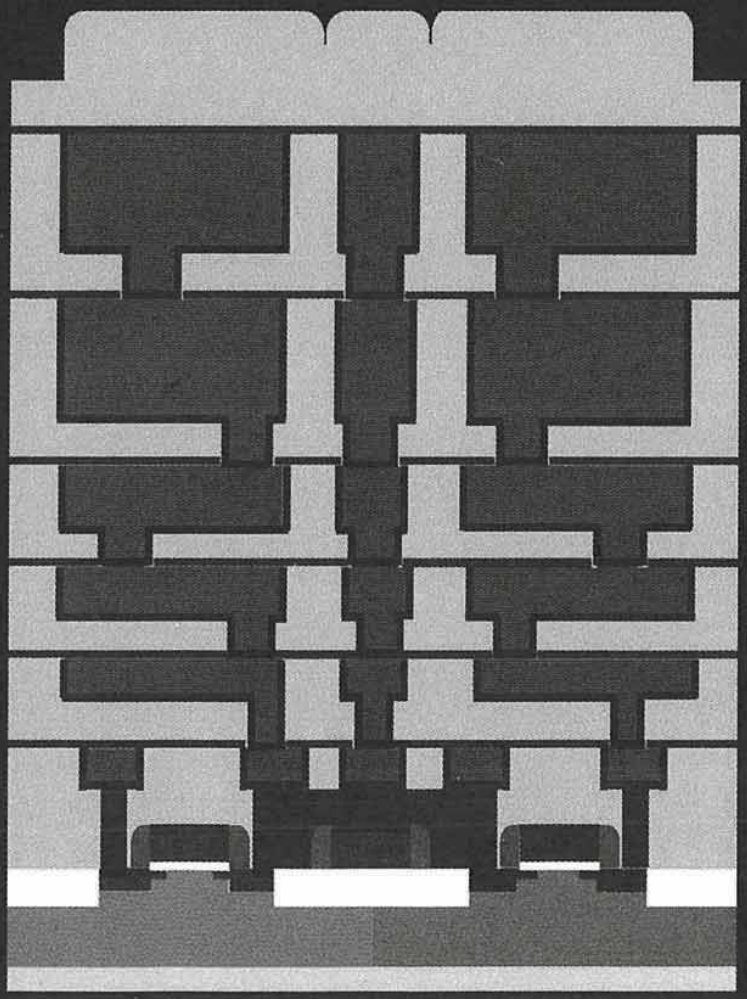
Silicon Processing

for the VLSI Era

Volume 4 - Deep-Submicron Process Technology



S. Wolf

Intel v. BeSang
IPR2023-00991
Exhibit 1015

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: S. C. Ross Wolf.

Copyright © 2002 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

Silicon Processing for the VLSI Era
Volume 4: Deep-Submicron Process Technology

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

ISBN 0-9616721-7-X

9 8 7 6 5 4 3 2 1

PRINTED IN THE UNITED STATES OF AMERICA

MASTER TABLE OF CONTENTS

PREFACE

Chap. 1 - THE EVOLUTION OF THE STRUCTURE OF MOSFETS 1

Chap. 2 - 300-MM SILICON WAFERS 17

Chap. 3 - GATE DIELECTRICS: THIN GATE OXIDES 75

Chap. 4 - HIGH-*k* DIELECTRICS 145

Chap. 5 - THE STRUCTURE OF DEEP-SUBMICRON MOSFETS: 181
Retrograde-Wells; Super-Steep Retrograde Channel Doping; Drain Engineering; Punchthrough-Control Implants

Chap. 6 - DEEP-SUBMICRON LITHOGRAPHY I: PHOTORESISTS 227

Chap. 7 - DEEP-SUBMICRON LITHOGRAPHY II: 259
OPTICS AND HARDWARE

Chap. 8 - CHEMICAL MECHANICAL POLISHING (CMP) 313

Chap. 9 - SHALLOW TRENCH ISOLATION (STI) 433

Chap. 10 - SILICON-GERMANIUM (Si-Ge) TECHNOLOGY 475
FOR HIGH-PERFORMANCE TELECOMMUNICATIONS ICs

Chap. 11 - SILICON-ON-INSULATOR (SOI) TECHNOLOGY 501

Chap. 12 - MULTILEVEL INTERCONNECTS FOR ULSI 573

Chap. 13 - POLYCIDES AND SALICIDES OF $TiSi_2$, $CoSi_2$ AND NiSi 603

Chap. 14 - LOW-*k* DIELECTRICS 639

Chap. 15 - DUAL-DAMASCENE INTERCONNECTS 671

Chap. 16 - COPPER INTERCONNECT PROCESS TECHNOLOGY 711

INDEX 795

Exhibit 1015 - Page 3 of 120

2.7 MINIMIZING PERFORMANCE-DEGRADATION CAUSED BY COPS THROUGH THE USE OF HIGH-PULL-SPEED SILICON 30

2.8 GETTERING FOR ULSI PROCESSES 32

2.8.1 Basic Gettering Principles
2.8.2 Extrinsic Gettering
2.8.3 Intrinsic Gettering
2.8.4 Gettering with Oxygen Precipitates
2.8.5 Summary of Gettering

2.9 STATUS OF SILICON WAFER TECHNOLOGY & FUTURE TRENDS 45

2.9.1 Epitaxy-Optimized Substrate (EOS) Wafers
2.9.2 Comparing the Grown-In Defect Characteristics of the "New" Wafers
2.9.3 Gettering Methods & Denuded Zone Formation Techniques Used with the "New" Wafers
2.9.3.1 "Magic DeNuded Zone"

2.10 FROM INGOT TO FINISHED WAFER: SLICING, ETCHING, AND POLISHING 52

2.10.1 Ingot Evaluation
2.10.2 Ingot Surface Grinding
2.10.3 Grinding Flats or Notches on the Ingot for Orientation Purposes
2.10.4 Sawing the Ingot into Slices (Wafers)
2.10.5 Laser Marking the Wafers
2.10.6 Lapping and Grinding the Wafers
2.10.7 Removal of Surface Mechanical Damage by Chemical Etching
2.20.8 Rounding the Wafer Edge (and Notch)
2.10.9 Edge Polishing of the Wafers
2.10.10 Chemical-Mechanical Polishing of the Wafers
2.10.11 Cleaning the Wafers
2.10.12 Depositing Epitaxial Silicon Layers on the Wafers
2.10.13 Shipping 300-mm Wafers: Its Impact on the Evolution of Wafer Shipping Boxes

2.11 SPECIFICATIONS OF SILICON WAFERS FOR VLSI 63

2.12 THE ECONOMICS OF SILICON WAFERS 66

REFERENCES 69

Chap. 3 - GATE DIELECTRICS: THIN GATE OXIDES 75

3.1 REQUIRED CHARACTERISTICS OF GATE DIELECTRICS FOR DEEP-SUBMICRON MOSFETS 76

3.2 THE STRUCTURE OF THERMALLY GROWN SiO_2 AND THE PROPERTIES OF THE Si/SiO_2 INTERFACE 80

3.2.1 The Microscopic Structure of Thermally Grown SiO_2
3.2.2 The Si/SiO_2 Interface
3.2.2.1 Interface Trap Charge
3.2.2.2 Effect of Interface Traps on IC Characteristics
3.2.2.3 Oxide Trapped Charge
3.2.2.3 Effect of Oxide Trapped Charge on Device Characteristics
3.3 DIELECTRIC BREAKDOWN IN SILICON DIOXIDE FILMS 90
3.3.1 Electron Trapping in Silicon Dioxide:
3.3.2 The Electric-Field Driven Model of Oxide Degradation (*E* Model)
3.3.3 The Current-Driven Model of Oxide Degradation (1/E Model)
3.3.4 The Hole-Trapping Model that Describes How Holes are Injected & Trapped in SiO_2
3.3.5 Comparing the Electric-Field Driven and Current Driven Oxide Breakdown Models
3.3.6 Time to Breakdown (T_{BD}) and Charge to Breakdown (Q_{BD})
3.4 LEAKAGE CURRENTS IN SiO_2 FILMS (TUNNELING PHENOMENA) 99
3.4.1 Fowler-Nordheim (FN) Tunneling (Tunneling Into Silicon Dioxide)
3.4.2 Direct Tunneling (Tunneling Through Silicon Dioxide)
3.5 MODELS OF THIN OXIDE GROWTH 104
3.6 SINGLE-WAFER TECHNOLOGY OF THIN OXIDE GROWTH 109
3.6.1 Rapid Thermal Oxidation Tools
3.6.2 Wet RTO Processes
3.7 NITRIDED & FLUORINATED OXIDES AS GATE DIELECTRICS 112
3.7.1 Oxynitridation of Silicon in N_2O
3.7.2 Oxynitridation of Oxides in N_2O or NO
3.7.3 Fluorinated Gate Oxides
3.8 PROJECTIONS OF THICKNESS LIMITS OF GATE OXIDES 121
3.8.1 Minimum Oxide Thickness Due to Defects and Tunneling
3.8.2 Minimum Oxide Thickness: Stress Induced Leakage Current (SILC)
3.8.3 Soft-Breakdown of Oxide Films
3.8.4 Impact of Polysilicon Depletion
3.8.5 Impact of Process Induced Damage
3.8.6 Summary of Oxide Thickness Projections
3.9 MEASURING THIN GATE OXIDES 135
3.10 MANUFACTURING THIN GATE OXIDES 137
3.10.1 Process Control Issues of Growing Ultra-Thin Gate Oxides
3.10.2 Thin Gate Dielectric Stacks Based on Silicon Dioxide
REFERENCES 139

Chap. 4 - HIGH-*k* DIELECTRICS 145

4.1 HIGH-*k* DIELECTRICS AS THE GATE DIELECTRIC IN MOSFETS 145
4.1.1 Integrating High-*k* Dielectrics into MOSFET Structures:
4.2 HIGH-*k* DIELECTRICS AS THE CAPACITOR DIELECTRIC IN DRAMS 150
4.3 HIGH-*k* DIELECTRICS FOR FeRAM APPLICATONS 154
4.4 TANTALUM PENTOXIDE (Ta_2O_5) 154
4.4.1 Tantalum Pentoxide as a DRAM Capacitor Dielectric
4.4.2 Tantalum Pentoxide as a MOSFET Gate Dielectric
4.5 BARIUM STRONTIUM TITANATE (BST) 162
AS A DRAM CAPACITOR DIELECTRIC
4.5.1 Deposition of BST Films by CVD
4.5.2 Microstructure Effects on the Electrical Properties of BST
4.5.3 BST Integration Issues
4.6 USE OF OTHER HIGH-*k* MATERIALS AS GATE DIELECTRICS 169
4.6.1 Atomic Layer Deposition for Depositing Thin High-*k* Dielectric Films
4.6.2 Aluminum Oxide (Al_2O_3) as a Gate Dielectric
4.6.3 Zirconium Oxide (ZrO_2) as a Gate Dielectric
4.6.4 Hafnium Oxide (HfO_2) as a Gate Dielectric
4.6.5 Praseodymium Oxide (Pr_2O_3) as a High-*k* Gate Dielectric

REFERENCES 177

Chap. 5 - THE STRUCTURE OF DEEP-SUBMICRON MOSFETS: 181
Retrograde-Wells; Super-Steep Retrograde Channel Doping; Drain Engineering; Punchthrough-Control Implants

5.1 WELL-FORMATION IN DEEP-SUBMICRON CMOS 181
5.1.1 Retrograde-Wells for Deep Submicron CMOS
5.1.2 Process Integration Issues Involving the Use of Retrograde-Wells
5.2 SUPERSTEEP RETROGRADE CHANNEL (SSR) PROFILES 191
5.2.1 Fabricating SSR Profiles for Sub-0.1 *m*m MOSFETs
5.3 SOURCE/DRAIN ENGINEERING IN DEEP-SUBMICRON CMOS 195
5.3.1 Parasitic Resistance of Deep-Submicron Source/Drains
5.3.1.1 Physical Meaning of R_{ac}, R_{sp}, and R_{SDE}
5.3.1.2 Comparing the Resistance of the SDE Region ($R_{sp} + R_{SDE}$) to R_{sat}
5.3.1.3 Parasitic Resistance of the Deep Contact Region of the Drain R_{co}
5.3.2 MOSFETs with Elevated Source/Drains

5.3.2 MOSFETs with Elevated Source/Drains
5.3.3 Shallow-Junction Formation for Deep-Submicron Source/Drains
5.3.3.1 Ultra-Low-Energy Implants
5.3.3.2 Transient-Enhanced Diffusion
5.3.3.3 Rapid Thermal Annealing of Shallow Junctions
5.3.3.4 Dopant Loss for Sub-keV Implants from Self-Sputtering, Out Diffusion, and Surface Oxide
5.3.4 Effect of the Overlap of the SDE Region and Gate Edge on MOSFET Performance
5.3.5 SDE Junction Lateral Abruptness
5.4 ANTI-PUNCHTHROUGH STRUCTURES IN DEEP-SUBMICRON CMOS 219
REFERENCES 223

Chap. 6 - ADVANCED LITHOGRAPHY I: DEEP-SUBMICRON RESISTS 227

6.1 CHEMICALLY-AMPLIFIED DEEP-UV RESISTS FOR OPTICAL LITHOGRAPHY 227
6.1.1 248-nm Photoresists
6.1.1.1 Amine-Contamination of DUV CA Resists
6.1.1.2 Current State of 248-nm Resist Technology
6.1.2 193-nm DUV Photoresists
6.1.2.1 Problems of 193-nm Resists
6.1.3 157-nm DUV Photoresists
6.2 ANTI-REFLECTIVE COATING (ARCs) 244
6.2.1 Organic BARCs
6.2.2 Inorganic Dielectric BARCs
6.2.3 PVD-Deposited BARCs
6.2.4 BARCs for Dual-Damascene Applications
6.3 PHOTORESIST PROCESSING SYSTEMS 251
REFERENCES 257

Chap. 7 - ADVANCED LITHOGRAPHY II: OPTICS AND HARDWARE 259

7.1 EXCIMER LASER DEEP-UV LIGHT SOURCES 259
7.1.1 KrF Excimer Lasers
7.1.2 ArF Excimer Lasers
7.1.3 F_2 Excimer Lasers
7.2 EXPOSURE TOOLS For DUV LITHOGRAPHY 265
7.2.1 Exposure Tools for 248-nm Lithography

7.2.2 Exposure Tools for 193-nm Lithography
7.2.3 Exposure Tools for 157-nm Lithography
7.2.4 Calcium Fluoride Optical Elements for DUV Lithography
7.2.5 300-mm Lithography Tools
7.2.6 Mix-an-Match Lithography
7.3 RESOLUTION ENHANCEMENT TECHNOLOGIES (RETs) 274
7.3.1 Off-Axis Illumination
7.3.2 Optical Proximity Correction (OPC)
7.3.3 Phase Shift Masks (PSM)
7.4 MASK ERROR FACTOR (MEF, or MEEF) 292
7.5 EXTENDING THE LIMITS OF OPTICAL LITHOGRAPHY 296
7.6 NON-OPTICAL (or NEXT GENERATION) LITHOGRAPHIC TECHNOLOGIES (NGL) 301
7.6.1 Extreme Ultra-Violet Reflective Projection Lithography (EUV)
7.6.2 Electron Beam Projection Lithography (SCALPEL and PREVAIL)
REFERENCES 309

