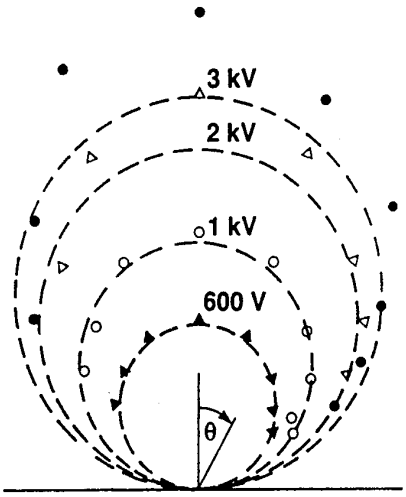


Fig. 11-38 Distributions of Fe atoms sputtered from a target by normal incidence Ar^+ ions with energies of 600, 1000, 2000, and 3000 eV. The dashed lines show best fit cos θ distributions, where θ is the angle of ejection.[40]

rate at the top corners will be higher than along the hole sidewalls or bottom. As a result, the top opening of the hole will begin to close off (i.e., an *overhang* will start to form), as shown in Figs. 11-39a and 11-39b (and in Fig. 6-53). Thus, as time goes on, for this case the deposition rate onto the sidewalls and bottom will continually decrease. If the overhang gets thick enough, the hole will close entirely, leaving a void or keyhole.

Even if the deposition does not continue long enough to completely close off the hole, the overhang will shadow the bottom corners of the contact hole, where the arrival angle for sputtered atoms is already the smallest (see Fig. 11-39b). Thus, the deposited film will be thinnest at the bottom corners of the hole. Since a barrier film must have a minimum adequate thickness along the entire bottom of the hole in order to function effectively, such thin spots at the corners can jeopardize the effectiveness of the barrier film (e.g., TiN).

Various methods have been explored to overcome this thinning of the TiN film at the bottom corners of contact holes, including the use of CVD TiN, as discussed in Chap. 6. However, modifications to the sputtering process have also been pursued, and methods of improving Ti/TiN sputtering to produce more robust barrier layers will be discussed. The first such technique involves the implementation of collimated sputtering of the Ti and TiN layers. This

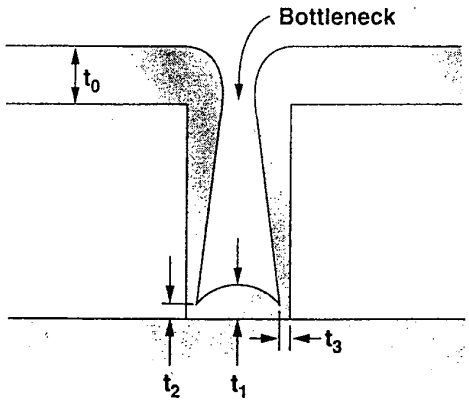


Fig. 11-39 Deposition of a sputtered film into a contact hole or via having a high aspect ratio. Development of a bottleneck due to the buildup of an overhang on the top corner is shown.

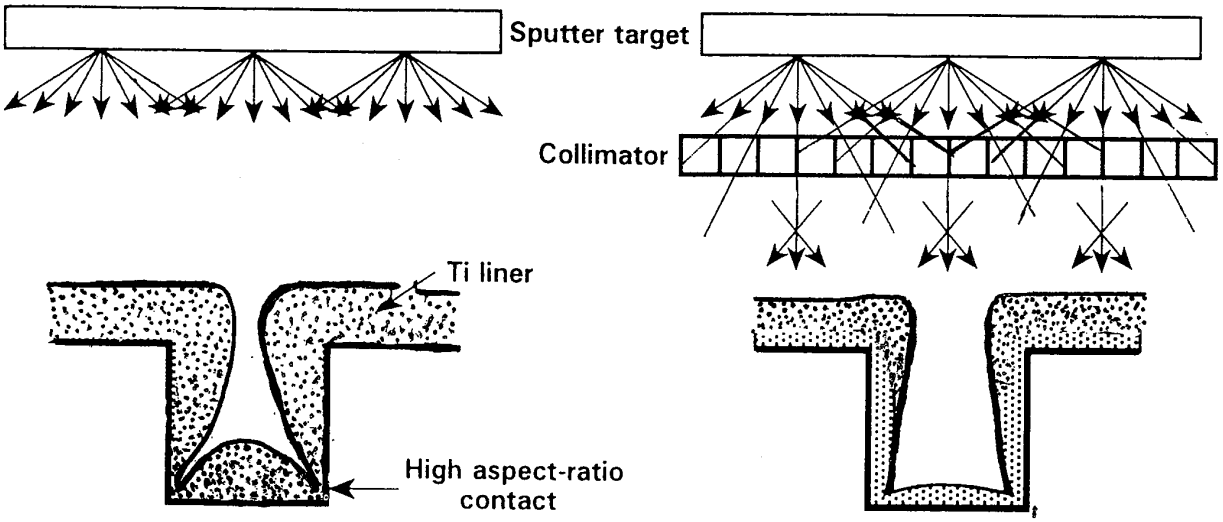


Fig. 11-40 Insertion of a collimator in the space between the target and the wafer, and its effect on the deposition of a sputtered film into a via.

concept has been implemented in ULSI sputtering systems using two different ways, namely by: 1) insertion of a *collimator* (i.e., a honeycomb baffle) into the space between the target and the wafer to trap obliquely-directed vapor (Fig. 11-40); and 2) carrying out the sputter process with a *long-throw-sputtering configuration*, in which the target and wafer are separated by a distance 2–4 times longer than in a conventional magnetron source (and in which the plasma is operated at a lower pressure). In the long-throw method, the sputtered atoms undergo few (or no) collisions in the low-pressure plasma, and travel in long-straight lines to the wafer. Thus, most of them arrive at angles close to perpendicular to the wafer surface. In both collimated sputtering methods a nearly normal incident flux of sputtered metal atoms strike the wafer surface, enabling effective deposition into hard-to-reach regions at the bottom of high-aspect-ratio features.

11.8.1.1 Sputter Deposition with Collimators: The use of collimators to improve sidewall and bottom coverage of high-aspect-ratio holes was first proposed in 1992.[41,42] A collimator is a metal honeycomb structure (typically made from Al), with an array of circular or hexagonal holes, and is electrically grounded (Fig. 11-41). Neutral atoms sputtered off of the target at a high angle are intercepted by the collimator, where they adhere to the sidewalls (vanes). As a result, only atoms ejected from the target at small angles are able to reach the wafer. These atoms have a much higher chance of being deposited on the bottom of the contact hole than those ejected at higher angles. The aspect ratio of the collimator openings can also be varied.

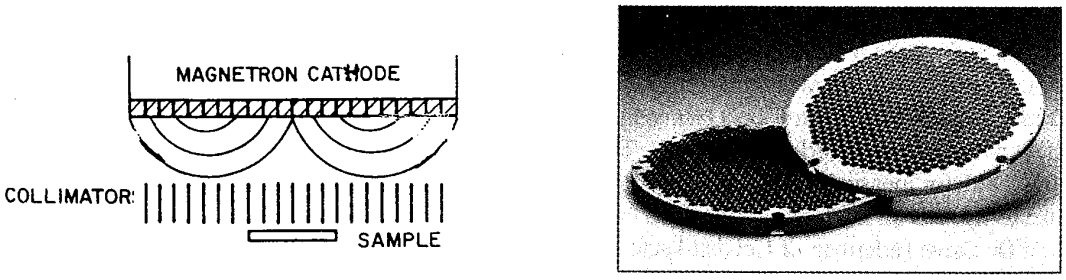


Fig. 11-41 Photograph of a collimator.

As this aspect ratio is increased (i.e., the collimator is made thicker for the same size opening), the collimation increases and the bottom coverage of the contact holes is improved. However, the deposition rate of the film on the wafer also decreases as more of the sputtered flux is intercepted. Typical collimators employ aspect ratios ranging between 1:1–3:1.[43] Figure 11-42a shows the bottom coverage of collimated sputtering compared to conventional (non-collimated) versus *contact-hole aspect ratio* for a 1:1 *collimator aspect ratio*.[42] Figure 11-42b shows a plot of the transmission of the collimator versus the collimator sputtering aspect ratio. The increase in bottom coverage is significant, and goes from <10% to > 25% for a contact hole with an aspect ratio of 2:1 However, this gain is obtained at the cost of a slower deposition rate. The sidewall coverage of the contact hole is also degraded by use of a collimator. Thus, use of collimated sputtering is counterproductive when an attempt is made to completely fill a contact hole (e.g., when trying to fill contact holes with Al), and hence is not used for this purpose. But for contact and barrier layer films (such as Ti and TiN), where bottom coverage is critical, collimation provides a simple extension of an existing process that is easily integrated into IC sputter technology. Collimated sputtering was therefore successfully adopted for 0.35 μm and 0.25 μm CMOS generations.

The use of a collimator does have some significant drawbacks. First, as prevoiusly noted, collimators cause a reduction in the sputter deposition rate. For fresh collimators with 1:1 aspect ratio openings, approximately 70% of the sputtered material is intercepted, but if the aspect ratio is increased to 1.5:1, nearly 90% is intercepted.[44] The percentage of metal intercepted also increases as more wafers are sputtered, since the openings of the collimator continuously close up as the film deposited on the collimator sidewalls gets thicker.[45] Since the deposition rate thus varies with time, this effect must be controlled in the manufacturing line to maintain a uniform barrier layer thickness from wafer-to-wafer as the collimator "ages" (e.g., by incrementally increasing the power dissipation during the sputter process to keep a constant deposition rate). Additionally, the collimator clogging has a radial dependence, with the center holes of the collimator clogging faster than edge holes, since the center holes are exposed to more flux from the target.[45] The high percentage of metal intercepted also increases the cost of the barrier film deposition step because the throughput of the sputtering tool and the target utilization are both significantly reduced. In addition, the collimator must eventually be replaced (e.g., once transmission has decreased to 50% of its initial value). The expense of the collimator, and the attendant tool downtime further raises the cost. The presence of the collimator also increases the potential for particle generation, as the brittle barrier layer films deposited on the collimator may eventually begin to flake off and shower the wafer below. (However, the collimator may also act to filter out particles suspended in the plasma, and this is an added benefit.) Controlling the temperature of the collimator is one method used to reduce particle generation (as well as treating the collimator surface to improve film adhesion and periodic collimator replacement). Finally, shadowing of the wafer by the vanes of the collimator produces a substantial ripple in the deposited film's thickness.[50] These drawbacks have caused other approaches to be pursued, including CVD of TiN (see Chap. 6) and ionized magnetron sputtering (see Sect. 11.8.3). Collimators have been offered in sputtering tools sold by Applied Materials, Inc, and Varian (Novellus Systems, Inc. has since acquired the Varian PVD business unit).

11.8.1.2 Long-Throw Collimated Sputtering: Another collimated sputtering technique that does not employ a collimator plate has also been pursued, namely *long-throw collimated sputtering*. In this approach, a much longer distance exists between the target and wafer than in a

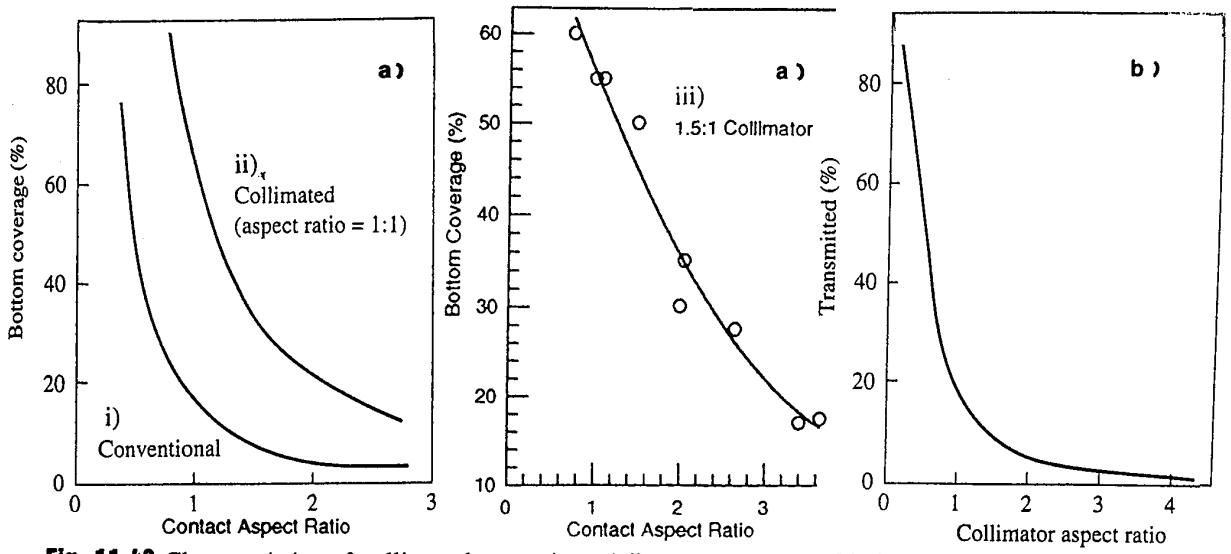


Fig. 11-42 Characteristics of collimated sputtering: a) Bottom coverage of holes versus the hole aspect ratio, for i) conventional magnetron sputtering, ii) collimated sputtering with an aspect ratio of 1:1, iii) collimated sputtering with an aspect ratio f 1.5:1[52] (© IEEE 1996); b) Deposition rate of Ti and TiN as a function of the aspect ratio of the collimator.[42] Reprinted with permission of VMIC.

conventional magnetron sputtering system (e.g., from 25–30 cm vs. about 5 cm). The plasma is also operated at a lower pressure (e.g., 0.1 mtorr[49]). Thus, the sputtered atoms undergo few (if any) collisions and arrive at the wafer by a straight-line (line-of-sight) trajectory after being ejec-ted from the target. The atoms ejected at large angles collide with and condense on the chamber walls (or shields), and only those ejected at small angles arrive at the wafer surface. In essence, the long-throw configuration behaves like a collimator with a single opening (i.e., the chamber itself). The long-throw systems provide a constant sputtering rate over the life of the target (e.g., only a drop of 0.5% in deposition rate occurs over 1000 wafers).[43] However, particle generation from the larger wall surfaces, asymmetry of the step coverage in holes away from the center of the wafer, and the difficulty of maintaining a plasma at lower pressures are other issues of concern. Long-throw sputtering systems have been marketed by ULVAC and Electrotech for depositing Ti and TiN barrier layer films. For 0.25 μm holes with aspect ratios of 3.0, ULVAC reported that bottom coverages of 40% can be routinely achieved.[43]

11.8.1.3 Ionized Magnetron Sputter Deposition: While the two collimated sputtering techniques described above produce adequate barrier layer films for contact holes and vias for 0.35 μm (and some 0.25 μm) CMOS technologies, they appear to fall short for other 0.25 μm (and smaller) contact holes and vias when their aspect ratio exceeds about 3:1. As shown in Fig. 11-42a, the bottom coverage in a hole with an aspect ratio of 3:1 is only 15%, even when a collimator aspect ratio of 1.5: is used.[52] An alternative to collimated sputtering, *ionized magnetron sputtering*, or *I-PVD* (also termed *ion metal plasma* [*IMP*] by Applied Materials), has been developed to deposit the next generation of ULSI barrier layers with PVD. In this method, a large fraction of the neutral sputtered flux (i.e., Ti atoms) are ionized while in flight between the target and the wafer. A bias on the wafer then results in an electrical acceleration of these ion species to the wafer surface with a controlled energy and direction (i.e., perpendicular to the wafer surface).[47,48,51]

Figure 11-43 shows how the large existing installed base of sputtering chambers could be retrofitted with an rf-coil that produces the secondary plasma needed to implement an ionized

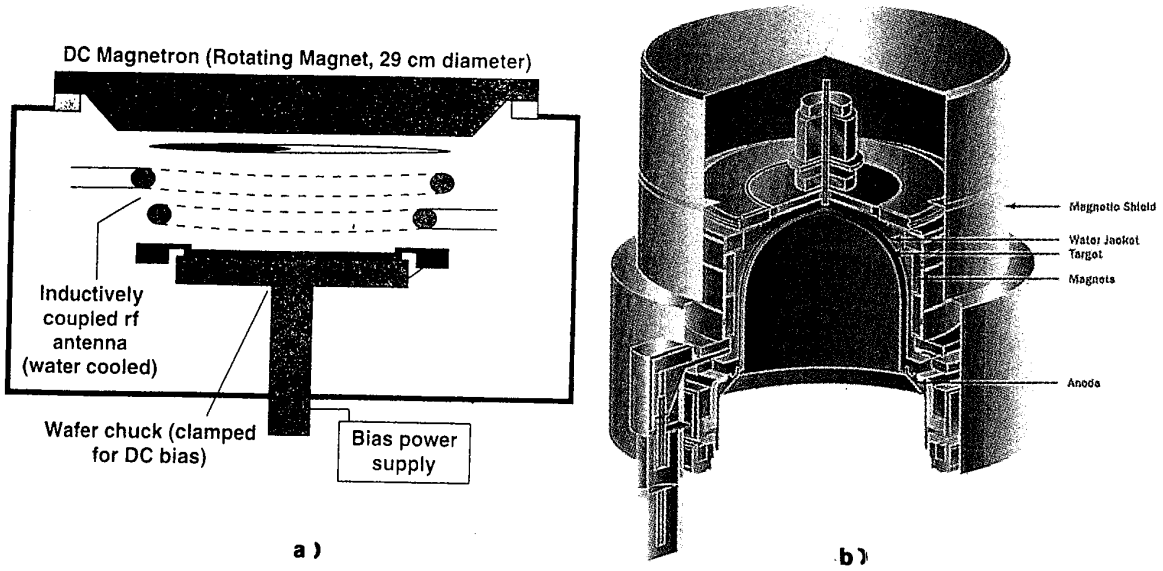


Fig. 11-43 a) Schematic of a magnetron sputter chamber fitted with an rf-coil which allows an ion magnetron sputtering process to be implemented. b) Cutaway drawing of the hollow-cathode source used in the Novellus Systems' INOVA® sputtering system. Courtesy of Novellus Systems, Inc.

magnetron sputter deposition process. The original magnetron source (powered with a dc power supply) remains in place so that neutral metal atoms are sputtered as before from the target. (The magnetron produces a plasma and dark space that is confined near the target surface.) In place of a collimator (but with virtually the same system geometry), an rf-powered (and water-cooled) coil is located within the chamber (Fig. 11-43). This coil inductively couples the energy from a second 13.56 MHz rf-power supply into a secondary plasma downstream of the magnetron plasma. By increasing the pressure in the chamber to 20–40 mtorr (up from the 5 mtorr of a conventional magnetron sputtering process), and by increasing the depth of the secondary plasma to 8–10 cm, the collisions of the sputtered atoms and their transport time through the gas phase is increased. This longer transport time, and the fact that the ionization potential of metal atoms (Ti) is much lower than that of Ar atoms, allows a much higher fraction of the metal atoms to be ionized by this secondary plasma (typically 50-70% of the Ti atoms are ionized). A medium-density secondary plasma is thus created in this region ($5x10^{11}$–10^{12} ions/cm^3). These ionized metal atoms can then be accelerated by a third negatively-biased dc or rf-power supply toward the wafer surface (at voltages ranging from a few to several hundred volts, depending on the specific application). The ion flux onto the wafer results in power dissipations ranging from about 25 watts to several hundred watts, which calls for some degree of wafer cooling.[51]

The bottom coverage of high aspect ratio holes is better than that observed in collimated sputtering. That is, 40–80% bottom coverage has been achieved on 0.18 μm holes with aspect ratios of up to 8:1.[52] Note that while the bottom coverage is high, the sidewall coverage is much thinner (although the film on the sidewall is continuous). However, for liner films, this is not a problem. Several other advantages of the IMP process should be noted. First, much better target utilization is obtained than with collimated sputtering, as almost all of the sputtered atoms arrive at the wafer surface (instead of being intercepted by the collimator). Sputtering tool throughputs are therefore also increased. Second, the high flux of metal atoms directed perpendicularly onto the wafer surface appears to re-sputter the atoms that would otherwise form an overhang at the top corners of the contact hole. This keeps the top opening of the contact hole from closing up

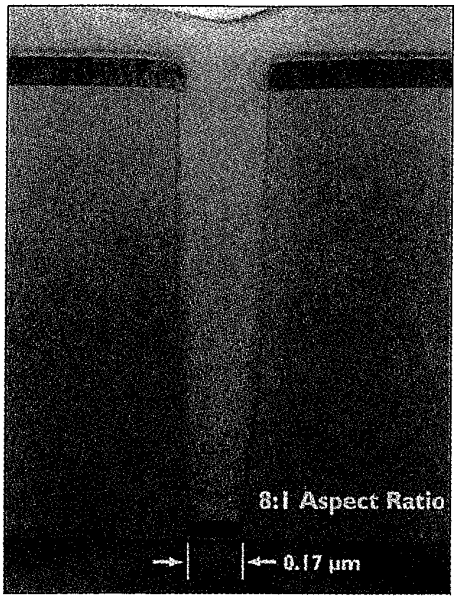


Fig. 11-44 SEM micrograph showing a hole with an aspect ratio of 8:1 deposited with a TiN layer using Applied Materials ionized metal plasma (IMP) process. Courtesy of Applied Materials.

and maintains the thickness of the film on the hole sidewalls at a small value. Figure 11-44 shows the coverage of the sidewall and bottom of a high-aspect ratio hole with IMP.

The *hollow-cathode source* used in the Novellus INOVA sputtering system (Fig. 11-43b) employs a similar technique to ionize sputtered atoms as they are transported to the substrate. A uniform plasma throughout the volume of the hollow cathode provides sputtering from a large fraction of the area of the target surface. This provides the benefit of high target utilization. A bias on the wafer chuck accelerates the ionized sputter atoms to the wafer in a direction perpendicular to its surface, allowing good coverage of high aspect ration holes.

11.9 FUTURE TRENDS IN SPUTTER DEPOSITION PROCESSES

Sputter deposition may largely be replaced by other deposition technologies such as CVD, or electroplating in some applications. However, many industry observers predict that sputter deposition will continue to be an important metal deposition process in ULSI fabrication for quite a while, and that novel solutions to its limitations will be developed to extend its life. For example, in the damascene process used to fabricate Cu interconnect structures (see Chap. 15), electroplated Cu is emerging as the dominant deposition process for depositing the bulk Cu layer. The role of sputtering has shifted to depositing the critical barrier layer (e.g., Ta, TaN, or TiN) and the Cu seed layer (sputtered Cu). For Al-based metallization, such novel approaches as ionized magnetron sputtering will likely be used to create barrier/liner layers in high-aspect-ratio contacts and vias.[53] The deposition of Ti and Co for salicide processes is also likely to continue to be performed by sputtering.

For sputtering to be used in future generations of IC technology, contamination and defect densities resulting from these processes will have to be reduced. Many defects are traceable to targets (arising from flaws in the target surface finish and other target material defects as surface abnormalities and target material porosity). All of these target defects can cause arcing in the plasma, which in turn produces large die-killing defects. Flaking of film material deposited on the shielding (or re-deposited on the target) will also have to be further reduced. Target vendors will have to incorporate internal raw material refining and advanced fabrication techniques to

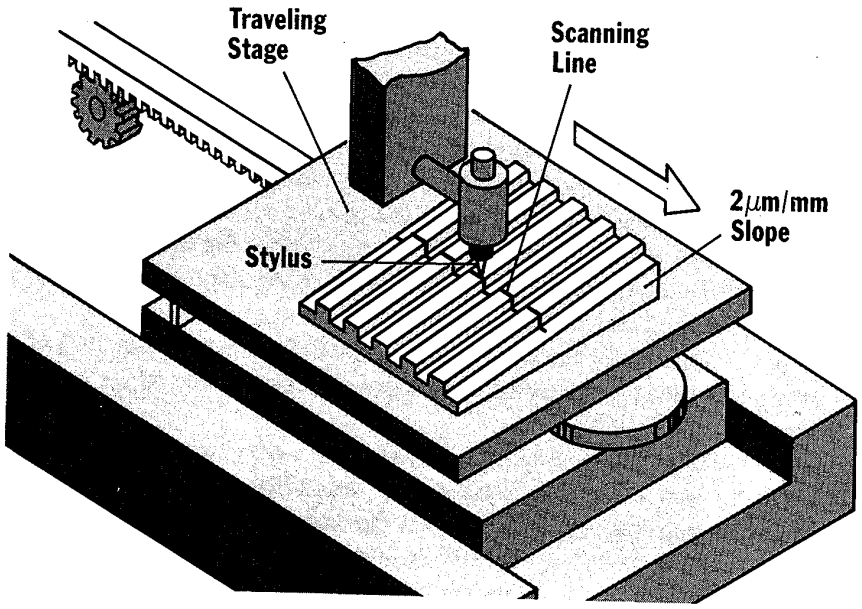


Fig. 11-45 Schematic drawing of a surface profilometer.

meet these requirements. Furthermore, since some materials as tantalum are rare, it will be necessary to be very conservative in terms of reuse and recycling of spent Ta targets, a practice that has not been applied to Al and Ti targets. Better target design has also recently increased target utilization from 30 to 50 percent, and work is continuing to improve this percentage.

11.10 METAL FILM THICKNESS MEASUREMENT

Metal films are generally deposited with sputter deposition without *in situ* monitoring. Instead, the film thickness is measuremed after the sputtering process is complete. Post-deposition measurements of metal film thickness (on flat areas of a wafer) are most commonly performed in one of three ways: 1) directly, with a surface profiling device (known as a *stylus profilometer*); 2) indirectly, by electrical measurement of the sheet resistivity of blanket-deposited layers;[54,55] and 3) indirectly, by measurement of short light induced sound pulses (picoseconds long) in opaque metal films (*photoacoustics*).[54,56] Examination of the thickness variation of metal films (and films of other materials) over topography on a wafer surface is usually accomplished by cleaving the wafer and inspecting the film cross-section with electron microscopes. For films with thicknesses >100 nm, a scanning-electron microscope (SEM) is adequate, but for thinner films transmission electron microscopy is required.

In the technique that uses the stylus profilometer, a step in the deposited film is first created, either by masking during deposition, or by a post-deposition etch process. Next, the profiling instrument draws a fine stylus along the surface containing the stepped film (Fig. 11-45). Whenever the stylus encounters a step, a signal variation (based on a differential-capacitance or differential-inductance technique), yields an indication of the step height. The information is displayed on a chart recorder or CRT. Three examples of commercial profiler instruments are the Veeco Dektak models, and the KLA-Tencor P-20 and HRP-200 profilers. Films less than 100 nm thick can be measured with such instruments, but films much thinner than 100 nm are normally measured with sheet resistivity techniques. Thin film measurements are more difficult with the stylus-type instruments because of vibration, surface roughness, and the criticality of

leveling the profilometer. Some recent surface profilers use atomic-force microscopy (AFM). Periodic calibration of stylus profilometers is also required, using calibration standards traceable to the National Bureau of Standards. If the instrument is to be used to measure films whose thickness differ greatly from the calibration standard, a reference can be created by etching a step in thermally grown SiO_2, and accurately measuring its height with an ellipsometer.

The traditional technique for measuring metal film thickness is an indirect method. The sheet resistance of the film is measured with a four-point probe, and the film thickness then inferred from a knowledge of the material's bulk resistivity. Details of this approach are discussed in more detail in Chap. 10. Some disadvantages to using this method exist. First, it requires the production of a monitor wafer, and this can consume as much as 10 percent of the production capacity of a multi-chamber cluster tool. Second, for very thin layers, the thickness of the film itself begins to influence the bulk electrical conductivity of the material. Then the simple relationship between film thickness and bulk resistivity breaks down. The thinnest films that can be accurately measured depend upon the metal itself. For Al, Ti, and Cu respectively, these minimums are 30–40 nm, 10–20 nm, and 100 nm.

The photoacoustic measurement technique, is a non-contact, non-destructive technique that employs a measuring spot size smaller than 8 μm, allowing it to be used on a test structure that is < 20 μm in diameter. Such test structures can be included in the scribe lines of product wafers, eliminating the need for monitor wafers. In one implementation, a picosecond sound pulse is launched vertically into the metal film. When reflected back, a slight change in the reflectivity occurs at the reflecting interface. A probe pulse, split off from the excitation pulse, is delayed, and the reflecting light from the probing pulse is measured. The time it takes for the reflecting pulse to come back (echo) to the top surface of the metal film is used to deduce thickness. This method can also measure metal stacks of up to six films, making the technique suitable for metal cluster tools. Rudolf Technologies and Active Impulse Systems are vendors that offer metrology tools employing the photoacoustic measurement technique.

REFERENCES

1. D. Pramanik and A.N. Saxena, "VLSI Metallization Using Aluminum and Its Alloys, pt. I," *Solid State Technol.*, Jan. 1983, p. 127, and pt. II, Mar. 1983, p. 131, and "Aluminum Metallization for ULSI," *Solid State Technology,* March 1990, p. 73.
2. J.L. Vossen, Ed., "Bibliography on Metallization Materials and Techniques for Silicon Devices," Thin Film Div., *American Vacuum Society*, 1974-82.
3. M. Hansen & A. Anderko, *Constitution of Binary Alloys,* McGraw-Hill, New York, 1958.
4. B. Chapman, *Glow Discharge Processes*, John Wiley & Sons, New York, 1980.
5. *The Book of Basics*, 3rd Ed., Materials Research Corp., no date, Orangeburg, N.Y.
6. J.L. Vossen and J.J. Cuomo, "Glow Discharge Sputter Deposition", in *Thin Film Processes*, Eds. J.L. Vossen and W. Kern, Academic Press, New York, 1978, Chap. II-1, p. 11-73.
7. W.D. Davis and T.A. VanderSlice, *Phys. Rev.*, **131**, 219 (1963).
8. L. Maissel, "Application of Sputtering to Deposition of Films", in L. Maissel and R. Glang, Eds., *Handbook of Thin Film Technology*, McGraw-Hill, New York, 1970, Chap. 4.
9. G.K. Wehner and G.S. Anderson, "The Nature of Physical Sputtering", in *Handbook of Thin Films*, Eds. L.I. Maissel and R. Glang, Chap. 3, McGraw-Hill, New York, 1970.
10. O. Almen and G. Bruce, *Nuclear Instrumentation Methods,* **11**, 257, (1961).
11. F. d'Heurle, L. Berenbaum, and R.A. Rosenburg, *Trans. Met. Soc.* (AIME) **242**, 502, (1968).
13. H.R. Koenig and L.I. Maissel, *IBM J. Res. Develop.*, **14**, 168 (1970).
14. H. S. Butler and G.S. Kino, *Phys. Fluids,* **6**, 1346 (1963).

15. R.K. Waits, "Planar Magnetron Sputtering," in J.L. Vassen and W. Kern, Eds., *Thin Film Processes*, Academic Press, New York, 1978, Chap. II-4, p. 131.
16. S.M. Rossnagel and H.R. Kaufman, *J. Vac. Sci. Technology,* **A5**, 88 (1987).
17. J.A. Thornton and A.S. Penfold, "Cylindrical Magnetron Sputtering," in *Thin Film Processes*, Eds. J.L. Vossen and W. Kern, Academic Press, New York, 1978, Chap. II-1, p. 11-73.
18. S.M. Rossnagel and H.R. Kaufman, *J. Vac. Sci. Technol.,* **A5**, 2276 (1987).
19. D.L. Smith, *Thin-Film Deposition: Principles and Practices,* McGraw-Hill, New York, 1995.
20. K. Skidmore, "Sputtering for Deposition & Etch," *Semiconductor International,* May 1988, p. 74
21. S. Rossnagel, "Glow Discharge Plasmas and Sources for Etching and Deposition," in *Thin Film Processes II,* Eds. J.L. Vossen and W. Kern, Academic Press, San Diego, 1991, Ch. II-1, p. 11.
22. F.T. Turner and D.J. Harra, "Advances in Cassette-to-Cassette Sputtercoating Systems," *Solid State Technology,* July 1983, p. 115.
23. *Basic Elements of the Semiconductor Capital Equipment Industry,* NationsBanc Montgomery Securities, San Francisco, CA, September 1998.
24. "The Endura 5500's PVD Metallization Process," *Semiconductor International,* 1990, p. 30.
25. E. Keller *et al.*, "Manufacturing Evaluation of an Endura 5500 PVD Multichamber System," *Solid State Technology*, February 1992, p. 71.
26. G. Birkmaier, A. Tepman, and H. Gilboa, "Ultrahigh Vacuum in Production Applications," *Semiconductor International*, April 1991, p. 108.
27. J. Klema, R. Pyle and E. Domangue, "Reliability Implications of N_2 Contamination During Deposition of Sputtered Al/Si Films," *Proceedings Intl. Rel. Phys. Symposium*, IEEE, New York, 1984, p. 1.
28. D. Friede and P. Lessard, "Improve Your Sputter Process by Better Water Vapor Pumping," *Semiconductor International,* August 1994, p. 95.
29. S. Lakeman, "Contamination Control in PVD Cluster Tools Through Mass Spectroscopy," *Semiconductor International,* October 1996, p. 114.
30. S.M. Rossnagel, "High-Vacuum-Based Processes: Sputtering," in D. Hoffman, B. Singh, and J.H. Thomas, Eds. *Handbook of Vacuum Science and Technology,* Academic Press, New York, 1998.
31. W.D. Westwood, "Physical Vapor Deposition," in *Microelectronic Materials and Processes*, Ed. R.A. Levy, Klewer Academic Publishers, Norwell MA, 1989, Chap. 4, p. 167-8.
32. M. Hill, "Magnetron Sputtered Titanium-Tungsten Films," *Solid State Technol.,* Jan.1980, p. 53.
33. S.M. Rossnagel, I. Yang, and J.J. Cuomo, "Compositional Changes During Magnetron Sputtering of Alloys," *Thin Solid Films*, **199**, 59 (1991).
34. W. Class and R. Hieronymi, "Measurement and Sources of Substrate Heat Flux Encountered with Magnetron Sputtering," *Solid State Technology,* December 1982, p. 55.
35. "Sputtering Tools: Still on Target?" *Semiconductor International,* August 1992, p. 45.
36. R.K. Waits, "Monitoring Residual and Process Gases in PVD Processes: The Importance of Sensitivity," *Microcontamination*, June 1997, p. 81.
37. "Tech Trends: Semiconductor Sputterers," *Microcontamination*, May 1991, p. 48.
38. D. Pramanik and V. Jain, "Barrier Metals for ULSI: Deposition and Manufacturing," *Solid State Technology*, January 1993, p. 73.
39. S. Berg *et al.*, "Effect of Contamination Reactions on the Composition and Mechanical Properties of Magnetron Sputtered TiN Coatings," *Surface Coatings Technology*, **49**, 127 (1992).
40. Y. Matsuda*et al., Jpn. J. Appl. Phys.*, **25**, 8 (1986).
41. S.M. Rossnagel, D. Mikalsen, H. Kinoshita, and J.Cuomo, "Collimated Magnetron Sputter Deposition," *J. Vac. Sci. Technol.* **A-9**, 261 (1991).
42. R.V. Joshi and S. Brodsky, "Collimated Sputtering of Ti/TiN Liners Into Sub-Half Micron High Aspect Ratio Contacts/Lines," *Proc. VMIC,* 1992, p. 253.
43. P. Burggraaf, "Straightening Out Sputter Deposition," *Semiconductor Internatl.,* Aug. '95, p. 69.
44. F.H. Baumann *et al.*,, "Application and Constraints of Collimated Ti/TiN Sputtering for 0.25 μm and Sub-0.25 μm Contact Structures," ," *Proc. VMIC,* 1993, p. 412.