Chap. 8 - CHEMICAL MECHANICAL POLISHING (CMP) 313

8.1 TERMINOLOGY AND EVOLUTION OF PLANARIZATION PROCESSES FOR ICS 313
8.1.1 Defining the Degree of Planarization
8.1.2 The Need for Dielectric Planarization:
8.1.3 Design Rules Related to Intermetal Dielectric-Formation and Planarization Processes
8.1.4 Step-Height Reduction of Underlying Topography as a Technique to Alleviate the Need for Planarization
8.1.5 Planarization through Sacrificial-Layer Etchback
8.1.5.1 Sacrificial-Etchback Process Problems
8.2 INTRODUCTION TO CHEMICAL MECHANICAL POLISHING (CMP) 322
8.3 THE TERMINOLOGY USED TO CHARACTERIZE CMP PROCESSES 325
8.3.1 CMP Removal Rate (RR)
8.3.2 The Degree of Planarization (DOP)
8.3.3 Within-Die Non-Uniformity (WIDNU)
8.3.4 Within-Wafer Non-Uniformity (WIWNU)
8.3.5 Wafer-to-Wafer Non-Uniformity (WTWNU)
8.3.6 Efficiency of Planarization (EOP)

8.5.1 The Mechanical Aspects of Silicon Dioxide Removal by CMP: (Preston's Law)
8.5.1.1 Model of WIWNU Effects in Rotary CMP Tools Based on Preston's Law
8.5.2 Models Which Describe Factors that Impact Oxide RR in CMP Through the Preston Constant
8.5.2.1 Impact on the Oxide RR from the Electrochemical Interactions Between the Slurry Particles and the Oxide Surface
8.5.2.2 Dependence of Oxide RR in CMP on Dielectric Hardness
8.5.3 Models Which Describe Chemical-Mechanical Phenomena that Give Rise to Planarization Phenomena Associated with CMP of Oxides
8.5.3.1 Model of the Mechanism That Produces Planarization of Steps in Oxide CMP
8.5.3.2 Pattern Dependence of the Planarization Rate in CMP Processes
8.5.4 Metal-CMP Mechanisms
8.6 CMP EQUIPMENT 352
8.6.1 CMP Polishing Tools:
8.6.1.1 CMP Polishing Tool Configurations
8.6.1.2 Rotary CMP Tools:
8.6.1.3 Orbital CMP Tools
8.6.1.4 Linear CMP Tools
8.6.1.5 Fixed Table CMP Tools
8.6.1.6 300-mm CMP Tools
8.6.2 Wafer-Carriers in CMP Polishing Tools (and the CMP Edge Effect)
8.6.2.1 CMP Edge-Effect
8.6.3 CMP Consumables (Slurries):
8.6.3.1 Slurries for Oxide-CMP
8.6.3.2 Slurries for Metal-CMP
8.6.3.3 Multi-Step Slurries
8.6.4 Slurry-Distribution Systems
8.6.5 Environmental, Safety, and Handling Issues of CMP Slurries
8.6.6 CMP Consumables (Polishing-Pads)
8.6.7 Pad-Conditioners
8.6.8 Slurry-Free Pads
8.6.9 Endpoint-Detection in CMP
8.7 CLEANING ISSUES IN CMP 392
8.7.1 Mechanisms of Particle Adhesion and Removal from Surfaces
8.7.1.1 Electrical Repulsion of the Particle and the Wafer Surface
8.7.1.2 Chemically Assisted Particle Removal
8.7.1.3 Mechanically Assisted Particle Removal
8.7.1.4 Steric Stabilization for Suspending Particles in a Solution

8.7.2 Practical Particle Removal Processes: Brush Scrubbing
8.7.2.1 Brush Scrubbing to Remove Particles in Post-Oxide-CMP Processes
8.7.2.2 Brush Scrubbing to Remove Particles in Post Metal-CMP Processes
8.7.3 Vibrational Scrubbing to Remove Particles (Megasonic)
8.7.4 Secondary (or Buff) Polish Procedure as a CMP Cleaning Step
8.7.5 Post-CMP Cleaning of Metallic Contaminants
8.7.6 CMP Cleaning Equipment
8.8 CMP METROLOGY 407
8.8.1 Detection of Defects Associated with the CMP Process
8.9 CMP POLISHER TOOL RELIABILITY 411
8.10 CMP SYSTEMS AND PROCESS INTEGRATION 412
8.11 CMP OF VARIOUS MATERIALS 413
8.11.1 CMP of Oxide Interlevel & Intermetal Dielectrics (ILDs & IMDs)
8.11.2 CMP of Tungsten
8.11.3 CMP of Copper
8.11.4 CMP of Aluminum
8.11.5 CMP of Polysilicon
8.11.6 CMP of Photoresist
8.11.7 CMP of Low-*k* Dielectrics
8.12 MISCELLANEOUS ISSUES CMP 421
8.12.1 Dishing and Erosion
8.12.2 Thickness Non-Uniformity Within a Wafer After CMP
REFERENCES 428

Chap. 9 - SHALLOW TRENCH ISOLATION (STI) 433

9.1 EARLY SHALLOW TRENCH ISOLATION STRUCTURES 435
9.2 STI-ENABLING PROCESSES 438
9.3 SHALLOW TRENCH ISOLATION FOR CMOS 439
9.4 DETAILS OF THE PROCESS FLOW TO FORM A BASELINE SHALLOW-TRENCH-ISOLATION (STI) STRUCTURE 446
9.5 ISSUES INVOLVED WITH STI PROCESS INTEGRATION 455
9.5.1 Issues Involved With Etching Trenches for STI
9.5.2 Corner-Rounding Techniques in STI Structures
9.5.3 Trench-Fill Dielectrics for STI
9.5.4 Dishing Problem Associated with the CMP of STI Trench Dielectrics.
9.5.5 Corner Engineering and Its Effects on MOSFET TURN-OFF Characteristics of MOSFETs Fabricated with STI
REFERENCES 472

Chap. 10 - SILICON-GERMANIUM (Si-Ge) TECHNOLOGY FOR HIGH-PERFORMANCE TELECOMMUNICATIONS ICs 475

10.1 INTRODUCTION TO HETEROJUNCTION BIPOLAR TRANSISTORS 475
10.2 HETEROJUNCTION BIPOLAR TRANSISTORS WITH LARGE-BANDGAP EMITTERS 480
10.3 HETEROJUNCTION BIPOLAR TRANSISTORS WITH SMALLER BANDGAP BASE REGIONS (Si-Ge BASE) 482
10.4 PROCESS TECHNOLOGY FOR FABRICATING SI-GE HBTS 484
10.5 PROCESS INTEGRATION ISSUES FOR BiCMOS TECHNOLOGY WITH Si-Ge HBT 486
10.6 HIGH-QUALITY SI-GE EPITAXIAL FILM GROWTH ON PATTERNED WAFERS 491
10.7 THERMAL STABILITY PROCESS INTEGRATION ISSUES 492
10.8 COMPATIBILITY WITH ION IMPLANTATION PROCESSES 492
10.9 COMPATIBILITY WITH THERMAL OXIDATION PROCESSES 493
10.10 VERTICAL SCALING OF Si:Ge HBTS 493
10.11 BASE-DURING-GATE AND BASE-AFTER-GATE BiCMOS PROCESS FLOWS 495
10.12 EXAMPLES OF THE MODULAR INTEGRATION OF Si-Ge HBTS INTO A BiCMOS PROCESS FLOW 497

REFERENCES 499

Chap. 11 - SILICON-ON-INSULATOR (SOI) TECHNOLOGY 501

11.1 WHAT IS SILICON-ON-INSULATOR (SOI)? 501
11.1.1 Isolation Issues in CMOS in Bulk Silicon
11.1.2 Isolation in SOI CMOS
11.2 THE "END OF THE ROAD FOR BULK CMOS" SCENARIO PREDICTED BY THE 1999 INTERNATIONAL TECHNOLOGY ROADMAP FOR SEMICONDUCTORS 505
11.2.1 Source/Drain Junction Scaling
11.2.2 Dopant Activation in Shallow Source/Drain Junctions
11.2.3 Doping Profile Abruptness in Shallow Source/Drain Junctions
11.3 WHY SOI? 510
11.3.1 The Path of Using SOI May Allow Extension of the ITRS Roadmap
11.4 HISTORY OF SILICON-ON-INSULATOR (SOI) TECHNOLOGY 516
11.5 SILICON-ON-INSULATOR DEVICES 518
11.5.1 Partially-Depleted Thin SOI MOSFETs (PD-SOIs)

11.5.2 Fully-Depleted Thin SOI MOSFETs (FD-SOIs)
11.5.3 Process Technologies Used to Fabricate Fully-Depleted SOI Devices
11.5.3.1 FD-SOI: Thinning Si-SOI Surface Layer by Thermal Oxidation
11.5.3.2 FD-SOI Fabrication by LOCOS Recess of the MOSFET Channel Region
11.6 FABRICATION TECHNOLOGIES USED TO PRODUCE SOI-STARTING-WAFERS 527
11.6.1 Silicon-on-Sapphire (SOS)
11.6.2 SOI by Separation by Implanted Oxygen (SIMOX)
11.6.2.1 High-Dose SIMOX
11.6.2.2 Low-Dose SIMOX
11.6.3 Wafer Bonding
11.6.3.1 Bond and Etch-back SOI (BESOI)
11.6.3.2 Hydrogen-Implantation-Induced Layer Splitting to Create SOI Substrates (SMART-CUT®)
11.6.3.3 NANOCLEAVE® Wafer Bonding SOI
11.6.3.4 ELTRAN® Wafer Bonded SOI
11.7 LATERAL ISOLATION IN SOI 550
11.8 SOI PRODUCT EVOLUTION 556
11.9 PROBLEMS ASSOCIATED WITH SOI 559
11.9.1 Floating Body Effects in Partially-Depleted SOI MOSFETs
11.9.1.1 Kink Effect in Partially Depleted SOI NMOSFETs
11.9.1.2 Transient I_D Effects
11.9.1.3 Effect of Building Partially Depleted SOI MOSFETs Without Body Contacts
11.9.2 Short-Channel Effects in SOI MOSFETs
11.9.3 Self-Heating Effects in SOI MOSFETs
11.9.4 Reliability Issues of SOI
REFERENCES 568

Chap. 12 - MULTILEVEL INTERCONNECTS FOR ULSI 573

12.1 EVOLUTION OF THE STRUCTURE AND MATERIALS USED IN IC INTERCONNECTS 575
12.1.1 Single-Level Metal Interconnects in Early ICs
12.1.2 Double-Level Metal Interconnects in Early Bipolar ICs
12.1.3 Double-Level Metal Interconnects in CMOS ICs (Partial Planarization of the ILD)
12.1.4 Three-Level Metal Interconnects in CMOS ICs (CMP of the ILD & W-Plugs)
12.1.4.1 Problems of W-plug/Al-line Interconnect Structures
12.1.5 Dual Damascene Cu/Low-*k* Interconnects

12.1.6 Other Geometrical Aspects of Multilevel ULSI Interconnects
12.2 PERFORMANCE REQUIREMENTS OF ULSI INTERCONNECT SYSTEMS 586
12.2.1 Propagation Delays in ULSI Multilevel Interconnects
12.2.1.1 Case When the *k*-Value of the ILD is Constant Throughout the Interconnect System
12.2.1.2 Case When the *k*-Values of the Interlevel and Intralevel ILD are Different
12.2.1.3 Minimizing the *RC* Delay in the Local Interconnect Region
12.2.1.4 Minimizing the *RC* Delay in the Global Interconnect Regions
12.2.2 Power Dissipation Issues in ULSI Interconnects
12.2.3 Crosstalk in ULSI Interconnects
12.2.4 Contact Resistance in ULSI Interconnects
12.2.5 Electrostatic Discharge Protection Issues of Interconnects
12.3 SUMMARY OF USLI INTERCONNECT ISSUES AND FUTURE TRENDS 599
REFERENCES 601

Chap. 13 - POLYCIDES AND SALICIDES OF $TiSi_2$, $CoSi_2$ AND NiSi 603

13.1 POLYCIDES AND SALICIDES 603
13.2 POLYCIDES FABRICATED BY CVD OF TUNGSTEN SILICIDE (WSi_x) 605
13.3 SALICIDE (SELF-ALIGNED SILICIDE) STRUCTURES 608
13.4 THE EFFECTS OF SALICIDES ON MOSFET PERFORMANCE 612
13.5 SALICIDES FORMED WITH TITANIUM SILICIDE ($TiSi_2$) 614
13.5.1 Reaction Kinetics of $TiSi_2$ Formation (and the Problem of *Bridging*)
13.5.2 The $TiSi_2$ Narrow-Line Effect
13.5.3 Proposed Process Modifications to Mitigate the $TiSi_2$ Narrow-Line Effect
13.5.4 Junction Dopant Redistribution and Silicon Consumption During $TiSi_2$ Formation
13.6 SALICIDES FORMED WITH COBALT SILICIDE ($CoSi_2$) 624
13.6.1 Advantages and Drawbacks of Cobalt Salicides
13.6.2 Details of the Cobalt Salicide Fabrication Process
13.6.3 Diode Leakage Issues Involving Cobalt Salicides
13.6.4 Sensitivity of Cobalt Salicide Films to Degradation by Oxygen Ambients: TiN Capping of Cobalt Salicide Films
13.7 EPITAXIAL $CoSi_2$ 630
13.8 $CoSi_2$ AS A SOURCE/DRAIN DIFFUSION SOURCE 631
13.9 NICKEL SILICIDE FILMS FOR SALICIDE APPLICATIONS 634
REFERENCES 635

Chap. 14 - LOW-*k* DIELECTRICS **639**

14.1 INTRODUCTION TO LOW-*k* DIELECTRICS 635
14.1.1 Spin-On versus CVD Methods for Forming Low-*k* Dielectrics
14.1.2 Silicon-Based versus Carbon-Based Low-*k* Dielectrics
14.1.3 Status of Low-*k* Dielectrics in 2000
14.2 DESIRED CHARACTERISTICS OF LOW-*k* DIELECTRIC FILMS 646
14.3 GENERAL PROCESS INTEGRATION ISSUES INVOLVING LOW-*k* FILMS 648
14.4 FIRST GENERATION LOW-*k* DIELECTRICS ($2.8 < k < 3.5$) 650
14.4.1 Hydrogen Silsesquioxanes (HSQ)
14.4.2 Spin-On Methyl Silsesquioxanes (MSQ)
14.4.3 Fluorinated Silicate Glass (FSG) Films
14.5 SECOND-GENERATION LOW-*k* DIELECTRICS ($2.5 < k < 2.8$) 655
14.5.1 2nd-Genereation Spin-On-Dielectrics: ($2.5 < k < 2.8$)
14.5.2 2nd-Generation CVD-Dielectrics: ($2.5 < k < 2.8$)
14.5.2.1 2nd Generation Low-*k* CVD-Films: Organo-Silicate Glasses
14.5.2.2 2nd Generation Low-*k* CVD-Films: *FlowFill* CVD Films
14.5.2.3 2nd Generation Low-*k* CVD-Films: Black Diamond & Coral CVD Films
14.5.2.4 2nd Generation Low-*k* CVD-Films: Parylene CVD Films
14.6 ULTRA-LOW-*k* DIELECTRICS: ($k < 2.3$) 663
REFERENCES 669

Chap. 15 - DUAL-DAMASCENE INTERCONNECTS **671**

15.1 DAMASCENE VERSUS SUBTRACTIVE INTERCONNECT STRUCTURES 671
15.2 DAMASCENE INTERCONNECTS AS ALTERNATIVE DEEP-SUBMICRON INTERCONNECTS 672
15.3 SINGLE-DAMASCENE INTERCONNECT STRUCTURES
15.4 INTRODUCTION TO DUAL-DAMASCENE INTERCONNECT PROCESSES 674
15.5 THE THREE DUAL-DAMASCENE PROCESS SEQUENCES 676
15.6 THE EVOLUTION OF DUAL-DAMASCENE FROM SiO_2 TO LOW-*k* DIELECTRIC STACKS 679
15.7 TRENCH-FIRST DUAL-DAMASCENE PROCESS FLOW 683
15.8 VIA-FIRST DUAL-DAMASCENE PROCESS FLOW WITH AN SiO_2 DIELECTRIC STACK 685

15.9 PROCESS INTEGRATION ISSUES OF THE VIA-FIRST DUAL-DAMASCENE PROCESS FLOW WITH A LOW-*k*-DIELECTRIC STACK 687
15.9.1 Via-First-Damascene Integration Issues: Lithography
15.9.1.1 Via-Patterning Lithography Issues
15.9.1.2 Trench-Patterning Lithography Issues – DUV Resist Poisoning
15.9.1.3 Trench-Patterning Lithography Issues – Trench CD Control Issues
15.9.1.4 Trench-Patterning Lithography Issues – BARC Veils
15.9.1.5 Bi-Layer Resist Process
15.9.2 Via-First-Damascene Integration Issues: Dielectric Etching
15.9.2.1 Via-Etching for Low-*k* Processes
15.9.2.1 Trench-Etching
15.9.2.1 Etching the SiN Barrier at the Bottom of Vias
15.9.3 Via-First-Damascene Integration Issues: Resist Stripping
15.9.4 Via-First-Damascene Integration Issues: Cleaning
15.9.4.1 Post-Dielectric-Etch-and-Strip Residue-Removal
15.9.4.2 Pre-Metal-Deposition Clean
15.9.4.3 Post-CMP Clean
15.9.4.4 Clean of Top-Surface of Cu Prior to Diffusion-Barrier-SiN-Deposition Process
15.9.4.5 Supercritical Fluids as Wafer Cleaners
15.10 VIA-FIRST DUAL-DAMASCENE DIELECTRIC STACK WITH NO EMBEDDED ETCH-STOP LAYER 708
15.11 PROCESS INTEGRATION ISSUES OF DUAL-DAMASCENE DIELECTRIC STACK WITH ULTRALOW-*k* DIELECTRICS 708
REFERENCES 709

Chap. 16 – COPPER INTERCONNECT PROCESS TECHNOLGY 711
16.1 WHY COPPER FOR DEEP-SUBMICRON IC INTERCONNECTS? 712
16.1.1 The Advantages of Copper Interconnects
16.1.2 Technological Challenges of Using Copper Interconnects
16.2 OVERVIEW OF COPPER PROCESS TECHNOLOGY 723
16.2.1 The Transition from Aluminum to Copper Interconnects
16.2.2 The Key Goal of Depositing Copper into Damascene Recesses
16.3 BARRIER LAYERS FOR COPPER INTERCONNECTS 728
16.3.1 The Impact of the Barrier-Layer Deposition Process on the Filling of Damascene Recesses with Copper
16.3.2 Introduction to the Technology of Depositing Barrier-Layers for Cu Interconnects

16.3.3 Ionized Magnetron Sputter Deposition
16.3.4 Hollow-Cathode Magnetron Sputtering Source
16.3.5 Self-Ionized-Plasma Sputtering Source
16.3.6 Material Properties of Tantalum and Refining Tantalum Metal For Semiconductor Applications
16.3.6.1 Material Properties of Tantalum
16.3.6.2 Refining and Forming Tantalum Metal For Sputtering Targets
16.3.6.3 The Properties of Tantalum (and TaN & TaSiN) as Cu-Barrier Layers
16.3.7 Other Candidate Materials for Cu-Barrier Layers
16.4 SEED-LAYER TECHNOLOGY FOR COPPER INTERCONNECTS 740
16.5 COPPER INTERCONNECT FILM DEPOSITION 747
16.5.1 Electrolytic-Plating of Copper
16.5.2 Filling High-Aspect-Ratio Recesses Without Voids Using a Cu Electrolytic Electroplating Process
16.5.2.1 The Electroplating Process Sequence
16.5.2.2 Additives in Plating Baths for Damascene Cu Interconnects
16.5.2.3 Models of How Additives in Plating Baths Impact "Bottom-Up" Filling
16.5.2.4 Electroplating Bath Control
16.5.2.5 Pulsed Plating
16.5.2.6 Copper Thickness Variation Within a Wafer After Electroplating
16.5.3 Post-Electroplating Cu Annealing and Recrystallization
16.5.4 Electrochemical Mechanical Deposition (ECMD™)
16.5.5 Electroless-Plating of Copper
16.5.6 CVD of Copper
16.6 WET-CHEMICAL ETCHING AND ELECTROPOLISHING COPPER 775
16.7 CONTAMINATION CONTROL IN COPPER TECHNOLOGY 778
16.7.1 Sources of Copper Cross-Contamination in a Wafer Fab
16.7.2 Effects of Copper-Cross Contamination at the Back-End-of-Line (BEOL)
16.7.3 Cu Removal from the Wafer Bevel/Edge and Backside
16.8 RELIABILITY OF COPPER INTERCONNECTS 784
16.8.1 Reliability of Copper Interconnects: "Missing-Metal" Defects
16.8.2 Reliability of Copper Interconnects: Plated vs. CVD Cu Films
16.9 ENVIRONMENTAL, HEALTH, & SAFETY ISSUES OF CU TECHNOLOGY 787
16.9.1 Copper Waste Handling
16.9.1.1 Handling Concentrated Waste from Plating and Wafer Reclaim
16.9.1.2 Handling Copper CMP Effluents
REFERENCES 789

INDEX **795**

PREFACE

With the dawn of the new millennium, IC technology-nodes have evolved beyond 0.18-micron. New processes also had to be developed to meet the needs of this *Deep-Submicron Era,* including: copper/low-*k* dual-damascene interconnects; high-*k* MOSFET-gate and DRAM dielectrics; 300-mm wafers; chemical-mechanical polishing (CMP); excimer-laser-based DUV-lithography; chemically-amplified (CA) resists; silicon:germanium (Si:Ge); and silicon-on-insulator (SOI).

The semiconductor industry, however, also experienced an economic decline of unprecedented severity at the start of the 2000 decade. It seams clear that to fuel the recovery, chip manufacturers will have to exploit the above innovations. Such processes are a path to achieving the performance goals of future IC technology-nodes that will be needed for manufacturing new and profitable electronic products.

This is the *first available book dedicated exclusively to describing such deep-submicron process technologies* (including those listed above). It is intended to provide semiconductor engineers and researchers with a comprehensive, state-of-the-art reference about these emerging and leading-edge processes. While many other books have been written about the "traditional" IC fabrication processes, none offer what this volume contains. It's sole focus is those techniques needed to *manufacture chips for 0.18-μm (and smaller) generations*. Thus, it belongs on the shelf of every microelectronic professional whose livelihood depends on having a thorough grasp of the latest advances in IC fabrication techniques. There is no faster way to gain this information. As with our other volumes, owners will find themselves referring to this text again and again!

Readers can still consult Vol.1, 2nd Edition for background material on basic IC fabrication modules, as this book (Vol. 4) does not repeat that information. Instead, it presents over 800 pages of new content about the process technologies of the 2000's. It also has over 900 references (with 200 of these being published since 2000), and more than 500 illustrations. It is written in the same lucid, user-friendly style that has been widely praised in the other volumes of *Silicon Processing for the VLSI Era.*

As with the previous volumes, this text was produced using an accelerated book-production schedule. By employing revolutionary publishing techniques made available in the 1990's, Lattice Press can turn a finished manuscript into a published text in less than 3 months (a task that takes most publishers of technical books 15-18 months). This allows our books to be released faster, and to contain the most-up-to-date information at the time of publication.

Exhibit 1015 - Page 17 of 120

SILICON PROCESSING FOR THE VLSI ERA is a multi-volume treatise designed to provide a comprehensive and up-to-date treatment of this important and rapidly changing field. The volume in hand is the fourth in the series. Volume 1 (*Process Technology*, 2nd Edition) was published in 2000 and deals with individual processes employed in the fabrication of silicon ULSI circuits (such as silicon wafering, thermal oxidation of silicon, CVD and PVD of amorphous and polycrystalline films, diffusion and ion implantation of dopants in silicon, microlithography, and patterning technology). Volume 2 (*Process Integration*) was released in 1990, and describes how the individual processes of Volume 1 are combined in various ways to produce MOS and bipolar VLSI and ULSI circuits. The task of integrating these various fabrication processes together is referred to as *process integration.* In Volume 3 (*The Submicron MOSFET,* published in 1995) the topics of submicron MOSFET device physics and the relationship between such device physics and submicron MOSFET fabrication are addressed.

A book of this length and diversity would not have been possible without the indirect and direct assistance of many others. From an indirect perspective, virtually all of the information presented here is based on the research efforts of countless numbers of scientists and engineers. Their contributions are recognized to a small degree in citations of their articles given in the references at the end of each chapter. The direct help was also provided by many people, and the resulting text is thus significantly better. The author expresses heartfelt thanks to those who gave their time, energy, and intellect.

Important direct help came from members of my immediate family. My wife, Carrol, read the manuscript for grammatical and stylistic correctness. My son, Ross, designed the cover and improved many of the graphics. My daughter, Jennifer, helped in marketing the book. Thus, it truly represents a family enterprise.

Stanley Wolf

For additional copies of the books contact:

LATTICE PRESS
Post Office Box 340
Sunset Beach, CA 90742 USA

www.latticepress.com

For your convenience an order form is provided on the final leaf of the book.
It can also be ordered from the **LATTICE PRESS** website.

Chapter 11

SILICON-ON-INSULATOR (SOI) TECHNOLOGY

In this chapter the characteristics of silicon-on-insulator (SOI) technologies fabricated with deep-submicron MOSFETs are described. The focus is on the SOI-devices themselves as well as on the process technologies used to fabricate both the starting SOI-wafers and the SOI ICs. An argument is also presented as to why SOI technology may provide a path for continuing to advance CMOS ICs for several more generations. That is, implementation of SOI may make it possible to exceed the limits facing bulk-silicon-CMOS in the near term.

11.1 WHAT IS SILICON-ON-INSULATOR (SOI)?

Silicon-on-insulator (*SOI*) circuits consist of ICs made from single-device islands dielectrically isolated from the substrate and each other. Each such island contains a single transistor, with the isolated transistor being either a MOSFET or a bipolar transistor. However, since about 90% of the ICs manufactured in 2002 are based on CMOS-technology, the discussion in this chapter will be restricted to SOI technology involving MOSFETs. Figure 11-1 shows a cross-section of a MOSFET fabricated on an SOI wafer that also uses STI as the isolation structure to provide lateral isolation. Note that in the case of a MOSFET fabricated on an

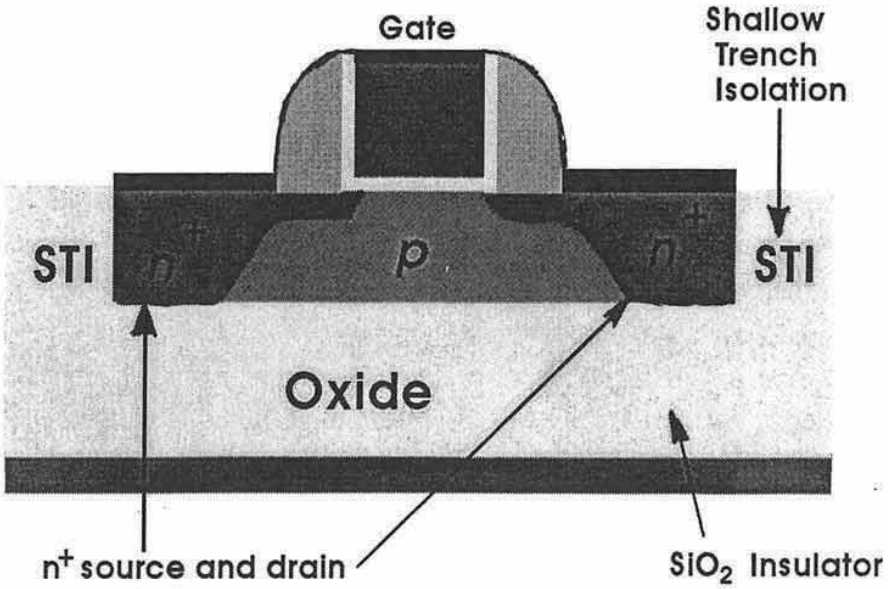


Fig. 11-1 Cross-section of a MOSFET fabricated on an SOI wafer.

Exhibit 1015 - Page 19 of 120

SOI wafer, a *buried oxide* (*BOX*) underneath the device provides vertical isolation, and a trench oxide (in this example), provides the lateral isolation. Since the trench oxide extends all the way through the silicon device layer atop the BOX (and a thermal oxide exists on the wafer top surface), dielectric material continuously surrounds each MOSFET on all sides. Thus, complete electrical isolation is provided for the MOSFET with respect to all other devices, and the substrate, on the wafer.

11.1.1 Isolation Issues in CMOS Fabricated in Bulk-Silicon Wafers

The electrical isolation requirements of CMOS ICs built in SOI are quite different from those encountered when CMOS ICs are fabricated in bulk-silicon. To see why, the circuit strategies and structures needed for isolation purposes in CMOS circuits on bulk-silicon must be examined. For reference purposes a single-MOSFET built in a bulk-silicon-substrate is shown in Fig. 11-2a (with various device structures and regions defined). Figure 11-2b shows two NMOSFETs built on a *p*-doped substrate, together with the parasitic-transistor that exists in the field-region separating the active devices. Semirecessed-LOCOS is used to produce the field-oxide (which is part of the device structure employed to maintain electrical isolation among such MOSFETs). Figure 11-2c shows an *n*-well CMOS structure fabricated in a *p*-type bulk-silicon wafer. This figure also shows the parasitic MOSFETs that are associated with that structure (as well as the *pn*-junction that separates the *n*-wells from the *p*-substrate). Again, a semirecessed-LOCOS field-oxide plays a role in maintaining electrical isolation among the MOSFETs of the CMOS structure. (It should be noted that shallow-trench isolation [STI] is used in place of LOCOS to form the field-oxide in CMOS technologies below 0.35-μm. Nevertheless, the isolation issues described here for CMOS on bulk-silicon remain the same, regardless how the field-oxide is formed). Finally, in Fig. 11-2d a cross-section of a CMOS IC structure is depicted, together with a schematic depiction of the parasitic-bipolar-transistors and associated lateral currents.