45. D.S. Bang *et al.*, "Dynamic Modeling of Collimator Clogging in Physical Vapor Deposition Systems," *Proc. 1994 VMIC Conference*, p. 554.
46. B.C. Butler *et al.*, "Enhanced Bottom Coverage of Sub-Micron Contact Holes Using a Novel Hi-Fill Ti/TiN Sputter Process," *Advanced Metallization for ULSI in 1994*, p. 277.
47. S. Rossnagel and J. Hopwood, *Applied Physics Letters*, **63**, 3285 (1993).
48. S. Rossnagel and J. Hopwood, "Metal Ion Deposition from Ionized Magnetron Sputtering Discharges," *J. Vac. Sci Technol.*, **B12**, 449 (1994).
49. *ibid.*, Ref. 21, p. 70-71.
50. C. Sorlic, "Advanced Process Simulation of Metal Film Deposition," *Sol. State Technol* June '95 p. 101.
51. S.M. Rossnagel, "Ionized Magnetron Sputtering for Lining and Filling Trenches and Vias," *Semiconductor International*, February 1996, p. 99.
52. G.A. Dixit *et al.*, "Ion Metal Plasma Deposited Titanium Liners for 0.25 μm/0.18 μm Multi-level Interconnections," *Tech Dig. IEDM*, 1996, p. 357.
53. A.E. Braun, "Sputtering Targets Adapt to New Materials and Shrinking Architectures," *Semiconductor International*, June 1998, p. 127.
54. R. De Jule, "Advances in Thin Film Measurement," *Semicond. Internat.* , May 1998, p. 52.
55. S.C.P. Lim and D. Ridley, "An Overview of Thickness Measurement Techniques for Metallic Thin Films", *Solid State Technology*, February 1983, p. 99.
56. C.J. Mrath *et al.*, "Ultrasonic Multilayer Metal Film Metrology," *Solid State Technol*, June 1997, p. 85.

PROBLEMS

1. Calculate the mean free path of a sputtered atom in an Ar glow discharge at 10 mtorr pressure. If the cathode-to-substrate spacing is 5 cm, how many gas-phase collisions is a sputtered atom likely to encounter as it traverses this distance?

2. Aluminum is bombarded by 1000 eV Ar ions. (a) What is the maximum energy that can be transferred to the target atom. (b) What is the maximum value that the mean energy of a struck target atom can attain if the threshold sputtering energy for the Ar-Si ion-target combination is 13 eV. The atomic number and mass of Al and Ar, respectively are $Z_{Al} = 13$ $M_{al} = 27$, $Z_{Ar} = 18$, $M_{Ar} = 39.44$

3. The sputtering yield of aluminum in Ar is known to be S = 1.2 at an ion energy of 600 eV. The measured rate of deposition in a specific sputtering system is 400 Å / min when the external electron current is $I_0 = 100$ mA. (a) Calculate the ion current, I, using the Townsend equation:

$$I = (I_0 \exp[\alpha d]) / (1 - \gamma[\exp\{\alpha d\} - 1]),$$

where, α = 6 ions /cm and γ, the secondary electron yield = 0.08 electrons per ion, and the inter-electrode spacing is d = 5 cm. (b) Find the new deposition rate when the ion energy is increased to 1000 eV, and S = 1.9, α = 3 ions /cm, and $\gamma = 0.1$ electrons per ion.

4. If a target is made of fused silica with a thickness of 2.5 mm, it has a capacitance of ~1 pF/cm^2. If the current density during rf sputtering of this target is 1 mA /cm^2 and the applied voltage is 1000 V, calculate the time required to charge this capacitance. This roughly represents the time interval during which a glow discharge can remain "on".

5. Explain why the higher secondary electron emission from insulators leads to slower sputter deposition of these materials.

Chapter 16

CMOS PROCESS INTEGRATION

In the first fourteen chapters of this book the focus was on the individual process modules employed in IC fabrication. This chapter will consider how these modules can be combined to define a set of manufacturing steps that becomes a recipe for fabricating monolithic silicon integrated circuits. The task of specifying such a sequence of individual process steps is called *process integration. Complimentary MOS* (or *CMOS*) technology is used to demonstrate some examples of such *process sequences* (or *process flows*). CMOS was chosen for our examples because it is the dominant silicon IC technology of the 1990's. Figure 16-1 shows that about 80 percent of all semiconductor devices and circuits manufactured in 1995 were CMOS ICs. By the end of the 1990s this percentage grew even larger. Since this text is intended to provide an introduction to *mainstream* silicon IC process technology, it is appropriate to use CMOS as the vehicle for illustrating process flows. Since CMOS technology was preceded by other MOS technologies, their evolution will also be briefly reviewed. How and why CMOS became the dominant semiconductor technology will be presented. Note that CMOS is so-named because it uses both *p*- and *n*-type MOS transistors in its circuits. Figure 16-2 depicts a CMOS inverter.

In the examples of CMOS process integration two CMOS process flows will be described. The first is a generic process flow for the generations of CMOS logic ICs extending from 1.2–0.5 μm. The second pertains to CMOS ICs having minimum feature sizes of 0.25 μm and smaller. For such *deep-submicron CMOS technologies* (i.e., $\leq$ 0.25 μm), several major changes occurred in the process flow and device architecture. Hence, the second process flow differs significantly from the first. Some of new processes used are: high-energy implanted retrograde wells, shallow-trench isolation, dual-doped polySi, $TiSi_2$ and $CoSi_2$ salicides, chemical-mechanical polishing, and formation of interconnects with damascene processing.

The purpose of this chapter is to present examples of how Si processes are integrated, resulting in the creation of fabrication sequences for CMOS ICs. A discussion of process integration involves the relationships between process technology and device physics. However, the operation of MOS devices is beyond the scope of this text, and is not presented here. Readers interested in more information on device physics should consult such texts as those by Chen,[1] Muller and Kamins,[2] and Tsividis[3] (as well as Vol. 3 of this series).

16.1 INTRODUCTION TO CMOS TECHNOLOGY

Integrated circuits built with MOS transistors were introduced in the early 1960s, and these were fabricated using PMOS technology. However, it was known that since electrons have higher mobility than holes, ICs fabricated with NMOS would have better performance than those made with PMOS. By the 1970s, the technological barriers that blocked NMOS fabrication were overcome (see Vol. 2, Chap. 5 for details), and NMOS became the dominant technology for MOS IC fabrication. NMOS remained the main MOS IC technology until the integration level of devices on a chip became too large (i.e., the late 1980's), when CMOS took over.

Exhibit 1014 - Page 151 of 218

NMOS is inexpensive to fabricate, very functionally dense and faster than PMOS. The earliest NMOS technologies required only 5 masking steps, including the pad mask.* On the other hand, NMOS logic-gates (e.g., inverters) consume dc power while in one of their two logic states. Therefore, at least some of the logic gates in an NMOS IC are always ON. Such circuits draw a steady current, even when being operated in the standby mode (when no signals are propagating through the circuit). Hence, as the number of logic gates on NMOS chips increased, the current being drawn (and hence the power dissipated) also rose. Although this was always a limitation of NMOS for such applications as space-borne or portable electronic systems, it did not represent a major drawback for most others. Such was the situation at the level of device integration that existed up to the mid-1970s. By the mid-1980's, with the dawning of the VLSI era (i.e., device counts on chips began to exceed 10^5), power consumption in NMOS circuits began to surpass tolerable limits. A lower-power technology was needed to exploit the VLSI fabrication techniques. CMOS represented just such a technology.[1]

From a quantitative perspective, the ascendancy of CMOS was the inevitable result of the two-hundred-fold increase in functional density and the twenty-fold increase in speed of ICs between 1968 and 1987. For example, in 1969, 256-bit SRAM circuits (e.g., the Intel 1101) used 12-μm design rules to create six-transistor SRAM cells; each cell occupied 20,600 μm^2. By 1998, SRAM cells using 0.25-μm design rules occupied less than 10 μm^2. (The memory access time in these respective memory circuits decreased from 1 μs down to 10 ns.) By 1998 the decreased device dimensions and the attendant increase in chip size allowed 16-Mbit SRAMs to be built. To take another example: The Intel 4004 4-bit microprocessor (introduced in 1971), had 2300 devices and was built with PMOS. The 8086 model, a 16-bit microprocessor introduced in 1978, had 29,000 devices and was built in NMOS, and the 80386, introduced in 1985, had 275,000 devices and was built in CMOS.

Chips dissipating in excess of 5 W of power can be housed in ceramic packages. But such packages are expensive. To be able to use the much cheaper plastic packages, the maximum power dissipation is limited to about 1 W. The Intel 8086 dissipated around 1.5 W of power

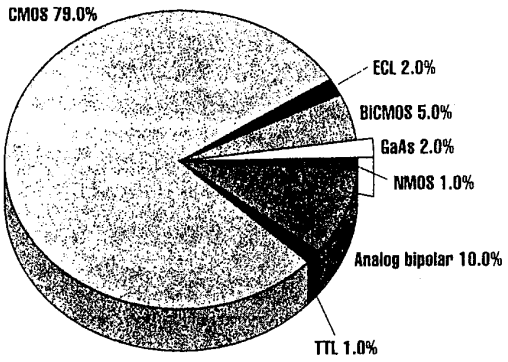


Fig. 16-1 Estimated market share of IC technologies in 1995. Reprinted with permission of Oxford Press.

* Early NMOS logic gates used butted contacts as well as enhancement-mode (E) transistors for both the driver and the load. Depletion-mode loads (D) and buried contacts came later. Hence, of the seven masks of the E-D NMOS process, only five were required in early NMOS technology. This resulted not in lower manufacturing costs per wafer, and also in higher yields. Thus, the cost of manufacturing NMOS circuits was brought even lower. See Vol. 2 for more details on this subject.

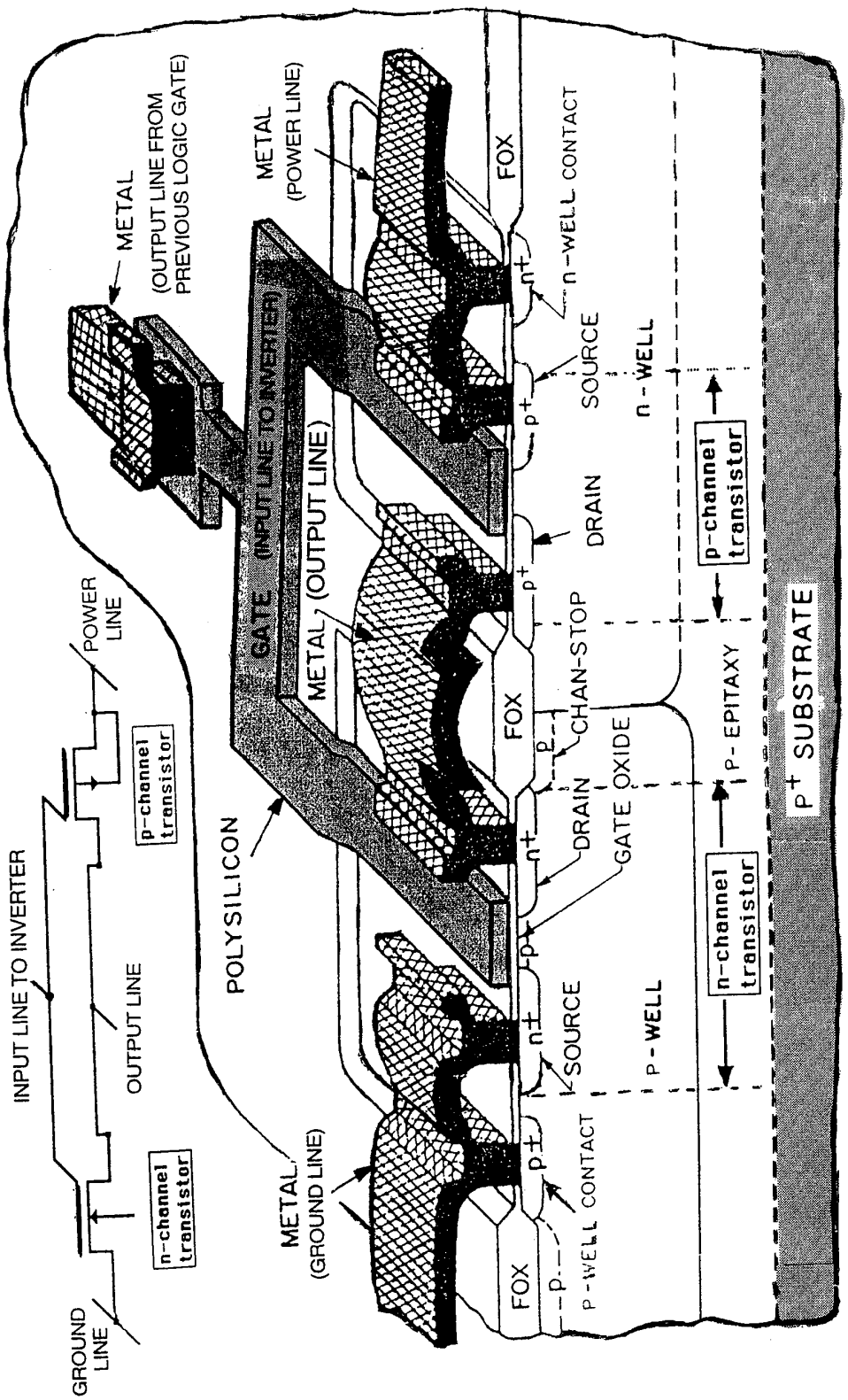


Fig. 16-2 CMOS inverter. (a) Circuit schematic. (b) Device cross section.

when operated at 8 MHz. By the late 1970s it was clear that chips would very soon be manufactured with power dissipation values that were unacceptably high. (Note that when the 8086 was later reissued in CMOS technology under the model number 80C86, its power consumption dropped to about 0.25 W.)

Unlike in an NMOS inverter, in a CMOS inverter (Fig. 16-2) only one of the two transistors is driven at any one time (if the inverter is not in a transient switching state). This means that a high-impedance path exists between V_{DD} and ground, regardless of the state the inverter is in. Hence, virtually no current flows between the power and ground lines, and almost no dc power is dissipated by the CMOS circuit. Thus, logic gates consuming only microwatts of standby power can be implemented with CMOS (Fig. 16-3).

The problem of power dissipation can also be considered from both a *chip* perspective and a *system* perspective. From the chip perspective, microprocessors of the 64-bit generation are built using CMOS and yet some still dissipate more than 15 W of power. If only NMOS technology was available, the power dissipation would probably be an order of magnitude larger! Now consider memory chips from the system perspective. Although a 1-Mbit DRAM may consume 120 mW of power in NMOS, it consumes even less (~50 mW) in CMOS. Since there may be hundreds of memory chips in a system (versus only a few microprocessors), the ramifications of lower power-dissipation at the system level are significant. Smaller power-supplies and smaller cooling fans are but two of these.

16.1.1 Historical Evolution of CMOS

Although CMOS is now the dominant integrated-circuit technology, during its first 20-years it was considered to be a runner-up for the design of MOS ICs. The pairing of complementary *n*- and *p*-channel transistors to form low-power IC logic gates was originally proposed by

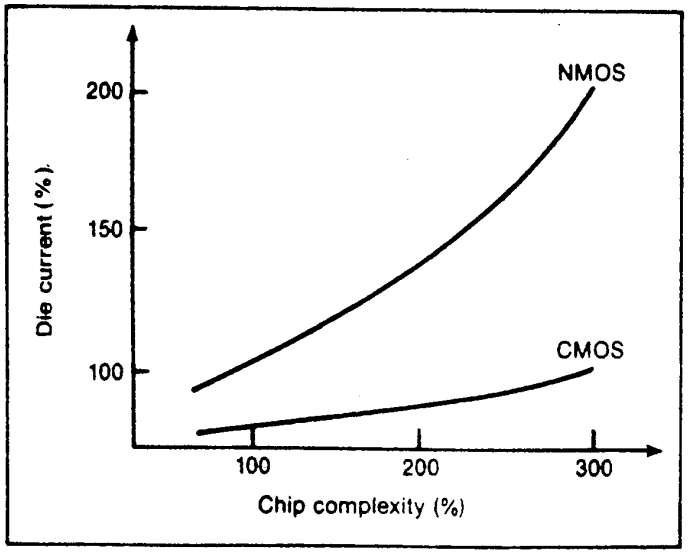


Fig. 16-3 As circuit complexity increases, NMOS power consumption rises to levels that eventually prevent further growth. In contrast, CMOS power consumption increases only slightly as the device count on a chip rises.

by Sah and Wanlass in 1963.[4] The first CMOS ICs were fabricated in 1966. Subsequent development of the technology was spearheaded by the RCA Corporation. The earliest volume commercial application of CMOS was its use as a logic-inverter in the frequency divider circuits of digital watches.

CMOS technology at that time had many disadvantages compared even to PMOS and, later, to NMOS. Drawbacks included significantly higher fabrication cost, slower speed, susceptibility to latchup, and much lower packing density. As a result, until the 1980s CMOS was limited to applications that mandated either the lower power dissipation of CMOS (e.g., watches and calculators), or it's very-high noise margin (e.g., radiation-hard circuits). Furthermore, the advances made in NMOS fabrication were not rapidly transferred to CMOS. For many years CMOS languished behind advanced Si-gate-NMOS and bipolar technologies. Except for the special applications mentioned above, it lay dormant for nearly two decades.

There were several reasons why this occurred. Recall that both *n*- and *p*-channel transistors must be fabricated on the same wafer in CMOS technologies. Obviously, only one type of device can be fabricated on a given starting substrate. To accommodate the device type that cannot be built on this substrate, regions of a doping type opposite to that present in the starting material must be formed. These regions of opposite doping, called *wells* (or sometimes *tubs*) are the first features to be defined on a starting wafer. In the earliest CMOS technologies, a single-well approach was used. A lightly doped *n*-type wafer was the starting material, and only *p*-wells were formed in the wafer. This was called *p*-well CMOS. At the time these circuits were being fabricated (late 1960s), the processes of ion implantation and local oxide isolation (LOCOS) had not yet been developed. Metal gates were still being used in MOS devices and control had not been gained over the large and quite variable positive oxide charges in the gate oxide. These issues made *p*-well technology the only viable technique for fabricating CMOS.

While PMOS enhancement-mode transistors could be successfully fabricated in a lightly doped (e.g., 10^{15} cm^{-3}) *n*-substrate with a V_T of about -2 V, NMOS enhancement-mode transistors could not be fabricated on lightly doped *p*-substrates because the V_T of NMOS metal-gate transistors on such substrates is negative. In addition, it was likely that parasitic channels would be established in the field regions between NMOS devices built on lightly doped substrates. These parasitic channels could not be reliably suppressed because of three other factors: 1) the oxide charge at the silicon/SiO_2 interface was large, and procedures for lowering it had not yet been discovered; 2) the segregation of boron at the field-oxide/silicon-substrate interface reduced the V_T of parasitic MOSFETs; and 3) the need to use relatively thin field oxides (to permit adequate step coverage of metal layers running over these oxide steps) also reduced the parasitic MOSFET V_T (see Vol. 2). Therefore, the only reliable way to manufacture enhancement-mode NMOS transistors for CMOS inverters without excess leakage was on regions with boron surface concentrations high enough to overcome these problems.

The *p*-well approach to building CMOS provided such regions, since the *p*-well had to be doped about ten times as heavily as the *n*-substrate for adequate control of doping in the well to be achieved. As a result, *p*-well technology became established in the companies that pioneered CMOS technology. Long after the problems of fabricating NMOS on lightly doped substrates had been solved (by controlling V_T with ion implantation, and the interface charge by annealing), most of these companies continued to use *p*-well technology to design new ICs.

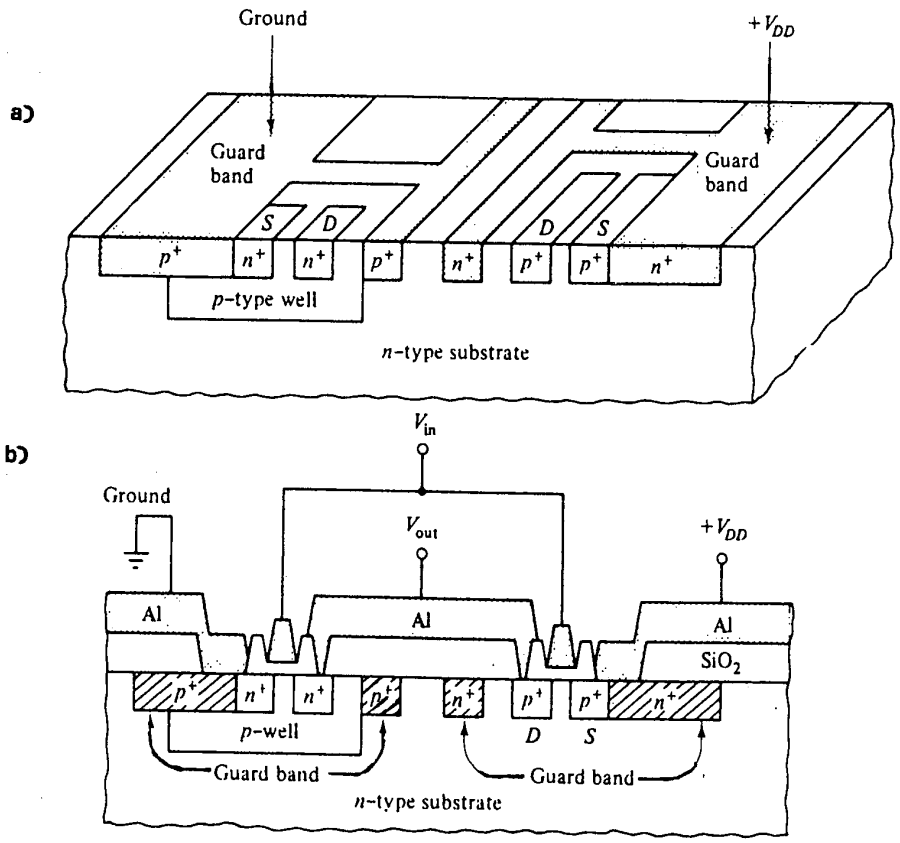


Fig. 16-4 CMOS structure with guard bands. A CMOS inverter circuit is depicted.[5] From W. C. Till and J. T. Luxton, *Integrated Circuits: Materials, Devices, and Fabrication.* Copyright 1982 Prentice-Hall. Reprinted with permission.

The packing-density limits of the early *p*-well CMOS technology, however, were responsible for its poor performance. The packing density was much worse than that of NMOS, primarily because the *n*- and *p*-devices had to be surrounded by guard rings (n^+ or p^+ diffusions that surround the device, as shown in Fig. 16-3)[5] to prevent the inversion of the field regions and latchup (see Vol. 2).

Before LOCOS isolation and ion implantation became available, guard rings were needed to provide adequately large values of V_T in the field regions. However, their use results in very large-area devices. Until oxide isolation and channel-stop implants were developed, interconnect capacitance and resistance severely degraded the speed performance.* Once guard rings were no longer needed, however, the devices could be brought closer together, and the speed of CMOS circuits improved dramatically.

When it finally became apparent in the late 1970s that the increases in power density and dissipation make it impossible to design future generations of circuits with NMOS,

* It should be noted that the fabrication of NMOS devices in heavily doped *p*-wells also increased the device junction-capacitances and reduced the magnitude of the drive current. This further decreased the performance of the circuits, but to a lesser degree than did the interconnect capacitance.

companies that had been stubbornly continuing to use NMOS for design and fabrication finally began to consider CMOS. It was natural for them to seek a technology that was compatible with the modern high-production-volume Si-gate-NMOS processes that they had successfully perfected. Since *n*-well CMOS (the other single-well CMOS type) offers near compatibility with such processes, and since it allows NMOS performance to be optimized (through their fabrication in the lightly doped *p*-substrate regions), it became the technology of choice for many companies that had formerly been manufacturing NMOS integrated circuits.[6]

Some time later, however, it became evident that neither *p*-well nor *n*-well would be the optimum choice for submicron CMOS. It was realized that *twin-well CMOS* would be more effective. As a result, many twin-well processes were subsequently developed. Examples of such twin-well CMOS process flows in Sects. 16.2 and 16.3 will be considered.

16.1.2 The Operation of CMOS Inverters

The CMOS inverter (for which the circuit schematic and a sample layout are shown in Fig. 16-2a and 16-5, respectively) uses enhancement-mode transistors for both the NMOS driver and the PMOS load transistors. (In enhancement-mode transistors the devices are *OFF* if the gate [or input] voltage is zero.) It can be seen that the gates of the two transistors in a CMOS inverter are connected, and serve as the input to the inverter. The common drains of each device are also connected to the inverter output. Assume the inverter is driving some load capacitance, C_L,

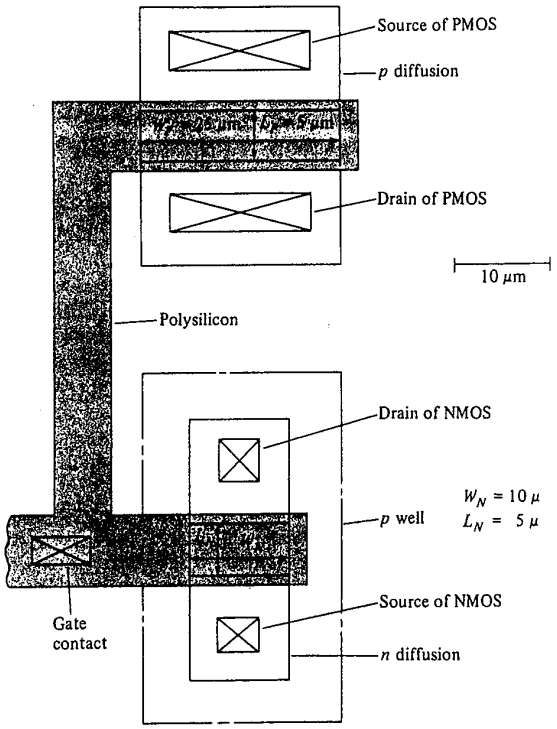


Fig. 16-5 CMOS inverter layout. From D. A. Hodges and H. G. Jackson, *Analysis and Design of Digital Integrated Circuits*, Copyright 1988, 2nd Ed., McGraw-Hill Book Co. Reprinted with permission.

(e.g., the input to another CMOS logic gate). Both the source and body of the NMOS transistor are connected to ground, while those of the PMOS transistor are connected to V_{DD} (e.g., 5 V).

The threshold voltage of the NMOS driver transistor V_{Tn} is positive (e.g., V_{Tn} = 0.6 V), while that of the PMOS load transistor V_{Tp} is negative (e.g., V_{Tp} = -0.6 V). Figure 16-6a shows the inverter's output voltage V_o as a function of the input voltage V_{in}. This curve is known as the *static voltage input/output characteristic*, or the *transfer characteristic*.

When V_{in} = 0, V_{GSn} = 0, the NMOS transistor is *OFF* (since V_{GSn} is <0.6 V). The PMOS transistor, however, is *ON*, since V_{GSp} = -5 V, which is much more negative than -0.6 V. Thus, when V_{in} = 0, C_L is charged to V_{DD} through the turned-on PMOS load transistor, and V_o = 5 V.

When V_{in} = 5 V, the NMOS transistor is turned *ON*, since V_{GSn} = 5 V (which is > 0.6 V). The PMOS transistor is turned *OFF*, since V_{GSp} = 0 V (which is more positive than -0.6 V). Consequently, when V_{in} = 5 V (V_{DD}), the output is connected to ground through the turned-on NMOS driver, allowing C_L to be discharged. Since the PMOS device is off, C_L will be completely discharged, and V_o will be 0 V.

The most important property of the CMOS inverter is that when the gate is sitting quiescently in either logic state (V_o = V_{DD} or 0), one of the transistors is *OFF* and the current conducted between V_{DD} and ground is negligible (i.e., it is equal to the leakage current of the *OFF* device). This feature can be seen in Fig. 16-6b, which plots the current through the inverter, I_{DD}, as a function of V_{in} (solid curve). The power dissipated in the static (or standby) mode is then determined by the product of the leakage current and the supply voltage. Since the leakage current of an MOS transistor being operated in cutoff is so small, very little power is consumed in the static mode. Another important feature is that the output voltage V_o swings all the way from V_{DD} to 0 as the inverter changes state. This switching characteristic is referred to as *swinging from rail to rail.*

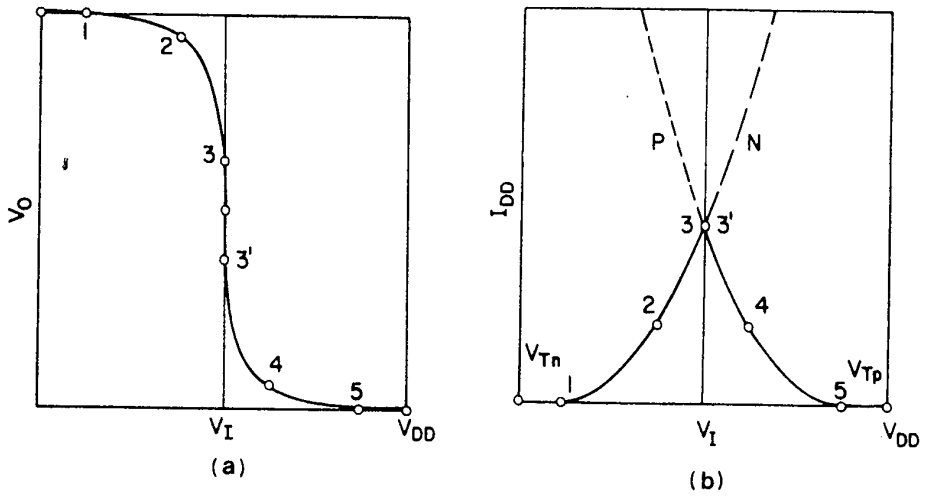


Fig. 16-6 (a) Output (V_o) versus input (V_I) voltage of CMOS inverter. (b) Current through inverter as a function of input voltage (solid curve); I-V characteristics of *n*– and *p*–channel transistors (dashed curves). The numbers correspond to different points on the inverter transfer characteristic.[8]

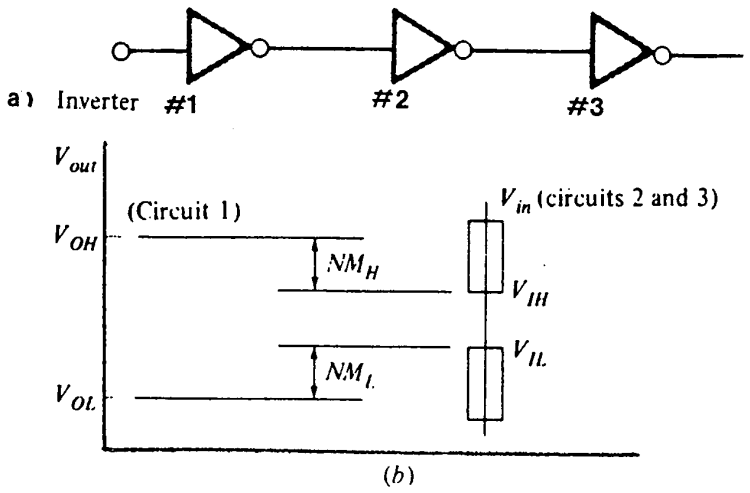


Fig. 16-7 (a) String of inverters connected in series. (b) Definition of noise margins.

Two operational features of the CMOS inverter must also be mentioned. First, as can be seen in Fig. 16-6b, significant current is conducted through the inverter only when both transistors are *ON* at the same time (i.e., during switching). Therefore, most of the power dissipation is due to the charging and discharging of C_L. In fact, it can be shown that the power dissipation is essentially given by $f C_L V_{DD}^2$, where f is the switching frequency. Because only a small fraction of the logic gates in an IC are likely to switch during any given clock cycle, CMOS power dissipation remains much smaller than in NMOS ICs.

Second, because the output voltage swings from rail to rail, the CMOS inverter inherently provides excellent noise margins. Noise margins are usually defined in terms of the *logic-gate threshold voltages* V_{OH}, V_{OL}, V_{IH}, and V_{IL} (which are *not* the same parameters as *device threshold-voltages,* see Fig. 16-7a). The noise margins NM_L and NM_H (Fig. 16-7b) are defined according to the following equations:

$$NM_L = V_{IL} - V_{OL} \tag{16.1a}$$

$$NM_H = V_{OH} - V_{IH} \tag{16.1b}$$

Briefly, the argument for Eq. 16-1a is that the logic gate (e.g., *Inverter 1* in Fig. 16-7a) should provide a maximum "low output" that is less than the maximum "low input" that the subsequent logic gate (e.g., *Inverter 2*) can accept without causing the output voltage of *Inverter 2* to be driven into the ambiguous portion of the transfer characteristic. The difference between V_{IL} and V_{OL} (the noise margin) then specifies how much noise voltage can be tolerated at the input node of *Inverter 2* before its output will be driven to a voltage value that is logically undefined.

In CMOS, V_{OL} approaches zero, and V_{OH} approaches V_{DD}. Because of the steep transition region in the transfer characteristic, V_{IL} and V_{IH} can be designed so that noise margins are on the order of $V_{DD}/4$ for expected process variations and operating temperatures. Thus, for a 5-V CMOS technology, V_{NML} can be 1.25 V, whereas in NMOS, V_{NML} is typically only 0.3 V.

For the performance of the logic gates to be maximized, the threshold voltages of the *p*- and *n*-channel transistors should be comparable (and, ideally, of equal magnitude). In addition, the threshold voltages should be as small as possible. The minimum values selected for V_{Tn} and V_{Tp} will be determined by the subthreshold leakage current, I_{Dst}.

16.2 PROCESS SEQUENCES FOR TWIN-WELL CMOS FOR THE GENERATIONS FROM 1.2 μm TO 0.5 μm

This section describes an 11 mask, twin-well CMOS process flow that is representative of process sequences used to fabricate CMOS ICs from 1.2 μm down to about 0.5 μm. The process flow is nevertheless useful as a vehicle for illustrating the general sequence of steps employed in CMOS fabrication. Twin-well CMOS technology has been generally adopted for CMOS generations below about 1 μm because this approach allows two separate wells to be formed in a lightly doped substrate region. The doping profiles in each well can thereby be tailored independently so that neither device will suffer from excessive doping effects.

16.2.1 Starting Material: All MOS technologies use silicon wafers with a <100>-orientation. Typically, in twin-well CMOS the starting material is also either a lightly *p*-doped wafer (*p*-bulk wafer, Fig. 16-8a), or a heavily *p*-doped wafer on which a thin, lightly *p*–doped epi layer is grown (Fig. 16-8b). The concentration range of the *p* doping in both the bulk wafer and the *p*–epi layer is $8x10^{14}$–$1.2x10^{15}$ cm^{-3}. The *p*-epi-on-p^+ wafers (see Chap. 7) provide several advantages including: 1) improved latchup protection (see Vol. 2); 2) gate-oxides with better reliability are grown on epi layers than on bulk-Si surfaces; and 3) improved gettering capability (see Chap. 2). Their chief disadvantage is higher cost. (In 1998, the price of a 200-mm bulk starting wafer was ~$100 US, while that of a 200-mm epi wafer was ~$180 US.) For this example process flow, assume *p*-epi-on-p^+ wafers are used (Fig. 16-8b). The exact doping concentration of the epi layer is chosen to provide the best overall set of the following device characteristics: low source/drain-to-substrate capacitance; high source/drain-to-substrate breakdown voltage; high carrier mobility; and low sensitivity to source-substrate bias effects. As feature sizes shrink, the epi-layer doping concentration value is expected to increase.