Using the IC device structures described above, the isolation issues involved in CMOS ICs can be undertaken. The first item to be noted is that the *pn*-junction separating a well from the substrate must be kept under reverse-bias to prevent current from flowing from the power-supply-line to ground through forward-bias of this junction. Second, inter-device isolation for devices within a well (*intra-well isolation)* or within the substrate regions, entails the use of a field-oxide to suppress spurious surface-leakage effects between active devices within a well. Third, a field-oxide and channel-stop implants are also used to prevent leakage

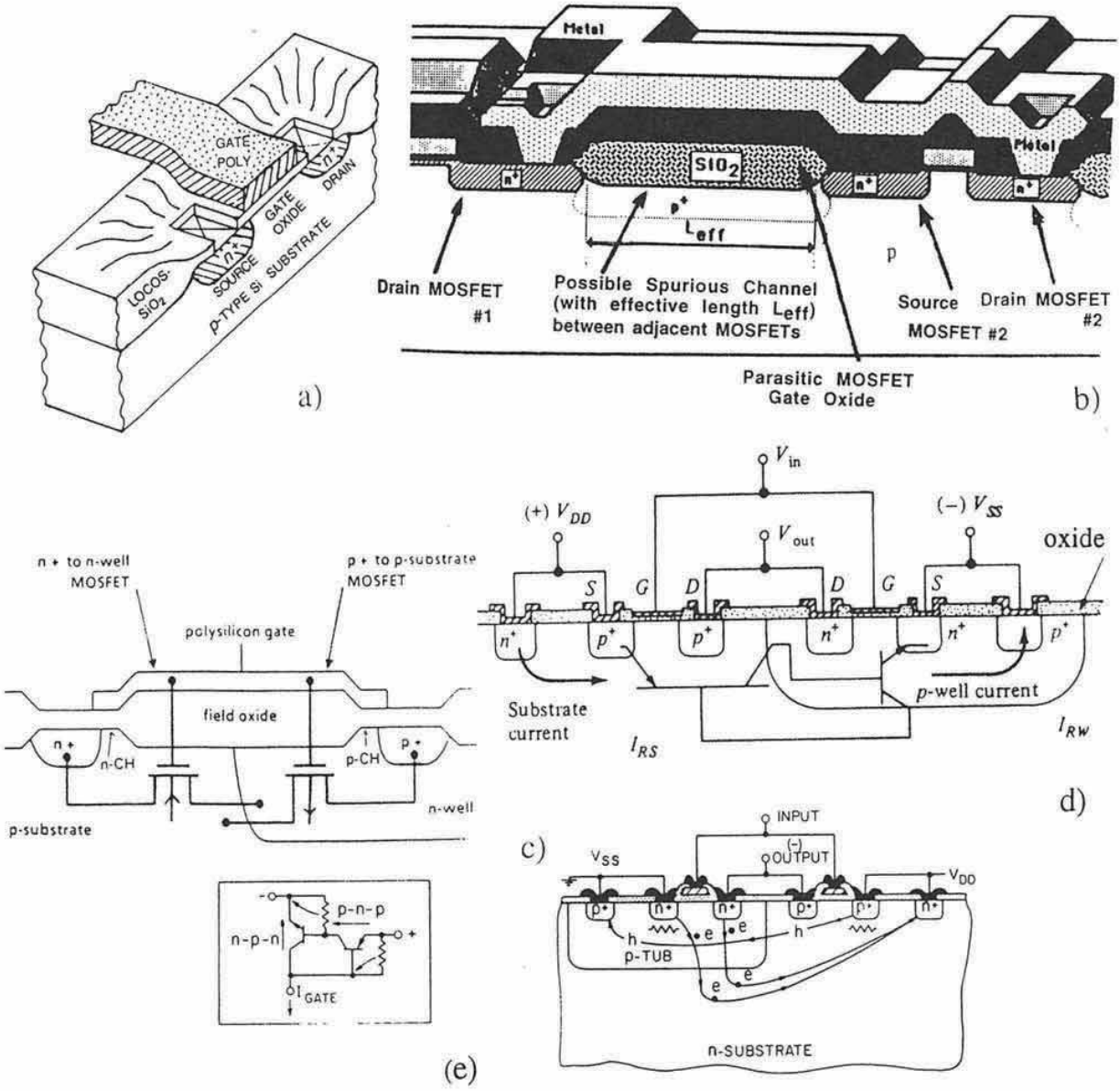


Fig. 11-2 a) A single MOSFET built in a bulk-silicon substrate, with various device structures and regions defined; b) Two MOSFETs (e.g., both NMOSFETs) built in a substrate having one type of doping (i.e., on a *p*-doped substrate), together with the parasitic transistor that exists in the field-region separating these MOSFETs. Semirecessed-LOCOS is used to produce the field-oxide (which is part of the device structure employed to maintain electrical isolation among such MOSFETs); c) An *n*-well CMOS structure fabricated in a *p*-type bulk-silicon wafer. Also shown in the figure are the parasitic MOSFETs that are associated with that structure, as well as the *pn* junction the separates the *n*-wells from the *p*-substrate; d) A cross-section of a CMOS-IC-structure is depicted, together with a schematic depiction of the parasitic-bipolar-transistors and associated lateral currents; e) The parasitic bipolar transistors that exist in CMOS structures can be *turned on*, which can cause the CMOS IC to go into a latchup state.

between devices among different wells (*inter-well isolation*). Yet, despite employing these isolation structures, intra-well- and inter-well-leakage still occurs to some degree in CMOS ICs made in bulk-silicon (and the problem gets worse as the devices are brought closer together). Note that Volume 3 can be consulted for a more detailed discussion of isolation issues in CMOS built in bulk-silicon wafers. In addition, the parasitic-bipolar-transistors that exist in CMOS structures can be *turned-on*, and this can cause the CMOS IC to go into a *latchup state* (Fig. 11-2e). That is, latchup can occur because two parasitic-bipolar-transistors are always simultaneously formed in the course of fabricating a CMOS structure. In bulk-CMOS, these two transistors are connected in such a way that if the current gain of the loop formed by these two devices is larger than unity, latchup can occur. Latchup can be triggered by different mechanisms, such as node- voltage overshoots, displacement-current, junction-avalanching, and pho-tocurrents. It is true that latchup is a well-understood phenomenon which can be dealt with in bulk-CMOS using proper processing and design methodologies. However, with the constant reduction in feature spacing, latchup immunity in bulk CMOS ICs remains a continuing challenge.

11.1.2 Isolation in SOI CMOS

The isolation issues of CMOS fabricated in SOI wafers are much less complex than those involving CMOS ICs in bulk-silicon. This is because each MOSFET in SOI is built in an island of silicon on an insulating substrate. Hence, all of the previously-mentioned possible isolation causes are eliminated. Let us explain this assertion in more detail.

First, the requirement of keeping the *pn*-junction that separates the CMOS well regions from the substrate under reverse-bias is eliminated. That is, this requirement no longer exists because a continuous oxide barrier exists in SOI between the well-regions and the substrate. Consequently, current cannot flow between the power and ground lines, even if the junction in question becomes forward-biased. Second, the possibility of leakage current between devices within a well (intra-well-leakage) and from well-to-well (inter-well-leakage) is eliminated. This is because the parasitic-field-region-MOSFET does not exist (as a result of the field oxide extending through the silicon film and joining the buried oxide). In other words, the channel between the source and drain of any parasitic-field-region-MOSFET is electrically interrupted by the oxide moat surrounding each MOSFET within an SOI IC. Consequently, leakage current due to: **1**) field-inversion of the channel of a parasitic-field-region-MOSFET; or

Table 11-1 ITRS Roadmap of the Required MOSFET Junction Characteristics

Year	2000	2001	2002	2003	2004	2005	2008	2010	2014
Gate size (nm)	120	100	85	80	70	65	45	32	22
S/D junction depths (nm)									
Contact	65–125	55–105	45–90	43–5	38–75	35–70	30–55	20–40	15–35
Extension/channel	36–60	30–50	25–43	24–40	20–35	20–33	16–26	11–19	8–13
SSRD channel	18–30	15–25	12–22	12–20	10–16	10–17	8–13	5–10	4–8
Extension doping									
Sheet resistance (Ω/sq.)	310–760	280–730	250–700	240–675	220–650	200–625	150–525	120–450	100–400
Resistivity (10^{-3}Ω-cm)	1.1–4.6	0.84–3.6	0.62–3.0	0.57–2.7	0.44–2.3	0.40–2.1	0.24–1.4	0.13–0.86	0.08–0.52
Lateral extension junction abruptness (nm/decade)									
For V_t-roll-off control	12	10	8.5	8	7	6.5	4.5	3.2	2.2
For spreading resistance control	4.8	4.1	3.4	2.9	2.7	2.4	1.25	0.8	0.5

2) parasitic-MOSFET subsurface-punchthrough, do not exist in SOI-technologies. Finally, the existence of vertical- (as well as lateral-) isolation within SOI ICs eliminates the possibility of latchup. That is, the current paths from the collectors of the parasitic-bipolar-transistors to the bases of the complimentary bipolar-transistors (which set up feedback paths in CMOS in bulk-silicon that cause the circuits to enter latchup), are interrupted by the vertical oxide of the SOI. Hence, latchup can no longer occur in SOI ICs.

Despite the fact that intra-well and inter-well leakage paths are interrupted in SOI CMOS structures, new *intra-device* parasitic transistor paths inherent in SOI CMOS structures can still produce leakage current, as is discussed in Sect. 11.7.

11.2 THE "END OF THE ROAD FOR BULK CMOS" SCENARIO PREDICTED BY THE 2001 INTERNATIONAL TECHNOLOGY ROADMAP FOR SEMICONDUCTORS

The *2001 International Technology Roadmap for Semiconductors* (ITRS-2001) indicates that planar CMOS transistors on bulk-silicon are expected to reach their limits at gate lengths of about 50-nm in the 2005-2006 time frame. That is, the ITRS-2001 analysis documents a clear "end-of-roadmap" event-horizon at the onset of IC device scaling to 50-nm gate-lengths.[1] The premise of the ITRS study was to predict, as far as possible, the characteristics of an extension of the industry's "standard technology," namely, planar-CMOS-transistors fabricated

on bulk-silicon wafers. The study concludes that many process and materials constraints combine to force this change in the technology path. That is, the conclusions drawn in this study are that sometime between 2005 and 2008 nearly all of the 30 or so transistor characteristics required for the extension of Moore's Law are listed as "no known solutions." This is a clear signal that IC transistor technology requires fundamental changes in the next half-decade. In this section the principal issues limiting *junction-formation* in sub-100-nm CMOS ICs on bulk-silicon will be detailed. As part of the discussion to be presented in the next section (Section 11.3), it will be described how some of these constraints can be removed or relaxed through the use of SOI. That is, by implementing CMOS-on-SOI a path may be provided to allow CMOS ICs to be fabricated beyond the 50-nm gate-length-limit faced by CMOS-on-bulk-silicon. An extension of as much as a decade beyond the "end-of-the roa-map" scenario posited by the ITRS-2001 may be possible by fabricating CMOS ICs on SOI-wafers.

The three principal challenges for forming shallow-junctions in CMOS on bulk-silicon wafers are as follows:

1. Limiting dopant-diffusion in the source/drain regions to shallow depths;
2. Activating a sufficient percentage of the dopants implanted into the source/drain regions to provide source/drain regions that exhibit sufficiently low contact-resistance and sheet-resistivities:
3. Obtaining source/drain junctions that have highly-abrupt vertical and lateral doping-profiles.

The ITRS-2001 values for these key transistor parameters, as a function of the device generation (which correlates to the gate-length of that generation and the expected time of its introduction) are given in Table 11-1.

11.2.1 Source/Drain Junction Scaling

As MOSFET gate lengths (L_L) are scaled, the other device structures (i.e., the gate-oxide-thickness and the junction-depth of the source/drain extension) are also scaled proportionally. To maintain such "proportional scaling" for CMOS when the gate-length decreases below 100-nm, the design rule used in the ITRS-2001 study was that the junction depth of the source/drain extension $X_{j(extension)}$ would scale according to: $X_{j(extension)} \sim 0.55 \times L_L$ within a range of ±25% (see Fig. 11-2a for a graphic that depicts the region referred to as the source/drain extension). Controlling junction depths to less than 55-nm for sub-100-nm MOSFETs will push the limits of sub-keV boron implantation and "spike" anneals.[2]

11.2.2 Dopant Activation in Shallow Source/Drain Junctions

In the ITRS-2001 model, it is expected that ways will be found to produce highly doped source/drain extension-regions having a sufficiently low resistivity (to reduce the sheet-resistance of these regions) at the same time the depth of the junctions of these regions is reduced. If this can be done (i.e., if the resistivity of the source/drain extension regions could somehow be reduced from the value of 10^{-3} Ω-cm [that is produced in the source/drain extension-regions of 0.15-μm NMOS transistors by using low-energy boron implantation {0.2-2.0-keV} and "spike" anneals] to say 10^{-4} Ω-cm) a sheet-resistance versus junction-depth curve (as plotted in Fig. 11-3 for the value of $\rho = 10^{-4}$ Ω-cm) could be obtained.[3] Overlaid on these sheet-resistance versus junction-depth curves (with ρ as the parameter), are the industry expectations of the value of ρ for: **1**) high-performance ICs; and **2**) cost-focus (i.e., lower-cost) ICs, as taken from the ITRS-2001. That is, the values of ρ that the ITRS-2001 roadmap expects in order to be able to fabricate CMOS ICs on bulk-silicon for smaller gate-length-devices, decreases from about 10^{-3} Ω-cm for NMOSFETs with gate lengths of 0.14-μm, down to

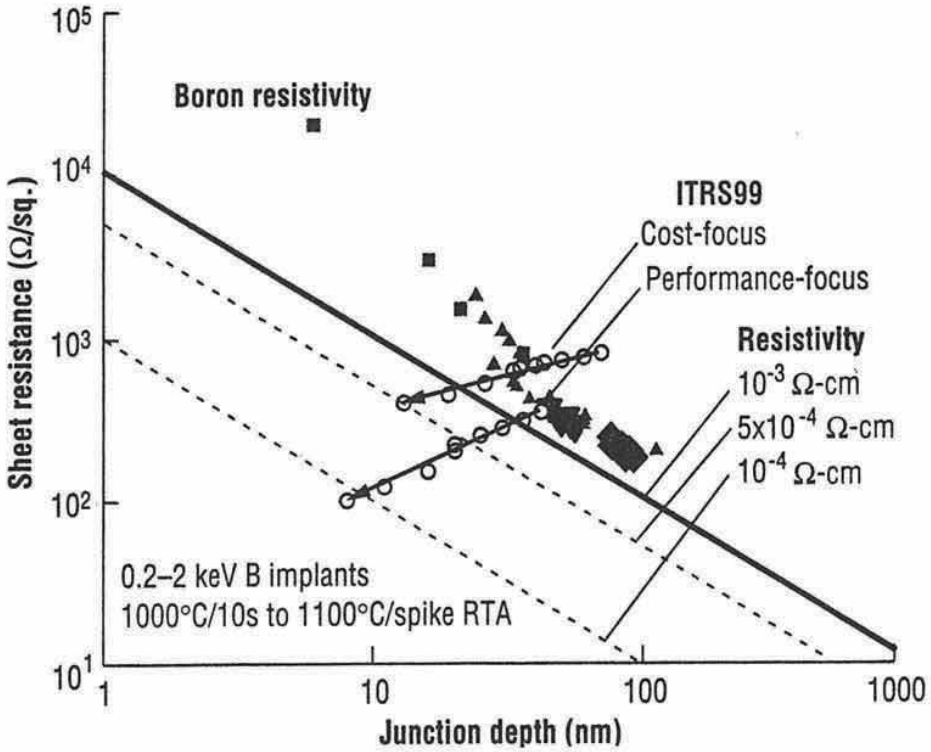


Fig. 11-3 Sheet-resistance and junction-depths for low-energy boron implants and spike-implant anneals, and industry expectations for S/D extension junctions (starting with 140-nm gates to 22-nm gates at the head of the arrows) for cost-focus and performance focus devices. The dopant resistivity for sub-keV boron implant and spike-RTP process (in CZ-Si) is limited by a "wall" at about 10^{-3} Ω-cm.

about 10^{-4} Ω-cm for NMOSFETs with 22-nm gate lengths (corresponding to a junction-depth of 8-nm and a sheet resistance of 100 Ω/sq). Some alternative doping technology may be required to achieve such low resistivity. Possible alternative doping techniques include: **a**) low-temperature (or selective) epitaxy; **b**) diffusion from a doped-glass source; **c**) diffusion from a low-temperature Si-Ge alloy; **d**) direct gas-source-diffusion in an RTP-system; **e**) gas-immersion-laser-diffusion (GILD); and **f**) planar-doping-layers using atomic-layer-epitaxy (ALE). Of these many possible processes, low-temperature-epitaxy may hold some key advantages, as it can also be done in a selective manner (which may possibly minimize masking steps). However, with selective epitaxy, separate hard-masks are required for *n*- and *p*-S/D-regions. On the other hand, GILD with projection masking (PGILD) does not require resists or hard-masking and may become an important doping technology.