16.2.2 Formation of Wells and Channel Stops: The wells are the first features to be formed in these generations of CMOS technology. A number of different procedures have been developed to form twin-wells. The most obvious method is to use two masking steps. Each blocks one of the two implant steps used to introduce the well dopants. A single masking-step procedure, however, was also developed. It is probably the one most commonly used, and it will be described here (see Fig. 16-9).

In this method, a single mask is used to pattern a nitride/oxide film that has been formed on the bare silicon surface (*Mask #1*). A *pad oxide* (which serves as stress-relief layer between the silicon nitride film and the silicon substrate) is first thermally grown to a thickness of 40–50 nm

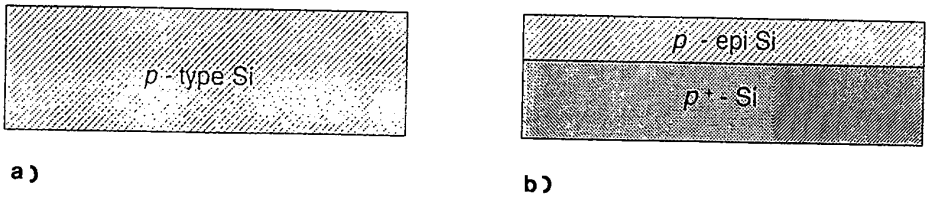


Fig. 16-7 Silicon starting substrates for twin-well CMOS: (a) *p*-bulk wafer; (b) *p*-epi-on-p^+ wafer.

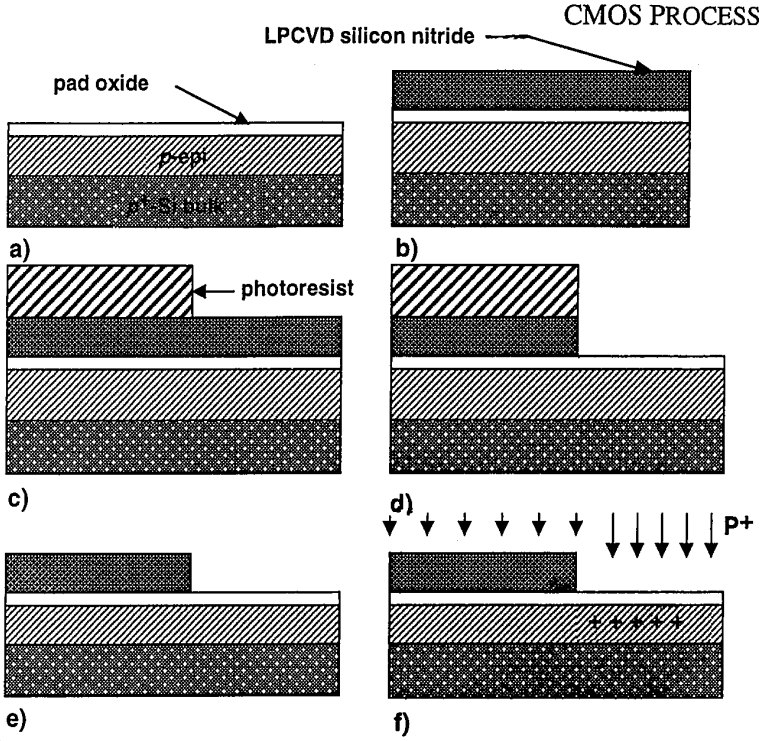


Fig. 16-8 Well formation in twin-well CMOS. (a) grow thermal pad oxide on p-epi-on-p^+-bulk starting wafer; (b) deposit LPCVD silicon-nitride film; (c) use *Mask #1* to define n-well regions; (d) etch nitride; (e) strip resist (f) implant n-well dopants (phosphorus).

(Fig. 16-8a). This is followed by an LPCVD of silicon nitride (140–200 nm thick- Fig. 16-8b). After the photoresist is applied, exposed and developed (Fig. 16-8c), the nitride is dry etched (Fig. 16-8d). The openings in the nitride film define the n-well regions. The required n-dopants are next introduced into these regions by ion implanting phosphorus (Fig. 16-8f). This implant is performed at 80 keV with a dose of ~$1x10^{12}$ cm^{-2}, using a medium-current implanter. The resist and unetched nitride regions (which define the p-well regions) act as an implant mask, preventing the phosphorus from entering the substrate in those regions. Next, a thermal oxidation step is carried out, using a wet oxide process (Fig. 16-9a). This grows an oxide to a thickness of 350–400 nm over the n-well regions. (Since nitride acts as a diffusion barrier, oxide grows only over the regions not covered by nitride.) The thickness of the thermal oxide is selected so it will block the subsequent boron implant used to form the p-wells. The nitride is next stripped using hot phosphoric acid (an etchant that does not attack the thermal oxide grown over the n–well regions). This leaves the silicon in the p-well regions covered only with the pad oxide, while the n-well regions remain covered with the thicker oxide. Thus, during the p-well implant (e.g., B at 50 keV, $1x10^{12}$ cm^{-2}, using a medium-current implanter), boron ions enter the silicon substrate only in the areas where the p–wells are to be formed (Fig. 16-9b). Next the well dopants are diffused into the substrate using a long, high-temperature drive-in step (e.g., 1100–1200°C for 3–20 hours). This is performed in an oxidizing ambient so that an oxide is also grown on the entire wafer surface during this step. At the conclusion of the drive-in step, the dopant concentration at the Si surface in the wells of 0.8-μm-CMOS is ~$1x10^{16}/cm^3$ for the p-well, and ~$3x10^{16}/cm^3$ for the n-well. The n-well might have a higher doping concentration to improve the punchthrough performance of the PMOS devices and to eliminate the need for a

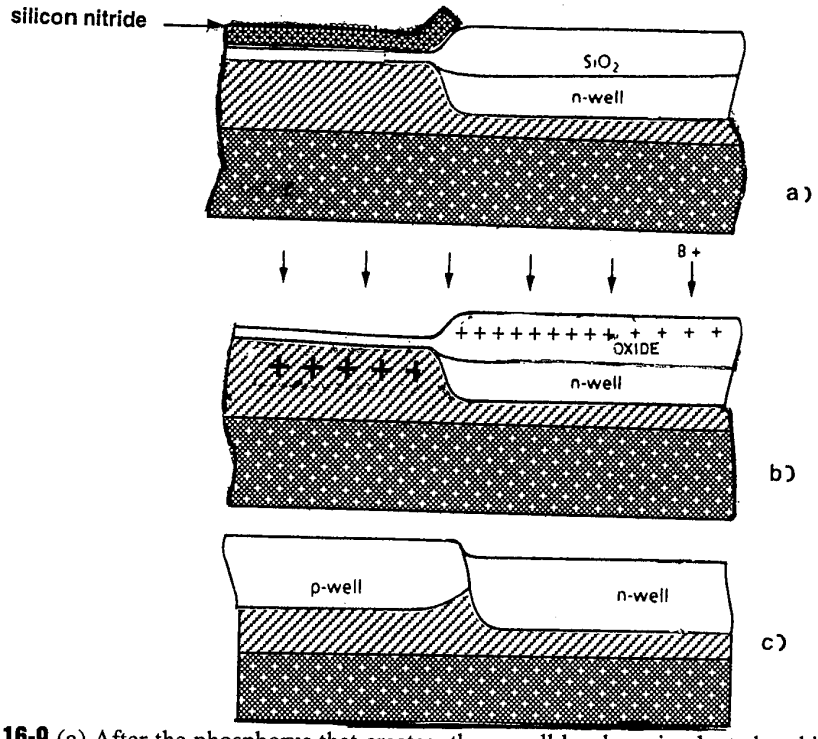


Fig. 16-9 (a) After the phosphorus that creates the *n*-well has been implanted, a thick oxide is grown, with the nitride preventing oxide form growing beneath it; (b) after stripping the nitride, a boron implant is performed that will form the *p*-well regions. The thick oxide above the *n*-wells prevents the boron from penetrating there; (c) wafer-surface topography after wells are driven-in and the oxide is stripped.

separate channel-stop implant step for the *n*-well. The oxide thickness is different over the *n*-well than the *p*-well, creating a step in the oxide film at the well boundaries. This oxide is nevertheless next etched away, but the step in the silicon remains behind (Fig. 16-9c). This step serves as an alignment mark to allow subsequent layers to be aligned to the well boundaries. (An alternative procedure that uses an additional mask can be used to create alignment marks. That is, deep pits can be etched into the wafer prior to the well formation step. Such marks provide improved registration accuracy for lithographic steps, and have become standard in deep-submicron CMOS technologies.)

It is important to observe that when wells are formed in this manner the well doping profile is at a maximum at the surface and then monotonically decreases with depth (see Fig. 16-10). The well doping concentrations are selected to be at least an order of magnitude higher than the epi-layer doping, so the expected variations in epi doping will not significantly affect the well concentration. The junction depth of the wells is typically ~2 μm deep.

16.2.3 Definition of Active and Field Regions: In conventional CMOS processes a standard LOCOS process (described in more detail in Vols. 2 and 3) is next used to define the active and field regions, and to create isolation structures between the active devices (Fig. 16-11). This is done by selectively growing a thick thermal oxide over the field regions, while keeping the active regions devoid of oxide. The steps that accomplish this begin with the growth of a pad oxide by thermal oxidation over the entire wafer to a thickness of 40–50 nm. Next, a blanket layer of silicon nitride (140–200 nm thick) is deposited by LPCVD (Fig. 16-11a). Then photo-

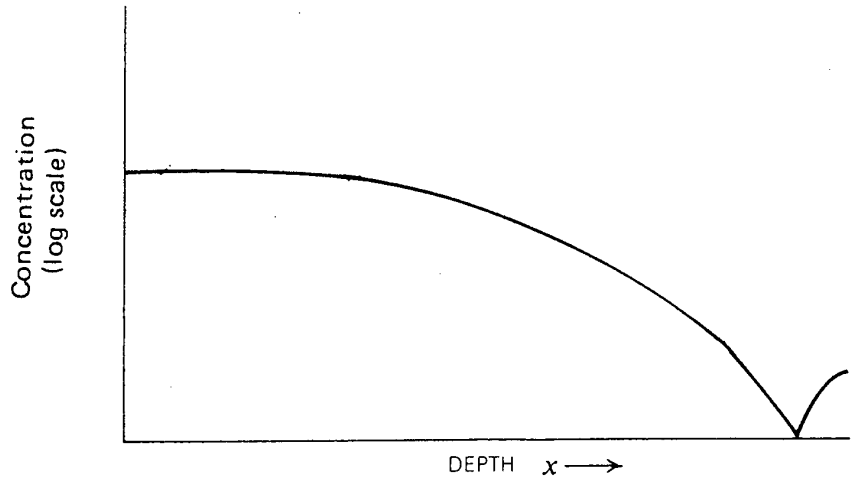


Fig. 16-10 Doping profile in a conventional well formed by a shallow implant and a drive-in step.

photoresist is spun onto the wafer and *Mask #2* is used to create the patterns that define the field and active regions of the circuit (Fig. 16-11b). After exposing and developing this pattern, the nitride is etched and the resist is stripped (Fig. 16-11c).

The next step creates the channel-stop dopant layer in the field regions of the *p*-wells. This is done by implanting a layer of boron atoms near the Si surface in the *p*-wells. This layer of implanted B atoms increases the threshold voltage of the parasitic PMOSFETs in the field regions of the *p*-wells. To prevent the boron from being implanted into the field regions of the *n*-wells, another mask is used. A layer of resist is applied prior to the boron channel-stop implant. *Mask #3* is used to form a pattern that covers the *n*-wells (Fig. 16-11d). Boron is then ion implanted into the surface (B with a dose of 10^{12}–10^{13} atoms/cm^2 at 40-80 keV, using a medium-current implanter). Boron only enters the silicon substrate in regions not covered by *Mask #3* or nitride (Fig. 16-11e).

A thick oxide (called a *field oxide*) is next grown selectively to a thickness of 400-700 nm over the field regions using a wet oxidation process (Fig. 16-11f). The combination of a thick field-oxide and channel-stop dopants beneath this oxide forms *parasitic MOSFETs* in the field regions whose threshold voltages that are too high to be turned on by normal voltages used to run the circuits. This keeps *active MOSFETs* in the IC electrically isolated from one another.

No oxide grows over regions covered with nitride (the active regions). Such a process of selectively growing field oxide has been named *local oxidation of silicon* (*LOCOS*). Observe, however, that some oxide does grow under the edges of the nitride film as a result of the oxidizing species diffusing laterally as well vertically. This forms a tapered oxide region at the edge of the active region called a *birds beak* (Fig. 16-12 and Chap. 8). As the active regions get smaller with device scaling, bird's beak begins to limit the use of LOCOS.

16.2.4 Threshold-Adjust Implantation Step: After the field oxide is grown, the nitride is stripped and the circuit is almost ready for carrying out the threshold adjust (or V_T-adjust) implant process and for growing the gate oxide. Prior to this, however, a problem caused by the previous field oxidation step of the LOCOS process must be rectified. During the wet oxidation process,

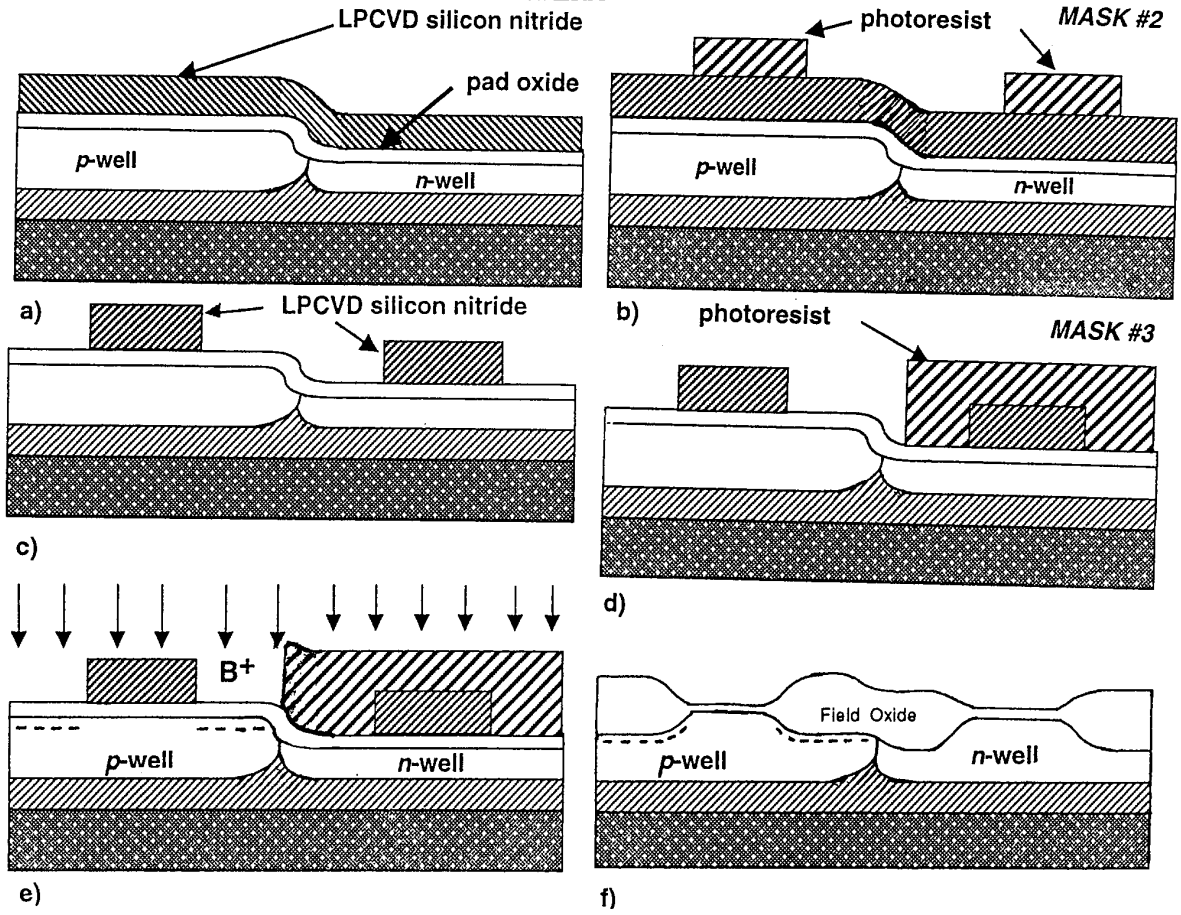


Fig. 16-11 Formation of the active and field regions using LOCOS isolation: (a) grow pad oxide and deposit silicon nitride film by LPCVD; (b) Use *Mask #2* to define active regions; (c) etch the nitride film; (d) Use resist and *Mask #3* to cover *n*-well regions; (e) implant B into the field-regions of the *p*-wells (channel-stop implant); (f) selectively grow field oxide with a wet-oxidation process, strip nitride.

a thin layer of nitride forms on the silicon surface of the active regions (near the edges of the nitride film). Such layers would interfere with the gate oxidation process, insofar as the gate oxide will not grow as rapidly over such regions (and hence it would be thinner there). This problem is called the *Kooi effect.* To overcome it a thin *sacrificial* (or *dummy*) *oxide* is next grown to a thickness of 50–100 nm. This also oxidizes the thin nitride layer so it can be removed with an HF etch. The V_T-adjust implant is usually performed with this dummy oxide layer in place, as the oxide serves another useful function. That is, it also acts as a screen oxide, preventing contaminants from the implantation process from entering the Si substrate.

The V_T-adjust implant step (Fig. 16-13a) sets the threshold voltage of both the NMOS and PMOS FETs, typically to values of 0.6 V and –0.6 V, respectively. The V_T-adjust implant process in CMOS technologies using only n^+ poly can be accomplished with either a single boron implant step (if the background doping is chosen correctly, see Fig. 16-13b), or with two separate boron implant steps (in which case an additional mask is required). A separate punchthrough-prevention implant for the NMOS devices is also usually needed, especially when the minimum device dimension decreases below about 1 μm (see Vol. 3). The punchthrough

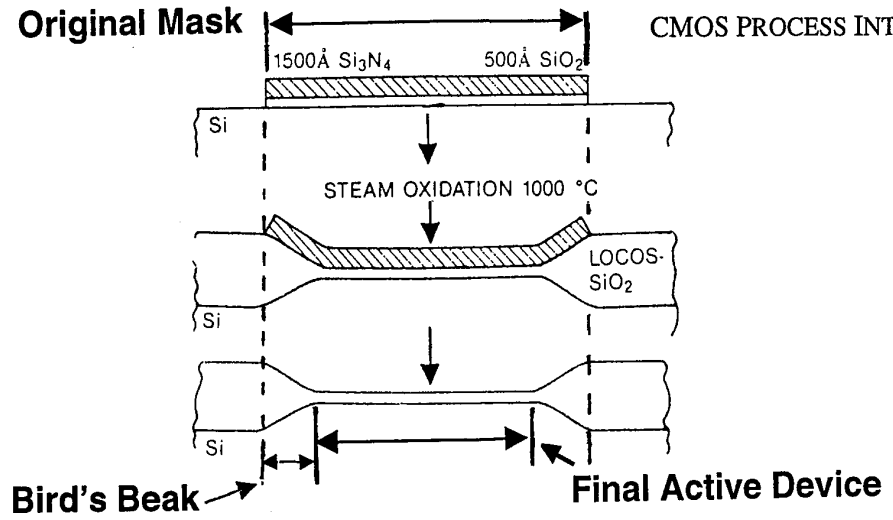


Fig. 16-12 "Bird's beak" in LOCOS isolation structure intrudes into the active regions of the wafer surface. This limits the use of LOCOS in deep-submicron IC process technologies.

implant can be performed sequentially using the same implanter, along with the V_T-adjust implant. Once the V_T-adjust implant step is completed, the dummy oxide is stripped with HF, leaving a residue-free Si surface upon which high-integrity gate oxides can be reliably grown.

16.2.5 Gate Oxide Growth: In the next step the gate oxide is formed. The growth of the gate oxide is a critical step. A defect-free, very thin (6-20-nm), high-quality oxide without contamination is essential for proper device operation. The gate oxide is grown only in the exposed active regions (actually, the field-oxide thickness is also very slightly increased as a result of the gate-oxide growth step). Since the drain current in an MOS transistor is inversely proportional to the gate-oxide thickness (for a given set of terminal voltages), the gate oxide is normally made as thin as possible – commensurate with oxide breakdown and reliability considerations. See Chap. 8 for details on thin gate-oxide growth.

16.2.6 Polysilicon Deposition and Patterning: Once the gate oxidation step is completed, a heavily *n*–doped polysilicon gate structure is fabricated. Polysilicon is the preferred gate material for several reasons. First, it can withstand the high-temperature steps required to form the source/drain junctions. Second, the polysilicon-SiO_2 interface is well understood and electrically stable. The gate formation procedure begins with the deposition of a 0.4–0.5 μm thick, undoped polysilicon film by LPCVD (Fig. 16-14a, and see Chap. 6 for details). This layer is then doped with phosphorus by ion implantation or chemical doping (Fig. 16-14b), producing a film with a sheet resistance of 20–30 Ω/sq While this sheet-resistance value is small enough for MOS circuits with gate lengths ≥ 2 μm, lower values are required for submicron ICs. In such applications, composite layers of refractory metal silicides and polysilicon (called *polycides*) are used to reduce the sheet resistance to ~1–3 Ω/sq. This allows the benefits of the polysilicon-SiO_2 interface to be preserved and the gate sheet resistance to be lowered. For the CMOS processes being discussed here, CVD WSi_x became the most popular material used to form such polycide structures (see Chap. 6 for details on CVD WSi_x deposition).

The gate structure and polysilicon interconnect structures are then patterned using *Mask #4* (Fig. 16-14c). Following exposure and development of the resist, the polysilicon (or polycide) film is dry-etched. This is a critical etch step for several reasons. First, due to the self-aligned

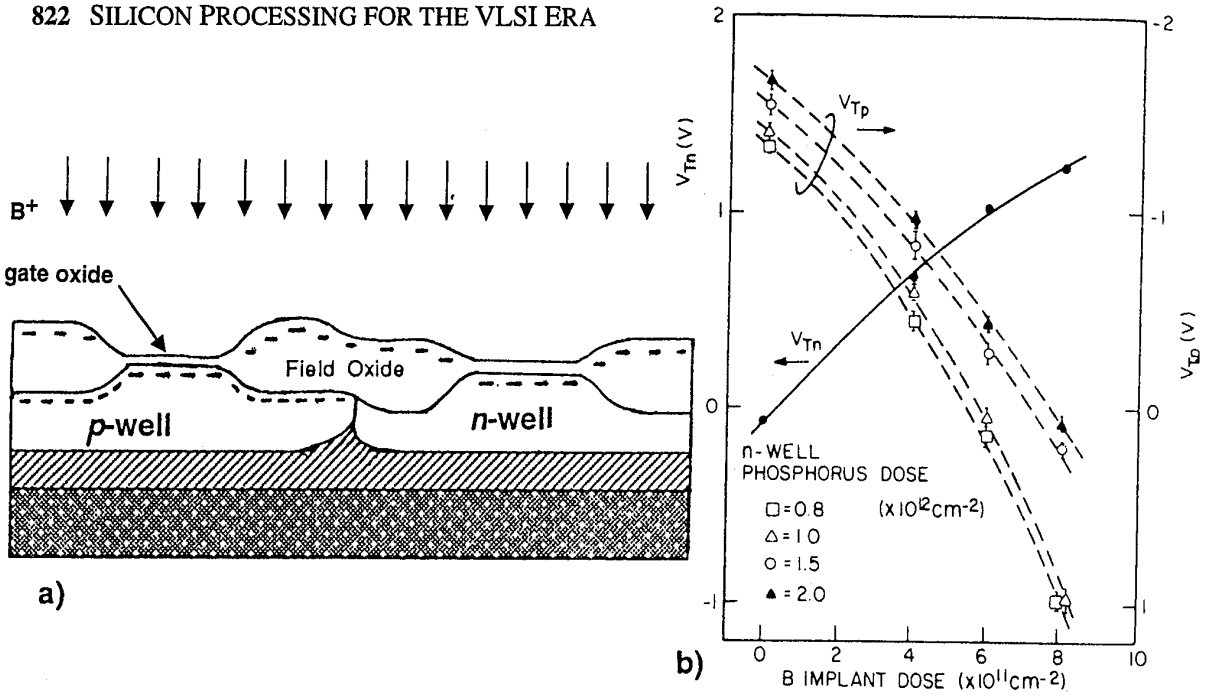


Fig. 16-13 (a) Threshold-voltage adjustment carried out using a single boron implant; (b) threshold voltages of NMOS (V_{Tn}) and PMOS (V_{Tp}) transistors as a function of boron dose. The CMOS structure uses an *n*-well implanted into a *p*-substrate (with doping level of $6x10^{14}$ atoms/cm^3). V_{Tp} results are shown for various implant doses of the *n*-well.[10] (© 1980 IEEE).

nature of silicon-gate technology, the channel length of the device depends on the width of the polysilicon line. Hence, the gate-length dimension must be precisely maintained across the entire wafer, and from wafer to wafer. If the gate length is too long, the drain current of the MOSFETs will decrease, slowing down the performance of the IC. If the gate length is too short, the source and drain may punch through. Second, the profile of the etched poly gate structure should be vertical. This will prevent variation of channel lengths by the penetration of the ions of the thinner regions of the gate sidewalls during formation of the source/drain regions by ion implantation (Fig. 16-14d). Third, to achieve the above goals, an anisotropic polysilicon etch process must be employed. This process, however, requires overetching to remove the locally thicker regions of polysilicon that exist wherever it crosses steps on the wafer surface. During the overetch time, areas of the thin gate oxide are exposed to the etchants. Thus, it is necessary to use a polysilicon etch process that is highly selective with respect to SiO_2.

After the polysilicon is etched, a step called *poly reox* is usually performed (16-14e). This grows a thin thermal SiO_2 layer on the polySi (on the order of 10 nm thick). The reasons for performing this step are discussed in Vol. 3, Chap. 9.

16.2.7 Formation of Source/Drain Regions: The next step in the process flow is the formation of source and drain (S/D) regions of the MOSFETs. Such regions are paths of current flow in the silicon between the metal interconnect lines and the channel of the transistor. As such, it is important that they have the lowest possible resistance. In addition, submicron MOSFETs require S/D junctions to be as shallow as possible to suppress such short-channel effects as *punch-through,* see Vol. 3. To obtain low resistivity, the S/D regions are doped as heavily as possible (typically using ion implantation, with a dose on the order of ~10^{15} cm^{-3}). The NMOS S/D

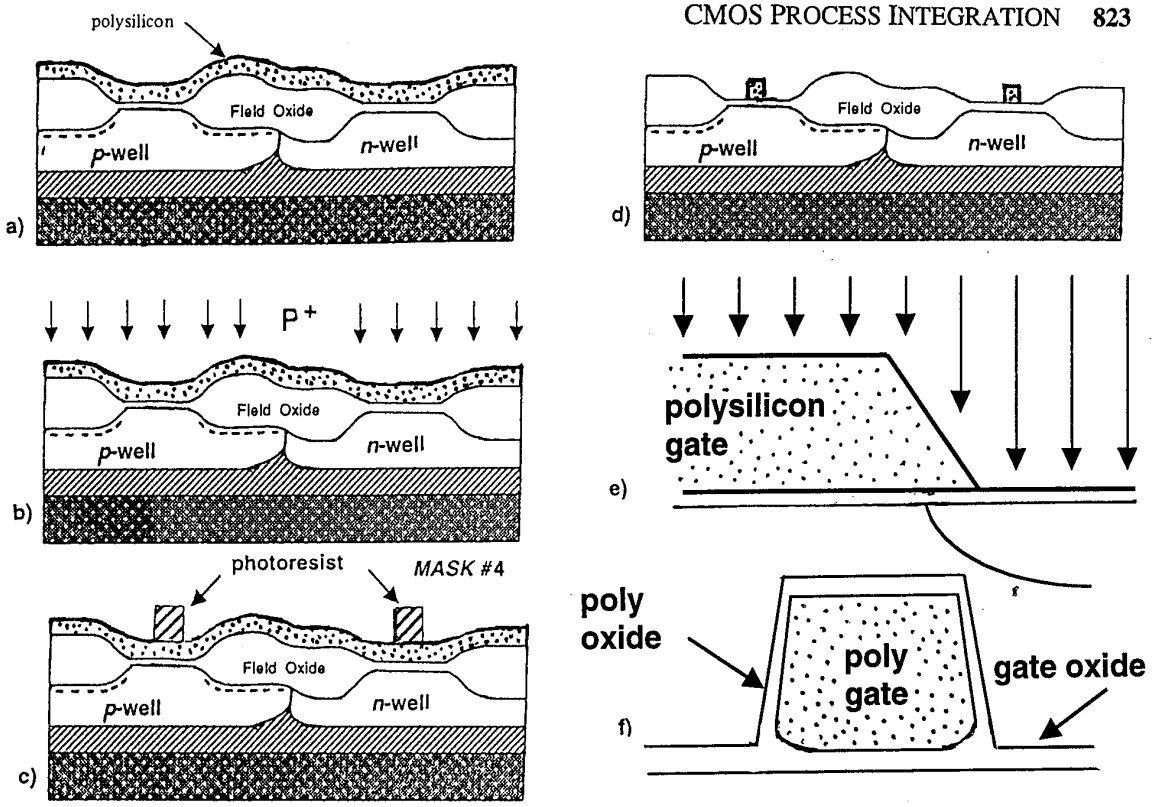


Fig. 16-14 n^+-polysilicon gate formation: (a) deposition of undoped polysilicon film by LPCVD; (b) *ex-situ* doping of the polysilicon with phosphorus by chemical diffusion of ion implantation; (c) Use *Mask #4* to define the polysilicon gates; (d) etch polysilicon and strip resist; (e) the etch process used to pattern the gate must yield poly features with vertical sidewalls (to maintain control of the effective gate length during the implant process that forms the source and drain regions); (f) *polysilicon re-ox* step grows a thin oxide on the poly by a short thermal oxidation process.

regions are doped with As, because it has a high solubility, low diffusivity, and a shallow projected range at the low ion implant energies that are used. Boron or BF_2^+ are used to dope the PMOSFET S/D regions. However, boron has a higher diffusivity than As. Hence, shallow S/D junctions in PMOSFETs are harder to achieve than in NMOSFETs. Since it is necessary to selectively implant source/drains to form n^+ regions for the *n*–channel FETs and p^+ regions for the *p*-channel FETs, this usually requires two masks (i.e., *Mask #5* and *Mask #6*, Fig. 16-15).

In submicron-CMOS processes, gate lengths become so small that *lightly doped drain* (*LDD*) structures must be used to minimize hot-electron effects, especially in NMOS devices. Thus, procedures are integrated into the CMOS process flow to fabricate such LDD structures. If these structures are used only for NMOS transistors, *Masks #5* and *#6* will still allow the sources and and drains of the two transistor types to be selectively implanted.* However, if LDD structures are needed for both PMOS and NMOS devices, two more masking layers may be needed.

* A procedure that requires only one masking step for creating both types of source and drain regions has also been reported. In this approach, the heavy boron implant is carried out non-selectively. The mask is then used to cover the PMOS device active regions. An *n*-type implant heavy enough to overcompensate for the boron implant in the active regions of the *n*-channel devices is used to form the sources and drains of the NMOS transistors.

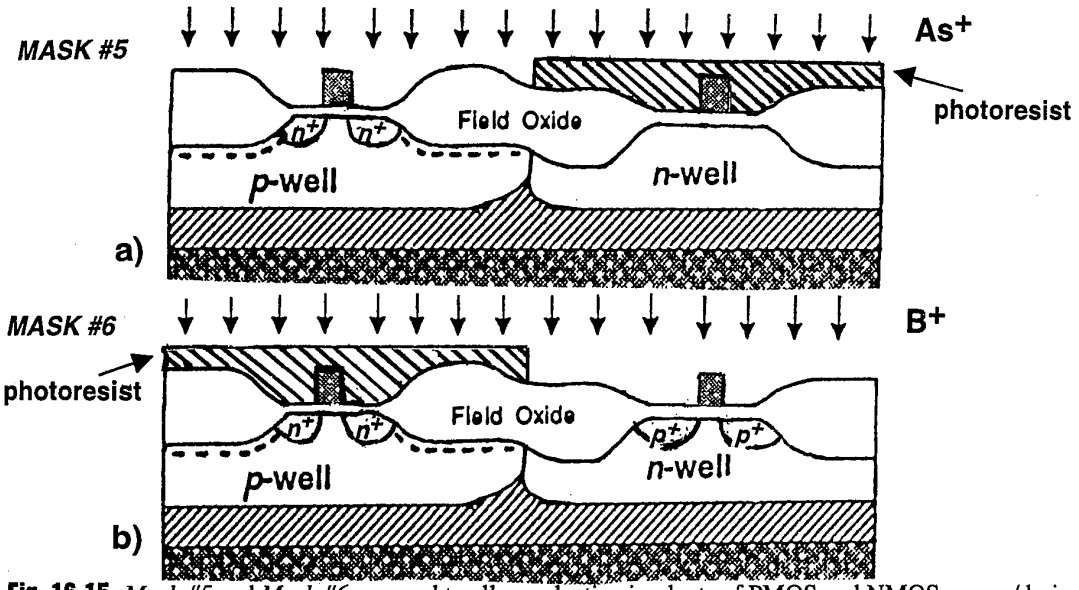


Fig. 16-15 *Mask #5* and *Mask #6* are used to allow selective implants of PMOS and NMOS source/drains.

The LDD structure is formed in the following way. The lightly doped regions of the source and drain are created with an ion implant step (Fig. 16-16c). A dose of approximately $3x10^{13}$–$3x10^{14}$ cm^{-2} dopant ions is implanted at low energy (30–50 keV). The implantation process causes the edge of these implanted ions to be automatically aligned to the edge of the gate (i.e., a *self-aligned process*). A conformal layer of dielectric material (usually SiO_2 or silicon nitride) is then deposited over the entire wafer. An anisotropic etch process is used to clear the oxide in the flat areas while leaving sidewall spacers on the sidewalls of the poly gates (Fig. 16-16e). These spacers cover and protect the regions beneath from the subsequent high-dose implant that forms the rest of the S/D regions. A dose of $1x10^{15}$–$5x10^{15}$ atoms cm^{-2} at 40–80 keV is used for this step (Fig. 16-16f).

16.2.8 Premetal Oxide Deposition and Contact Formation: Following formation of the source and drain regions, a doped *premetal dielectric* (*PMD*) is deposited by CVD. (This layer is also known as an *interlevel dielectric* or *ILD*.) Contact windows are opened in this dielectric layer to allow electrical connections to be made between Metal 1 and the following structures: 1) source/drain contact regions: 2) gate contacts; 3) substrate-contact regions; and 4) well-contact regions. A CVD process is used to deposit doped SiO_2, about 1 μm thick (Fig. 16-17a), onto the wafers (see Chap. 6 for details of this process). The dopant in the SiO_2 is either phosphorus (in which case the material is referred to as *phosphosilicate glass* or *PSG),* or both phosphorus and boron (making it *borophosphosilicate glass* or *BPSG*).

The doped CVD glass layer plays several roles in the fabrication and operating aspects of the circuit. First, it acts as an insulating layer between polysilicon and Metal 1. Second, it reduces the parasitic capacitance between Metal 1 and the substrate. Third, the addition of phosphorus to the glass makes the layer an excellent getter of Na ions (contamination by Na can destabilize the V_T of an MOS device). The PSG (or BPSG) binds otherwise mobile Na atoms within the doped-glass layer, preventing them from reaching the gate oxide and altering V_T. Finally, the dopants in the glass make it viscous at elevated temperatures (1000–1100°C for PSG, 850–950°C for BPSG, see Ch. 6). This allows the layer of glass to be flowed after it is deposited (Fig. 16-17b).

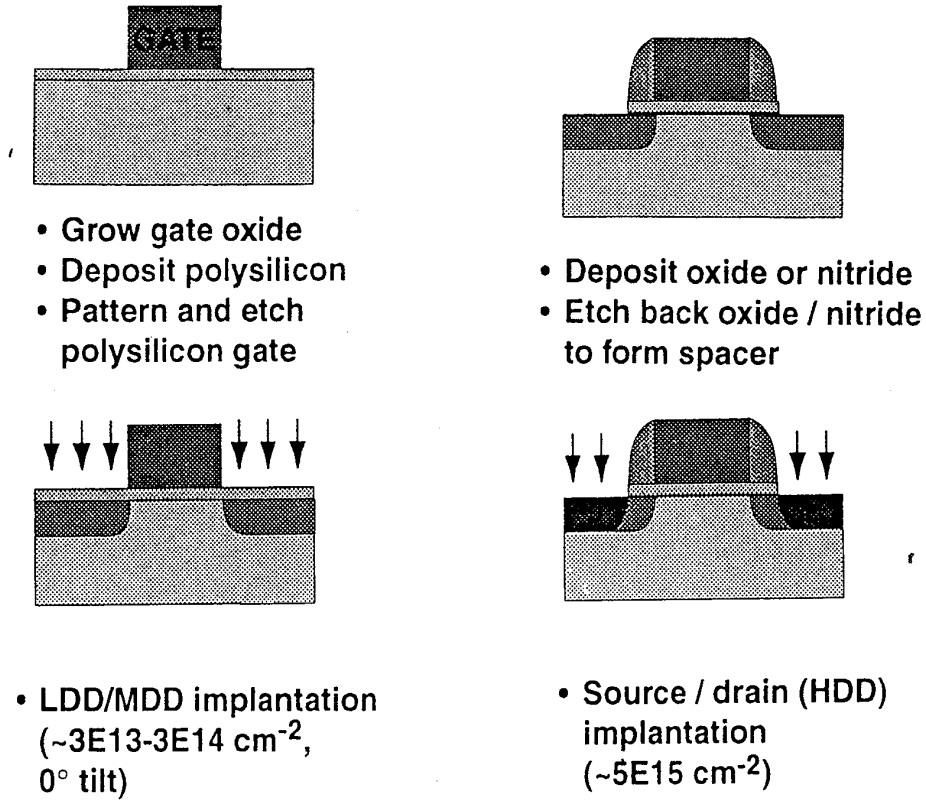


Fig. 16-16 Basic process to form an LDD NMOSFET: (a) after poly etch; (b) LDD implant; (c) after spacer material deposition and etch; (d) n^+ implant.

Through this procedure, a rounding of the glass contours and smoothing of any sharp steps is achieved. This produces better step coverage of the metal (which is deposited next) over the otherwise severe wafer topography. The high-temperature glass-flowing step also serves as an annealing process that activates the dopants implanted to create the source/drain junctions.

Contact openings are next created by a lithography-and-etch step (Fig. 16-17c). *Mask #7* is used to define contact patterns in a photoresist film. A dry-etch process is then used to open the contact windows through the interlevel dielectric to the underlying polysilicon and the source/drain regions in the silicon.

This contact-opening step can be critical, as the contact size and its alignment to underlying patterns limit the minimum size of the device. The source/drain regions must also be large enough for the contact to fit, with an allowance for alignment tolerance. If the contact opening exposes a part of the substrate, the drain or source will be shorted to it. Likewise, if the contact opening overlaps the both the source/drain and the gate, a short will be created between them.

After the contact etch is completed and the resist is stripped, the doped CVD glass is subjected to another flow step. This procedure rounds the corners of the top of the oxide windows so metal step coverage into the contact windows is improved (called *reflow* of the contact windows, see Chap. 6). It may still be too difficult to deposit metal into these contact holes with adequate step coverage. A number of techniques have been studied to improve the coverage of Metal-2 into the vias (Chap. 15). In Fig. 16-17d, the step coverage of Metal-1 into

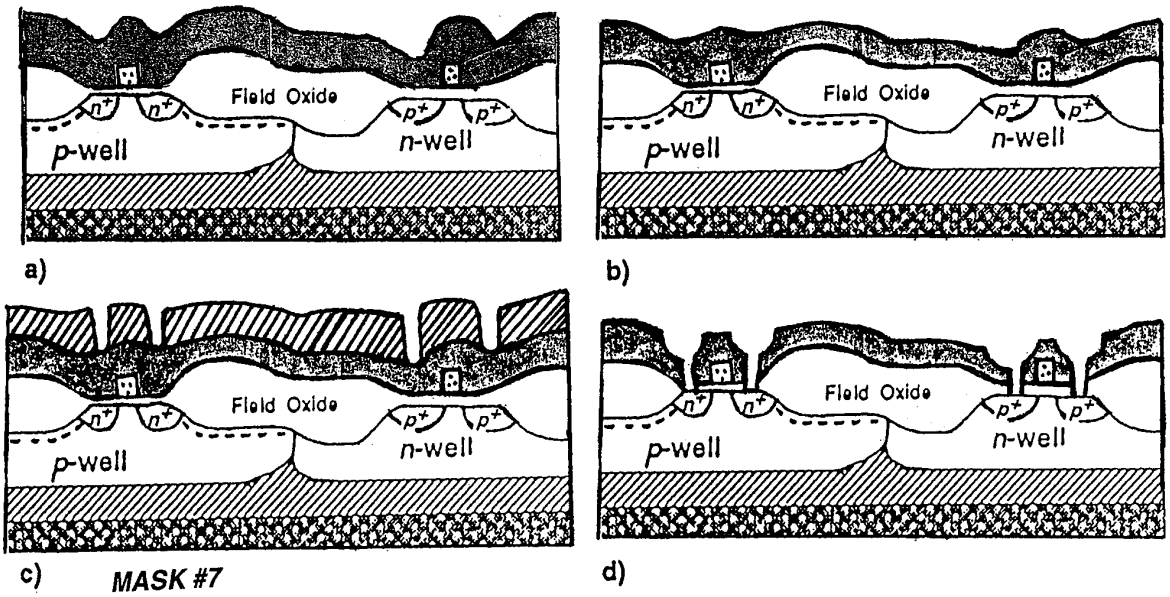


Fig. 16-17 (a) Deposit interlevel dielectric BPSG layer (ILD) by CVD; (b) flow BPSG. (c) use *Mask #7* to define contact holes; (d) etch contact holes with "wine glass" etch and strip resist.

the contact hole is facilitated by the use of a *wine-glass* etch process. The contact hole is formed with a two-step via-etch process (i.e., using an isotropic etch process to etch half of the ILD thickness, and an anisotropic etch step to etch the remaining half). Reflow may still be carried out after such a two-step etch.

16.2.9 Metal Deposition and Patterning: After the contacts have been opened, the metallization layer is deposited. Because the metal layer is highly conductive, it is used whenever possible to interconnect circuit elements and to carry large amounts of supply current. The metal interconnect lines that are fabricated must have sufficient thickness, width, and step coverage to keep the current density in each line below the value that could produce electromigration failure (Vol. 2, Chap. 4). In addition, the space between adjacent metal lines must be kept large enough that the lines never touch, even under worst-case process variations.

Although evaporation was the method employed to deposit Al in the early days of MOS, it has generally been replaced by sputtering. The change was made because Al alloys with tightly controlled compositions became the materials of choice for the metal layer. Sputtering allows alloys to be deposited with much better compositional fidelity (see Chap. 11).

The metal alloy eventually chosen for NMOS was Al:1wt% Si. Silicon is added to the aluminum film to prevent spiking of the contacts during subsequent annealing steps (see Chap. 11). In CMOS, however, such Al:Si alloys were replaced with a metallization system that includes a diffusion barrier (typically Ti/TiN, see Chap. 15), and Al:Cu alloys for the main interconnect layer (Fig. 16-18a). Dry etching is used to pattern the interconnect films using *Mask #8* (Fig. 16-18b).

If a single level of metal is used in the CMOS process, a sintering step occurs after the metal has been patterned. This step brings the metal and the n^+ and p^+ regions in the silicon into intimate contact. Such intimate contact between the metal and the heavily doped Si regions

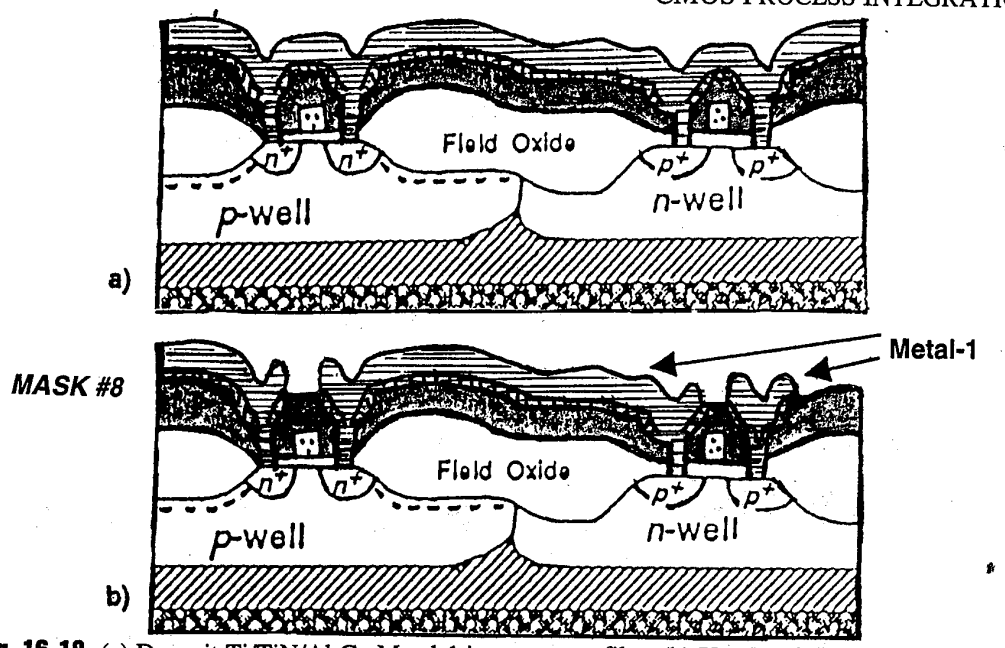


Fig. 16-18 (a) Deposit Ti/TiN/Al:Cu Metal-1 interconnect film. (b) Use *Mask #8* to define Metal-1 pattern and then etch the metal with a dry-etch process.

establishes a low-resistance ohmic contact. The annealing process exposes the wafer to a 375–500°C temperature in an H_2, or $N_2 + H_2$ (5%) ambient for about 30 minutes. This step is also used as the annealing process for reducing the interface trap density in the gate oxide that was introduced by earlier processing steps (see Chap. 8).

16.2.10 Intermetal Dielectric Deposition and Via Patterning: If a two-level-metal interconnect structure is used, an *intermetal dielectric* (*IMD*) must be deposited to electrically isolate the Metal-1 layer from the Metal-2 layer (Fig. 16-19a). A variety of CVD processes for depositing such dielectric films have been developed (see Chap. 6).

Deposition of this layer may make the wafer topography once again too severe to allow Metal-2 films to be deposited with adequate step coverage. One of the planarization processes discussed in Chap. 15 may need to be used to overcome this difficulty.

If such techniques are successfully implemented, the wafer topography will be relatively planar, and any steps on its surface will be gently sloped rather than severely vertical. It is also necessary to open vias in the intermetal dielectric layer so an electrical connection can be established between Metal-2 and Metal-1 at desired locations (*Mask #9*). Metal-2 must be deposited with adequate step coverage into these vias, but reflow is not possible because Al is present on the wafer (Metal-1). Thus, in this example, a champagne glass etch process is again used to etch the vias (Fig. 16-19b).

16.2.11 Metal 2 Deposition and Patterning.: The processing issues of depositing and patterning (*Mask #10*) the Metal-2 layer (Fig. 16-19c) are discussed in Chaps. 11, 14, and 15.

16.2.12 Passivation Layer and Pad Mask: Finally, a *passivation* (or *overcoat*) layer, such as CVD PSG or plasma-enhanced CVD silicon nitride (or both), is put down onto the wafer surface (Fig. 16-19d). This layer seals the device structures on the wafer, protecting them from contaminants and moisture. It also serves as a scratch protection layer.

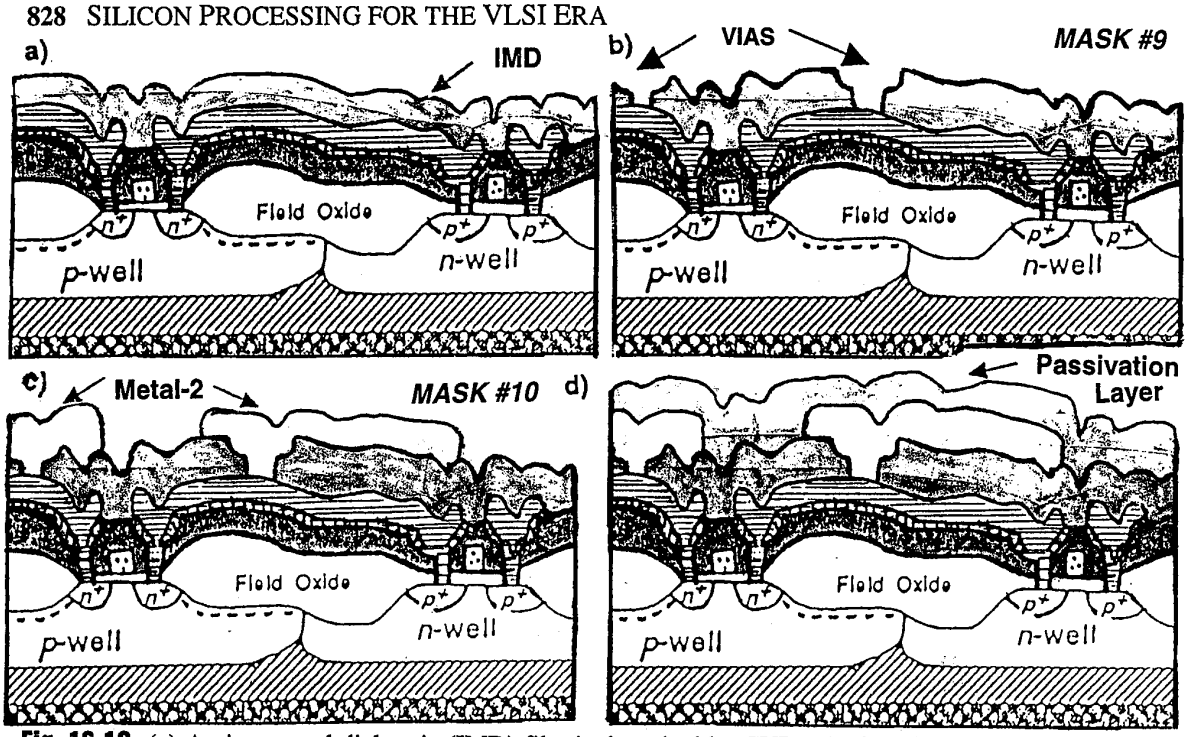


Fig. 16-19 (a) An intermetal dielectric (IMD) film is deposited by CVD. A planarization process such as resist etchback, or spin-on-glass, may be used to smooth this surface. (b) Vias are defined in the IMD film with *Mask #9* and an etch step. (c) Metal-2 is deposited and defined with *Mask #10* and a metal etch process. (d) A final passivation layer is deposited on the wafer by CVD.

Openings are etched into this layer so that a set of special metallization patterns under the passivation layer is exposed. These metal patterns are normally located in the periphery of the circuit and are called *bonding pads* (Fig. 16-20). Bonding pads are typically about 100 x 100 μm in size and are separated by a space of 50 to 100 μm. Wires are connected (bonded) to the metal of the bonding pads and are then bonded to the chip package. In this way connections are estab-

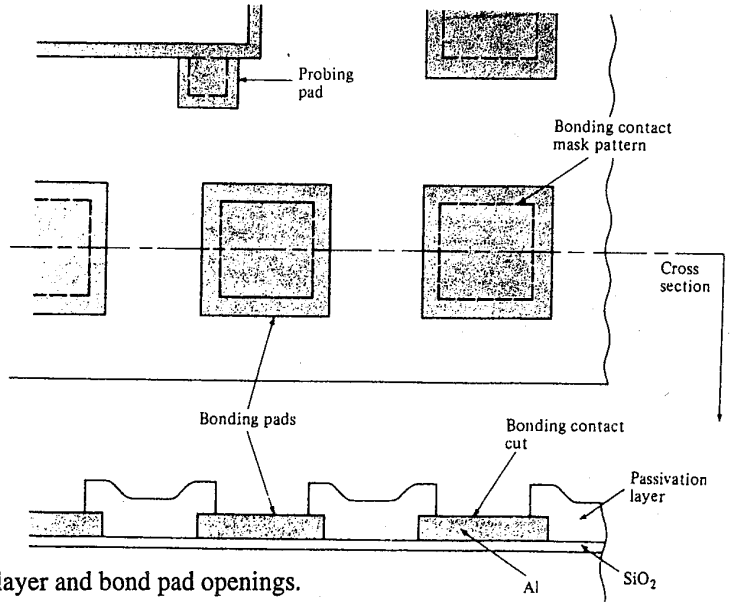


Fig. 16-20 Passivation layer and bond pad openings.

lished from the chip to the package leads (see Chap. 17 for details on chip bonding).

The bond pad openings are created by patterning the passivation layer with *Mask #11*. If a PSG layer is used, the phosphorus (2-6 wt%) in the glass not only causes it to act as a getter for Na but also prevents the film from cracking. Care must be taken to ensure that not more than 6% phosphorus is incorporated into the PSG, as this can cause corrosion of the underlying metal if moisture enters the circuit package (see Chap. 11). When silicon nitride is used, care must be taken to ensure that the deposited nitride film exhibits low stress (either tensile or compressive), so it will not crack. Cracking would compromise the sealing capability of the film.

16.3 PROCESS FLOW FOR 0.25 μm CMOS

An example process flow for 0.25 μm CMOS (and perhaps for the 0.18 μm generation as well) will be presented in this section. (At this book's writing, the process flows for 0.25 μm CMOS, and below, were not yet "standardized," and many possible variations of the flow postulated here existed.) This flow is "compiled" from the literature, where a number of 0.25 μm CMOS technologies are reported, including those in Refs. 11, 12, 13, 14. Since the literature does not provide all the details of such flows, some of the information here is necessarily conjectural. Despite these issues, the 0.25 μm flow presented here should be a reasonable representation of those found in actual use.

A number of general characteristics of 0.25 μm CMOS process technology should be noted. First, the lithography technology used for the critical masking layers is DUV optical lithography, which is based on 248-nm KrF laser light sources (and *i*-line lithography perhaps being used for the non-critical layers). Second, the power supply voltages of the systems using 0.25 μm CMOS is 2.5 V, and for 0.18 μm CMOS is estimated to be 1.8 V. This is a deviation from the 5-V power supply used from the mid-1970s's and down to when CMOS reached 0.65 μm. However, when CMOS reached 0.5 μm, the gate oxide thickness had been scaled so much that the electric field across the gate oxide exceeded the value that would permit reliable operation for the designed life of the systems (e.g., 10 years). To ensure adequate device lifetime, the power supply voltage had to be changed for 0.5-μm CMOS to 3.3 V, and again at 0.25-μm CMOS to 2.5 V. (This is an example of how the limits of process technology impact the electronic systems specifications.) Third, the number of masking layers is likely to exceed 20, especially if 5 or more interconnect levels are needed. Finally, 0.25-μm CMOS was initially being manufactured on 200-mm wafers. It is likely that 0.18-μm CMOS will also use such wafer sizes. However, it appears that 0.15-μm CMOS, or 0.12-μm CMOS will enter large scale production on 300-mm wafers.

As mentioned in the introduction, there are a number of process modules that had to be fundamentally changed as CMOS technology was scaled below about 0.35 μm. Figure 16-21 demonstrates this quite clearly.[13] The order of carrying out some of the steps has also changed. Why these changes in the process flow were necessary will be indicated.

16.3.1 Starting Material: The starting material for deep-submicron CMOS is predicted to shift more to p-epi-on-p^+ wafers (Fig. 16-7b), because the gate-oxide-integrity is better on epi than on bulk wafers. However, the higher cost of epi compared to bulk wafers remains a concern, especially for such low-price commodity chips as DRAMs. Less expensive alternatives to epi are being studied. One process that shows great promise is the implantation of a high dose of

Generation	0.5um	0.35um	0.25um	0.18um	0.15um	0.13um
Isolation	LOCOS		STI			
Well	Diffused		Retrograde			
Gate	WiSix n+ Polycide		Dual Poly (n+/p+) Salicide			
Gate Oxide	Single Oxide		Dual/Triple Oxide => Oxynitride			
Gate Litho	I-line		248nm, Hi NA		PSM	193nm
Silicide	None	TiSix	TiSix => CoSix			
Contact/Via	Etchback		CMP			
BEOL Dielectric	SOG	Oxide	Low-k			
Metal	Al		Al =>Cu Damascene			

Fig. 16-21 The process modules used for successive submicron-CMOS generations.[13] © IEEE (1998).

boron deep below the surface of a bulk wafer, using high-energy ion implantation. This step creates a heavily doped B buried layer that acts in the same way as the heavily doped substrate of the epi wafer.[16] Good latchup protection and GOI are reportedly achieved by this approach. The cost for carrying out this step is 10–20% that of depositing an epi layer on a bulk wafer.

16.3.2 Formation of the Shallow-Trench-Isolation Structure: The isolation structures are formed next. Note that not only is a different type of isolation structure formed (*shallow trench isolation*, or *STI*) instead of LOCOS, but it is formed prior to the well structures. STI replaces LOCOS because STI has no bird's beak and provides a planar surface for further processing.[15]

The sequence of steps for forming STI structures begins with the growth of a pad oxide, followed by an LPCVD nitride layer. Then resist is applied and *Mask #1* is used to pattern the STI trench openings. The nitride and pad oxide are etched first. Then the trench is anisotropically etched (Fig. 16-22a) to a depth of 400–500 nm (e.g., using an etch gas mixture of $HBr/Cl_2/O_2$). The trench etching process should yield smooth trench sidewalls with angles of 70–85°, rounded bottom-corners (to minimize stress), and a residue-free silicon surface after etch (see Fig. 16-22f). After the trenches are etched the resist is stripped and a thin thermal oxide is grown on the trench walls (Fig. 16-22b). Next, a CVD dielectric film is used to fill the trench. This layer also covers the areas of the wafer where the nitride remains (Fig. 16-22c). A CMP step is used to polish back the dielectric layer until the nitride is reached (the nitride acts as a CMP-stop layer, Fig. 16-22d). The dielectric material is densified at about 900°C. Then the nitride is stripped, leaving the STI structure in place (Fig. 16-22e).

16.3.3 Formation of Retrograde Wells and Carrying Out of V_T-Adjust Implants: The well formation sequence occurs next. In addition, the V_T-adjust implants and the punchthrough implants are often carried out on the same ion implanter and during the same pumpdown as the well-formation implants. This process of performing multiple implants during the same pumpdown is termed a *chained implant.* The well formation step employs high-energy ion implantation to locate the well dopant peak ~1–2 μm below the Si surface. The implanted dopants will

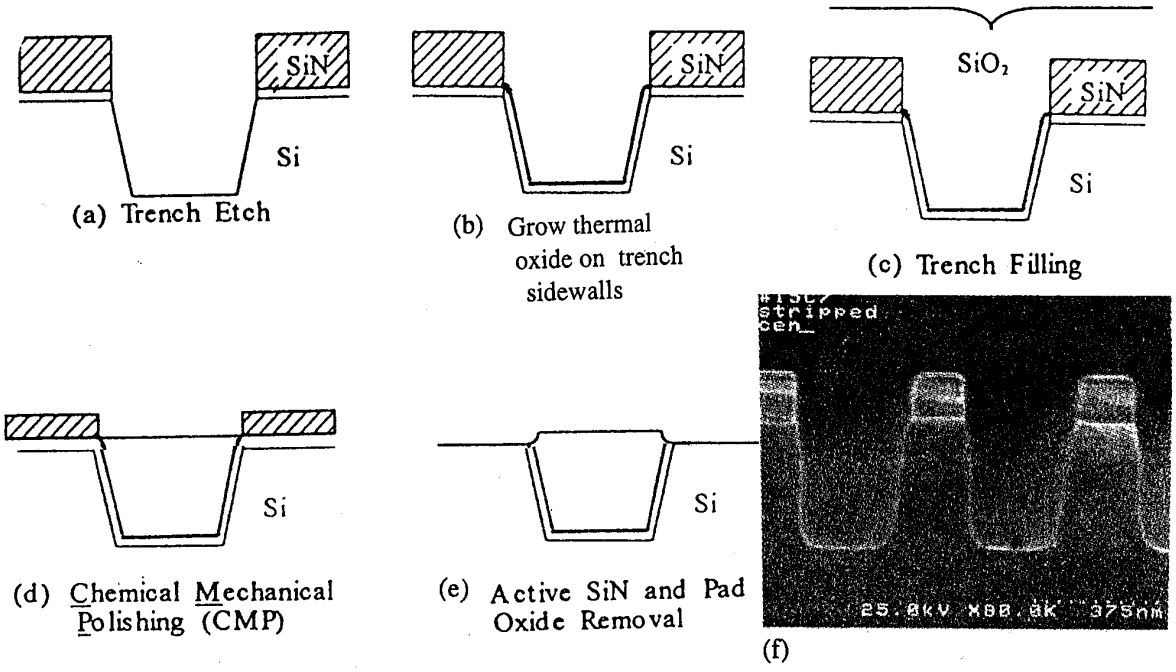


Fig. 16-22 Process sequence used to form shallow-trench isolation structures (STI): (a) etch trench in silicon substrate; (b) grow thermal oxide trench liner to improve trench Si/SiO_2 interface (and for trench corner rounding); (c) fill trench with CVD oxide; (d) use CMP to etch back the oxide to the nitride layer; (e) Strip nitride to leave the STI structure; (f) SEM of trench profile with 87° trench sidewall slope, and top and bottom trench corner rounding. Courtesy of Lam Research Corp.

have a *retrograde* concentration profile toward the surface from this peak (Fig. 16-23). Such a doping profile has several device performance advantages (discussed in Vol. 3). Since the dopants are placed deeply enough by the implant itself, there is no need for the long, high-temperature drive-in step required previously. Instead, only a lower-temperature anneal (carried out in a furnace at 900°C for 30 min), or a very-short, high-temperature RTP anneal (1100°C for 30 sec), is required to activate the implanted dopants. However, the high-energy ion beam used in the implantation process (500 keV–1 MeV) requires the use of a high-energy ion implanter and very thick resist films (~5 μm). Such thick resists are needed to prevent high-energy ions from reaching the Si surface in regions covered by the resist.

The well-formation process is begun by growing a layer of sacrificial oxide on the wafer. After applying the thick resist layer, *Mask #2* is applied to open the patterns for where the *n*-wells will exist. Then a high-energy P implant is carried out (700–850 keV, $1x10^{13}$ cm^{-2}). The peak of the P implant is ~1 μm below the Si surface (Fig. 16-24b). Both the V_T-adjust implant for the PMOSFETs (using P, since the dual-doped poly will allow the use of surface-channel PMOS devices – see Vol. 3), and the punchthrough implant are carried out in this implanter in the same pumpdown. Then the resist is stripped, and a new coat of thick resist is applied. *Mask #3* is then used to pattern the *p*-well openings. A high-energy B implant is then carried out (450–500 keV, $1x10^{13}$ cm^{-2}, Fig. 16-24c). A retrograde profile is again created, with the B peak concentration residing at 1-2 μm below the surface (Fig. 16-25). The V_T-adjust implant and the

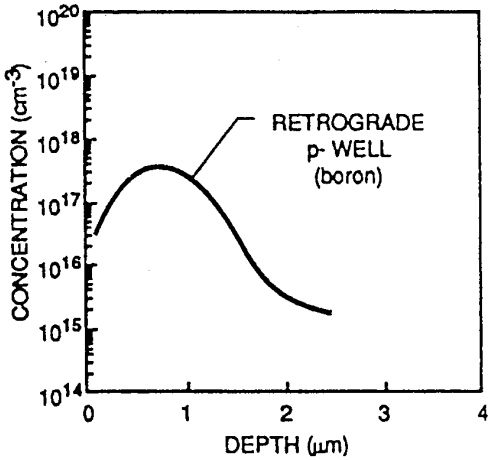


Fig. 16-23 Example of the doping profile of a retrograde *p*-well (boron) implant (400 keV).

punchthrough implant for the NMOSFETs (now using B) are also carried out during this "chained-implantation" procedure.

As discussed in Chap. 10, V_T-adjust implants with retrograde doping profiles have recently been introduced. A high-mass dopant of the same conductivity type is substituted for the ordinary dopant (e.g., indium for boron in NMOSFETs, antimony for phosphorus in PMOSFETs). The higher mass of the dopant reduces its diffusivity, which significantly decreases dopant redistribution during subsequent processes. Thus, the desired as-implanted profile is kept more intact. Such high-mass dopant V_T-adjust implants end up having a reduced

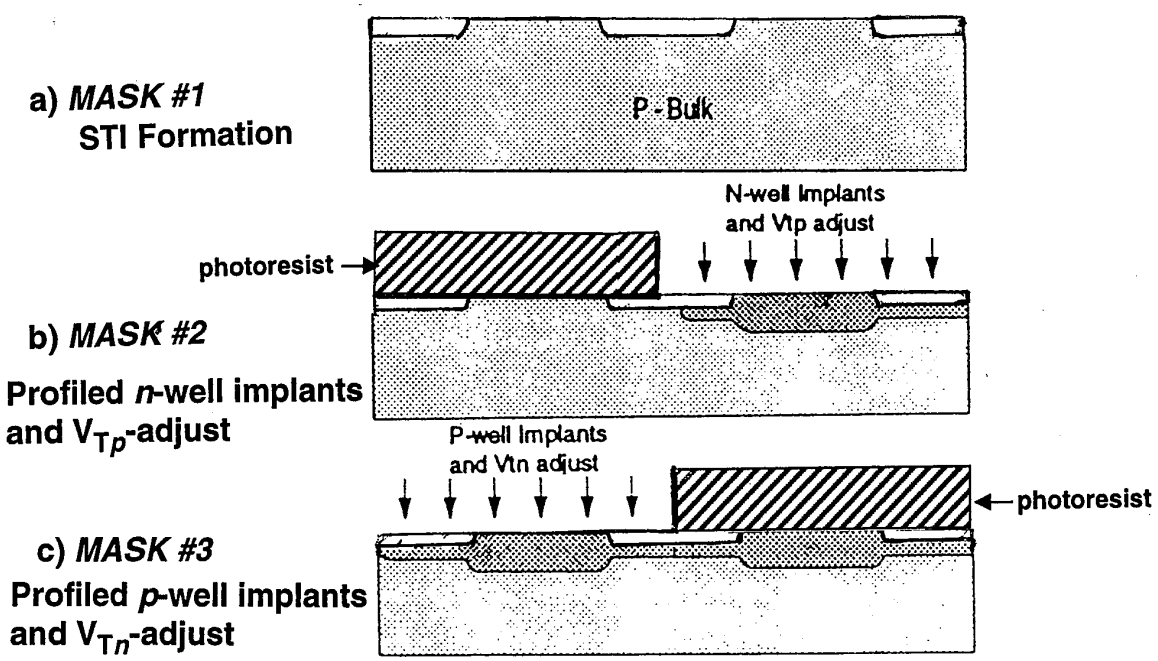


Fig. 16-24 (a) *Mask #1* patterns the shallow-trench isolation features; (b) *Mask #2* defines the *n*-well regions. A high-energy phosphorus (P) implant creates the retrograde *n*-wells, and a lower energy P (or Sb) implant is used to adjust V_{Tp}; (c) *Mask #3* defines the *p*-well regions. A high-energy boron (B) implant creates the retrograde *p*-wells, and a lower energy B (or In) implant is used to adjust V_{Tn}.

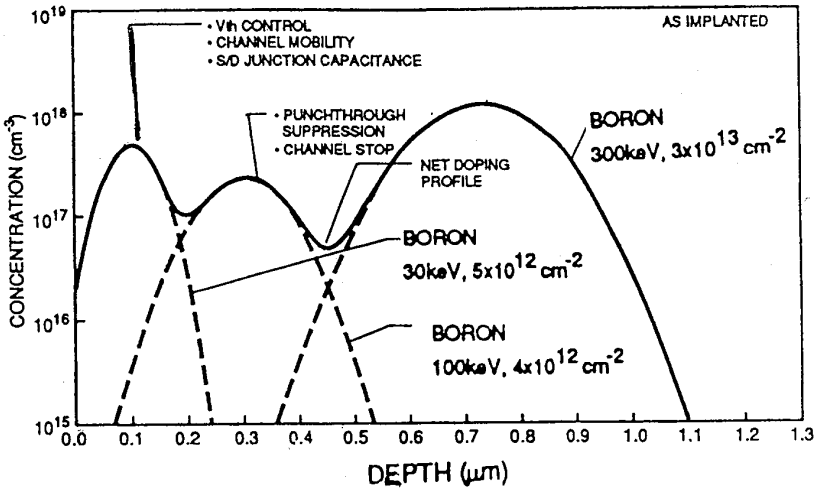


Fig. 16-25 Doping profile in the *p*-well showing the as-implanted: (a) retrograde *p*-well implant (B, 300keV); (b) the anti-punchthrough implant (B, 100 keV); and (c) the V_{Tn} implant (B, 30 keV).

doping concentration (~$5x10^{16}$ cm^{-2}) at the interface while at the doping peak (about 80 nm below the oxide interface) the concentration is ~$1x10^{18}$ cm^{-2}. These profiles have been named *super-steep retrograde* (SSR) channel profiles (Chap. 10, Fig. 10-39). Such SSR profiles allow deep-submicron MOSFETs to be fabricated in which short-channel V_T-rolloff and drain-induced barrier lowering (DIBL) effects are significantly suppressed (see Vol. 3).

16.3.4 Formation of the Gate Oxide: The gate oxides of deep-submicron CMOS will continue to scale, with gate oxides for logic CMOS predicted to get thinner than 30 nm (see Chap. 8). The gate oxide, however, will have to prevent boron (from the p^+-poly) from penetrating into the Si substrate. The addition of nitrogen to the gate oxide has been reported to combat this problem, and so gate dielectric layers may be more like oxynitrides than SiO_2.[13] Different techniques have been investigated for incorporating the nitrogen into such films (see Chap. 8), including annealing the as-grown SiO_2 film in N_2O[18] or NO,[13] implanting N into Si and then growing SiO_2,[19] or exposing the SiO_2 to a high-density nitrogen discharge and then performing a post-nitridation anneal.[20]

16.3.5 Formation of the Gate-Stack Structures: Dual-doped polysilicon (i.e., n^+-doping of NMOS poly gates and p^+-doping of PMOS poly gates) is needed for 0.25-μm CMOS technologies (and smaller). This permits both the NMOS and PMOS devices to be fabricated as surface channel FETs (see Vol. 3), a necessity for being able to control short-channel effects.