At this time, however, it is not clear whether any of the above methods of reducing the value of ρ in the source/drain extension regions of NMOSFETs will be able to reduce ρ down to 10^{-4} Ω-cm. In fact, the resistivities of junctions formed by the present day "best methods" of sub-keV boron implants and spike anneals appears to be limited by the "activation-wall" to a value no smaller than 10^{-3} Ω-cm.[4]

11.2.3 Doping-Profile Abruptness in Shallow Source/Drain Junctions

To permit sub-100-nm MOSFETs to exhibit the performance characteristics expected in the ITRS-2001 roadmap, the total source-to-drain resistance of such MOSFETs must be kept sufficiently small.[5] Controlling this total source-to-drain resistance-value in deep-submicron MOSFETs requires that special attention be paid to the contact-resistance and spreading-resistance at the junction of the source extension and the channel. As shown in Fig. 11-4, parasitic-resistances exist in the path between the metal-to-Si substrate interface and the region in the device where the actual transistor action begins (see also Sect. 5.4, Chap. 5).

In the planar MOSFET, the device-region in which transistor-action occurs is the channel-region under the gate. The current flowing from the metal-interconnect lines to the channel-region must be transported from the metal-Si contact to the edge of the channel through the source or drain regions (Fig. 11-4). In the MOSFET, current enters the contact perpendicular to the wafer surface, and then travels parallel to the surface in order to reach the channel. The total parasitic-series-resistance, R_S, of the current path from the contact to the edge of the channel can be modeled by the sum of four components,

$$R_S = R_{co} + R_{sh} + R_{sp} + R_{ac} \qquad (11.1)$$

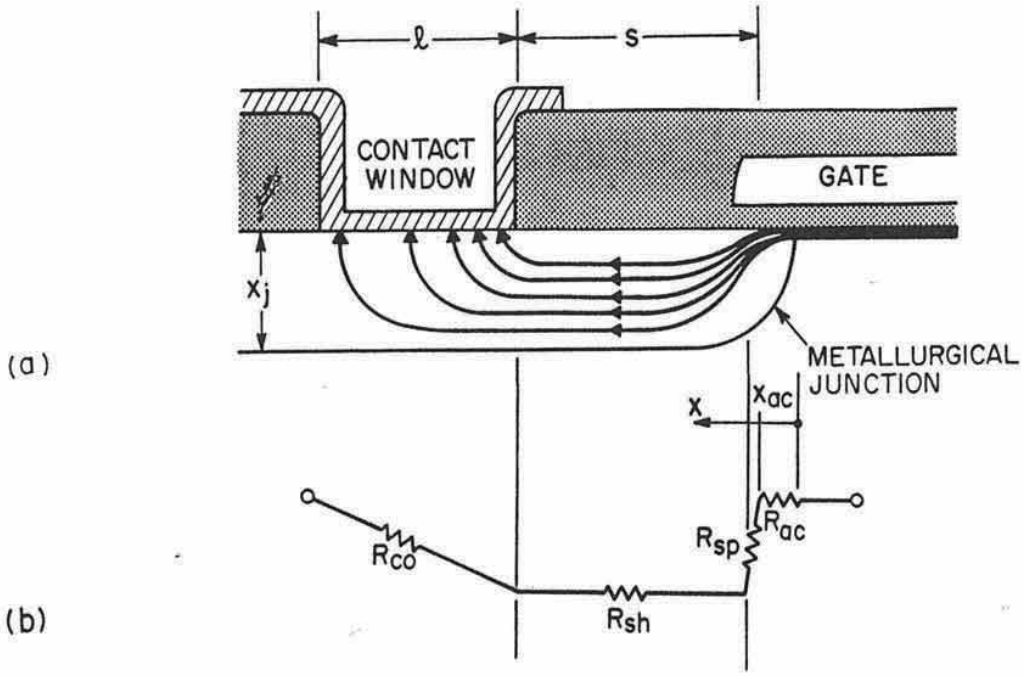


Fig. 11-4 Schematic drawing showing: a) current paths within the source and drain regions, of a planar MOSFET and; b) their corresponding representative series resistance components.[5] (© 1979 IEEE)

where R_{co} is the *contact-resistance* between the metal and the source/drain region, R_{sh} is the *sheet-resistance* of the bulk-region of the source and drain, R_{sp} describes the *resistance of the current lines crowding near the channel end of the source*, and R_{ac} is the *accumulation-layer resistance*. Qualitatively, for a non-abrupt junction, R_{sp} and R_{ac} arise because, after leaving the channel, the current does not immediately spread into the bulk of the source-drain regions (the parasitic-resistance-effect in these bulk regions is R_{sh}). Thus, R_{sp} and R_{ac} exist because the accumulation-layer conductance is much higher than the semiconductor conductance in the vicinity of the junction (see again Sect 5.3).

The key issue for sub-100-nm MOSFETs is the abruptness of the doping-profile at the lateral edge of the source extension and its effect on R_{sp} and R_{ac}. The ITRS-2001 study calls for lateral-doping-profiles that decrease into the channel region at a slope of ~3 nm/decade in concentration, for 100-nm gates. Obtaining a lateral-doping-profile abruptness of less than 3-nm/decade presents severe challenges for bulk silicon doping by implantation and rapid thermal annealing. As-implanted profiles for 0.2-keV boron into pre-amorphized silicon have a vertical profile steepness of 2.5-nm/decade.[6] The key S/D-extension profile in the lateral direction, still remains to be characterized, and will be further broadened by thermal annealing in actual transistors.

11.3 WHY SOI?

Silicon-on-insulator (SOI) technology has long been used in many specialized applications, such as radiation-hardened or high-voltage integrated circuits. It is only in recent years (since about the mid-1990s), however, that SOI has emerged as serious contender for low-power, high-performance applications.[7] The power consumption, P, in CMOS ICs is given approximately by:

$$P \sim C V^2 f + V I_{leakage} \quad \textbf{(11.2)}$$

where C is the total capacitance, V is the supply voltage, f is the frequency, and $I_{leakage}$ is the standby-leakage-current. Thus, in low-power applications the supply voltage must be reduced to decrease both terms of this equation. This power-supply reduction has had to follow the evolution in battery-voltage and capacity. The SEMATECH roadmap target has been to reach 0.9V by the end of the century. However, at such a low voltage, the IC performance is also reduced due to a smaller value of I_D in MOSFETs. As will be seen, however, the use of SOI allows the drain current to be significantly increased, even at such low voltages. SOI therefore also becomes appropriate for emerging system-on-chip ICs used for smart power, as well as for micro-electromechanical systems (MEMS), mixed-signal, and integrated-optics applications.

SOI also provides a path to permit the scaling of ICs to go forward. That is, to continue to increase performance and keep up with Moore's Law, several cumulative approaches have been adopted. The evolution of IC technology in the near future which uses this strategy is summarized in Fig. 11-5. This diagram indicates that integrating copper wiring and SOI will represent the next step toward the continued improvement in IC performance. Here we list the advant-

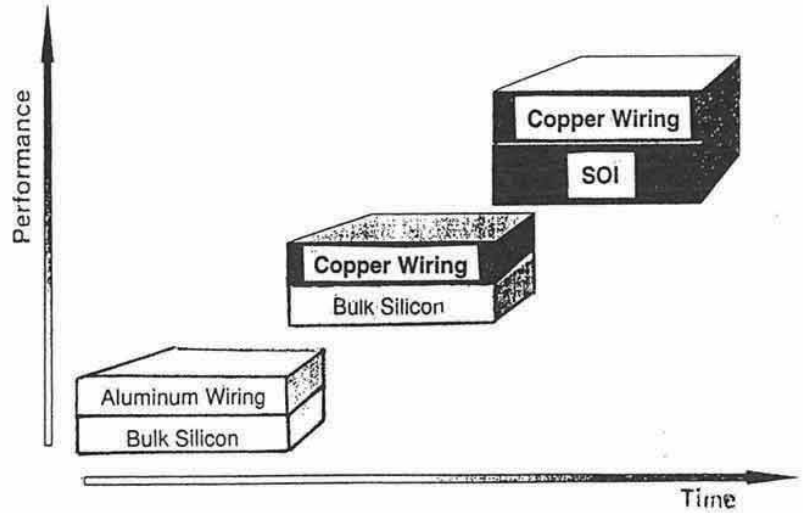


Fig. 11-5 Technology migration for future IC progress.

ages offered by CMOS-onSOI which contribute to such performance improvement.

1. **Improved Speed-Performance.** The speed performance of SOI ICs is enhanced by 20-35% compared to the same circuits fabricated in CMOS on bulk-silicon if they are operated at the same voltages. (The reasons for this increase in speed performance is due to a number of factors that are listed later as other advantages of SOI.) Thus, SOI circuits built using the design rules of "generation-n" of some CMOS technology in bulk silicon will operate as well as the ICs of "generation-n + 1" of CMOS-on-bulk-silicon. This also implies that if a roadblock occurs in some upcoming generation of CMOS on bulk silicon, by having the capability to produce these ICs in SOI, it will be possible to fabricate ICs that allow that roadblock to be exceeded for at least one generation (if not more).

2. **Reduced Power-Consumption.** If CMOS-ICs in bulk-silicon and CMOS ICs in SOI are operated to give the same speed-performance, there is a 35-70% reduction in power-consumption in the CMOS-on-SOI ICs. This advantage provides CMOS-on-SOI a huge competitive edge for low-voltage/low-power applications operated with a one-battery supply. Figure 11-6 depicts this performance-versus-power tradeoff issue.

3. **Reduced Junction-Capacitance.** One reason for the increased performance of CMOS-on-SOI is that the junction-capacitance of the MOSFETs is significantly reduced. For example, a depletion-capacitance exists at the drain-substrate junction, and its value is a function of the

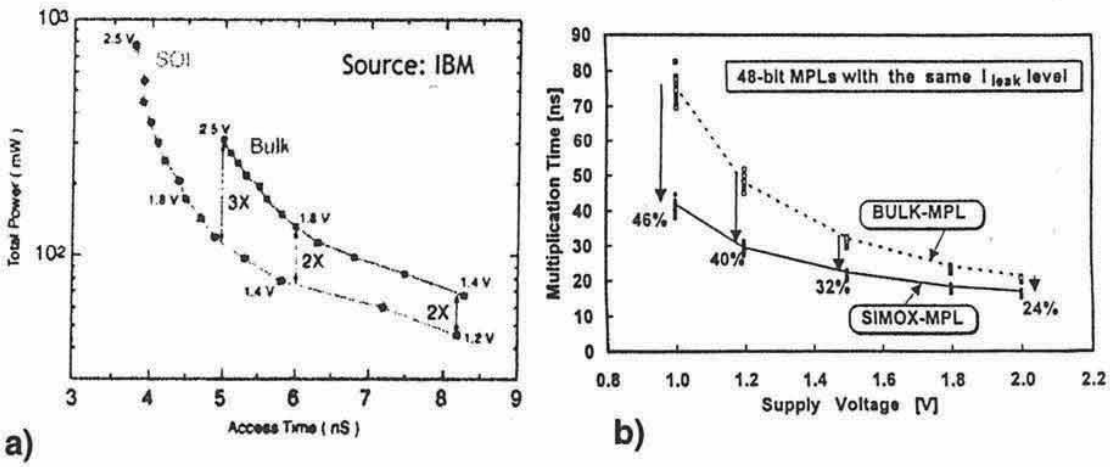


Fig. 11-6 a) Power drain and access time for bulk and SOI CMOS for drive voltages ranging from 1.2 V; b) Comparison of 48-bit-multiplier performance between bulk and SOI using a 0.25-μm CMOS technology.

Exhibit 1015 - Page 29 of 120

substrate doping-concentration. The higher the substrate doping-concentration, the higher the capacitance. Modern submicron circuits increase the substrate doping-concentrations with each new technology generation, which contributes to an increased junction-capacitance. However, in SOI circuits, the maximum capacitance between the junctions and the substrate is the capacitance of the buried-insulator. This capacitance is proportional to the dielectric constant of the capacitance material. Silicon dioxide (which is used as the buried insulator), has a dielectric constant $k_{SiO_2} = 3.9$ which is three times smaller than that of silicon $k_{si} = 11.7$ Therefore, a source/drain junction located on a buried-oxide gives a parasitic-capacitance approximately three times smaller than that of a bulk-junction. Since the SOI buried-insulator-thickness does not scale with device scaling, parasitic-capacitances do not necessarily increase as SOI technology progresses.

As a result of the above phenomena, and depending on the circuit layout, the total capacitance of a CMOS SOI-IC can be reduced by 15-50% compared with the same IC fabricated in bulk-silicon. In fully-depleted SOI (FD-SOI) MOSFETs (described in Sect. 11.5.1), the source/drain-regions butt-up against buried oxide (BOX), and thus no *pn*-junction exists at that device-region boundary (and hence there is no depletion-region that needs to charged and discharged there). Even in partially-depleted SOI devices (PD-SOI), the depletion-region of the junction of the source/drain-regions extends to the BOX, and does not increase in volume beyond that point once this depth is reached. Thus, even in PD-SOIs the maximum junction-capacitance of the source-to-body, C_{jSB}, and drain-to-body, C_{jDB}, is limited to a smaller value than that existing a comparably-sized MOSFET in bulk-silicon. The wiring-capacitance (i.e., associated with interconnect lines running over the field regions of an IC) is also slightly reduced by the presence of the BOX in SOI wafers.

4. Latchup Elimination. As noted in Sect. 11.1.2, the phenomenon of latchup is eliminated in CMOS-on-SOI ICs and the techniques needed to suppress latchup in bulk CMOS are unnecessary. This also means that the use of epitaxial-wafers with a heavily-doped-substrate to control latchup in some applications is no longer needed.

5. Reduction of Interwell-Isolation-Spacing. Due to the elimination of the possibility of latchup, and the ability to retain adequately low intra-well and inter-well leakage-currents due to the dielectric isolation, it is possible to reduce the isolation-spacing design rules, compared to those needed in

CMOS ICs in bulk-silicon (especially the n^+-p^+ spacing). In addition, the contact-overlap-spacing can be smaller. Thus, the packing density of the CMOS SOI-IC is increased, allowing circuits with greater functional-density to be fabricated on the same area of silicon. There is also no need for high-energy implantation of well-dopants.

6. **Reduced Soft-Error-Sensitivity.** SOI reduces the substrate-collection-volume for alpha-particle- and cosmic-ray-generated electron-hole tracks, hence lowering soft-error-sensitivity. In particular, SOI- technology for ECL-bipolar/CMOS ICs is preferred because of the high sensitivity of the ECL peripheral-circuitry to radiation.[8]

7. **Improved MOSFET Turn-Off Characteristic (Steeper Subthreshold-Slope).** The subthreshold slope, S_t, of MOSFETs built on SOI-wafers is much steeper than that of MOSFETs in bulk-Si-wafers. Subthreshold-slopes as small as 62-mV/decade have been observed in SOI MOSFETs.[9] The small value of S_t implies that an SOI-MOSFET can be turned-off with a smaller gate-voltage swing. This allows faster switching in MOSFETs, along with the use of lower threshold-voltages. A lower threshold-voltage allows lower power-supply-voltages to be used, and an increase in the drive-currents of SOI-MOSFETs compared to MOSFETs built in bulk-silicon.