The polysilicon layer is deposited undoped by LPCVD, to a thickness of 300-400 nm. But the selective doping of the appropriate poly structures is not carried out immediately. (Instead, it is performed later, simultaneously with the formation of the heavily doped regions of the source and drain junctions by ion implantation.) *Mask #4* is used to pattern the poly layer, and the gate structure is etched with an anisotropic dry-etch step (Fig. 16-26a).

16.3.6 Formation of the Source/Drain Junctions: After the poly gate structure is etched, the formation of the source/drain regions is undertaken. These regions in the substrate have to provide a low-resistance path from the contact to the channel edge. They are formed in two steps. In the

first, so-called *source/drain extensions* (SDEs) are formed. In the previous flow these parts of the S/Ds were termed the *lightly doped areas* of the source and drain. Since they were formed by an ion implantation step, the doping in these regions is characterized in terms of the implant dose (typically $3x10^{13}$–$3x10^{14}$ cm^{-2}). In deep-submicron FETs, the doping concentration is increased in the SDE areas above the values used in the first process flow presented earlier. If the SDEs are also formed by ion implantation, the doping in them can again be characterized by the implant dose, which in the present process flow is raised to about $1x10^{15}$ cm^{-2}.

However, the primary function of the SDE is no longer to act as a region that reduces the local electric field for the purpose of reducing hot-carrier effects (because this problem is reduced by the smaller power supply voltages used). Instead, the SDE has two other principle roles that must simultaneously be met: 1) to provide a low resistance path from the contact to the channel edge; and 2) to have extremely shallow junction depths (e.g., 50–100 nm for 0.25 μm CMOS, and 36–72 nm for 0.18 μm CMOS), so that short-channel effects can be effectively suppressed. It is imperative that a very high doping concentration be established within such a shallow structure so that the parasitic series resistance of the SDE becomes as small as possible. Unfortunately, forming a shallow junction at the wafer surface with a high doping concentration is a challenging task, especially for B-doped regions. Several techniques are being developed to try to achieve the above goals. Ion implantation at very low energies followed by an RTP anneal has been the most widely used process for forming these drain extension regions (see Sect. 10.6). Other techniques being investigated for this process include *gas immersion laser doping* ([*GILD*] - see Chap. 8), *plasma immersion doping* (see Chap. 10), and out-diffusion from a layer of deposited doped oxides (e.g., B from a layer of BSG), as described in Refs. 21 and 22.

Regardless of the technique being used, two masks are needed (i.e., *Mask #5* and *Mask #6,* shown in Fig. 16-26b and 16-26c). This allows the NMOS and PMOS source/drain extension (SDE) regions to be selectively doped. If ion implantation is being used, a "halo implant" (to control punch-through) can also be performed within the same pumpdown as the SDE implant step. This "halo implant" is also self-aligned with the channel edge: and it puts dopants at a depth beneath those dopants implanted into the SDE regions. "Halo" implants can be performed either with a 0° tilt angle or a high (30°) tilt angle (see Chap. 10).

The spacer formation is carried out next. Typically LPCVD silicon nitride is used as the spacer material. The deposited nitride layer is then anisotropically etched-back, to form the spacer. At that point the heavily doped, deeper part of the source/drain regions are formed. This is done with ion implantation, using two masks, so the NMOS S/D regions are formed with one implant, and the PMOS S/D regions with the other (*Mask #7* and *Mask #8,* as shown in Figs. 16-27a and 16-27b). Doses of $1–5x10^{15}$ cm^{-2} are used for this step. Note these regions have deeper junction depths than the SDEs (Fig. 16-28). The poly gates are also doped simultaneously by these implants. A single rapid-thermal annealing step is used to activate the implanted dopants in both the SDE and the heavily doped S/D regions.

16.3.7 Salicide Formation: A *salicide* (self-aligned silicide) structure is fabricated next. It consists of a metal silicide formed atop the lines of polysilicon that make up the gates and the local interconnects (akin to the polycide structure described in the first flow). However, in this case it is also formed on the source-drain regions. The silicide on the source/drain regions

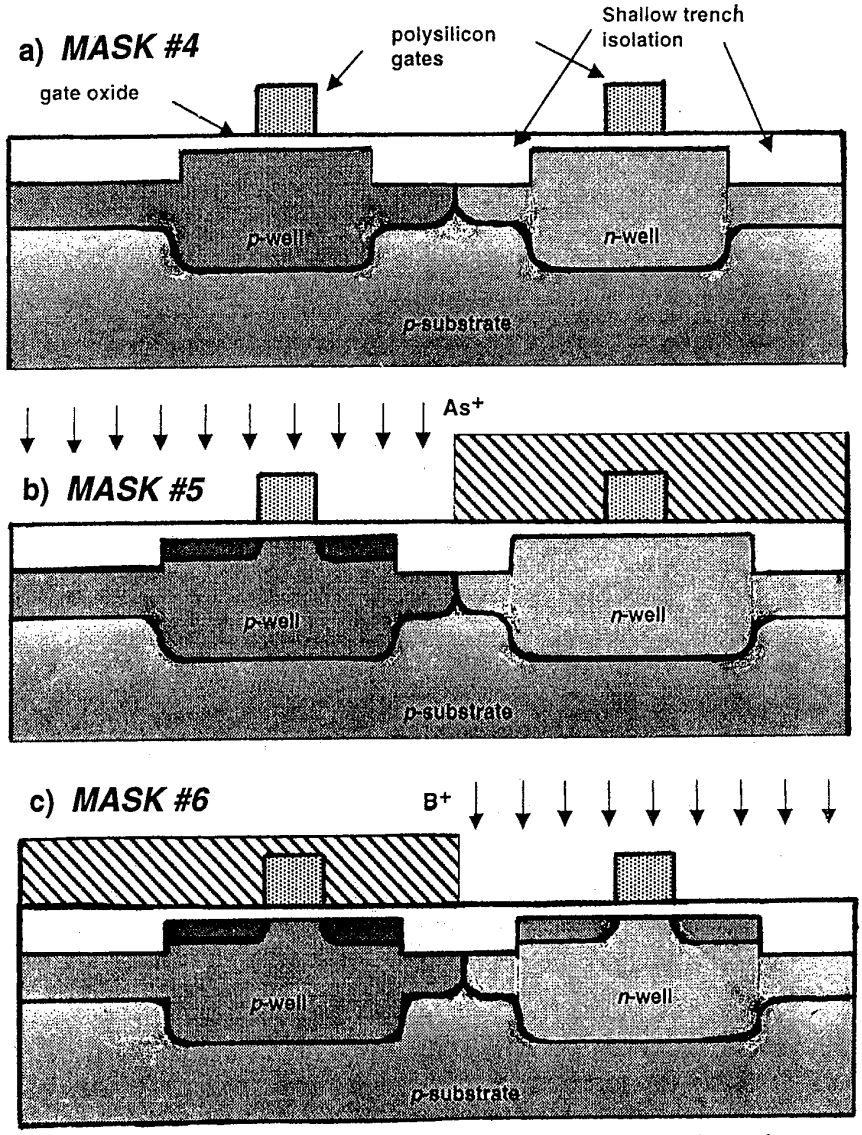


Fig. 16-26 (a) *Mask #4* is used to define the polysilicon gates; (b) *Mask #5* is used to cover the *n*-well regions, so that the NMOS SDEs can be selectively doped with an As implant; (c) *Mask #6* is used to cover the *p*-well regions, so that the PMOS SDEs can be selectively doped with a B implant.

reduces the sheet resistance of the path between the metal contact and the channel edge. The spacers formed on the sidewalls of the poly lines (in the course of fabricating the two parts of the source/drain regions) are now used once again in the salicide formation sequence. Two metals have been used to form such salicides, Ti and Co. They both form silicides with resistivities of 10–15 $\mu\Omega$-cm. Ti-salicide structures were the first ones implemented, but Co has become more popular as the dimensions have been scaled toward 0.18 μm.[23]

The sequence to form the salicide begins with the removal of any oxide on the source/drain regions and the poly lines (Fig. 16-29). Then, Ti or Co is blanket deposited by PVD (sputtering) onto the wafer. A lower temperature thermal step (~650°C) is next used to react the metal and the Si exposed on the substrate and the poly lines. (Use of an RTP step, in which oxygen is

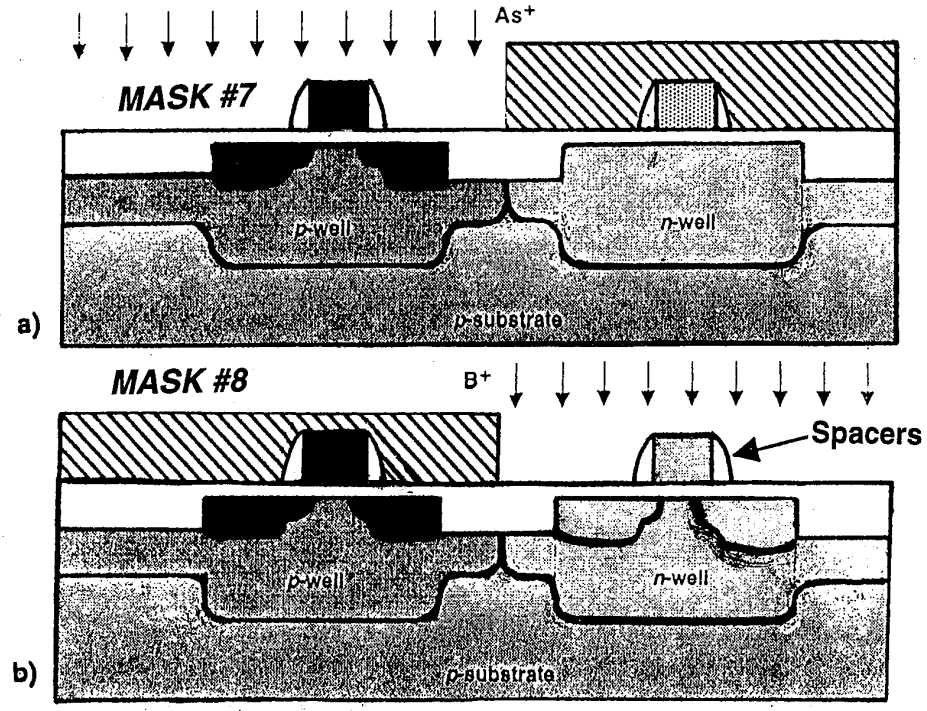


Fig. 16-27 (a) *Mask #7* is used to cover the *p*-well regions, so that the heavily doped NMOS S/D regions can be selectively doped with an As implant; (b) *Mask #8* is used to cover the *n*-well regions, so that the heavily doped PMOS S/D regions can be selectively doped with a B implant.

excluded from the ambient during the anneal, is a key aspect of this process.) This forms a metastable C49 phase of $TiSi_2$. The Ti that lies on either the field oxide or the nitride spacers is, in principle, not converted to $TiSi_2$ during this step. The C49, however, is a metastable, high resistivity material, not yet suitable for the interconnection task. The wafer is then etched in a bath of H_2O:H_2O_2:NH_4OH that selectively removes only the unreacted Ti, leaving behind the C49 $TiSi_2$. A second thermal step using RTP is carried out, this time at a temperature greater than 750°C. This converts the C49-phase $TiSi_2$ to the more stable, lower resistivity (~10 $\mu\Omega$-cm) C54-phase $TiSi_2$. The thermal steps are carried out by RTP and in an N_2 atmosphere. Details of this sequence are found in Vol. 2, Sect. 3.9.1.1.

When the polysilicon lines become smaller than about 0.5 μm, the formation of $TiSi_2$ becomes difficult. The transformation of the C49 phase to the C54 phase becomes the limiting factor. On such small lines, the density of C54 grains is low, and only a small fraction of the $TiSi_2$ consists of the C54 phase, because it can grow only from the few nuclei available. This problem could be solved by heating the $TiSi_2$ above 750°C to accelerate the nucleation of the C54 phase. Unfortunately, thin $TiSi_2$ agglomerates at high temperatures, rendering this solution unworkable. Since $CoSi_2$ does not exhibit this problem, it has started to replace $TiSi_2$. $CoSi_2$ transformation occurs at a lower temperature (600 to 700°C). Thus, full formation of low-resistance $CoSi_2$ is achieved before the silicide agglomerates. The Co is also transformed to $CoSi_2$ by a two-step process. In the first, Co and Si are reacted to form CoSi at 450°C. After selectively etching the unreacted Co away, a second thermal step at 700°C reacts the CoSi with more Si to form $CoSi_2$. An even lower resistance cobalt-salicide structure can be obtained by

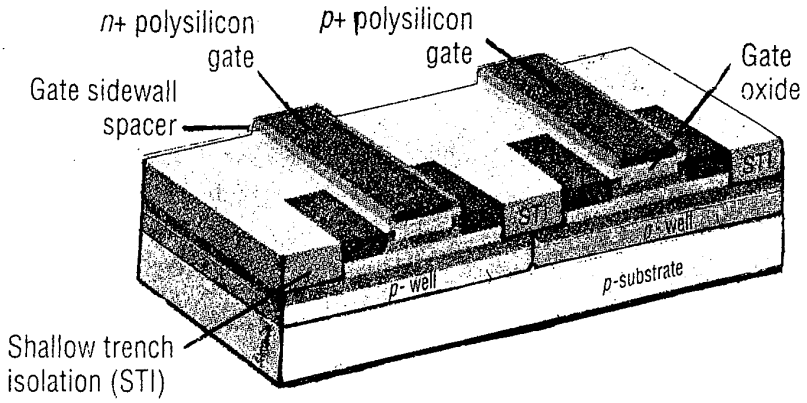


Fig. 16-28 Perspective drawing of a 0.25-μm CMOS structure just prior to the salicide formation process. It depicts the STI isolation structures, the wells, the gates, and the S/D regions. Courtesy of Eaton Corp.

depositing a 20-nm cap layer of TiN onto the Co film before the first thermal step. This TiN cap cuts the sheet resistance from ~10 Ω/sq to about 5 Ω/sq by protecting the Co from oxidizing during the first silicidation anneal.[25]

16.3.8 Premetal Dielectric Deposition and Contact Formation: The premetal dielectric layer is deposited next, much as was done in the first flow. To fill the narrow gaps between adjacent gate structures without voids, a number of dielectric processes may be used including such "sandwich" films as: (1) (PECVD-TEOS/Ozone-TEOS/PECVD-TEOS); or (2) (PECVD-TEOS/SOG/PECVD-TEOS), or an HDP-CVD dielectric film.[26] However, for planarization purposes, instead of performing a flow and reflow, this layer is polished by CMP to provide a flat topography. *Mask #9* is used to pattern the contact hole openings in the premetal dielectric layer. The contact dry-etch process produces vertical-walled contacts, which must be filled (Fig. 16-30). These contact holes can be filled in various ways, including with CVD-W. In that case, a liner film is deposited first (e.g., TiN deposited by sputtering or CVD).

16.3.9 Interconnect Structure Formation: The technologies for fabricating the interconnect structures used in 0.25-μm CMOS were discussed in Chap. 15. A variety of possible materials and processing approaches can be used to implement a particular multilevel interconnect structure. Figures 16-31 and 15-61 show examples of the kinds of multilevel interconnect structures that are used in deep-submicron ICs.

A number of different materials and technologies can be incorporated when establishing the fabrication sequence for such structures. The exact combination chosen will depend on the IC manufacturer and the application that the chip will be serving. The following are some materials and processes that are used to manufacture such structures: (a) Ti/TiN diffusion barriers; (b) Al:Cu films deposited by sputtering; (c) TaN/Cu seed layers (deposited by sputtering) for Cu interconnects; (d) electroplated Cu films; (e) CMP of the dielectric and metal layers; (f) W-plugs (deposited by CVD); (g) low-k intermetal dielectrics; (h) second- or third-generation spin on-glasses (SOG) for gap filling; (i) use of high-density plasmas (HDP) to deposit gap-filling

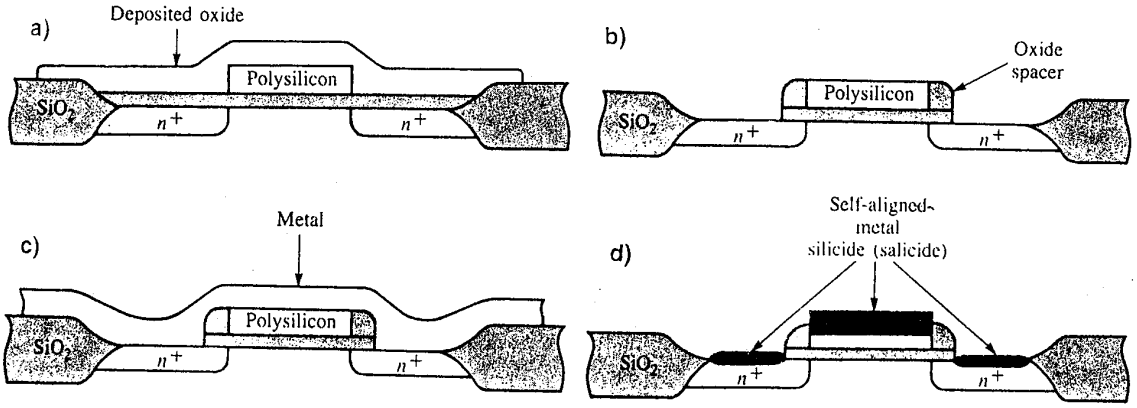


Fig. 16-29 *Salicide* (self-aligned silicide) processing steps.[24] (a) Oxide (or nitride) is deposited following polysilicon definition; (b) structure after etching back, leaving sidewall spacers on the polysilicon; (c) a metal film (Ti or Co) is deposited over the structure and heated. Where the metal is in contact with the silicon substrate, or the polysilicon gate, silicides form; (d) the unreacted metal is selectively etched away, leaving silicide automatically aligned to the gate and source/drain regions. (© 1984 IEEE).

dielectric layers; and (j) single damascene or dual-damascene interconnect structures. Some interconnect structures may use Al films for some interconnect levels, and Cu films for others.[27]

An example of a 0.18-μm CMOS technology using six levels of metals was described in 1998.[28] Both contacts and vias were filled with CVD-W plugs (preceded in each case with a PVD-Ti/CVD-TiN liner film). To make the plugs, the CVD-W was blanket-deposited and then etched back by CMP. The metal interconnect lines consisted of multi-layer films of Ti/TiN/-Al:Cu/TiN to achieve low resistance, good electromigration resistance, and low via resistance. The intermetal dielectric is HDP oxide doped with 5.5% fluorine (which reduces the dielectric constant to 3.5 from the 4.2 for undoped-CVD-SiO_2 films) and is planarized by CMP.

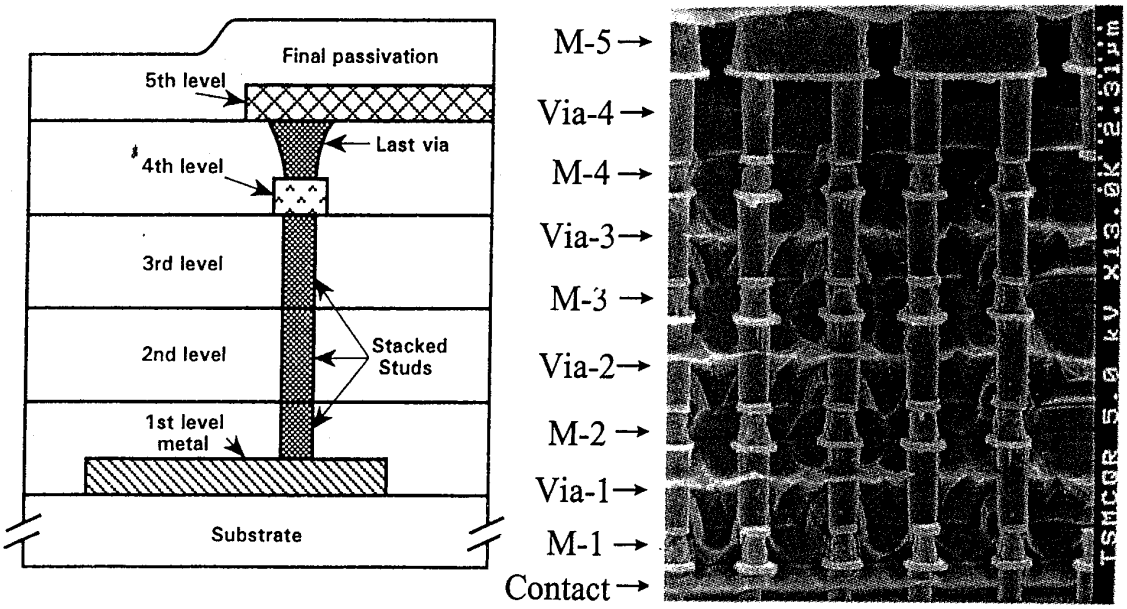


Fig. 16-30 Stacked vias in a 5-metal interconnect, showing filled vertical-walled vias.[15] (© 1998 IEEE).

REFERENCES

1. J.Y. Chen, *CMOS Devices and Technology for VLSI*, Prentice-Hall, Englewood Cliffs, N.J., 1990.
2. R.S. Muller and T.I. Kamins, *Device Electronics for Integrated Circuits*, 2nd. Ed., New York, John Wiley & Sons, 1986.
3. Y.P. Tsividis, *Operation and Modeling of the MOS Transistor,"* McGraw-Hill, New York, 1987.
4. F.M. Wanlass and C.T. Sah, *IEEE Internat. Solid-State Circuits Conf.*, February 1963.
5. W.C. Till and J.T. Luxton, *Integrated Circuits: Materials, Devices, and Fabrication,"* Prentice-Hall, Englewood Cliffs, N.J., 1982.
6. R. Chwang "CHMOS: An *n*-Well Bulk CMOS Technology for VLSI," *VLSI Design*, 4th Q, 1981, p. 42.
7. B. Hoeflinger and G. Zimmer, in J. Carrol Ed., *Solid-State Devices 1980, Institute of Physics Conf. Series 57*, from the 10th European Solid-State Device Research Conference.
8. L.C. Parrilllo *et al.*, "Twin-Tub CMOS -II," *Tech. Dig. IEDM*, 1982, p. 706.
9. E. Kooi and J.A. Appels, in *Semiconductor Silicon, 1973,* H.R. Huff and R. Burgess, Eds. P. 860, The Electrochemical Society Symposium Series, Princeton, NJ (1973).
10. T. Ohzone *et al., IEEE Trans. Electron Dev.*, **ED-32**, 1789, (1980).
11. W.S. Chang *et al.*, "A High-Performance 0.25 μm CMOS Technology: I–Design and Characterization," *IEEE Trans. Electron Dev.*, **Vol. 39**, April 1992, p. 959.
12. W.S. Chang *et al.*, "A High-Performance 0.25 μm CMOS Technology: II – Technology," *IEEE Trans. Electron Dev.*, **Vol. 39**, April 1992, p. 967.
13. M.S.C. Luo *et al.*, "A 0.25 μm CMOS Technology with 45 Å NO-nitrided Oxide," *Tech. Dig. IEDM,* 1995, p. 691.
14. T.C. Holloway *et al.*, "0.18 μm CMOS Technology for High-Performance, Low Power, and RF Applications," *Dig. of Tech Papers, 1997 Symp. on VLSI Technology*, p. 13.
15. J.Y.C. Sun *et al.*, "Foundry Technology for the Next Decade," *Tech. Dig. IEDM* 1998, p. 321.
16. L. Rubin, "Buried Layer Substrates: Economically Enhancing Device Performance," *Solid State Technology*, June 1999, p. 95.
17. L. Peters, "Choices and Challenges for Shallow Trench Isolation," *Semiconductor International.,* April 1999, p. 69.
18. Y. Okada and P.J. Tobin, "Hot-Carrier Degradation of LDD MOSFETs with Gate Oxynitride Grown in N_2O," *IEEE Electron Dev. Letts.*, **EDL-15**, July 1994, p. 233.
19. C.T. Liu *et al.*, "High Performance 0.2 μm CMOS with 25Å Gate Oxide Grown on Nitrogen Implanted Silicon," *Tech. Dig. IEDM*, 1996, p. 499.
20. S.V. Hattangady *et al.*, "Ultrathin Nitrogen-Profile-Engineered Gate Dielectric Films," *Tech. Dig. IEDM,* 1996, p. 495.
21. P. Packan et al., "Modeling Solid Source Boron Diffusion for Advanced Transistor Applications," *Tech. Dig. IEDM,* 1998, p. 505.
22. J. Schmitz et al., "Ultra-Shallow Junction Formation by Outdiffusion From Implanted Oxide," *Tech. Dig. IEDM*, 1998, p. 1009.
23. J.A. Kiitl *et al.*, "Salicides and Alternative Technologies for Future ICs," *Solid State Technology,* June 1999, p. 81.
24. C.Y. Ting, "Silicide for Contacts and Interconnects," *Tech. Dig. IEDM*, 1984, p. 110.
25. T. Yamazaki *et al,*, "21 psec Switching 0.1 μm-CMOS at Room Temperature using High-Performance Co Salicide Process," *Tech. Dig. IEDM*, 1993, p. 906.
26. S. Nag *et al.*, "Low-Temperature Pre-Metal Dielectrics for Future ICs," *Solid State Technology,*

September 1998, p. 69.
27. M. Iragashi *et al.*, "The Best Combination of Aluminum and Copper Interconnects for a High Performance 0.18 μm CMOS Logic Device," *Tech. Dig. IEDM*, 1998, p. 829.
28. S. Yang *et al.*, "A High Performance 180-nm Generation Logic Technology," *Tech. Dig. IEDM*, 1998, p. 197.

PROBLEMS

1. Assume you were asked to design a MOS capacitor to act as a test structure for monitoring the threshold voltage of a silicon-gate NMOSFET. Formulate a detailed process flow of the fabrication steps needed to manufacture such an *n*-channel MOS capacitor.

2. Compare retrograde-well CMOS and conventional-well CMOS. What are the most important differences in the process technology and the finished device structure? List four advantages that retrograde well technology provides when compared to conventional-well CMOS technology.

3. (a) Why is the <100> orientation preferred for the Si wafers used to fabricate MOSFET ICs? (b) What are the disadvantages of using a field oxide that is too thin in CMOS ICs? (c) List three functions that the BPSG interlevel dielectric layer performs in CMOS ICs?

4. The following are some of the new technologies used in ULSI CMOS technologies (i.e., technologies in which the minimum feature size is 0.25 μm, or smaller. (a) shallow trench isolation; (b) dual-doped polysilicon; (c) $TiSi_2$ salicides; (d) shallow source/drain junctions; (e) copper interconnects; (f) low-*k* dielectrics. List at least two advantages and two problems associated with each of these technologies.

5. In an *n*-well 1 μm CMOS IC technology, the starting material is a *p*-type Si wafer with <100> orientation. The wafer may be either lightly doped bulk *p*-type Si, or heavily *p*-doped bulk Si with a lightly *p*-doped epi layer on its surface. In either case the resistivity of the Si near the surface is about 8 Ω-cm. The well doping in the *n*-well near the Si surface is $3x10^{16}/cm^3$. The gate material of the MOSFETs is a heavily *n*-doped polysilicon film. The source and drain regions of the NMOSFETs are formed with an As implant of $1x10^{16}$ As ions/cm^2 at 50 keV. A channel implant with $8x10^{11}$ boron ions/cm^2 is performed through a gate oxide that is 15 nm thick. (a) What would the threshold voltages of the NMOSFETs and PMOSFETs be if the devices were fabricated without implanting the channel with the boron ions? (b) Calculate the threshold voltage that the NMOSFETs and PMOSFETs would exhibit as a result of implanting the boron ions into the channel? (c) Draw a doping profile along the coordinate perpendicular to the surface and passing through (i) the channel region; and (ii) the drain region of both types of FET.

6. Discuss each of the following drawbacks of LOCOS technology: (a) bird's beak encroachment; (b) oxide thinning in narrow field-oxide openings; (c) boron encroachment; (d) non-planarity; (e) process complexity; (f) sensitivity of bird's beak encroachment on pattern geometry.

Chapter 17
ASSEMBLY and PACKAGING for ULSI

The packaging of IC chips (dice) is a multi-disciplinary technology, encompassing a multitude of assembly and packaging issues.[1,2] Factors such as the costs of IC packaging, the impact of the package on the circuit and system performance, and on the reliability of the chips they house and the systems they inhabit, makes an awareness of packaging technology important. Since chip packaging is the final procedure carried out in the manufacture of integrated circuits, it is also the last topic covered in this text. The discussion begins with the steps involved in the assembly of ICs, and concludes with IC package technology. Figure 17-1 shows the assembly and packaging steps typically performed on a chip after the completion of the wafer fabrication steps. This sequence of test-and-assembly procedures is often referred to as "back end" processing.

17.1 WAFER SORT

Following the conclusion of the wafer-fabrication process comes an important testing step referred to as *wafer sort.* The wafers are tested to determine which of their die are functional. It is a kind of report card on the fabrication process. In this test, the intact wafers are mounted on a vacuum chuck. The wafer is then aligned so the thin electrical probes of the tester can contact each of the bonding pads of one of the chips. With the aid of a computer-controlled test system, appropriate electrical signals are applied to these pads, and the chip is tested for functionality. On completion the wafer is stepped so that the next chip can be tested. When all of the die have been tested, those that failed are traditionally marked with a drop of ink from an *inker* (and are eventually discarded). Those that functioned correctly remain unmarked (Fig. 17-2). Inkless probing is now being implemented, where the computer in the tester remembers the failed die and they are by-passes them at die-attach.

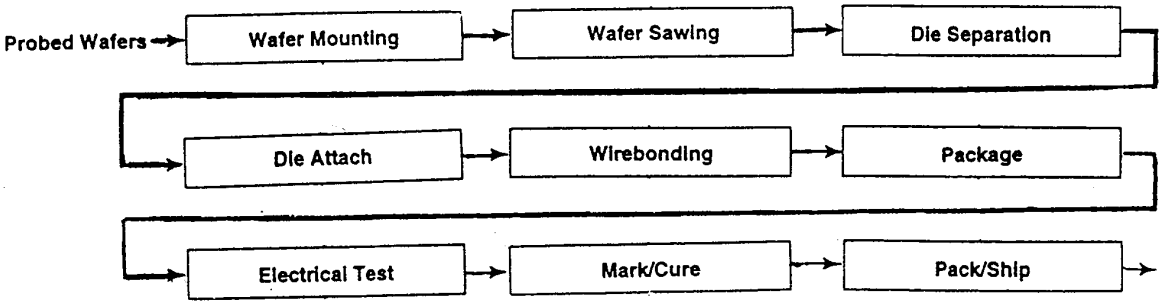


Fig. 17-1 Sequence of conventional assembly and packaging steps.

Exhibit 1014 - Page 185 of 218

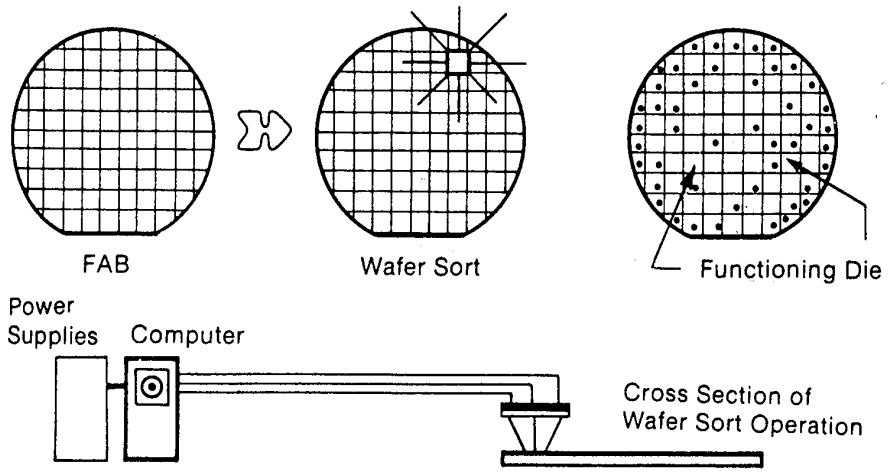


Fig. 17-2 Wafer sort.

17.2 WAFER BACKSIDE PREPARATION

The backside and/or thickness of a wafer may have to be modified to carry out subsequent assembly and packaging steps. Two types of wafer backside/wafer thinning processes are commonly utilized:

1) *Backside Lapping*: The backside of a wafer is often abraded or mechanically polished after IC fabrication has been completed. This procedure removes process contaminants and/or deposited layers that may interfere with the ohmic contact that may need to be established between the die backside and the package substrate. Such a lapping step is also used to prepare the backside for a subsequent metal deposition, and/or to thin the die so that it can subsequently fit into a shallow package cavity. Silicon die can be reduced with such a grinding process to 0.25–0.35 mm in thickness (from an initial thickness of 0.6–0.75 mm). To thin die down even further (e.g., to the range needed for use in ultra-thin packages [0.1–0.25 mm], such as the 180 μm thickness needed for ICs used in Smartcards), other methods, including wet- and plasma-etching, have been developed.[3]

2) *Backside Metal Deposition*: A thin film of gold may be deposited on the back side of the wafer to facilitate attaching of the die to the package. This type of die-attach process is discussed in the section dealing with *Hard Solders that form Eutectic Die Bonds.*

One, both, or neither of these steps may be performed on the wafers to prepare them for later operations. These processes may also be carried out before, or after, wafer sort.

17.3 DIE SEPARATION

After the die on the uncut wafers have been subjected to preliminary testing for functionality (wafer sort), the wafer is cut apart (a process also known as *dicing*), so that each functional die can be mounted into its own package. The industry standard for die separation uses a circular diamond-impregnated dicing saw rotating at a speed of 20,000 rpm to cut through 100% of the wafer. The vast difference between the hardness of the diamond crystals embedded in the saw blade (Knoop hardness of 7000) and that of the silicon wafer (Knoop hardness of 1000) enables the blade to easily penetrate the silicon and make a clean cut. Earlier die separation techniques included partial scribing (with a diamond-tipped scribing tool, or a pulsed laser beam), followed

by a secondary mechanical breaking step. Problems of chipping and cracking during the breaking step led to the abandonment of these earlier techniques, except for such brittle materials as GaAs. In addition, the adoption of high-speed automatic die-attach equipment requires the die sides to be straight and have exactly square edges. Such edge characteristics are reliably produced by diamond sawing. The sawing operation itself consists of many sub-steps. Typically, one dicing saw can support three to five die bonders that, in turn, can keep 20 wire bonders supplied. Saws generally range in price from $150,000–$180,000.

The wafer surface is designed for the die separation step by leaving enough room between adjacent die to allow for the width of the saw blade during cutting (*kerf width*). These regions are referred to as *scribe lines* or *scribe streets.* They range in width from 60 μm to 150 μm, depending on the die design. To keep the sawing operation simple, some processes deliberately maintain bare silicon in the scribe lines throughout the fabrication cycle. However, on most wafers now being processed the scribe lines contain test patterns, and this presents a more difficult sawing problem (since the saw blades must cut very dissimilar materials). For example, the glass ILD and passivation layers on the surface of the silicon in the scribe lines can cause sporadic, uncontrollable chipouts and cracks emanating from the kerf. Metallization in the scribe lines causes clogging of the blade, reducing its life by as much as 50%.

Prior to dicing, the wafer is mounted onto a sticky, flexible tape (e.g., Mylar) that is itself attached to a rigid frame. The tape continues to hold the die after the sawing operation and during transport to the next assembly step. The tape is about 0.1 mm thick, and the saw cut penetrates about one third of this thickness. Adhesives that hold the wafer onto the tape are available with different adhesion levels to accommodate various die sizes. Combination wafer-tape/die-attach adhesives have been developed to allow additional process simplification. The frames are easier to handle than individual wafers and can be marked to identify the wafer being carried.

The tape itself is attached to the rigid frame with uniform and controllable tension by machines especially designed for this task. The wafer is then mounted to the tape. This procedure must be accomplished without bubbles trapped between the wafer and tape. In addition, the wafer must be precisely positioned with respect to the frame to facilitate repeatable automatic alignment of the wafer to the dicing saw.

The frames containing mounted wafers are then placed into a cassette to be loaded into the dicing machine (Fig. 17-3). After such loading, each frame is moved from the cassette to an alignment station, where the wafer is aligned with the cutting blade. The alignment must be accurate so the saw cut occurs in the center of the scribe street and not in an area of the die. Automatic alignment systems have been developed as part of the move toward implementing automation of the entire sawing process. Split-field alignment systems reportedly work well if the scribe lines are bare silicon. If other materials or test patterns exist in the scribe lines, more complicated pattern recognition systems may need to be used. Alignment should be monitored during the sawing operation so that if tension on the tape shifts the wafers on the chuck (causing the blade to offset and cut out-of-street), this shift can be corrected. Other possible process errors that should be guarded against include: a) saw cuts that are too wide or cause excessive chipping; b) saw cuts that are too deep or too shallow; and c) broken saw blade.

The sawing process itself has several important variables to be considered when operating an efficient sawing system, including material type, depth of cut, kerf blade exposure, diamond particle size, desired throughput, and cooling system design. For silicon wafers, the saw blade

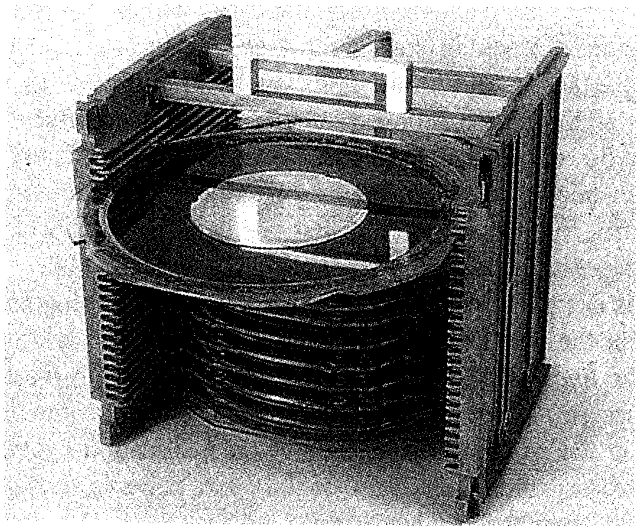


Fig. 17-3 Wafers are mounted on mylar tape frames in preparation for sawing.

is 25 μm thick, and 0.4 to 1.25 mm in *exposure* (the distance from the edge of the wheel hub to the edge of the blade, see Fig. 17-4). The diamond fragments are electroplate-bonded to the saw blade and are from 3 to 6 μm in size. The depth of the cut should extend approximately 0.03 mm into the 0.1-mm-thick mounting tape. That is, if the nominal cut depth in the tape is less than 0.03 mm the wafer may not be 100% sawed, but if the nominal cut depth is significantly deeper, the tape could tear when it is stretched after sawing (making the die unliftable from the tape). Optimum feed rates for "through" cuts have been reported to be 5–10 cm/second. Effective cooling of the diamond blade is also essential. Cooling systems deliver coolant during sawing to the cutting interface. The coolant reduces friction between the cutting edge and the wafer, and keeps the kerf clear of cutting debris.

After sawing, the wafers are moved to a cleaning station. Here, colloidal silicon produced in the sawing operation is removed. If this powdered silicon was allowed to dry on the wafer

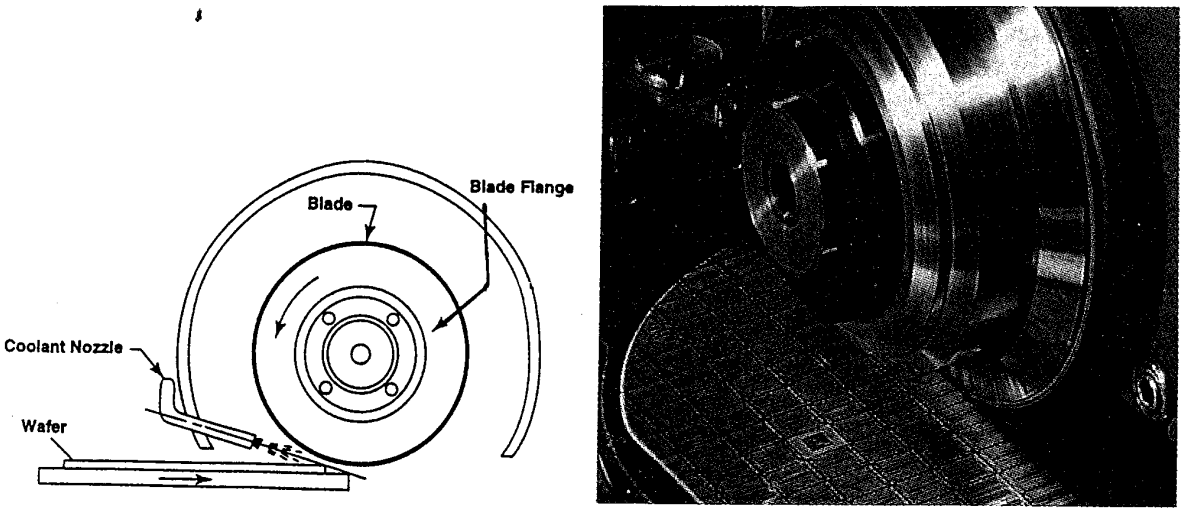


Fig. 17-4 The wafer sawing operation.

surface, bonding pads could become contaminated with silicon dust, creating wire bonding problems. Also such residue could collect in the kerfs, producing die-attach pickup problems. The ideal cleaning process consists of adequate jet pressure during sawing, followed by wafer scrubbing after the sawing process. The post-saw cleaning step must be done quickly, before the wafer dries. To accomplish this, advanced sawing systems contain integrated cleaning stations. After the sawed wafers are cleaned, and the die are still attached to the tape, the tape is usually stretched so the distance between adjacent die is increased. This step is necessary when using some types of automatic die bonders to allow them to pick up the die from the tape.

17.4 DIE ATTACH

After the wafer has been sawed, the backside of the now-separated, functional die must be attached to a package or substrate, and this procedure is known as *die attach* or *die bonding*.[4] The quality of this adhesive bond is critical to the long term stability of the assembled parts. The choice of die-bonding technology depends on tradeoffs between various packaging issues such as: 1) the thermal stresses caused by mismatches between the temperature-coefficient-of-expansion (TCE) of the chip and packaging materials; 2) power dissipation; 3) electrical requirements; and 4) cost. The connection between the die and the substrate must be in the form of a total intimate contact, without any voids that would lead to crack-causing stresses in the die. The intimate contact is also necessary to assure adequate heat dissipation during later circuit operation, and to provide a firm foundation that will give sufficient support during subsequent wire bonding operations. The increased use of larger die sizes has introduced new challenges for die-attach technology. This results from even greater stresses and thermal management problems that are associated with larger die. It is also important that the die-attach operation be carried out in a cost effective manner, which today implies the use of high-speed automatic die-bonding equipment. This discussion first presents some of the reliability issues involved in the die attach step, and then information on the materials that connect the die to the package is given. Finally, the equipment used to perform the bonding operation is described.

17.4.1 Die-Attach-Related Reliability Issues

The following reliability issues arise in connection with the die-attach process: a) corrosion of the die resulting from the die-attach step; b) inadequate heat dissipation resulting from a poor die bond; c) disbonding failures; and d) die cracking. The problem of corrosion is addressed in the later section dealing with conductive-epoxy die-attach materials. The chip disbonding failure and die cracking failure mechanisms are considered here.

Chip disbonding failures, which lead to complete separation of the die and package are usually detected during centrifuge testing of completed packages. The most important causes of such failures are a high degree of voiding under the die, or the lack of chemical bond formation between the die-attach material and the die (or between the die-attach material and the package substrate). These can result from: a) die backside or solder oxidation (see section on *Hard Solders that form Eutectic Die Bonds* for details); b) the formation of brittle intermetallic phases (see again same section); c) package substrate metallization failure; and d) insufficient amount of die-attach material.

Die cracking is a problem primarily manifested by large die in which crack propagation occurs after a number of thermal cycles are experienced by the die-bonded chip. Die cracking represents a serious reliability hazard since the chip undergoes heat cycling during assembly

(i.e., during die-attach, wire bonding, and encapsulation). In addition, power cycling of devices occurring during normal circuit operation also results in thermal cycling. The cracks have been found to originate at microcracks which in turn first form on the silicon surface, and then propagate under stress until the die fractures. The microcracks occur as a result of damage to the Si surface during backside preparation and wafer sawing. Stresses that cause the microcracks to propagate have been found to arise from voids at the die-bond interface, especially when the voids occur at the edges and corners of the die (open-edge voids). Based upon this die-cracking model, it has been recognized that a void-free, uniform die-attach process is required for large ULSI chips.

Note that the force of impact of the electrical probes during the previous wafer sort process is another common source of die cracking. Some force is needed to break through the thin aluminum oxide formed on the bond pads, allowing good electrical contact between the probes and the pad metal. Sometimes this force is excessive, causing die cracking.

17.4.2 Die-Attach Materials

Normally the materials used to connect chips to packages are electrically conductive, so the low-resistance electrical connection required in some applications can be made to the back of the die. Since electrical conductors are good conductors of heat, they also provide a good thermal path between the die and the rest of the package. There are a variety of materials used to perform the die-attach function, including: a) hard solders that form a eutectic die bond; b) silver-filled specialty glasses; c) conductive epoxies; and d) conductive polyimides. The selection of the optimum die-attach material for a particular application is determined by the hermeticity requirements of the package, thermal dissipation, the heat being transferred through the die bond, and the thermal expansion coefficient (TCE) of the chip carrier media. Eutectic bonds and silver-filled glasses are typically used in ceramic packages, and organic polymers (conductive epoxies and polyimides) are commonly used in plastic packages.

17.4.2.1 Hard Solders That Form Eutectic Die Bonds: The die bond that is formed by creating a Au-Si eutectic from the heating of a hard solder (usually in the form of a Au-Si thin sheet) and the backside of the die, is regarded as the industry standard for the die bonding process. If properly formed it offers the highest reliability in assembled parts and allows for a relatively fast die-attach process. Eutectic die bonds are also not potential sources of ionic contamination. Offsetting these advantages are the high cost of the process, the high temperature needed to form the bond (which results in high thermal stresses on the chip), and difficulty in automating this process.

Although the melting point of pure silicon is ~1420°C, when some other elements are mixed with it, the mixture has a lower melting point. The *eutectic* is the composition of such a mixture that will result in the lowest melting point when the materials are combined. For example, in the application under discussion, the eutectic of gold and silicon has a composition of 97.15% Au and 2.85 % Si by weight. This mixture melts at 363°C, and this low melting point can be exploited to produce a bond between a silicon die and a gold-coated package substrate. This is accomplished by placing these two entities in contact and heating them to a temperature above the eutectic temperature (e.g., 390–420°C). The package substrate is typically either a metal *leadframe* made of a Cu or Fe-Ni alloy, or a ceramic substrate consisting of Al_2O_3. Note that the melting point of this process must be higher than the temperature of the next operation (usually wire bonding), to avoid remelting the eutectic created during die attach.

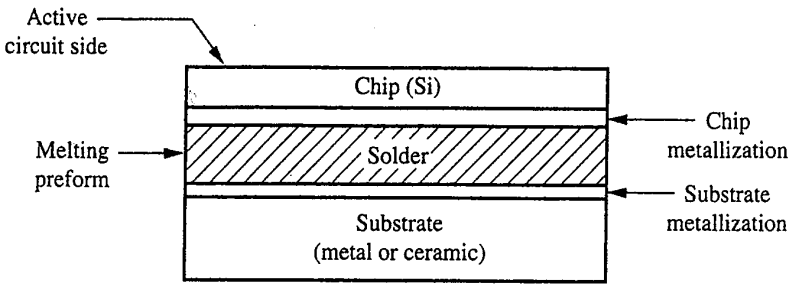


Fig. 17-5 The basic structure of a silicon device chip-bonded with a metal preform.

The details of this die-attach process are as follows: The backside of the die is usually coated with a thin layer of gold (150–250 nm), to improve the wetability of the silicon surface. (Note it has been reported that since Si diffuses rapidly through gold [even at room temperature], when it reaches the surface it can form a thin layer of SiO_2. Such SiO_2 layers degrade the wetability and the bonding characteristics of the die surface, and the problem increases with time after the gold layer is deposited. Hence, use of a refractory-metal diffusion-barrier film underneath the backside gold layer [e.g., Ti or Cr] has been reported as a way to minimize this problem.)[5]

A thin sheet (<0.05 mm) of the hard solder (known as a *preform*) slightly smaller than the size of the die, and consisting of Au-2%wt Si, or pure Au, is placed on the gold-coated package substrate (Fig. 17-5). The die is placed on the preform, and a heating block heats the substrate above the eutectic temperature. As the temperature of the die rises, the gold atoms from the deposited layer diffuse rapidly into the silicon. When enough gold has diffused to satisfy the eutectic composition, a very thin liquid film forms at the interface. The eutectic formation will continue until one of the reacting materials is used up. Since the silicon supply is much larger than the available gold, silicon from the backside of the die dissolves until the liquid reaches a saturation composition (which is greater than the eutectic composition, if the temperature is above the eutectic point). Upon cooling, the mixture freezes and a solid bond is formed, consisting of *hypereutectic phases* (silicon-rich eutectic microstructure), in which mechanical interlocking of the gold and silicon phases are the key to the adhesion of the die to the substrate.

Since the silicon saturates the liquid solder during the process, and the silicon in the liquid is as chemically reactive as pure silicon, if any oxygen is present it will react with the silicon to form SiO_2. Such an oxide surface layer can degrade the quality of the bond. Thus, a mechanical scrubbing step is included after the liquid layer has formed. The scrubbing step (e.g., moving the die in a semicircular manner for ~5 sec) has the effect of displacing any oxide regions and other contaminants away from the interface. In addition, since the preform also melts, it provides an additional wet surface that facilitates the scrubbing action of the process. Finally, it has been recognized that the package cavity should be flooded with copious amounts of an inert gas during the eutectic die-attach process to minimize the tendency of the silicon present in the liquid Au solder to oxidize. Pure gold preforms have also been recommended for large die, since the silicon in the Au-Si preforms also reacts to form SiO_2. This reportedly causes Au-Si preforms to melt less uniformly than pure gold preforms.

Although a highly reliable bond can be made with this process, the cost of the gold preforms can be prohibitively high, especially for inexpensive high volume ULSI (e.g., those encapsulated in plastic packages). Therefore, other die-attach materials have been developed for such applications. In addition, some of the problems associated with human error during the eutectic die-attach process have been resolved by the introduction of automatic equipment developed for the process. However, there are still applications in which a manual procedure is needed, especially when assembling small quantities of expensive special circuits. In some applications in which a hard-solder die-attach bond is desired, but which also require lower temperatures and more ductile chip-bonding (i.e., "soft" solders), Pb-Sn-based solders are used.

Bonding *large die* with hard solders also poses some difficult problems. In particular, as mentioned earlier, oxidized die surfaces are not wetted by the solder. With large die the scrubbing step is less likely to displace the oxidized regions at all of the corners or edges of a die. This increases the likelihood of edge-void formation under large die. Such voids have been identified as causes of the stresses that lead to die cracking during thermal cycling. Alternative materials for large die are therefore replacing the hard solders, even in ceramic packages.

17.4.2.2 Silver-Filled Specialty Glasses for Die Attach: Die attach is also being performed with special glass compositions (*lead borate* glasses), that exhibit low glass-transition temperatures (temperature at which the material changes from a hard and relatively brittle condition to a viscous or rubbery state). The glasses are filled with silver to provide adequate electrical and thermal conductivity. Currently available glasses become molten and form an "oxide bond" to silicon at 400–450°C (bond formation also requires exposure to these temperatures for several minutes). After processing, the bond microstructure consists of an interlaced network of silver in a sintered, glassy matrix. It is mandatory that the bonding process be conducted in oxidizing ambients, as the lead borate glasses reduce to metallic lead in reducing atmospheres. This would interfere with the oxide bond mechanism, and would result in die bonds that exhibit poor thermal stability. In addition, the condition of the backside of the wafer must allow Si to be available to form the oxide bond. Thus, any film on the wafer backside acting as a diffusion barrier that prevents Si from diffusing to the surface will severely degrade adhesion. Such films must be removed. An example is a layer of silicon nitride left inadvertently on the wafer backside.

The silver-filled-glass die-attach process offers significant cost advantages over Au-Si technology, while creating a bond without using an organic adhesive material. Furthermore, such processes offer the possibility of a void-free die-bond interface with excellent thermal stability. On the other hand, the glass materials require solvents and binders for processing purposes, and the problem of complete solvent removal is similar to that involved with organic adhesives. Nevertheless, it has been reported that for large die, such glasses are replacing the Au-Si eutectic technology in ceramic packages. The firing temperature is too high for use with plastic packages.

17.4.2.3 Epoxy Die-Attach Adhesives: Epoxy materials are well known for their superior adhesive properties. They have become popular for many die-attach applications, especially with chips mounted in plastic packages. Epoxies are thermo-setting polymers, meaning that they crosslink when heated. Epoxies are much less expensive than the gold-based hard solders (e.g., for a 150-mil-square chip, a gold preform may cost ~15¢ versus ~0.3¢ for an epoxy bond). They are soft and flexible. They lend themselves well to automation and can be processed at lower temperatures than the hard solder materials. Such die-attach epoxies are usually silver-filled (~70% by weight) to provide a good thermal contact between chip and package. Epoxies,

however, may also exhibit the following limitations for this application. First, the cured epoxies have a limited thermal stability, which means that their bonding strength degrades under aging at elevated temperatures (>200°C), and they also exhibit weight loss at elevated temperatures. Thus, they are not compatible with other assembly processes that require temperatures of 300°C or more (due to outgassing effects). The outgassed products are also generally corrosive in nature and thus present a fundamental problem in high reliability applications (as well as applications in which the chips must be operated at elevated temperatures). Second, when bonding large die, it is very difficult to remove all of the solvents from the adhesive without leaving behind voids at the die-bond interface. Such voids cause degraded electrical and thermal conductive properties of the die bond, as well as mechanical stresses in the chip that can cause cracking. (To help eliminate the voids from solvent outgassing, a new generation of epoxies has been developed. These newer products have either a low-solvent content, or are "B"-staged [i.e., partially cured, and in thin-film self-supporting form].) Third, early generations of epoxy materials for die-attach contained high levels of sodium and chlorine. If moisture penetrated the circuit package, the Cl ions produce HCl, which causes corrosion. The demands of the semiconductor industry for materials with much lower concentration levels of these ions led to the development of alternate epoxy chemistries for such applications. An unwanted characteristic of these newer epoxies, however, is a more sluggish reaction rate (i.e., high throughput requires a fast-curing epoxy).

The uncured epoxy material is usually in paste form, which can be dispensed onto the die paddles by a nozzle to ensure that the required bond-line thickness is created without voids. The chip is pressed into the paste immediately after it is dispensed. The backside of the die is prepared by sand blasting to provide a rough surface, without any metallization layer. The die is placed onto the epoxy-coated substrates by a high-speed automatic machine, and the bond is formed without the need for a scrubbing step. The automatic chip-attach operation proceeds at a speed of about 1 to 2 sec per chip. The bond is formed as the epoxy is cured, typically at temperatures of 125–175°C for 1 to 2 hours. The adhesion of epoxies to Si and metallic substrates is primarily due to Lewis acid-base interactions. Silane coupling agents may be included to enhance the bonding to silicon surfaces. This implies that rougher surfaces will produce better adhesion, and that a native oxide present on the wafer back surface is expected to produce high adhesion (quite unlike the conditions required for forming a eutectic die bond!).

For applications that use large die sizes and packages with higher pin counts in a diversity of configurations (e.g., thin small outline packages [TSOPs], thin quad-flat packages [TQFPs], and large ball-grid-array packages, described in a later section), conventional die-attach adhesives run into problems. "Popcorn cracking" of plastic packages can occur with surface mount devices. This is due to the additional stress from the rapid vaporization of absorbed moisture in the epoxy during reflow soldering. Such stresses may exceed the bond strength of the bonding material. If this occurs, the result is delamination and sometimes cracking of the plastic package. (The same basic principle pops kernels of corn when heated, hence the term *popcorn cracking*.) A resin based on cyanate ester has been developed that reduces or eliminates popcorn cracking.[6] While it absorbs about the same amount of moisture as do epoxies, the moisture cannot readily escape when the cyanate ester resin is heated.

17.4.2.4 Polyimide Adhesives for Die Attach: Polyimides are polymers that can withstand higher temperatures than epoxies. They have been developed as an alternative die-attach material. They can tolerate temperatures of >450°C before suffering decomposition. Polyimides can also be

formulated with lower extractable ionic contaminant concentrations and less moisture content than epoxies, two desirable properties that reduce corrosion problems. On the other hand, polyimides are known to absorb moisture (which may subsequently be released in the hermetic package cavity), and thereby still cause moisture related failures. The material is usually supplied in the precursor form (polyamic acid), and this is converted to a polyimide by heating through an imidization reaction. This imidization reaction releases water. Control is needed to complete the imidization step without trapping moisture in the cured film. The die-attach process, however, usually does not allow sufficient curing to ensure a proper imidization reaction with currently available materials. Therefore, die-attach adhesives that are applied in polyamic-acid form generally have not achieved the full potential capable of polyimide adhesives.

Two approaches to overcoming the aforementioned limitations of conventionally-cured polyimides have been developed: 1) using partially-cured modified polyimides; and 2) using fully imidized thermoplastic polysiloxaneimides. The *partially cured polyimde* is supplied already screened and in solid form, onto a film that emulates the conventional wafer dicing tape described in the previous section on die separation. This tape is applied to the dicing frame, and the wafer is applied to the tacky surface of the partially-cured polyimide adhesive (which has the same area as the wafer). Since this polyimide is a partially-cured solid with most of the solvent driven-off, it can allegedly form a void-free die bond, as well as provide full coverage of the die backside. This approach is designed to avoid crack-causing stresses in large die, as well as to provide a reproducible thermal connection to the package. By combining the dicing-adhesive step with the die-attach adhesive step, assembly process throughput is also increased.

The *fully-imidized thermoplastic polysiloxaneimide* is supplied as a silver-filled *thixotropic paste* (materials that are gel-like when at rest, but fluid when agitated) in a polyether solvent. This solvent has a lower boiling point than other solvents, and does not complex with the imide linkage. Since the adhesive is fully imidized, it requires no "curing" and its chemistry reduces moisture pickup to levels below those of conventional polyimides. The die bond is formed in the "hot-melt" fashion, much as in the Au-Si eutectic process, except at lower temperatures. The die

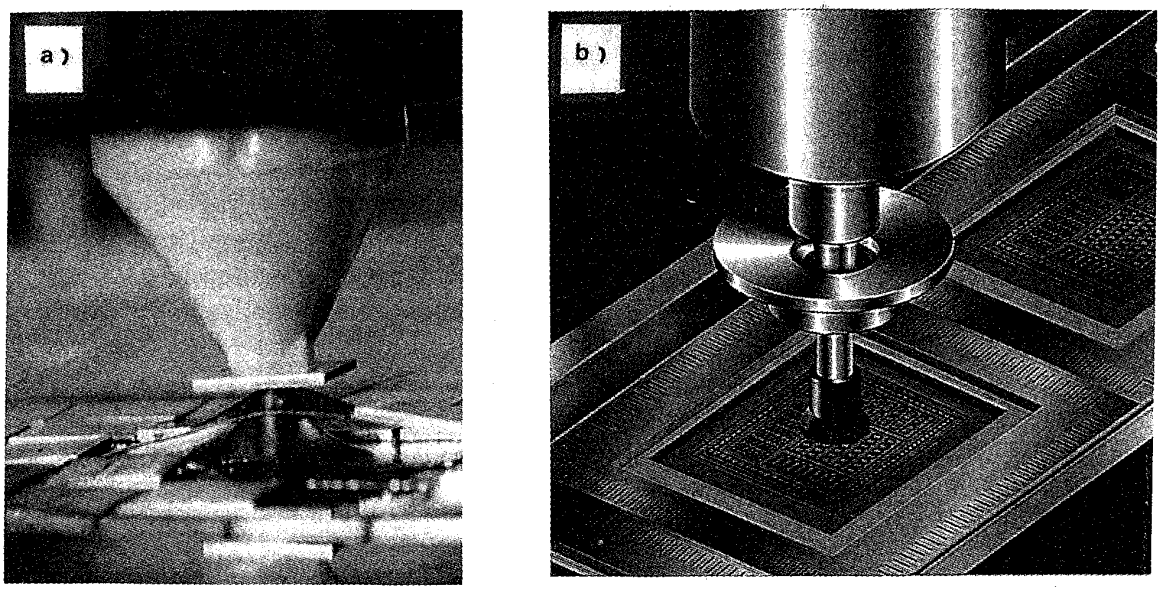


Fig. 17-6 The die-attach operation. (a) Photograph of a die being lifted from the mylar tape carrier. (b) Vacuum collet placing a die on the package for die attach.

is placed on the hot adhesive with a force of 150–200 grams for up to two minutes at a temperature of 230°C. The thermoplastic nature makes this adhesive amenable to remelting for hybrid circuit repair or rework, but it also limits the maximum wire bonding temperature. That is, a 220°C maximum thermosonic wire-bonding temperature limit is recommended.

17.4.3 Die Attach Equipment

Equipment used in the process of die attach has evolved from operator controlled manual systems to the fully-automatic systems in standard use today. Systems capable of die bonding up to 4000 to 6000 parts/hr are now available, particularly for epoxy attachment. In such die bonders the wafers are presented in sawed-wafer form attached to an adhesive Mylar tape. The die-bonder has a pick-and-place unit, and the pickup arm then moves into position over the die to be attached. The arm is lowered so that a vacuum collet can pickup the die (Fig. 17-6). At the same time that the collet lifts the die, an ejector needle pushes the die from underneath, causing the die and the adhesive Mylar tape to disengage from one another. Some collet designs make contact with the upper edges of the die, while others have a rubber tip that makes contact only with the top surface of the die as it is lifted. Pattern recognition systems enable the machines to recognize chipped die or rejected die. The die to be bonded is then placed on the package substrate, onto which has previously been placed the die-attach material. Die placement pressure is applied by the bond arm, and can be programmed to accommodate different die sizes. The die-bonder also contains a chip-attach material supply station and a lead-frame loader and unloader.

Automatic die bonders must have the following characteristics. First they must perform the die attach process for a variety of applications. For example, new package types are continually being developed, especially to accommodate large die. Thus, the equipment may need to handle a wide range of die sizes. Second, the yield of the die-attach process must be high, and the long term reliability of the bonded parts must also be preserved. Third, the equipment must operate with a reliability of more than 90% uptime. Fourth, the equipment should be flexible enough to accommodate a wide range of product types, and be able to be converted quickly from one product type to another. Finally, the equipment needs to be easily integrated into a computer-controlled factory-automation system. Die bonders in 1999 ranged from $80,000–$250,000 for fully automated systems.

17.5 BOND-PAD TO PACKAGE CONNECTIONS

After the chip has been attached to the package substrate (or leadframe), electrical connections must be made between the bonding pads of the chip and the inner leads of the package (which are themselves connected to the package pins).[7] Wire bonding has been, and will likely remain, the most common technique for making the connections between the chip input/outputs (I/Os) and the package (especially for chips with up to about 224 I/Os). It is estimated that in 1996 over $4x10^{12}$ wire bonds were formed in the fabrication of ICs. In this section wire bonding is discussed first. The other two methods for making such connections are tape-automated bonding (TAB), and flip-chip bonding. They are discussed in subsequent sections.

17.5.1 Wire Bonding

The wire bonding process is carried out after the die-attach step. Flexible wires are attached, one at a time, from bonding pads on top of the chip to the package. One of three methods is used to

join the wire to the chip and substrate: *thermocompression (T/C) bonding*; *ultrasonic (U/S) bonding*; and *thermosonic (T/S) bonding*. T/C and T/S bonding can produce either a ball or wedge bond, while U/C bonding is only used to make wedge bonds. The materials used for wire bonding are either Au or Al because these metals are highly conductive and ductile enough to withstand the deformation during the bonding steps. They also bond well to the metallizations used on both the chip and package. In addition, the technology of manufacturing fine Au and Al wires (typically 25–30 μm in diameter) is mature, and metallurgically stable bonds are formed by each process. Wire bonds with Au use a ball bond at one end and a wedge bond at the other. These are most widely used today. Wire bonding with Al wire typically uses two wedge bonds.

The *T/C wire-bonding process* with Au wire begins by first making a ball bond to the chip. (Fig. 17-7). The use of the ball-and-wedge bonding technique has the advantage that it allows the subsequent wedge bond to be placed anywhere on a 360° arc around the ball bond. The ball (about 2 or 3 times the diameter of the wire) is formed on the end of wire (Fig. 17-7b) with either a small hydrogen torch or an electric spark. As the wire tip is melted, the molten liquid is pulled into a spherical shape by gravity and surface tension. (Gold also forms balls of uniform size, ball shape, and surface conditions.) The ball formed in this manner is then pressed against the metal film on the bond pad by a tool called the *capillary* (which is kept heated to a temperature of 300–350°C). Bonding is accomplished by simultaneously applying heat and mechanical pressure to cause plastic flow of the two metals in contact. The plastic flow produces clean metal surfaces at the interface, which enhances interdiffusion of the two joining surfaces, and allows a weld to form between them. The typical metallurgy of this bond is an Au-wire ball and an Al film on the bond pad of the chip. After this ball bond is formed, the capillary rises to the loop height position and is then moved over the site of the wedge bond. During this movement wire is fed through the capillary, allowing the loop to be formed. The capillary is then lowered and the wire (without a ball) is pressed against the package metallization, producing a wedge-shaped bond that has a gradual transition into the wire. The capillary then rises, and the flame breaks the wire (leaving it wedge bonded to the package). At the same time it forms a new ball. The metallurgy of the wedge bond is typically Au-wire to Au or Ag bond-pad metal on the package. The strength of the bond is tested on other machines by either the ball-shear or wire-pull test (Fig. 17-8). The capillary is made of alumina, tungsten carbide, or other refractory materials. Gold wire is well suited to T/C bonding* because it deforms easily under the bonding capillary at elevated temperatures, exposing clean metal, without forming any oxide that would inhibit joining. Such Au-wire ball-and-wedge bonding is widely used in plastic packaging operations. The wire bonders used in the early days of the IC industry were manually operated and quite labor intensive. These were replaced by fully automated ball- and-wedge wire bonders with pattern recognition and computer control. Such bonders now run at speeds of 6 wire bonds per second (or 18,000 bonds per hour!). Automated wire bonders in 1999 ranged in price from $75,000 to more than $250,000.[9]

In *ultrasonic* (U/S) *bonding*, unheated wire is moved over the bonding site and is then pressed onto the surface by a stylus ("bonding wedge"). While the wire is firmly clamped between the bonding tool and the bond pad, a burst of ultrasonic vibrations is applied to the

* Gold wire, however, has two disadvantages. First, it is much more expensive than Al wire. Second, it forms an alloy with Al that can reduce the reliability of the bond. This alloy has a purplish color and is known as the "purple plague." Since such alloying requires at least 200°C, it is avoided by keeping the temperature of processes below this value after the Au-Al bonds are formed.

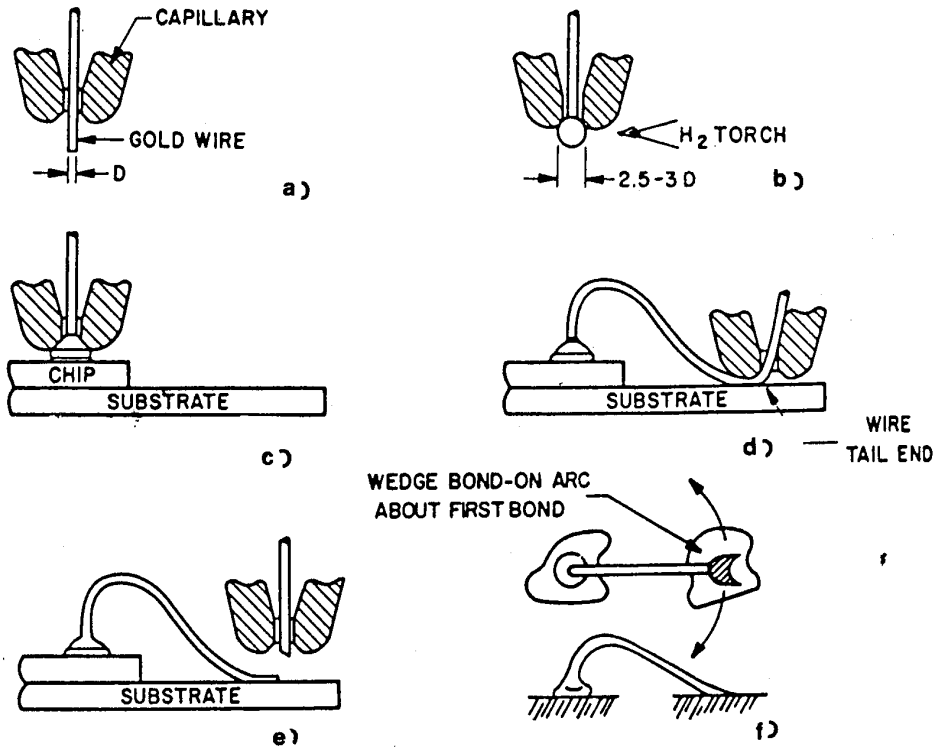


Fig. 17-7 Thermocompression bonding (T/C). Thermosonic ball-wedge bonding of a gold wire. a) Gold wire in capillary. b) Ball formation accomplished by passing an H_2 torch over the end of the wire. c) Bonding accomplished by simultaneously applying a vertical load on the ball while ultrasonically exciting the wire (the chip and substrate are heated to about 150°C). d) A wire looped and a wedge bonded under load and ultrasonic excitation. e) A wire broken at the wedge bond ready to go to step a). After S. Sze, Ref. [8].

wedge. The combination of pressure and vibration accomplishes a metallurgical cold weld between the wire and pad metallization. Aluminum or Al-alloy wire is typically employed in U/S wedge bonding, as no satisfactory method has yet been developed to form Al balls (i.e., the oxidation of Al wire at high temperatures makes it difficult to form a good ball). In U/S bonding, however, the wire direction needs to be the same as the direction of the applied vibration, making this wedge-and-wedge bonding technique less flexible than the ball-and-wedge bond. On the other hand, U/S bonding is carried out at room temperature. Aluminum U/S wire bonding is frequently used in hermetically sealed ceramic packages.

Thermosonic (T/S) *wire bonding* combines the best features of the T/C and U/S techniques. In T/S bonding, Au-wire ball-and-wedge bonds are also formed with the aid of heat and pressure. In addition, a burst of ultrasonic vibration (at 60 kHz) is also applied to each bond to augment the metal joining. This permits a much lower bonding temperature to be used than in T/C bonding (e.g., 150–200°C vs. 300–350°C). The T/S bonding offers an advantage when using devices and die-attach materials that cannot withstand the temperature of T/C wire bonding, and hence has largely replaced T/C bonding.