8. **Reduced Leakage-Current Arising from Carrier-Generation in Junction Depletion Regions.** The ability to make the source/drain junction-areas of SOI-MOSFETs smaller by approximately two decades, also decreases the volume of the depletion-regions of these junctions. Thus, less leakage - current arising from the generation of carriers in these regions is produced, which reduces the standby leakage-current of SOI-ICs to a comparable degree. The negligible leakage current makes SOI ideal for low-power ICs.

9. **Compatibility with Bulk-Si CMOS Process Tools & Processes.** The SOI process-technology is compatible with the conventional process-technology used to fabricate CMOS ICs in bulk-silicon. Thus, the tools and facilities used to fabricate CMOS-on-bulk-silicon can also be used to fabricate CMOS-on-SOI. The SOI process-sequence is even somewhat simpler than that used to fabricate CMOS on bulk-silicon. That is, there is no need for the well-formation process in SOI-ICs, nor any need for channel-stop-implants to suppress lateral-surface-leakage between devices due to field-inversion.

10. **Advantages for System-On-Chip (SOC).** A high-resistivity substrate under the BOX can be used, which can significantly reduce the device

capacitance and the loss of passive components, such as inductors and transmission-lines, commonly used in analog- and rf-devices.[10] Therefore, SOI offers advantages over bulk-Si to realize a system-on-a-chip by integrating high-performance digital and analog/rf circuits on the same chip.

11. Increased Drain-Current. The drain current of SOI-MOSFETs, I_D, is also increased due to the significantly lower doping that can be used in the channel. The ability to tolerate lower-channel-doping occurs because of the natural punchthrough suppression arising from the backgating effects of the insulator interface.

12. Improved Circuit Reliability Due to Elimination of Junction-Spiking. When contact is made to a MOSFET in bulk-CMOS, unwanted reactions between the metal or the silicide can take place, leading to the problem of "junction-spiking" (and the attendant rise in uncontrolled leakage currents). While barrier materials normally control this problem, the potential for its occurrence is still there in bulk-CMOS. If MOSFETs are fabricated in thin SOI material, the source and drain junctions extend to the buried insulator. In that case, no leakage will be produced if some uncontrolled metal-silicon reaction occurs.

11.3.1 The Path of Using SOI May Allow Extension of the ITRS Roadmap

Many of the challenges for CMOS transistors in bulk-silicon discussed in Sect. 11.2 are relaxed or removed when SOI-wafers are used instead of bulk-silicon wafers, and SOI technologies offer new options for transistor design. With the 20-50% increase in transistor speed that can be achieved by switching to SOI, the performance of ICs can be increased without the cost of a shift to a higher lithography resolution. SOI "thin-body" CMOS transistors provide excellent control for the off-current (leakage) characteristics that limit the endurance of battery-operated computer and communication devices. Furthermore, soft-error-immunity and radiation-hardness of SOI devices is seen as a key requirement for expansion of high-data-rate network-servers that form the core of Internet and other global data-transfer-links.

In addition, SOI-wafers are opening a path for the extension of planar CMOS transistor design by perhaps a decade beyond the "end-of-roadmap" limits foreseen in the ITRS-2001 study for 50-nm gates on bulk-silicon. Operational SOI transistors have been studied with gate lengths as small as 25-nm and Si channel thicknesses of 1-nm.

TABLE 11-2 COMPARISON OF KEY ISSUES FOR SUB-100-nm CMOS FOR BULK-SILICON AND FULLY-DEPLETED SOI MOSFETS

Critical Issue	Bulk silicon	SOI
Junction depth	RTP/(LTP)	Set by Si-SOI film
Dopant activation	Thermal budget on vertical and lateral diffusion	Lateral diffusion limit
Junction abruptness	Critical vertical and lateral abruptness limits	Carrier injection still requires abrupt lateral junction in thin-body
Contacts	Extended S/D	Extended S/D
Off-current	Requires halo	Dual gate
Threshold control	SSRD (super-steep retrograde doped channels) or SiGe gate	SiGe gate
Well doping	High-energy implant	No well

To begin, it is easier to control the junction-depths of the source/drain extensions in SOI devices. In bulk-silicon, obtaining shallow-junctions requires careful control of thermal budgets, with the use of rapid-thermal-annealing-"spikes" or laser-annealing to limit both the vertical- and lateral-dopant-diffusion, while obtaining high degrees of dopant-activation and adequate defect annealing. With fully-depleted, thin SOI (FD-SOI) MOSFETs (discussed in Sect. 11.5.2), the source/drain-extension junction-depth is determined by the Si-SOI-layer thickness. This means that the thermal-budget-constraints on vertical-junction-depth are essentially removed for fully-depleted-SOI CMOS devices. Second, the ability to activate dopants in the source/drain-regions of CMOS in bulk-silicon is limited by the thermal-budget on vertical- and lateral-diffusion effects, while in FD-SOI MOSFETs only the lateral-diffusion sets a limit on this thermal-budget (a potentially less difficult problem to solve). Third, in CMOS-on bulk-silicon both the vertical- and lateral-abruptness of the source/drain-extension doping-profile needs to be controlled, whereas in FD-SOI, only the abruptness of the lateral-doping-profile of the source/drain extension-region needs to be controlled. Fourth, the well-doping in CMOS-on-bulk-silicon needs to be carried out with high-energy ion-implanters, but in CMOS-on-SOI, no wells even need to be formed. Finally, SOI-technologies can be developed to include new device structures, such as dual-gate and high-channel-mobility (with

Si:Ge-CMOS), that provide new tools for transistor designers to increase signal speed while decreasing off-current leakage-levels. The transistor characteristics and process tradeoffs for bulk-silicon- and SOI-wafers are summarized in Table 11-2. The process challenges shift to controlling lateral-diffusion and obtaining abrupt-junction-interfaces.

One of the critical processes that will be needed to allow thin-SOI-wafers to serve as the substrates for sub-50-nm CMOS is the ability to fabricate thin-SOI wafers with high-levels of control over the SOI-Si device-layer-thickness, electrical-properties, and interface-characteristics. The Smart-Cut® and Nanocleave® processes described in Sects. 11.6.3.2 and 11.6.3.3, respectively, are two approaches that may allow these goals to be met.

The timeline for the use of the various types of "thin" SOI MOSFETs can be linked to the projections made by the ITRS-2001 (Fig. 11-7). The thickness of the Si-to-SOI surface layer (T_{Si}) will shrink according to the approximate design rules of $T_{Si} \sim 2 \times X_{j(S/D\text{contact region})}$ for partially depleted SOI MOSFETs (see Fig. 11-2a, which shows what is meant by the terms *"S/D contact-region"* and *"S/D extension region"*), and $T_{Si} \sim 1 \times X_{j(S/D\text{extension})}$. Up to the year 2000 timeframe, production SOI-ICs were being manufactured using PD-SOI.

For the time between 2001 to 2005 (i.e., for MOSFETs with gate lengths of 0.15-μm down to 0.07-μm), the Si-SOI layer thicknesses being used is expected to cause the type of SOI FETs to be a mix of partially-depleted (PD) CMOS for mainstream logic and ASIC devices, and fully-depleted (FD) SOI for high-performance processor and server/router devices.

For CMOS devices with gate-lengths in the range of 50-nm (beginning in about 2006), the multitude of process and performance issues described in the ITRS study will preclude the use of planar-CMOS on bulk-Si as well as in partially-depleted SOI. By migrating mainstream transistor structures to fully-depleted SOI, it may also be possible to extend CMOS circuit technologies a full decade beyond the "end-of-roadmap" for planar-CMOS on bulk-silicon. "Thin-body" SOI, with gates sizes at 20-nm on Si-SOI layers of ~10-nm are now at the threshold of research devices. If scaling continues at the historical rate of a 0.7 shrink every three years, these devices would be in production by about 2015.

11.4 HISTORY OF SILICON-ON-INSULATOR (SOI) TECHNOLOGY

Silicon-on-insulator technology has been available since the late 1960's, and has been used since that time for military integrated-circuit applications. However, it has taken more than 30 years for the first mainstream commercial SOI devices

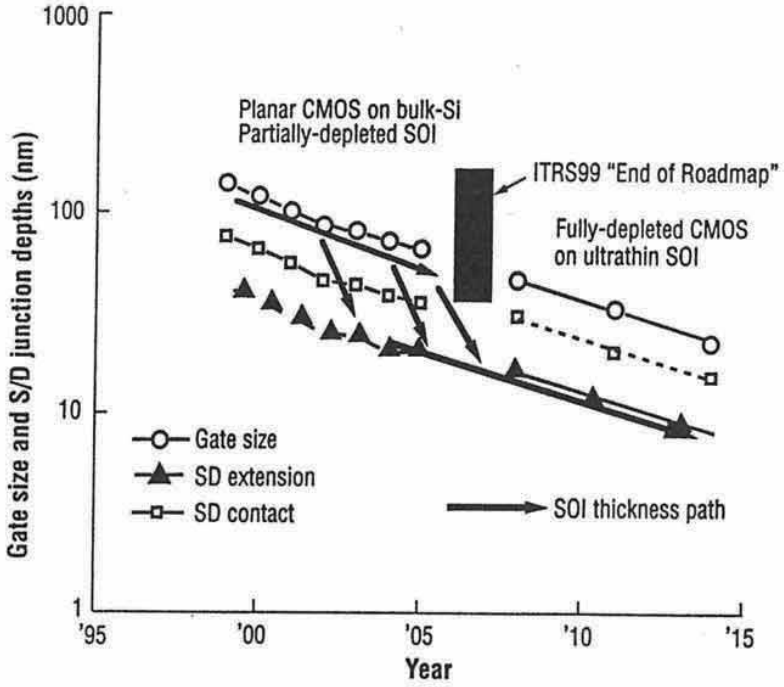


Fig. 11-7 Timeline of Si-SOI layer thickness linked to junction depth and gate size predictions from ITRS-2001.

to be offered. As described in Sect. 11.3, 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 have become more important, a keen interest in SOI technology has been re-kindled among mainstream device suppliers.[41]

Currently there are a number of useful approaches for making SOI wafers. The first group includes the "legacy" SOI technologies of *dielectric isolation* (see Vol. 2, Chap. 2, for more information on the dielectric-isolation-technology) and *silicon-on-sapphire* (*SOS*) that were developed in the 1960's, and that even today, still find limited application. (As we will note, an improved SOS technology is currently being used to fabricate high-performance rf-devices.) The more modern SOI-technologies are based on either a buried-insulator-layer

created by ion-implantation technology (e.g., *Separation by IMplanted OXygen* [*SIMOX*]), or by a *wafer bonding* (*WB*) technique. (The insulating substrate for SOS wafers is sapphire [i.e., very pure single-crystal, 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.[42] When devices were first built on SOI-wafers, small islands of silicon were formed (typically by etch techniques) on top of the insulator. In modern SOI ICs, either LOCOS or STI is used to complete the dielectric isolation. Individual devices are then fashioned in the islands and these devices are then interconnected in a conventional manner. SOI-technology has been the subject of intense research and development in the late 1990s. A number of symposia that address SOI technology are held on a regular basis.[44,45] For more information on the evolution of SOI circuits, see Sect. 11.8.

11.5 SILICON-ON-INSULATOR DEVICES

The MOSFETs built using SOI-substrates are categorized in a number of ways. The first such grouping differentiates between so-called *thick-SOI-MOSFETs* (built on SOI-substrates whose Si layer is thicker than 1 micron - i.e., > 1.0-μm), and *thin-SOI-MOSFETs* (in which this layer is less than 1 micron thick – i.e., < 1.0-μm). The applications for "thick" SOI include power-IC switches, rf bipolar-signal-processors, and a wide variety of MEMS and optoelectronic devices. Thick-SOI devices were also historically used for a variety of military applications (e.g., radiation-hardened ICs) and for such commercial-applications as high-voltage and high-power devices. The principal SOI-material being presently used for VLSI and ULSI CMOS signal, logic, and memory applications is "thin" SOI (50-1000-nm thick [0.005-1-μm]. The "thick" SOI market segment, at only $50 million/year is a very small piece of the $110 billion/year commercial, mainstream CMOS IC market (and is not likely to grow significantly in the future). Hence, we will not further deal with such devices, but instead will focus on "thin" SOI devices in the rest of this section.

Within the category of "thin" SOI devices, there are a number of additional subcategories, namely *partially-depleted SOI-MOSFETs* and *fully-depleted SOI-MOSFETs*. The thickness of the silicon-device-layer of the SOI-structure can be chosen so that the SOI-MOSFETs will operate in one or the other of these modes. Each of these SOI device types will be discussed separately in more detail. Note once again that one of the ways that MOSFETs fabricated in SOI differ from the MOSFETs fabricated in a bulk (i.e., non-SOI) substrate is that the

channel region (or body) of the SOI-MOSFET is not automatically connected to any external terminal (i.e., intrinsically the body of the SOI MOSFET is electrically floating). This aspect of SOI-MOSFETs (a floating body), leads to so-called *floating-body-effects,* which can be detrimental to device performance. Floating-body effects in SOI-MOSFETs will be described in more detail later.

11.5.1 Partially-Depleted Thin-SOI-MOSFETs (PD-SOI)

Partially-depleted, thin-SOI-MOSFETs (PD-SOIs) are given this name because when the maximum gate-bias is applied to the MOSFET, the depth of the resultant depletion-region in the channel of such devices is shallower than the thickness of the silicon layer (Fig. 11-8). Note that "thick" SOI MOSFETs are also operated as partially-depleted MOSFETs.

As mentioned earlier, the body of SOI-MOSFETs is intrinsically floating. That is, unless a contact to some circuit potential is explicitly provided, the potential of the SOI-body is floating (Fig. 11-9a). However, if the SOI-MOSFET body is grounded (i.e., by independent body contacts or by body-source ties), partially depleted SOI MOSFETs behave very much like bulk-Si MOSFETs, and most of the standard $I_D(V_{GS},V_{DS})$ equations and design concepts apply. If body-contacts are not supplied, so-called *floating-body effects* arise, leading to potentially detrimental consequences (including the introduction of a *kink* in the dc I-V characteristics, a reduction in breakdown-voltage, and the causing of hysteresis and instability during dynamic operations). The issue of floating-body

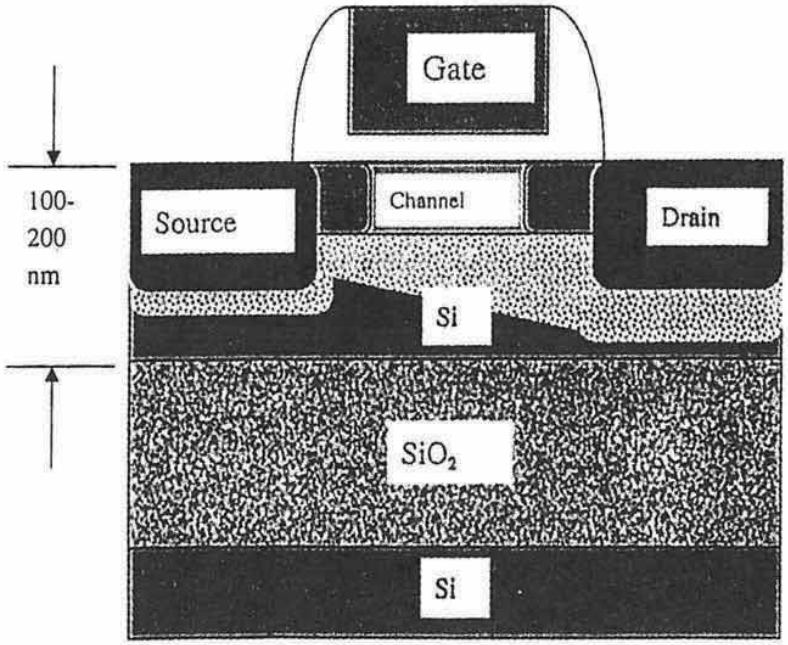


Fig. 11-8 Partially-depleted transistor in a Si layer over an isolation layer.

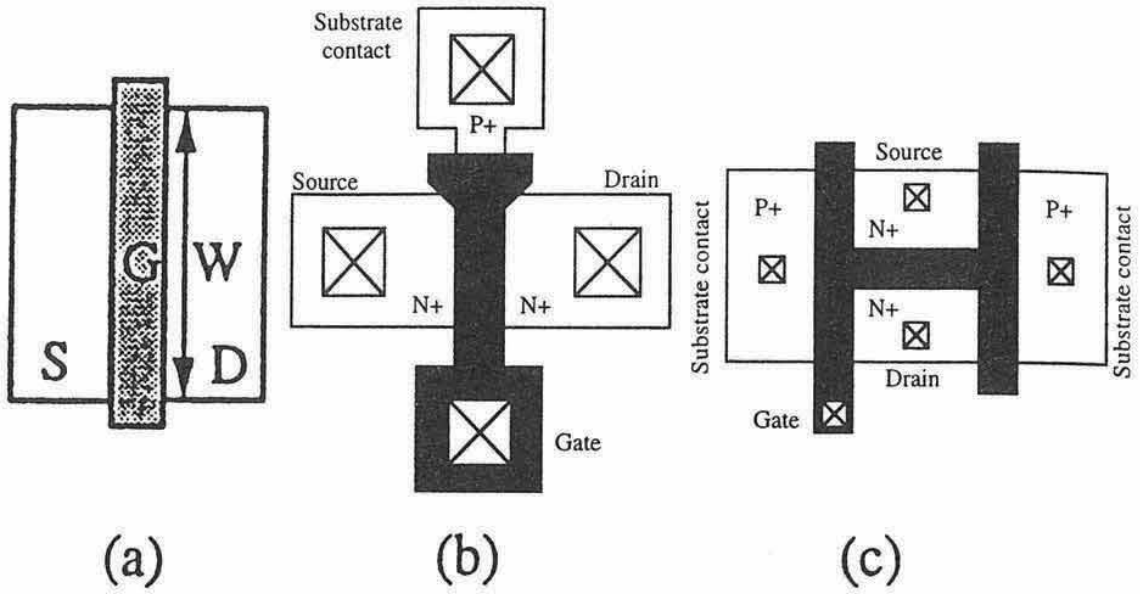


Fig. 11-9 a) Top-down view of an SOI-MOSFET in which no provision is made for contacting the body. Unless a contact to some circuit potential is explicitly provided, the potential of the SOI-body is floating; b) MOSFET with body-contact made on one side of the MOSFET channel; c) An "H-gate" SOI-MOSFET (i.e., a body-contact is made on each side of the MOSFET channel).

effects in partially depleted SOI MOSFETs is described in further in Sect. 11.9.

Note that body- and substrate-contacts were routinely used when "thick" SOI devices were fabricated, as shown in Fig. 11-9b. To avoid the possible device-performance-degradation that floating-body-effects could cause, early partially-depleted, thin-SOI-MOSFETs also used either separate body-contacts or source-body ties, again as shown in Fig. 11-9b. However, since the body-region of the MOSFET beneath the channel can exhibit a high resistance, a single body-contact tied to ground may not be adequate. In some applications a body contact is made on both sides of the MOSFET channel, forming a so-called "H-gate" (Fig. 11-9c).[11] Figure 11-10 shows how a body-contact is made to an SOI-NMOSFET both from a top view, and from a cross-sectional view. Figure 11-11 shows the layout of a DRAM sense-amplifier-circuit fabricated in SOI, in which a body-to-source-tie is employed.

While body-contacts (and body-source-ties) can be used to avoid the problems associated with the floating-body-effect, these structures consume valuable silicon-circuit-area. That is, the global use of body-contacts on every SOI MOSFET in a circuit may severely degrade the density (and hence also the

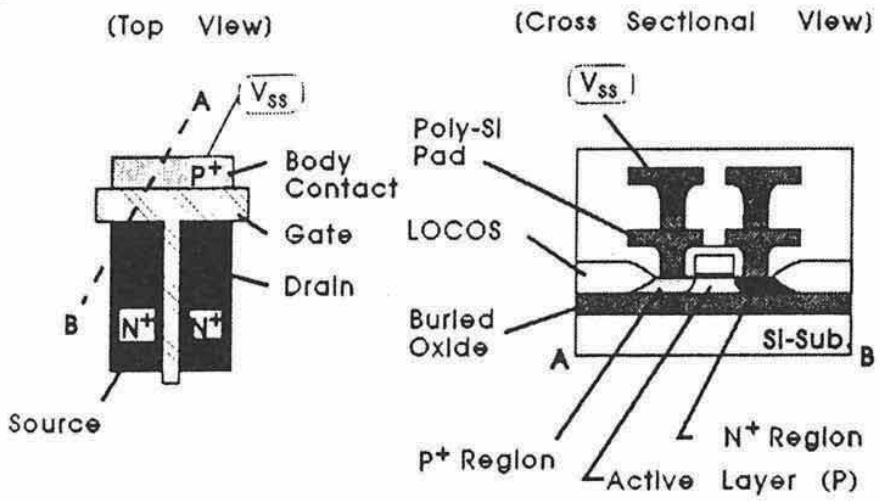


Fig. 11-10 A top-down and diagonal cross-sectional view of an SOI-NMOSFET with a body-contact made on one side of the MOSFET channel.

performance), of a large-scale IC design. As a result, in more recent SOI ICs built using partially-depleted thin SOI MOSFETs, attempts were made to find ways to eliminate the need to use such body-contacts. As shall be seen in Sect. 11.9, it has been demonstrated that in some circuit applications, such body-contacts may not need to used. That is, the device-performance-degradation caused by floating-body-effects may be small enough to allow the elimination of body-contacts from such SOI circuits.

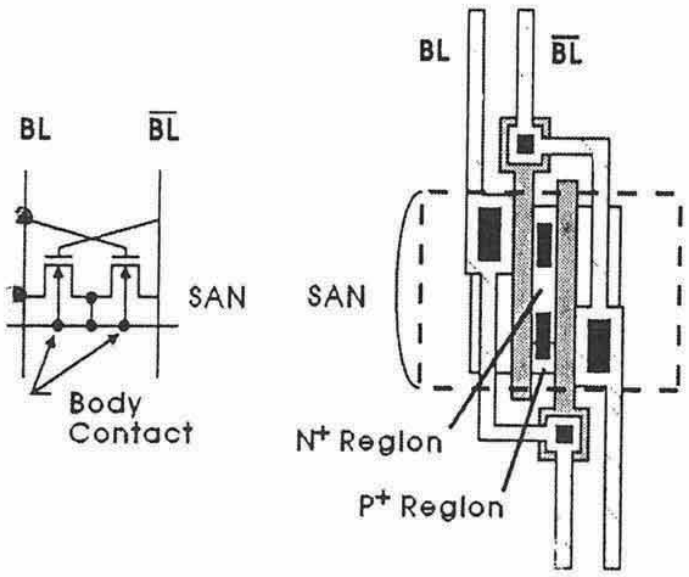


Fig. 11-11 Layout of a SOI-NMOSFET DRAM sense amplifier incorporating a body contact.

When CMOS circuits are built with partially-depleted thin-SOI-MOSFETs (and are powered with the same voltages), switching-speeds are observed to increase by 20-30% compared to equivalent circuits in bulk-silicon devices. Alternatively, partially-depleted SOI devices can be operated at lower voltages and lower battery-power-drains with a significantly-smaller speed-penalty than bulk-silicon devices. As noted earlier, since partially-depleted SOI-MOSFETs operate electrically in essentially the same manner as bulk MOSFETs, an immediately useful advantage is that circuit designs from bulk-silicon devices can be imported into the SOI environment with only a modest investment in circuit redesign. As of 2000, the SOI circuits being offered commercially were largely based on partially-depleted SOI-MOSFETs.

11.5.2 Fully-Depleted Thin SOI MOSFETs (FD-SOIs)

In *fully-depleted, thin-SOI-MOSFETs* (FD-SOIs) the depth of the depletion-region in the channel of the transistor extends entirely through the thickness of the surface-silicon-layer of the SOI-structure (Fig. 11-13b), even when there is no-bias applied to the gate. (In addition, the junction-depth formed in the source/drain-regions in fully-depleted SOI-structures is also as deep as the Si-surface-layer-thickness. Hence, there is no *pn*-junction at the [source/drain]/BOX interface, and thus there is no depletion-region at that interface to contribute to the source/body- and drain/body-capacitance.) This also implies that there is less overall junction-surface area than in a CMOS-structure formed in bulk-silicon (which results in reduced leakage-currents and junction-capacitances). In the FD-SOI case, the depletion-charge is thus constant and cannot extend further into the body as the gate-bias is increased. As a result, the application of a gate-bias results in more inversion-layer-charge for the same gate-bias as in a bulk-Si MOSFET. This has several beneficial effects, including: **1**) drain current is increased; and **2**) subthreshold-slope becomes steeper (Fig. 11-12). However, the $I_D(V_{GS},V_{DS})$ equations used to model bulk MOSFET and PD-SOI MOSFETS are no longer valid, and new I-V relationships need to be used for FD-SOIs. Although circuit-designs need to take full account of the special conditions of operation in an SOI layer, the highest gains in circuit-speed, reduced-power requirements, and the highest level of soft-error-immunity are realized with fully-depleted SOI-devices. That is, FD-SOI devices exhibit the following advantages over bulk and partially-depleted SOI-MOSFETs: **1**) sharper subthreshold slope;[12] **2**) higher drive-current; **3**) higher-transconductance; **4**) reduced parasitic-capacitance; and **5**) reduced short-channel-effects.[13] The latter benefit is gained because some of the charge that would otherwise be needed to support the

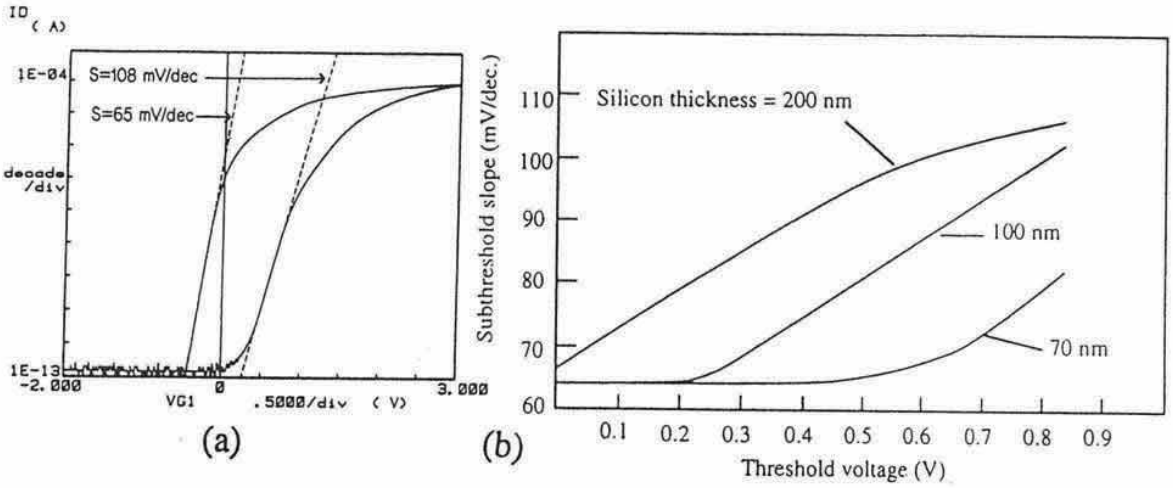


Fig. 11-12 a) Log-plot of drain-current I_D versus gate-voltage in SOI-MOSFETs made with two different silicon film thicknesses: 90-nm (left curve) and 450-nm (right curve). The inverse-subthreshold-slope, S_t (in millivolts per decade) is indicated. V_{DS} = 1 V; b) S_t as a function of threshold voltage for various SOI film thicknesses (*n*-channel device).

channel depletion-region-charge is instead produced by charge in the substrate under the BOX. However, as of 2001, fully-depleted SOI MOSFETs were not yet being widely implemented in commercially available ICs.

The reasons fully-depleted SOI MOSFETs were not being used in commercial production by 2001 is that these devices must meet the requirement that the silicon-film-thickness must always remain well below the depletion-region-width in the channel, during all modes of the MOSFET operation (i.e., from when the device is OFF through inversion). This dictates the use of a low threshold-voltage-value, and the ability to keep the Si-SOI surface-layer-thickness well below the depletion-region-width under conditions of thickness variation. That is, it must be possible to control the Si-surface-layer-thickness to within 5-10-nm.[14] This imposes severe limitations on the manufacturability of the device. (Use of partially-depleted SOI MOSFETs in an SOI-circuit, on the other hand, alleviates the constraint on V_T and its sensitivity, allowing the channel-doping-profile to be tailored to meet any desired V_T value, thus easing the manufacturing problem.) Nevertheless, by the end of 2001 this situation was starting to change. Intel announced a 50-nm FD-SOI structure, and said that they would use FD-SOI MOSFETs to build ICs from 2005 and beyond.[66] In addition to FD-SOI, the Intel MOSFETs incorporate elevated source/drains (see Sect. 5.3.2).

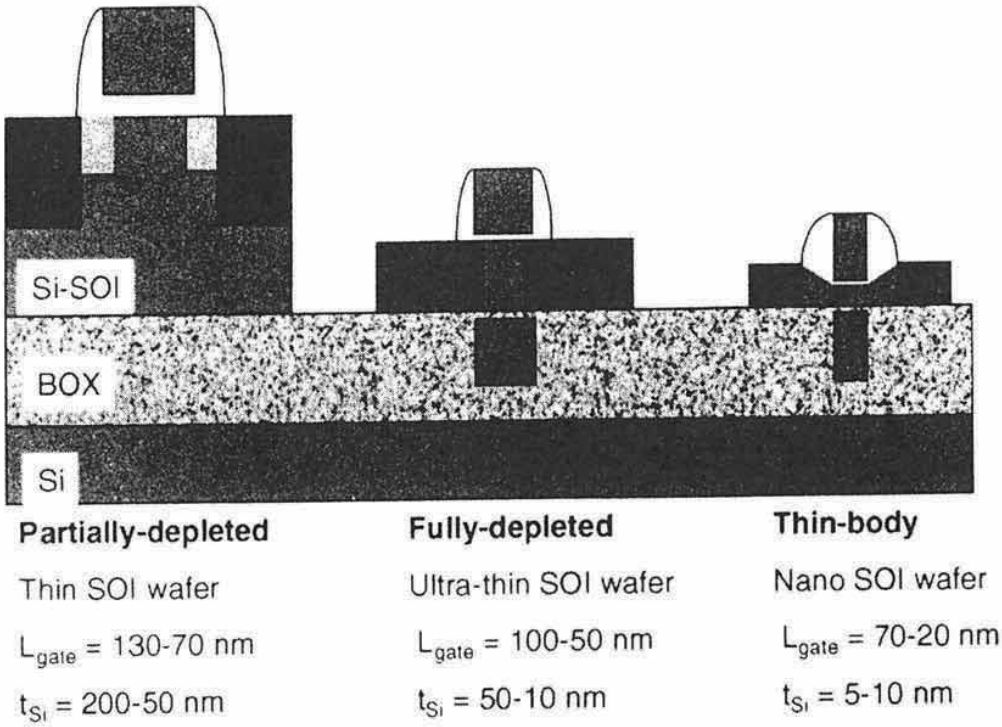


Fig. 11-13 Schematics of partially-, fully-depleted, and "thin-body" (double gate) SOI MOSFETs.