One drawback of wire bonding is that even after a wire bond is formed on the Al pads of a chip, a fairly large area of the pad remains exposed. Aluminum is a highly chemically reactive metal, and without its protective passivation layer (SiO_2 or Si_3N_4). It reacts readily even with pure water. In the other interconnection methods (i.e., TAB or flip chip) a thin interfacial metal sandwich layer (e.g., Cr/Cu/Au) is first deposited over the exposed Al bond pad. Subsequently,

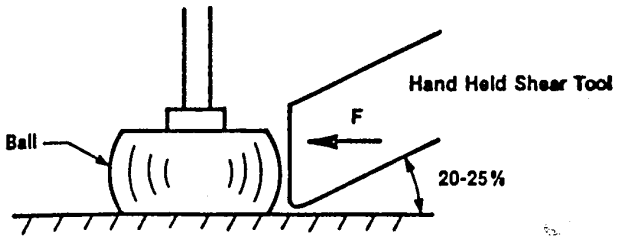


Fig. 17-8 Ball shear test setup, showing the manual shear probe in use.[22]

a gold (Au) bump (in TAB), or a solder ball (in flip-chip) is placed over the pad. This protects the corrosion sensitive Al. ICs assembled with the latter methods can exhibit reduced device failure rates from corrosion over devices assembled with wire bonding.

17.5.2 Wire Bond Spacing

As the number of external connections to an IC increases, a larger number of bond pads on the chip are needed. Since the number of such pads is increasing faster than the chip size, the bond pad *pitch* (bond pad center-to-center distance) must decrease, especially if the bond pads are along the periphery of the chip (as is the case with wire bonds). In memory chips, the pads are located along only two facing sides of the chip (or in the chip center for LOC packages, see Sect. 17.10.1). In logic and microprocessor chips the I/O count is increasing and the active device size is decreasing. Thus, the interconnect area represents a large fraction of the total Si area. In some cases, it may even determine the chip size, as extra space is needed around the active area for additional pad placement. The minimum wire bonding pitch has been steadily declining, and by 1998 had decreased to 50–80 μm (with 23 μm wires). The wire span from the chip to the package is typically 1 to 4 mm. Longer wires could lead to shorts to adjacent wires, as longer wires are more likely to droop or deform. In plastic packages, longer wires can also deform due to resin flow during the molding operation. To allow more wire bonds to be made on a chip, while meeting the minimum pad pitch and wire length requirements, staggered bond pads and package terracing are sometimes employed as a solution (Fig. 17-9).

17.5.3 Tape Automated Bonding (TAB)

For chips that require high pin counts (e.g., more than about 208 pins), wire bonding becomes undesirable, as the chip size must grow to accommodate the space for the wire-bond pads. An alternative chip-pad to package-pad connection method called *tape-automated bonding* (TAB) has been developed to handle such applications. In the TAB method, the wires of wire-bonding are replaced with finely patterned thin metal leads (usually made of Cu foil, plated with Au or Sn). The TAB method involves bonding silicon chips to patterned metal-on-polymer tape by thermocompression bonding (prior to the die-attach procedure). Usually, these leads are formed by depositing the metal film onto a flexible strip of polymer tape. Then this film is patterned by lithographic techniques similar to those used in the IC fabrication process. This forms the desired lead pattern, and the result is a strip of tape containing many individual lead systems (Fig. 17-10). The most popular way of connecting these leads to the chip involves the creation of bumps on the chip bonding pads (formed by electroplating Au). The tape is moved by sprockets until one of the lead systems is positioned exactly over the chip and the leads are then thermally bonded to these bumps. Note, however, that variations to the TAB process described above have also been used commercially (see Ref. 12).

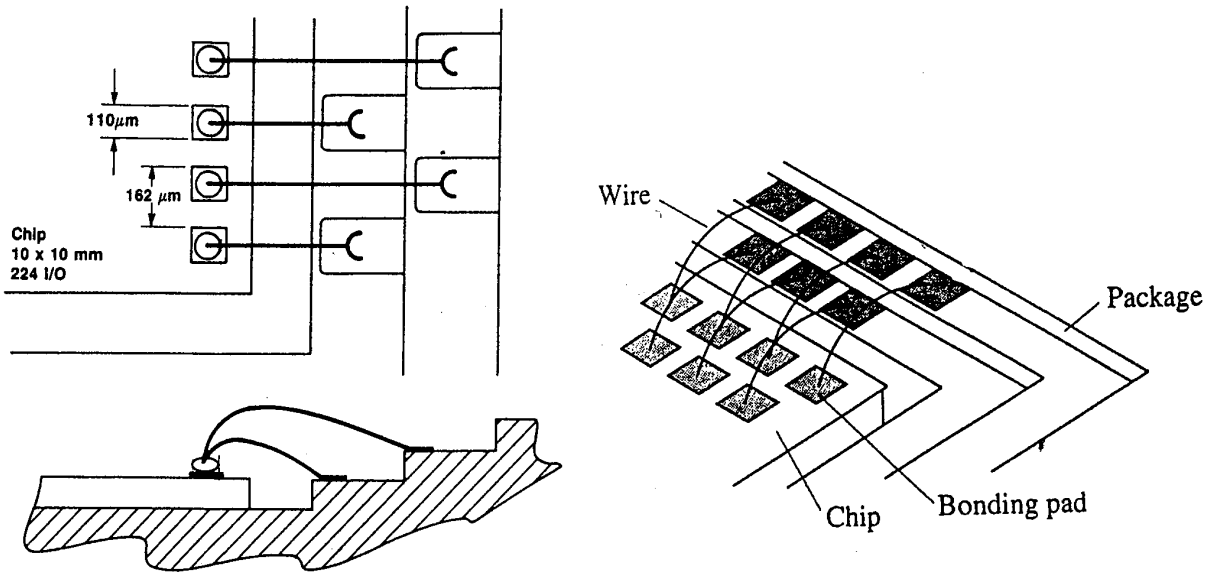


Fig. 17-9 Bond pads staggered on different levels. Chip and bonding pad configuration.[11] © ISHM.

The fabrication of the bumps on the bond pads is a multi-step process (Fig. 17-11). First a multi-layer metal film (consisting of Cr, Cu, and Au layers) is sputter-deposited in blanket form onto the wafer surface (which is covered with a passivation oxide layer, except for the bond pad. Lithography is then used to create patterns in a photoresist film that will act as a 25–30 μm-thick mask for the Au electroplating step. That is, openings in the resist are created over the bonding pads, where electroplating is then used to deposit Au bumps, 10-30 μm thick. Next, the resist is stripped. Finally, to complete the bump formation, the diffusion-barrier film is etched.

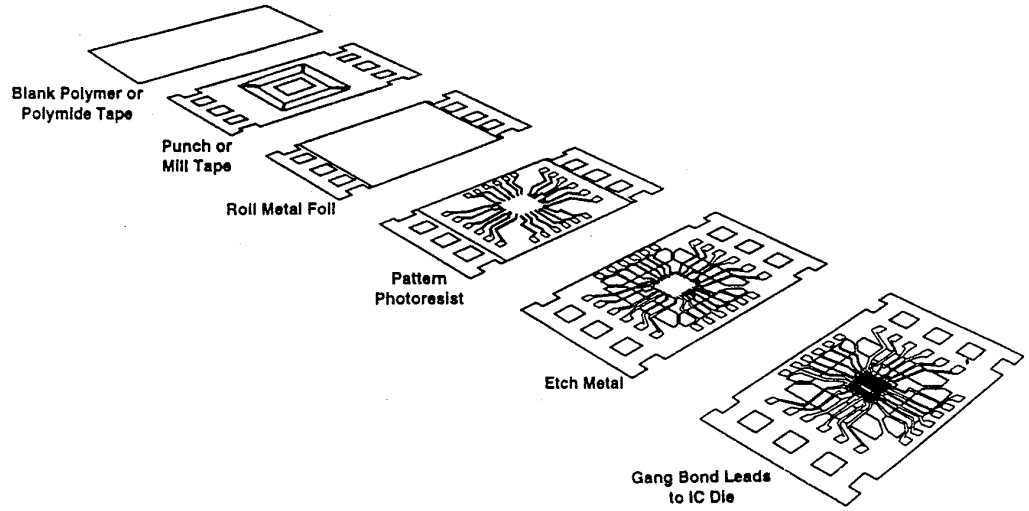


Fig. 17-10 Manufacturing sequence for tape-automated bonding (TAB). Reprinted with permission of Integrated Circuit Engineering Corporation.

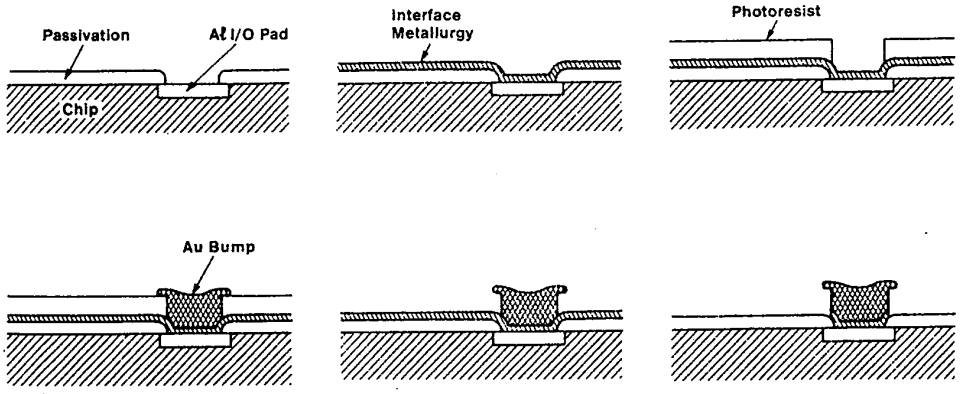


Fig. 17-11 Gold bump processing for TAB.[13] Reprinted with permission of Solid State Technology.

The Au bumps protect regions of the barrier layer film underneath them from being etched during this step.

The bonding of the metal leads to the bumps on the chip is done simultaneously by thermocompression (Fig. 17-12). That is, a heated *thermode* (300–360°C) is pressed against the leads, which are each pressing against a bump on the chip. The heated thermode applies a load of about 10 kg/mm^2 for 1–2 seconds to form the bond between the lead and bump. This step is known as *inner-lead bonding*. At this point, the chip has been attached to the tape that carries the lead pattern. Next, the chip and lead pattern are encapsulated (while they are still attached to the strip of tape), usually by plastic molding (Fig. 17-13). The packaged chips are separated from the continuous tape that carries the leads, to form an individual packaged part. These packaged parts are finally attached to a *printed wiring board* (*PWB*) by a process called *outer-lead bonding* (Fig. 17-14). That is, the packaged chip with the leads in place is transferred to the PWB, where the leads are attached to the land patterns on the board (usually by a lower temperature bonding process, such as *pulse-thermode reflow*, to protect the PWB laminate).

In addition to allowing bonding pads with smaller pitches than the ones needed for wire bonding, TAB offers the advantages of speed (since all the bonds are formed at once), and the ease of automation offered by the tape and sprocket system. Since the leads have no loop (as do

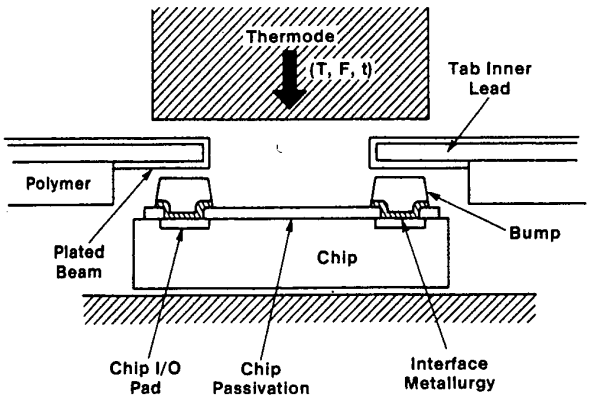


Fig. 17-12 TAB inner lead bonding.[2] Reprinted with permission of Van Nostrand Reinhold.

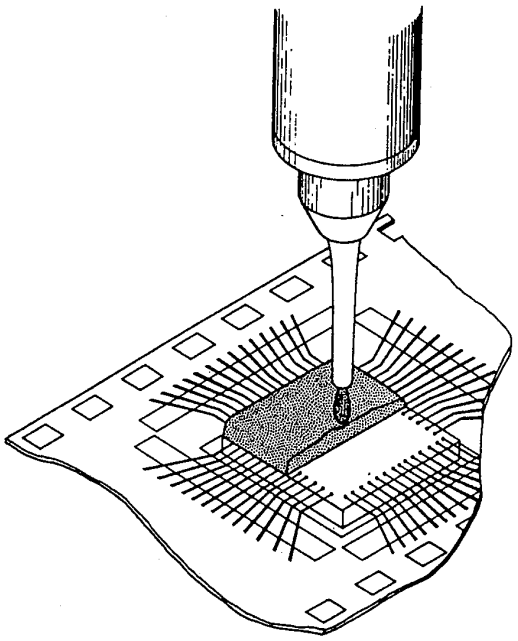


Fig. 17-13 TAB encapsulation.[2] Reprinted with permission of Van Nostrand Reinhold.

wire bonds) TAB also offers a lower profile, which makes TAB applicable for smart-cards, watches, read/write head circuitry, flash memory modules, and the like. The disadvantages of TAB are: 1) it requires a special tape design for each different chip design; 2) there is less infrastructure to provide testability of packaged TAB parts and for mounting the packaged TAB parts to the next levels of packaging; and 3) higher cost compared to wire bonding.

17.5.4 Flip-Chip Bonding

In *flip-chip technologies* solder bumps are fabricated directly on the Al bonding pads of the chip. The chip is then flipped face down and aligned to the package or substrate. The bumps are

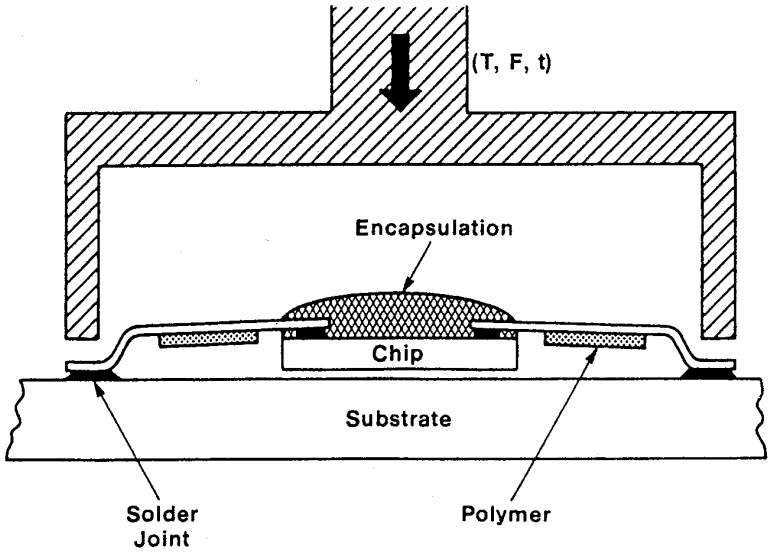


Fig. 17-14 TAB outer lead bonding.[2] Reprinted with permission of Van Nostrand Reinhold.

bonded directly (and all at once) to the package (or substrate) pads by reflowing the solder bumps. The process was introduced by IBM in 1964, and they called it C4 for *controlled collapse chip connection.*[14] The advantages of flip-chip bonding are: 1) the entire chip surface can be covered with solder bumps (making it an *area-array interconnect*). In other words, bonding locations are not limited to the chip perimeter, thus more I/O capability is provided than by a perimeter-interconnect on a die with the same size; and 2) the very short lengths of the chip-to-package interconnect paths minimizes their inductance.

As in TAB, the bumps are formed while the chip is still in wafer form. The bump formation process represents an added expense to overall wafer fabrication cost. However, since this process eliminates the cost of wire bonding, other production costs are reduced. Furthermore, the bump metallurgy reseals the area over the bonding pads that otherwise remains open after wire bonding. Thus, flip-chip technology hermetically seals the chip without the need for a package. This attribute is being exploited for so-called "packageless" attachment of die to substrates (discussed in Section 17.10.2).

The solder bumps most widely used to attach chips to alumina ceramic substrates have been high lead solders, especially 95 Pb:5 Sn. This solder melts at the relatively high temperature of 315°C, which permits other lower-melting-point solders to be used in subsequent module-to-card, or card-to-board packaging level processes without remelting the flip-chip bonds. A multi-layer film of metal (e.g., Cr-Cu-Au) is sandwiched between the Al chip pad and the solder bump to prevent the solder from interdiffusing into the Al. The Cr-Cu-Au film forms a cap over each of the Al bonding pads, and the size of each cap is also restricted by sequentially evaporating the Cr-Cu-Au film through a mask (Fig. 17-15a). The Pb:Sn is next evaporated onto the chip, again through a mask. The local area of each Pb:Sn layer is slightly larger than the area of the cap. Heating the wafer in an H_2 ambient at 350°C melts these localized Pb:Sn layers. The surface tension of the liquid solder causes the film to recede from the oxide surface, and to form a solder ball on top of the Cr-Cu-Au cap. The diameter of the base of these spherical bumps is determined by the dimension of the Cr-Cu-Au cap. This process is called *ball-limiting metallization* or BLM (see Fig. 17-16). Note that bumps made of conductive epoxy, which allow

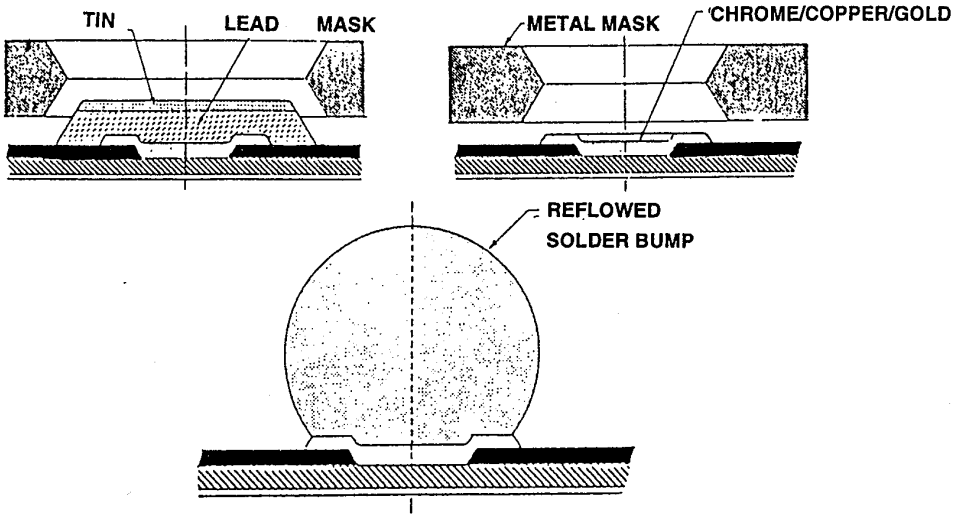


Fig. 17-15 Fabrication of solder bumps for C4 technology.[15] © ISHM 1974.

lower temperature processing than solder bumping, have been pioneered by the Polymer Flip-Chip Corp.[16,17]

Following the reflow of the solder bumps that connect the chip to the substrate, an underfill adhesive is formed between the chip and the substrate.[18] Such underfilling is necessary to reduce the stress on the solder bumps during temperature cycling. When the underfill adhesive is present, it shoulders a large fraction of the stress burden, and the stress on the bumps is significantly reduced. This decreases their tendency to fail through solder-bump stress fatigue.

The underfill adhesive (an epoxy or polyimide) is dispensed as a bead along the perimeter of the bump-attached chip. The chip is heated to 70–100°C and this lowers the viscosity of the underfill epoxy so that capillary force wicks it into the space underneath the die. Once flowed into place, the epoxy is cured to its solid form by an additional heat-curing step.

The term "flip chip" refers only to the method of attachment. No indication is made about what the chip is attached to. That is, if the chip is attached to a package it is called *flip-chip in package* (FCIP) and this method applies to single chip packages and multi-chip modules (MCMs). FCIP is used when the electrical performance of wire bonds is inadequate or when the number of I/Os is too large for wire bonding. Use of FCIP together with a ball-grid array package (see Sect. 17-10) would represent a packaged chip with a very large I/O count. For example, the Intel Pentium II microprocessor is put into plastic and ceramic BGAs and MCMs with flip-chip attach. *Direct chip attach* (DCA) refers to the direct attachment of a chip to a printed circuit board (also called *flip-chip on board*). *Flip-chip DCA* (FC-DCA) bypasses the attachment of the chip to the package.

Flip-chip attachment, however, has some limitations, including the following:

1. The infrastructure for manufacturing, testing, and handling flip-chips from semiconductor vendors by 1998 was not yet mature. Yet, forecasts were that by 2001 the number of chips requiring FC bumping will have increased to more than 2.5 billion. Furthermore, by 1998, flip-chip foundries were beginning to emerge.
2. A lack of standards exists for these procedures, due to the limited adoption of the flip-chip technology by the late 1990's.

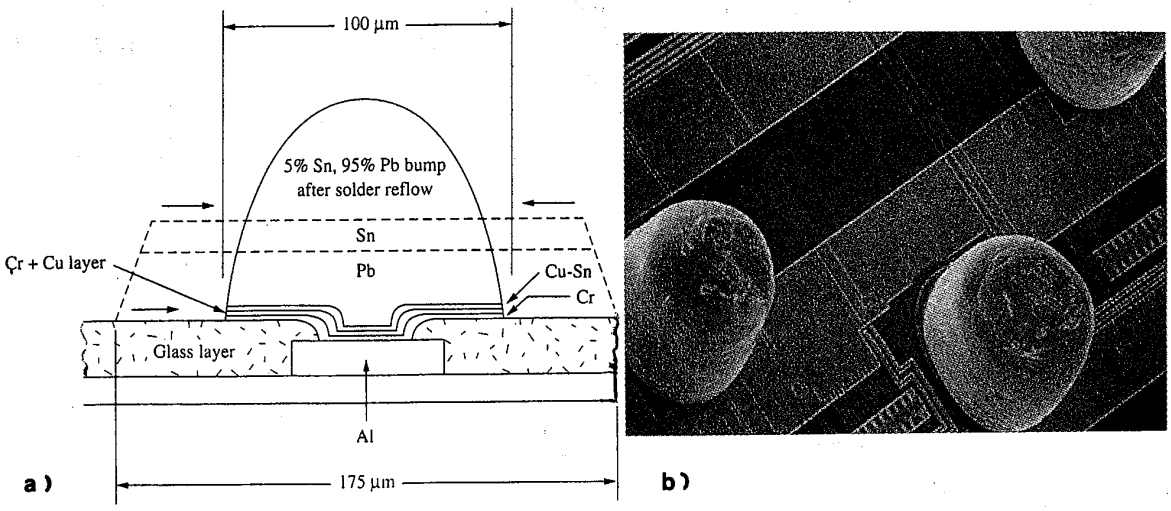


Fig. 17-16 (a) Cross section of a solder bump used in flip-chip technologies and the deformation caused by reflow-soldering operations; (b) SEM view of the C4 solder bumps on a chip.[14] Reprinted with permission of IBM Corporation.

3. Some of the manufacturing issues are also difficult. For example, visual verification of bump placement is not easy, making die placement accuracy very important. Thermal cycle fatigue of the solder joints is also a problem, especially if an underfill material is not used. The viscosity of the underfill epoxy material, however, changes with time, making the under-fill wicking process vary over the space of a day or shift. Another problem is that most ordinary Pb solders contain trace quantities of radioactive elements (i.e., which emit alpha particles that can cause soft errors in memory circuits). Solders made from lead that emits low levels of alpha particles are available for applications that cannot tolerate high soft error rates.

17.6 INTRODUCTION TO CHIP PACKAGES

Integrated circuit chips (or die) are usually encapsulated in an IC package prior to their being installed into electronic systems.[19] IC packages perform four key functions:

1. They provide a sturdy set of leads that allow an IC to be connected to the system in which it will operate. These leads, however, should not significantly degrade the performance of the chip housed by the package.
2. They provide physical protection for the chip against breakage or scratching.
3. They provide environmental protection for the chip against damage from chemicals, moisture, or gases.
4. They dissipate the heat generated by the chip during operation. Chips generating large quantities of heat require additional measures to be incorporated into the package design.

In order to achieve the above objectives, packages are designed to have the following characteristics: 1) low lead capacitance and inductance; 2) material compatibility; 3) good thermal conductivity; 4) good hermetic integrity; 5) ease of manufacture; 6) low cost; and 7) stress levels that will not harm the chip or package. ICs in packages are also easier to handle and test than bare die. They make it easier to compensate for the mismatches in coefficients of thermal expansion among the materials used in the construction of electronic systems. Chip packages, however, also impact the cost, reliability, and sometimes even the performance of the IC (and maybe even the system in which the IC is being used). Reduction in the prices of electronic systems have primarily been driven by the shrinkage of the IC devices themselves, but improvements in packaging have also helped decrease these costs. Packaging cost is important because it can account for a significant fraction of the total cost of the packaged IC. This cost, as measured by cost per pin, has increased from about 1 cent ($US) per pin in small-scale and medium-scale ICs (SSI and MSI), to about 10 cents per pin for ULSI. For ULSI devices in ceramic packages, the package cost may exceed the chip cost by more than 2 times.

In the upcoming sections IC packages will be described. The discussion begins by comparing hermetic (ceramic) and non-hermetic (plastic) packages. Then various IC package types (i.e., through-hole versus surface mount, and specific package configurations) will be analyzed. Finally, so-called "packageless" techniques for ICs are described.

IC package technology is generally divided into two categories, namely: *hermetic* and *non-hermetic* (i.e., plastic) packages. In hermetic packages the chip is housed in an environment isolated from the external world by a vacuum-tight enclosure. The package material is usually ceramic-based, and the two most widely used hermetic-package types are *ceramic packages* and *glass-sealed refractory (cerdip) packages.* Such packages are used in high-reliability

applications and/or when the heat generated by the chip is too great for a plastic (i.e., non-hermetic) package. Since hermetic packages are much more costly than plastic, the applications that employ them must be able to justify such cost penalties. For example, almost all microprocessors (which dissipate large amounts of power and have high average selling prices [ASPs]) go into ceramic packages.

In non-hermetic packages, the chip is not completely isolated from the outside ambient, as moisture or other substances can penetrate the polymer material (an epoxy resin) that encapsulates the chip. Such plastic packages, however, are much cheaper than the hermetic packages, and their reliability and performance characteristics have significantly improved since the early days of the IC industry. Hence, they have become by far the most widely used IC package type (in terms of the number of chip packages produced each year). Note, however, that because ceramic chip packages are much more expensive than plastic packages, the sales volume of ceramic packages constituted roughly two-thirds of the total packaging market in the early 1990's. The package industry is a multi-billion dollar industry worldwide, and in 1998, 62 billion IC packages were produced.

17.7 PACKAGING TECHNOLOGY I: HERMETIC PACKAGES (CERAMIC and CERDIP)

17.7.1 Ceramic Packages

A typical hermetic package is shown in Fig. 17-17. The base of the package is made of a multi-layer ceramic material, and the chip sits in a cavity of this base. The layers of ceramic material are formed by casting thin sheets of alumina* from a slurry consisting of alumina powder, solvent, and resin binder. After these sheets have dried they are cut to size. Cavities and via holes (i.e., holes that extend through the ceramic sheet so that interconnections can be made to the inside of the package) are then mechanically punched into the sheets. Next, custom wiring paths are screen printed with a slurry of a refractory metal powder (tungsten or molybdenum) onto some of these sheets (Fig. 17-18), and the via holes are filled with metal. Refractory metals must be used for these conductors because of the high temperatures required for sintering the alumina. Several sheets (up to 30, in some cases) are then stacked together and placed in a press where they are laminated together under pressure and moderate temperature.

Next, this structure is fired at high temperature (1600°C) in a reducing atmosphere (to prevent oxidation of the metal circuit lines). This firing operation completely burns out any binder or solvent remaining in the solidified ceramic sheets. It also densifies the ceramic and metal components to form a single monolithic structure (which constitutes the package base). Note that the wiring paths formed on the surface of some of the ceramic sheets are now sandwiched between layers of ceramic material, and they "daylight" at the edges (sides or bottom) of the package. These metal paths are now an integral part of the package material. However, after the firing operation, the areas where such metal is exposed on the exterior of the package must be plated (e.g., with nickel) for protection against corrosion, and to allow attachment of external leads to the package. For *leaded packages*, external pins are bonded to the metal regions that penetrate the package surface. For example, in the *dual-in-line* ceramic

* Note that the modest thermal conductivity of alumina has made other materials (beryllia [BeO] and aluminum nitride [AlN]) attractive for applications requiring better thermal performance, despite their much higher cost.

packages (DIPs), leads made of a Fe-Ni-Co alloy (called *Kovar*) are brazed to the metal along the sides of the package (Fig. 17-19), while in *pin-grid-array* ceramic packages, leads are attached to the bottom of the package base (Fig. 17-20). These leads are then plated (e.g., with Au over Ni) for bondability and environmental protection. In *leadless packages*, external metal patterns are formed on the outside of the package. At that point, the package is ready for die-attach, wire-bonding, and package sealing.

After the chip-to-package connections are made, a package-sealing process completes the hermeticity. A metal or ceramic lid is placed over the cavity and it is hermetically sealed to the base. The most widely used method for this sealing process is welding, because it is a rapid, reliable, and highly reproducible technique. In welding, high-current pulses produce local heating of 1000–1500°C, fusing the lid (or the plating thereon) to the package. Since this heating is localized, damage to internal components is avoided.

17.7.2 Glass-Sealed Refractory Packages

Glass-sealed refractory packages offer the advantages of hermetic sealing but at significantly lower cost than ceramic packages (albeit, resulting in a somewhat less reliable package). These packages consist of two ceramic pieces (a *base* and a *cap*), cut to size and formed with cavities. A paste made from a low-melting point glass (i.e., having a high Pb content), is screened onto the periphery of the mating surfaces of both ceramic pieces. A stamped metal lead frame is embedded and aligned in the paste on the base (with the leads extending beyond the edge of the base). The base with the lead frame is passed through a furnace at 450°C to temporarily soften the glass, so that the frame becomes rigidly seated in the glass when it cools. A chip is then die-bonded to the package base and wire bonds are made to the lead frame. The ceramic cap is next placed over the assembly and the two ceramic pieces are pressed together and heated to about 400°C. This causes the glass on the base and cap to flow together and form a hermetic seal. This glass not only forms the seal, but also holds the lead frame (which is sandwiched between two ceramic pieces) and the ceramic pieces together. Note that since Au-Al intermetallics form at the temperatures needed to flow the glass (~400°C), Al instead of Au wires must be used for wire-bonding in such packages. There are two drawbacks of glass-sealed

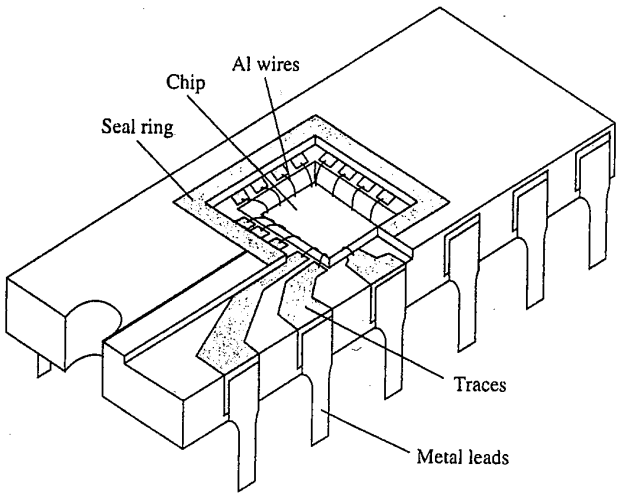


Fig. 17-17 A typical ceramic package.[20] © IEEE 1979.

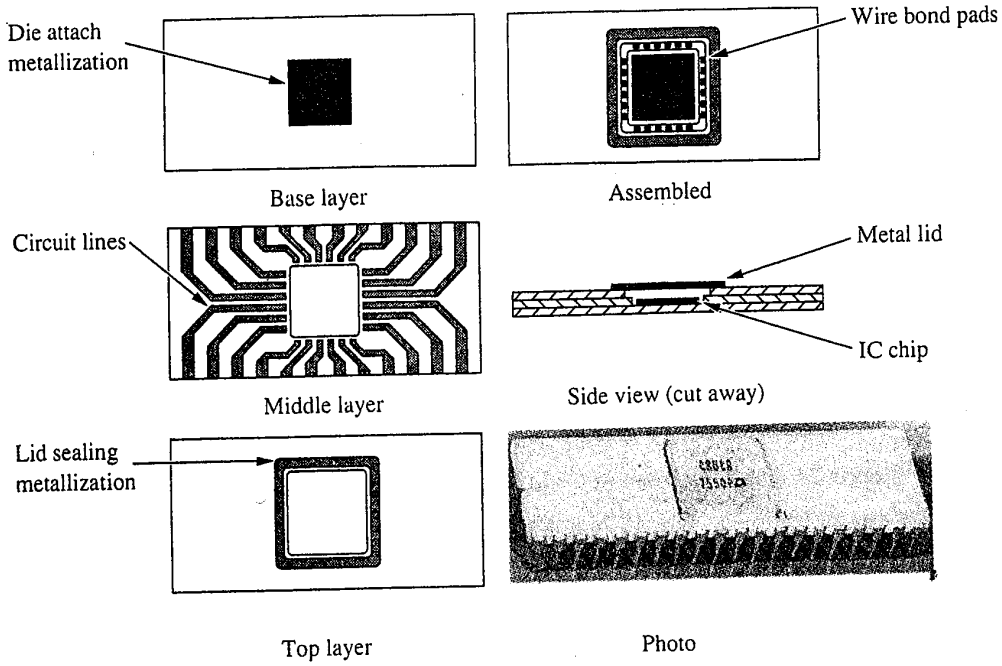


Fig. 17-18 The parts of a ceramic dual in-line package.[1] Reprinted with permission of McGraw Hill.

refractory packages. First, moisture is released by the glass during the cooldown after closure of the package has been made. Such moisture then remains trapped inside the sealed package. Second, residual stresses that remain in the glass after it cools can cause cracks. Such cracks may destroy the hermeticity of the package. Both through-hole-type packages (ceramic DIPs, or *CERDIPs*) and surface-mount-type packages (CERQUADs) are available in this technology.

17.8 PACKAGING TECHNOLOGY II: PLASTIC PACKAGES

As noted earlier, the most widely used packaging technology involves encapsulating the IC in molded plastic, with more than 90% of ICs in the late 1990's being assembled this way. The dominance of plastic packages is primarily due to their much lower cost. However, this advantage is traded off against the limitation of reduced reliability. Plastic does not completely isolate the chip from the environment, and moisture or other substances can penetrate the plastic and cause corrosion of the metals on the chip. Nevertheless, the reliability of plastic packages has increased to the degree that they can satisfy many ULSI applications (especially consumer

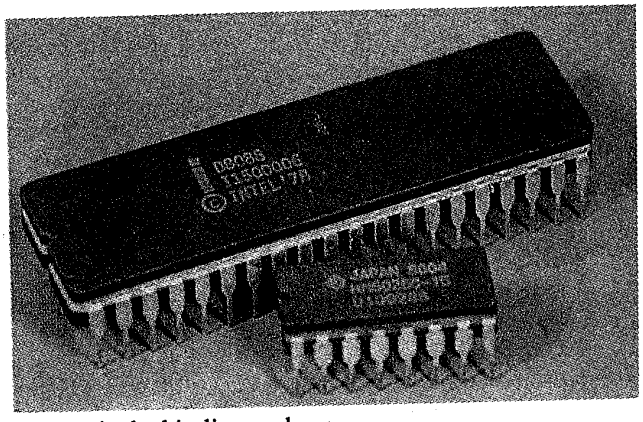


Fig. 17-19 An assembled ceramic dual in-line package.

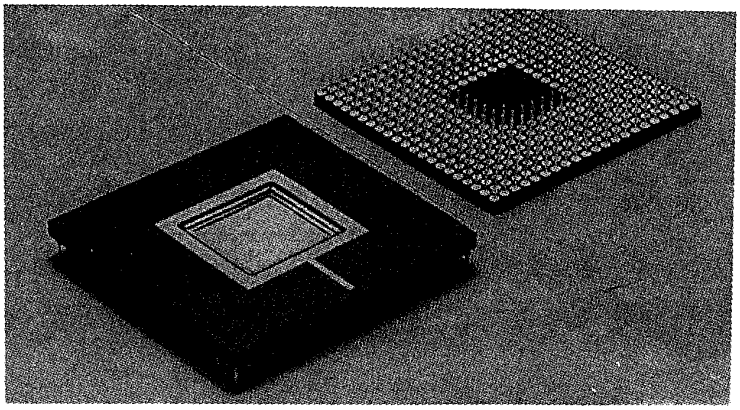


Fig. 17-20 Ceramic pin grid array (PGA) package. a) Cavity up configuration. b) Cavity-down configuration.[1] Reprinted with permission of McGraw-Hill.

products). For example, in the late 1970's it would take less than 1000 hours to induce a failure in a plastic packaged IC by an accelerated stress test. By the late 1980's the parts could survive 5000 hours under the same test conditions. Such parts operate reliably for much longer periods of time under normal operating conditions (e.g., >10 years). It may now be more appropriate to characterize plastic packages as being "less hermetic." A detailed description of molded plastic package technology is given in Ref. 21.

Figure 17-21 shows a typical plastic DIP (see Sect. 17.9.1 for a description of DIPs). This type of package consists of a lead frame and plastic molding material. The lead frame is made from a thin (0.25-mm thick) chemically milled or stamped sheet of a Fe-Ni or Cu-based alloy. The most commonly used lead frame material is Alloy 42 (42%Ni-58%Fe). The lead frame is, in effect, a skeleton around which the packaging material is formed. But it also provides the external leads in the assembled package. The chip is die-attached to the lead-frame paddle, and the bond pads on the chip are connected to the radiating lead frame terminals by gold wire bonds. The lead frame is then encapsulated with a thermoset epoxy resin, which produces a rigid plastic body that completely surrounds the chip. The most common method of such encapsulation is transfer molding, as is discussed in the following paragraphs. After the plastic encapsulation is completed, the leads are bent down to the configuration shown in Fig. 17-21.

The most widely used materials for forming the rigid body of the plastic packages are epoxy novolac-based molding compounds. These compounds typically consist primarily of novolac epoxy resin (25-30%); inert filler material (68–70%); and flame retarder (2%). The remaining small percentage of material consists of an accelerator, curing agent, mold-release agent, and colorant. The epoxy resins originally used for IC packages had high concentrations of chlorine (up to 3%), but by the late 1970's this concentration was reduced to less than 30 ppm. The fillers are used to reduce the thermal coefficient of expansion (TCE) of the compound and to increase the thermal conductivity of the plastic. Powders of fused silica are fairly common filler materials. Mold-release agents prevent the cured compounds from sticking to the mold cavity walls, facilitating easy removal of the plastic package. The following characteristics are desired of the plastic molding compound: low resin viscosity; good wetting of chip components; good adhesion to chip components; short cure time; low TCE; good dimensional stability; and low moisture permeability.

Transfer molding is the standard technique for encapsulating ICs in plastic packages. The process begins by loading the lead frames (carrying the die-attached and wire-bonded chips) into the mold *cavities* of a transfer-molding press (Fig. 17-22). These cavities define the size,

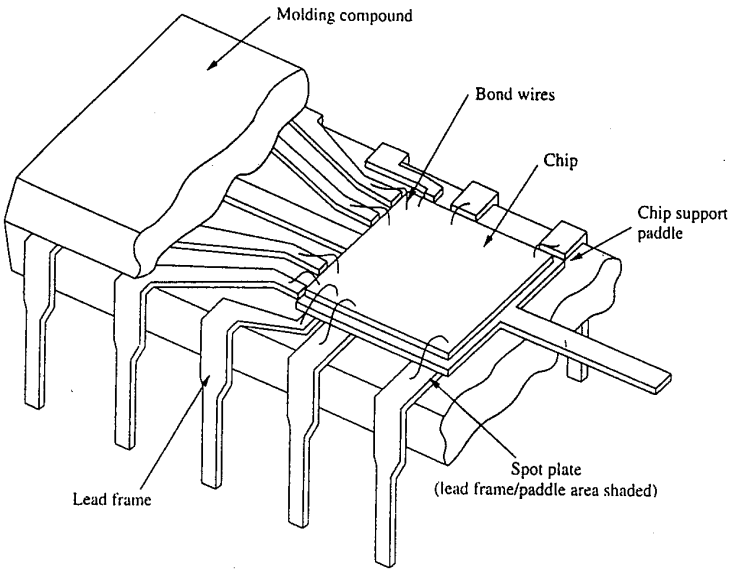


Fig. 17-21 A typical plastic package.[22] © IEEE 1981.

shape, and surface finish of the plastic IC package. The mold has two halves (top and bottom), that open to receive the lead frames. Theses are closed during the molding cycle. A mold has multiple cavities (up to 400 for DIPs). A solid resin pellet is put into a *pot* at the center of the mold, where it is heated until it melts. When the heated resin has been liquefied to the correct viscosity, a hydraulically driven piston forces it from the pot, through *runners*, to the mold cavities. The resin flows into each cavity at a controlled rate determined by small openings called *gates*. (Since the bond wires are long and fragile, care must be taken to prevent them from being damaged or deformed by the flowing resin. This is accomplished by controlling the flow velocity and the viscosity of the resin to within certain ranges.) Once the resin has filled the cavity and surrounds the lead frame and chip, curing within the mold takes place (typically for 1 to 5 minutes at 175°C and at high pressure [e.g., 6 MPa]). During curing, some of the mold-release agent is also extruded to the plastic/mold cavity interface, forcing a lubricity between the mold and molded part. After the molding process is completed, the mold is opened and the pack-ages are removed. The mold is then cleaned of any remaining resin from the pot and runners.

Post-cure storage at 175°C (i.e., after the packages are removed from the mold) is needed to complete the epoxy resin crosslinking. The packages are then subjected to a *trim-and-form* procedure. That is, the outer leads of DIPs and flat-packs are made with a *tie bar* (which keeps the leads from being bent during the packaging process). The trimming process simultaneously cuts away the tie bar and trims the leads to their proper length. In the forming process, the leads are bent into their final position. A *deflashing* step is next performed to remove excess molding material from the package body. (*Flash* on molded plastic parts is that small amount of excess material which flows out through the parting line where the mold halves meet.) Deflashing is typically done by a physical abrasion process, on a machine similar to a sand blaster, with plastic beads as the abrasive.

Once completed, the package must be identified with key information, in a process called *marking*. Typical information coded on the package is the product type, device specifications, date of manufacture, lot number, and where it was made. The main methods of marking are ink printing and laser inscription. Laser printing is especially well-suited for plastic packages.

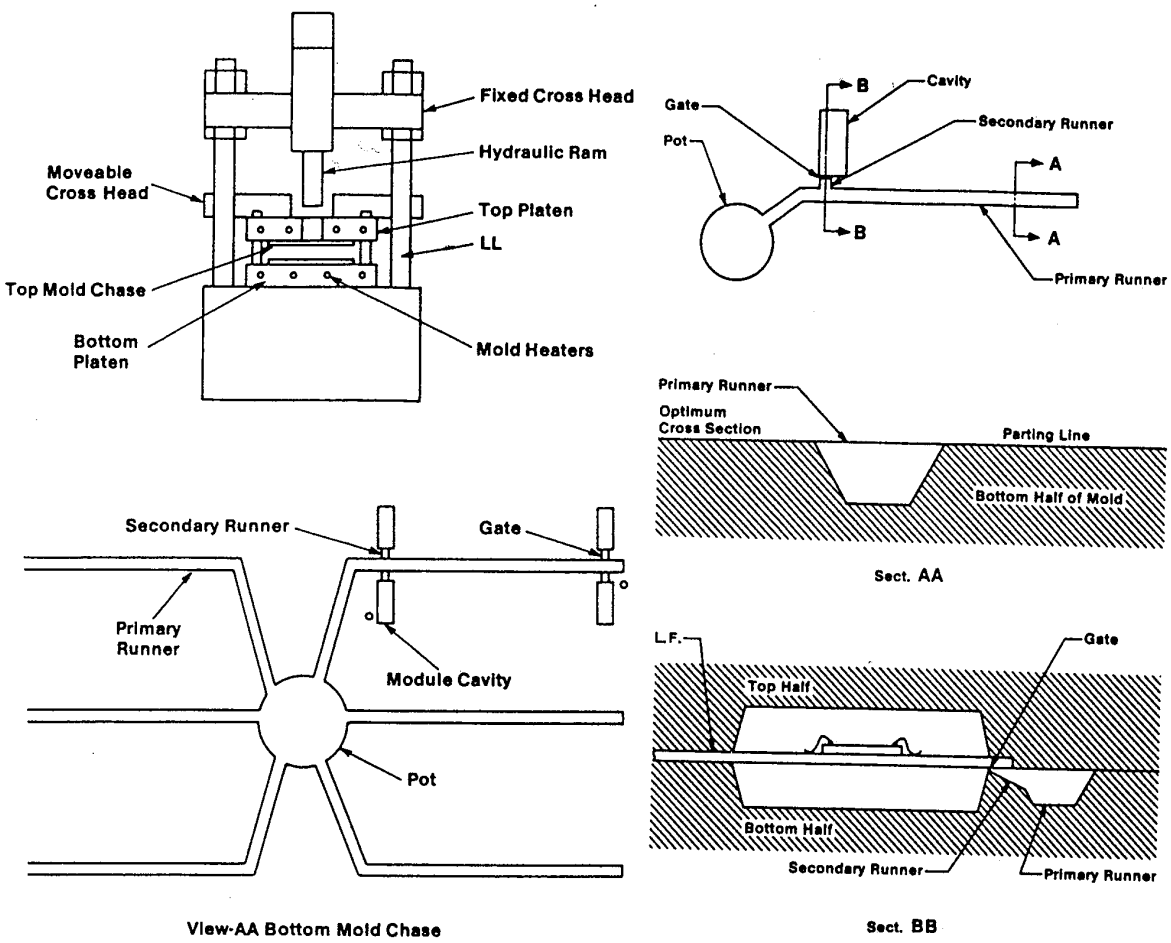


Fig. 17-22 The transfer molding process for plastic IC packages. (a) Transfer-molding press schematic. (b) Bottom mold chase. (c) Runner design, in detail. (d) Cavity and runner cross sections.[1] Reprinted with permission of McGraw-Hill.

17.8.1 Plastic Package Reliability Problems

The failure mechanisms inherent to plastic packaging are those due to mechanical stresses, corrosion, and moisture induced failures. The stresses that cause IC failure arise from epoxy shrinkage during cure, and TCE mismatches between the molding compound, lead frame, and chip materials. When the resins cure, they also shrink in volume by about 3 to 5 per-cent. Silicon has a TCE of $2.6 \times 10^{-6}/^{\circ}C$. The TCE of a typical filled-epoxy resin is $24 \times 10^{-6}/^{\circ}C$. As the package cools from 175°C to room temperature, a temperature differential of over 100°C occurs, creating a large strain mismatch. The resulting stresses can shear and completely lift off wire bonds. Measures to reduce these stresses include the use of compliant materials between the plastic and the chip (such as polyimides), or the use of special-ly formulated low-stress molding compounds. (See Ref. 2 for details).

The epoxy packing material also contains filler particles, which are non-spherical (e,g,, they are triangular, or pointed instead). As a result they can exert a tremendous downward force on the die, causing the passivation layer to crack. A 5-μm thick polyimide layer is therefore applied

to the die to act as a stress-relief layer which can deform to accommodate forces exerted by the filler particles. This radially dissipates such forces so they are not transmitted to the wafer surface. If no polyimide coating is used, severe cracking of the passivation layer is likely.

When the resin cools, it shrinks and exerts compressive forces around the chip, wires, and lead frame. This forms a compressive seal around the chip and lead frame. However, this seal can be penetrated by moisture or other substances to cause corrosion of the chip metallurgy. Water is transported to the chip from outside the plastic package along two avenues: 1) the interface between the lead frame and plastic; and 2) by diffusion through the plastic molding material itself. The former problem can arise when the plastic and lead frame materials separate to some degree. In a humid environment, moisture and chemical components from the PWB can migrate along the leads, climb across the wire bonds, and finally arrive at the chip metallization on the die surface. Any chlorine that is present at the chip surface can also react with the moisture to form HCl, which will corrode the Al metal on the pads. Additional information on plastic package reliability can be found in Ref. 23.

17.9 ULSI PACKAGE TYPES

ICs are also used in electronic systems. The level of packaging in such systems above that of the chip package, is the printed wiring board (PWB). The two ways IC packages are attached to the PWB are: 1) they are soldered into holes created in the PWB (Fig. 17-23); and 2) they are soldered to the surface of the PWB. Thus, two types of IC packages are needed, depending on whether *through hole* (TH) or *surface mounting* (SM) of the IC packages is required by the PWB. Hermetic and plastic packages are available for both TH and SM PWB applications.

17.9.1 Through Hole (TH) IC Packages

There are a number of the most popular TH package configurations. The earliest TH IC package was the *dual-in-line package* (DIP). It became extremely popular due to its low cost and ease of use. In fact, it has become one of the icons of the IC industry. DIPs are still widely used, and hermetically sealed ceramic DIPs (Fig. 17-19) and plastic DIPs (Fig. 17-21) are available. DIPs are thick, sturdy packages with two rows of outer leads coming out of the sides of the package and bending down. While they are available with 4 to 80 leads, DIPs become less satisfactory for ICs with larger pin counts.

The *pin grid array* (PGA) has been developed to house such large I/O circuits on TH PWBs, and PGAs are available in ceramic or plastic package form. The leads come out of the bottom of the package in the form of pins (Fig. 17-20), and these pins are arranged in an x-y array. PGAs can have hundreds of pins. The chip is attached in a cavity that is formed in either the top or bottom of the package body. TH packages offer the advantages of easier placement for soldering and stronger solder joints at the PWB level. However, they restrict PWB design flexibility, and limit the chip mounting density. This is because there is a limit to the ability of drilling fine through holes. Other TH package types include the *single-in-line package* (SIP, Fig. 17-24a) and the *zig-zag-in-line package* (ZIP, Fig. 17-24b).

17.9.2 Surface Mount (SM) IC Packages

Figure 17-25 shows some of the many SM package configurations that have been developed. Although TH technology was developed first, by 1998 about 80% of IC packages being used were surface mounted. SM packages enhance PWB design flexibility because it is not necessary

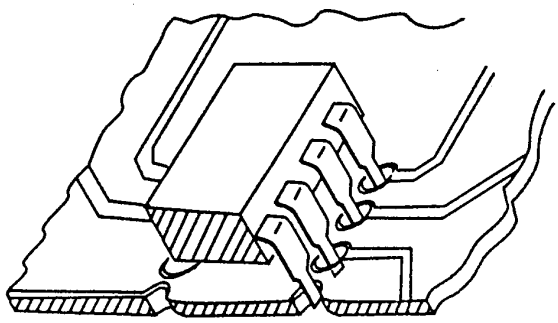


Fig. 17-23 DIP through-hole assembly.

to drill through holes in the PWB. Finer pitches of the PWB circuit traces are possible because the pitch is limited by the etching technology of Cu foil, and not by the drilling of through holes. Many of the first SM packages were leadless ceramic packages mounted onto organic PWBs. Unfortunately, they had a tendency to cause cracked solder joints and broken packages. As a result, SM packages now use either *J-leads* (Fig. 17-26a) or *gull-wing leads* (Fig. 17-26b). Where a smaller number of package leads are required, J-lead packages are often preferred. The solder joints of the J-leads are formed beneath the package, and hence they need less PWB space than do gull-wing-lead packages. The *small-outline-J-lead* (SOJ) and the *quad-flat-J-lead* (QFJ) are examples of such packages.

On the other hand, J-lead bonds are more difficult to fabricate with high yields at fine pitches. Thus, for circuits that have large pin counts, gull-wing-lead packages are more advantageous. Gull-wing-lead joints are easier to form and inspect. Thus, finer spacings between the joints allow circuits with larger pin counts to be accommodated with smaller packages. The *small-outline package* (SOP) and the *quad flat package* (QFP) are examples of gull-wing SM packages. The *plastic quad flat pack* (PQFP) is an inexpensive version of the QFP. *Thin small outline packages* (TSOP) have extended flat leads on two sides of the package (Fig. 17-26c & d).

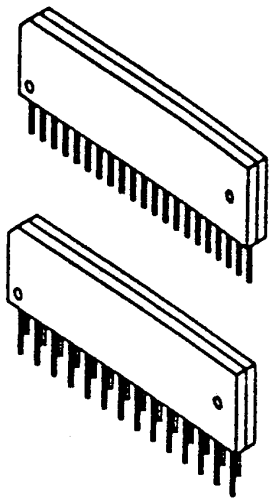


Fig. 17-24 a) Single-in-line package (SIP) and the zig-zag-in-line package (ZIP).

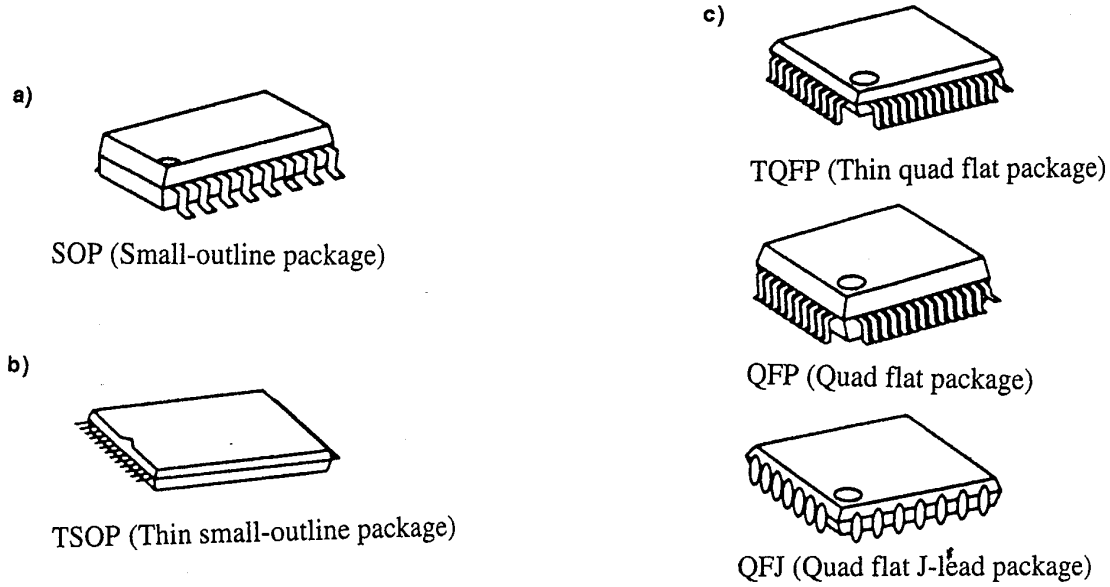


Fig. 17-25 Some of the many surface mount packages that have been developed: a) Small outline package (SOP;) b) Quad flat pack (QPF); c) Thin small outline package (TSOP).

17.10 TRENDS IN ULSI IC PACKAGES

QFP and PQFP configurations become cumbersome for circuits with I/O counts above about 208 pins. To accommodate chips needing larger lead counts than this, QFPs must either increase in size (to where they take up too much board space, or degrade circuit performance), or the package lead pitch must be decreased to the point where assembly yields fall precipitously.[23] There has thus been a move toward developing area-array packaging, with the *plastic ball-grid-array package* (PBGA) as the bellwether of this format. The ball grid array package is fundamentally a re-engineered pin-grid-array package, with the pins replaced by balls. The balls are usually 0.8-mm Pb:Sn solder bumps placed on the back surface of the package at a spacing of 1-2 mm. The bumps are attached to the package after the plastic molding process.

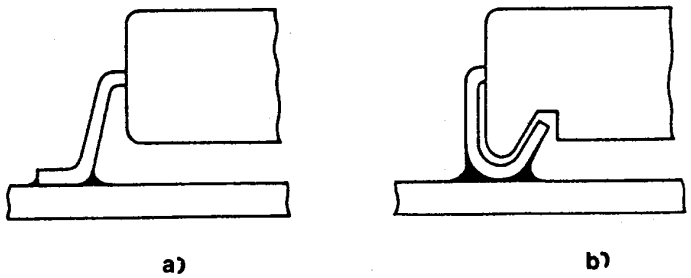


Fig. 17-26 a) SM package with J-leads. b) SM package with gull-wing leads.

Another new family of IC packages followed on the heels of the PBGA. They are collectively referred to as *chip-scale packages* (CSP, Fig. 17-27).[23,24] Their name stems from the fact that they have such a small form factor (i.e., the footprint of a CSP is no larger than 1.2x that of the chip itself). CSPs appear to offer a path toward packages in which the cost will approach the 1 cent/lead cost of the SSI and MSI package era. Their small size is also an aid to improving the package performance. CSPs can be mounted to low-cost PWBs with standard SM techniques. In 1998, CSPs were tagged as the fastest growing segment of the packaging market.

A number of different approaches are being developed to produce chip-scale packages. Some use wire bonding, others TAB, and still others flip-chip techniques to interconnect the chip and package. Perimeter-leaded CSPs are suited for small lead count packages (e.g., 50–100 leads). High I/O CSPs use an area array (e.g., a BGA) for the external lead arrangement. It is predicted that such packages may eventually handle up to 1000 I/Os. Some packaging suppliers offer processing services that prepare a finished wafer for their package type (e.g., they fabricate the TAB or flip-chip bumps on the wafer.) Another innovation is the use of a flexible *interposer* (the electrical inter-connection structure between an IC and its package). In Tessera Corp's μBGA CSP, the interposers flexibility decouples the IC thermal expansion from the package substrate, an important aspect of the package reliability.[15]

17.10.1 Packages for Memories and Microprocessors

Different IC product types have varying packaging requirements. Memory chips, for example, require low-cost packages, because they have become commodity items, in which the overall manufacturing cost must be made as low as possible. Thus, memory chips are almost invariably housed in plastic packages, and SM packages also predominate. As noted earlier, memory chips require a relatively small number of I/O pins, and that the bonding pads on the chips are usually placed on oppositely facing edges of the chip. A variety of dual-sided SM package types are therefore used for memory chips, including SOJ, SOP, and TSOP packages. A lead-on-chip (LOC, see Fig. 17-28) has also recently been developed for such applications.

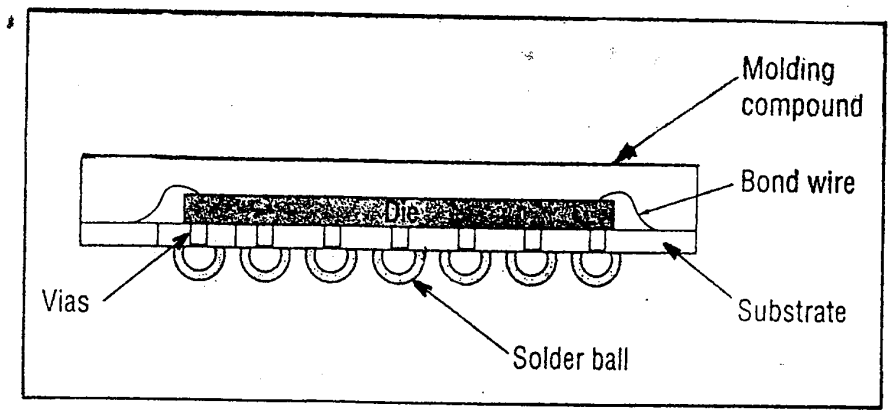


Fig. 17-27 Example of a chip-scale package (CSP).[17] Reprinted with permission of Solid State Technology, published by PennWell.

Memory chips have quadrupled in density every three years since their inception in 1970. By 1998, 64-MB DRAMs were commercially available. With the shrinking of many electronic systems in which memory chips are used (e.g., laptop computers), it has become necessary for the packages which house the memory chips to remain about the same size (despite the fact that the size of the chips themselves have grown with each generation). Memory package advances have allowed this requirement to be met by increasing the ratio of the memory chip-to-package area. In 16 MB DRAMs, this ratio was about 70%.

Logic and microprocessor (μP) ULSI ICs, on the other hand, generally require a much larger I/O pin count than do memory chips. Circuits with gate counts of 1 million transistors require packages with more than 700 pins. (For example, the Intel Pentium II has 500 pins.) For ICs that require such a large number of pins, pin-grid-array packages (PGAs) are used for TH applications, and ball-grid-array packages (BGA) for SM applications. For logic and microprocessor chips with up to 300 pins, QFP (and PQFP) packages predominate. For applications that require a low profile, low-profile QFP (LQFP) and thin QFP (TQFP) are used.

The increase in the operating speeds of the logic and μP chips has also raised the power consumption of these ICs. In CMOS logic circuits, the power consumption is directly proportional to the clock operating frequency. By 1998, some state-of-the-art μPs consumed as much as 30 watts. Package designs which permit such a large amount of power to be effectively dissipated must be used (including more elaborate cooling techniques).

17.10.2 Packageless Technologies

Despite the fact that IC packages perform an important set of tasks in electronic systems, they have also been viewed by some as merely a necessary evil. It has been argued that the IC package adds no value to the product, but instead enlarges the physical presence of the IC and robs it of some of its performance potential. As a result, some systems designers have sought to rid themselves of the IC package by interconnecting ICs directly to PWBs. In fact, such *direct chip attachment* (*DCA*) methods as flip-chip (or C4), and beam-leaded chip technologies were introduced in the mid-1960's by IBM and AT&T, respectively. Yet, until recently the overall advantages offered by IC packages prevented such "packageless" approaches from being widely implemented. By the late 1990's, however, the costs of IC packages for ULSI had risen precipitously. Since packageless technology appears to be a method in which costs can be reduced, interest in them has revived.[26] In fact, this approach has been adopted for such applications as watches and smart cards. Since rework of failed circuits is not yet feasible in packageless applications, this attachment technology is usually reserved for mature, high-yielding circuits, rather than seeking known-good-die (KGD) for attachment.

Proponents of CSP pointed out that packaging procedures are not actually eliminated by such techniques as FC-DCA. Instead, the packaging steps are performed *in situ* on the substrate where the die is flip-mounted and is then either encapsulated or underfilled. Thus, the costs of packaging are not completely circumvented. CSPs were described by their proponents as offering even lower costs than packageless technologies.[28] Other experts, however, contended that FC-DCA is the ultimate IC package solution, and that CSPs are just stepping stones to FC-DCA. By 1998, a FC-DCA technology from Flip Chip Technologies called *Ultra CSP* was offered at a cost of about 0.7 cents/pin. These packaged chips reportedly demonstrated reliable operation in demanding automotive applications.[16] It seems, in any case, that CSP and FC-DCA will both be rapidly adopted for electronic systems beyond the year 2000.

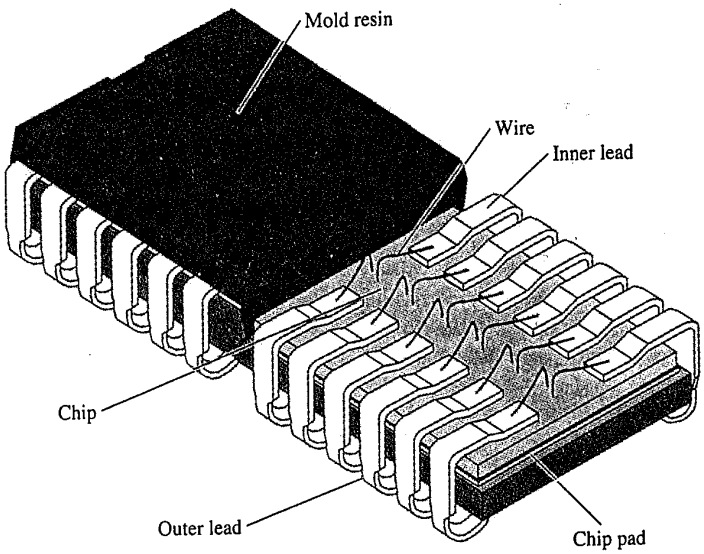


Fig. 17-28 A typical lead-on chip (LOC) package. Tips of the lead frame extend above the chip center where the bonds are made.[27]

REFERENCES

1. *Principles of Electronic Packaging*, Eds. D.P. Seraphim, R. Lasky, and C.-Y. Li, McGraw-Hill Book Company, New York, 1989.
2. *Microelectronic Packaging Handbook,* Eds. R.R. Tummala and E.J. Rymaszewski, Van Nostrand Reinhold, New York, 1989.
3. C. McHatton and C.M. Gumbert, "Eliminating Backgrind Defects with Wet Chemical Etching," *Solid State Technology,* November 1998, p. 85.
4. J. Baliga, "Package Styles Drive Advancements in Die Bonding," *Semiconductor International,* June 1997, p. 101 & Dec. 96. p. 89.
5. R. Shukla and N. Mencinger, "A Critical Review of VLSI Die-Attachment in High-Reliability Applications," *Solid State Technology,* July 1985, p. 67.
6. I. Chien and M.N. Nguyen, "Pursuing the Perfect Paste: A Novel Cyanate Ester Die Attach Adhesive for Plastic Packaging," *Solid State Technology*, November 1994, p. 87.
7. J. D'Ignazio, "Wirebonding's Reign Continues," *Semiconductor International* June 1996, p. 117.
8. C.A. Steidel "Assembly Techniques and Packaging," in *VLSI Technology*, S.M. Sze p. 551, McGraw-Hill, New York, 1983.
9. R. Iscoff, "High Volume Wire Bonding," *Semiconductor International*, November 1994, p. 73.
10. G.G. Harman, "The Microelectronic Ball-Bond Shear Test: A Critical Review and Comprehensive Guide to Its Use," *ISHM*, 1983.
11. G. Dehaine *et al.,* "Crossing the 200 Wire Barrier in VLSI Bonding," *Proc. ISHM* Symp. pp. 353-356, September, 1984.
12. N.G. Koopmans, T.C. Reiley, and P.A. Totta, "Chip-to-Package Interconnections," in *Microelectronics Packaging Handbook*, R.R. Tummala and E.J. Rymaszewski, Eds., Ch. 6, p. 409-435, Van Nostrand Reinhold, New York, 1989.
13. G. Oswald and W.R. deMiranda, "Application of Tape Automated Bonding Technology to Hybrid Circuits," *Solid State Technology*, pp. 33-38, 1977.
14. P.A. Totta and R.P. Sopher, "SLT Device Metallurgy and Its Monolithic Extension," *IBM Journal of Research & Development*, **13**, May 1969, p. 226.
15. D.L. Brownell and G.C. Waite, "Solder Bump Flip-Chip Fabrication Using Standard Chip and Wire Integrated Circuit Layout," *1974 ISHM Conf. Proc.,* p. 77, September 1974.
16. R.H. Estes, "Flip-Chip Packaging with Polymer Bumps," *Semicond. Internatl*, February 1997, p. 103.

17. P. Burggraaf, "Chip Scale and Flip Chip: Attractive Solutions," *Sol. State Technol.*, July 1998, p. 239.
18. A.J. Babiarz, "Key Process Controls for Underfilling Flip Chips" *Sol. State Technol.* April 1997, p. 77.
19. G. Murakami, "Packaging Industry: Road Map to the Future," *Semicon/West Proc.* p. 99-123, 1994.
20. S. Prokop and D.W. Williams, "Chip Carriers as a Means for High-Density Packaging," *IEEE Trans. Components, Hybrids, Manuf. Technol.*, **3**, 297 (1978).
21. P.V. Robock and L.T. Nguyen, "Plastic Packaging," in *Microelectronics Packaging Handbook,* R.R. Tummala and E.J. Rymaszewski, Eds., Ch. 8 pp. 523, Van Nostrand Reinhold, New York, 1989.
22. J.R. Howell, "Reliability Study of Plastic Encapsulated Copper Lead Frame Epoxy Die Attach Packaging System," *Proc. Internat. Reliabil. Symp.* 1981, p. 104.
23. D.A. Jeannotte, L.S. Goldmanm, and R.T. Howard, "Package Reliability," in *Microelectronics Packaging Handbook*, R.R. Tummala and E.J. Rymaszewski, Eds., Ch. 5, p. 225-359, Van Nostrand Reinhold, New York, 1989.
24. J. Baliga, "Flip-Chip Packaging: Prepare for the Ramp-Up," *Semicond. Internatl*, March 1998, p. 87.
25. T. DiStefano and J. Fjelstad, "Chip Scale Packaging Meets Future Design Needs," *Solid State Technology*, April 1996, p. 82.
26. J. Fjelstad, "Trends in Low Cost IC Packaging," *Semiconductor Internat.*, Dec. 1996, p. 73.
27. N. Ueda "Studies of a New LOC Structure Package," *Proceedings of the VLSI Packaging Workshop,* III-4, 1993.
28. "Chip-on-Board are Packageless Solutions," *Semiconductor International*, December 1996 p. 73.

PROBLEMS

1. Name the four functions of a chip package
2. Define a hermetic package, and give an example of a hermetic and a nonhermetic package
3. There are five stages in IC manufacturing at which yield is calculated: 1) process line yield, Y_t (the fraction of wafers that the survive up to pad mask opening); 2) parametric test, Y_p; 3) wafer sort, Y_s (preliminary functional test); 4) assembly Y_a; and 5) final functional test, Y_f. If the total yield $Y_{Tot} = Y_t Y_p Y_s Y_a Y_f$, calculate Y_{tot} if the yield at each stage is: (a) 0.8; (b) 0.9, (c) 0.95, and (d) 0.99.
4. What chip size is required to place 352 pads (I/Os) on a pad-limited chip? Assume the chip is square, the bonding pad pitch is 120 μm, and the chip size is defined by the loci of bonding-pad centerlines.
5. A new circuit design is estimated to require a die which is 5 mm x 8 mm and will be fabricated on a wafer 200 mm in diameter. The process is achieving a defect density of 2 defects/cm^2, and the manufacturing cost is $350 per wafer: (a) What will be the cost of the final product if testing and packaging adds $1.60 to the completed product? (b) The circuit design could be partitioned into two chips rather than one, but each die will increase in area by 20% in order to accommodate the additional pads and I/O circuitry. If the testing and packaging cost remains the same, what is the cost of the two-chip set? Base your answers on the yield equation given in Problem #3 in Chap. 5.
6. If a package cavity has a volume of 0.1, and the water vapor trapped in the cavity after the package is sealed is present at a concentration of 5000 ppm by volume. How much silicon (in grams) is required to react with all of this water? Assume atmospheric pressure and room temperature within the cavity.

PROBLEMS