In FD-SOIs some of the gate-charge is balanced by charge in the depletion-region of the channel, but additional charge will arise in the substrate beneath the BOX layer. This induced charge in the substrate will have the effect of trying to establish an inversion layer *along the interface between the Si surface layer and the BOX* (i.e., the substrate beneath the BOX acts as a *back-gate*). As a result, a coupling exists between the potentials on the top-gate and the back-gate, and the electrical characteristics of the channel along the back-gate vary with bias applied to the top-gate. For example, leakage along the back-gate-channel will be caused by the application of a bias on the top-gate. In fact, when the Si-surface-layer of an FD-SOI becomes too thin, the top-gate eventually can no adequately longer turn-off the FD-SOI MOSFET. In such cases it may be necessary to fabricate these FD-SOI MOSFETs with a dual-gate structure and a thinned body-region (as discussed in Sect. 11.5.3.2 and as shown schematically in Fig. 11-13c). Such dual-gate FD-SOI-devices offer the prospect of excellent control of the off-current of small channel-length SOI-MOSFETs.

A dual-gate FD-SOI has both a top-gate and bottom-gate (Fig. 11-13c). The channel is sufficiently thin so as to be fully-depleted and an inversion-layer is established at both the top and bottom gates. To minimize short-channel-effects the channel-thickness must be considerably smaller than the gate-length.[48] For gate lengths smaller than 50-nm, this will require thin silicon layers of 20-nm or

less. The potential transconductance of a dual-gate device is approximately twice that of a bulk-device because of the presence of two gates. Also, the subthreshold slope, S_t, is potentially much improved over a conventional MOS device. Several theoretical studies have shown that the silicon thickness for a single-gate FD-SOI device must be about 1/2 that of a dual-gate device for equivalent short-channel effects.[48]

However, by 2000, dual-gate FD-SOIs were still in their infancy, and were far from being ready for consideration as devices in commercial ICs. That is, there are major problems in the practical manufacturing of dual-gate FD-SOIs. In a horizontal configuration, a gate must be fabricated below a thin silicon layer and accurate alignment of the top-gate over the bottom-gate must be achieved. All of these steps are very difficult.

11.5.3 Process Technologies Used to Fabricate Fully-Depleted SOI Devices

As noted earlier, one reason it is difficult to manufacture FD-SOI is that it must be possible to fabricate very thin Si-SOI surface-layers and to control their thickness to within 5-10-nm. Here some of the process sequences that have been developed to try to meet these requirements are listed.

11.5.3.1 FD-SOI Fabrication by Thinning the Si-SOI Surface Layer by Thermal Oxidation: An early process for fabricating FD-SOI devices was reported in 1992.[15] In this work a 220-nm-thick Si-SOI layer formed with a SIMOX process was the starting substrate. This layer was thinned to a 40-nm-thickness by carrying out a blanket thermal-oxidation of the Si-SOI layer and then etching off the oxide in an HF etch step. A LOCOS isolation-structure was used to provide lateral isolation. After forming the gate, a titanium salicide process was used to form $TiSi_2$ on the source-drain regions and on the polysilicon gate. The silicon in the source-drain regions was completely converted to $TiSi_2$ so that the silicide layer lies on top of the buried-oxide. A silicidation that fully consumes the SOI layer in the source/drain regions was chosen for ease of control. Figure 11-14 shows a cross-section of the finished device.

11.5.3.2 FD-SOI Fabrication by a LOCOS-Recess of the MOSFET Channel-Region: As the thickness of the Si-SOI layer (T_{Si}) has to become even thinner than the 40-nm-thick Si layer used in the FD-SOI device of the previous section, the parasitic resistance of the source/drain regions of such devices becomes too large to obtain the required device performance characteristics (even if these regions

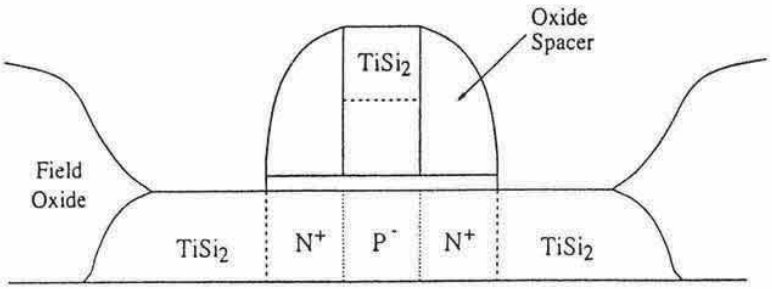


Fig. 11-14 Cross-section of the ultrathin SOI transistor with silicided source/drain and gate.

are silicided). An approach which allows the S/D regions to remain thicker while reducing the thickness of the body has thus been developed. In this method, the body region of the FD-SOI device is thinned by using LOCOS process, as shown in Fig. 11-15.[16] In one example, a SOI wafer with a T_{Si} of 195-nm was the starting material. A LOCOS process was used to thin the body region of the FD-SOI device to a 110-nm thickness. MESA active islands were formed by selectively etching back the Si-SOI to the BOX layer. These MESA islands were then oxidized to provide the lateral isolation.[17] Figure 11-16 shows an SEM of a completely fabricated FD-SOI device.

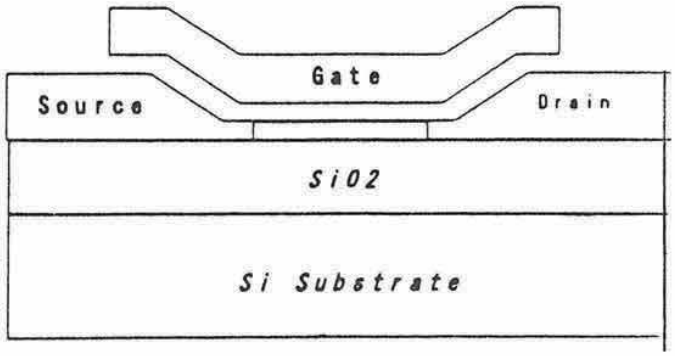


Fig. 11-15 Cross-section of a LOCOS-thinned "thin-body" SOI MOSFET.

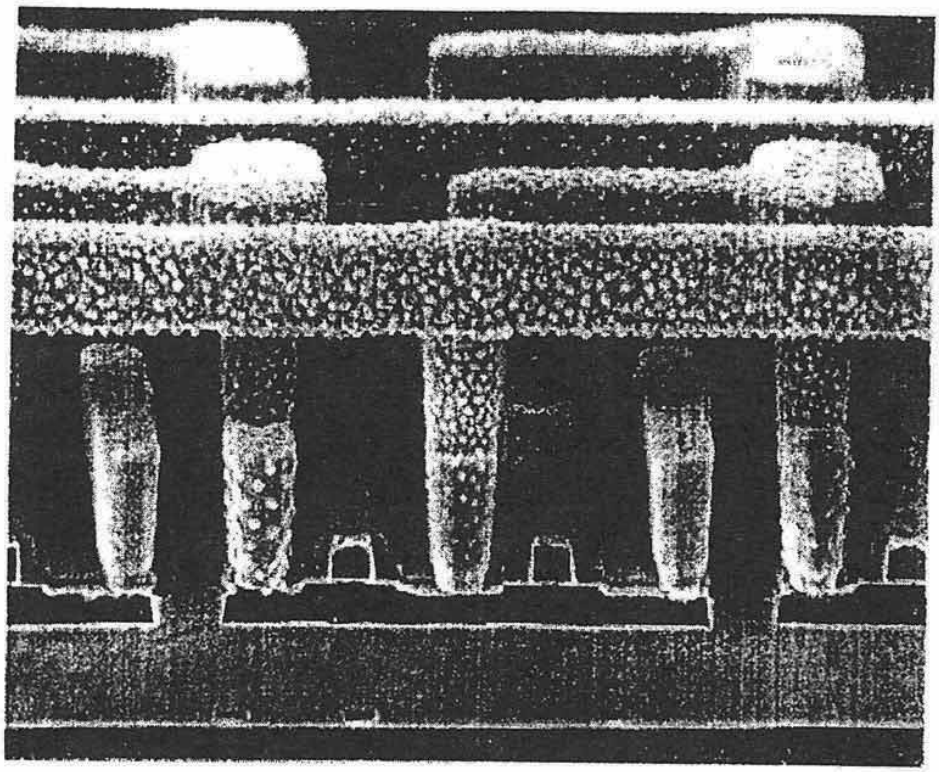


Fig. 11-16 SEM of completely fabricated SOI-MOSFET.

11.6 FABRICATION TECHNOLOGIES USED TO PRODUCE STARTING SOI-WAFERS

There a number of methods for fabricating starting SOI-wafers. Here those being currently used to make SOI wafers will be discussed. The annual production of SOI wafers in 1998 was only 10-million square-inches of silicon wafer. However, it is predicted that the total SOI-wafer-demand will grow by a compound annual growth rate (CAGR) of 55% between 1999 and 2005. In 2005, this should make the production of SOI wafers reach 187-million square-inches. The SOI merchant-wafer market registered $100 in revenue in 1999 (this did not include the captive SOI wafer production by such companies as IBM).

11.6.1 Silicon-on-Sapphire (SOS)

The work on silicon-on-sapphire (SOS) dates back to the 1960s when there was interest in this technology for its well-known immunity to radiation damage. Sapphire is insulating, and so circuits built in SOS substrates have high speed and excellent device-isolation. That is, unlike in other SOI approaches (where a buried oxide-layer is used to provide vertical isolation), the entire sapphire substrate is an insulating material. This reduces microwave losses and makes SOS an ideal candidate for the integration of rf-circuits for wireless

communications. Furthermore, the high thermal-conductivity of sapphire facilitates thermal dissipation through the substrate, which alleviates the self-heating effects of the other SOI technologies as described in Sect. 11.9. Device isolation can be easily achieved through simply removing the conducting silicon between the transistors by etching down to the insulating sapphire-substrate. SOS circuits were historically used for radiation-hardened applications. There are several excellent review articles on this subject and the reader is referred to them for additional details.[18,19]

Since SOS technology is limited to fabrication of MOS devices, a (100) orientation of the crystalline-silicon film is desirable. This orientation has been hetero-epitaxially grown on several different sapphire (i.e., crystalline alumina) orientations, specifically (0112), (1012), and (1102). The lattice-parameter mismatch between the Si-epi layer and the sapphire-substrate is sufficiently large to cause the generation of misfit dislocations, stacking faults, and edge dislocations in the Si-epi layer. The defect-concentration is heaviest near the substrate/epi interface and decreases as the film grows.[20] Figure 11-17 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 MOSFETs* (which are defined in Sect. 11.5) and these must be built in a Si-epi films only 0.1–0.3-μm thick. These films can be highly faulted, and require additional processing to improve their crystalline quality.

The accepted way to improve the film quality is the technique of amorphization (by implanting a high dose of Si into the deposited silicon

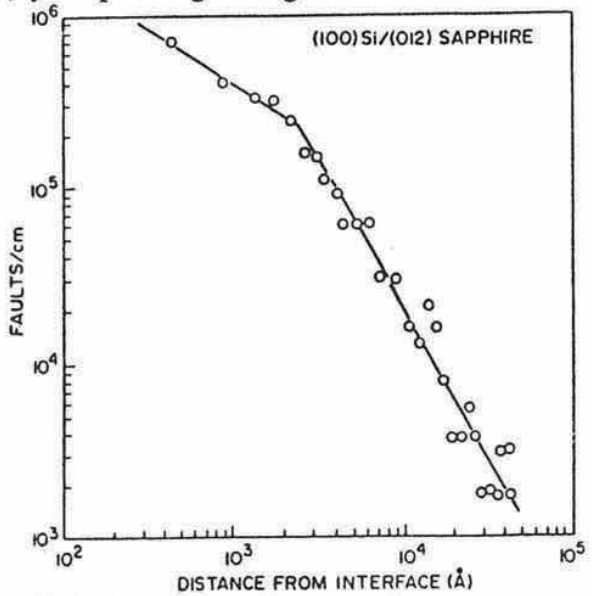


Fig. 11-17 Stacking-fault-density as a function of the distance from the silicon/sapphire interface in an MOS structure.[21] Reprinted by permission of the American Physical Soc.

epitaxial layer) followed by *solid-phase-epitaxial regrowth* (*SPE*).[22] In *thick SOS films* only a single amorphization is required. That is, an ion implantation step causes 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 approach, the Si-region near the sapphire-interface is first amorphized by a high-energy ion-implantation step, and the surface-regions are used as a seed for regrowth.[23] This is followed by a lower-energy ion-implantation step that then amorphizes the surface-region of the Si. The previously re-crystallized bottom layer is now utilized as the seed to re-grow 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.[24] 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*,[25] and Munteanu *et al.*[26]

11.6.2 SOI by Separation by Implanted Oxygen (SIMOX)

A widely used approach to manufacture SOI wafers is that of *Separation by IMplanted Oxygen*, or SIMOX. This process was first proposed by Izumi of NTT in 1978 (and who also coined the acronym SIMOX).[27] The SIMOX technology

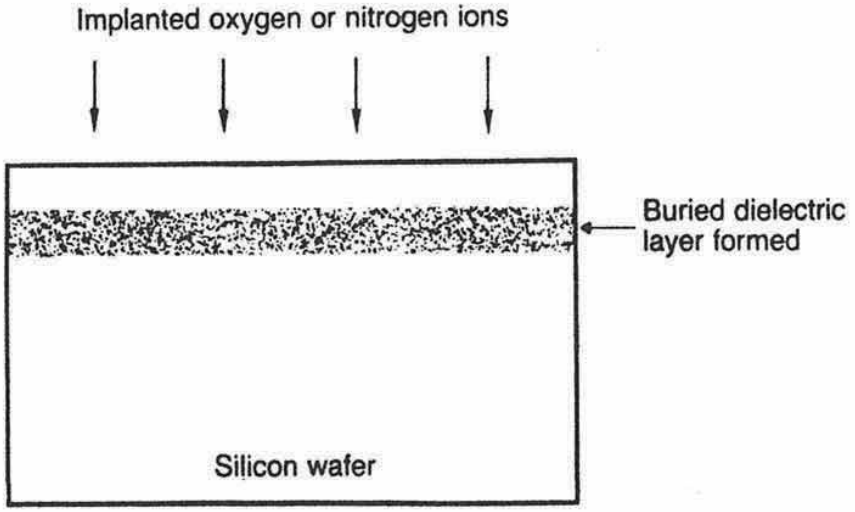


Fig. 11-18 Implanted oxygen ion process for the formation of a SIMOX wafer.

is based on the ion-implanting a high-dose of oxygen ions (O^+) into a silicon wafer (~$1.4x10^{18}$ O^+ ions/cm^2), such that the peak concentration lies about a half micron beneath the silicon surface. After implantation these wafers are subjected to a high-temperature-anneal which forms a continuous stoichiometric subsurface-layer of silicon-dioxide (Fig. 11-18). This *buried-oxide* (or BOX) layer electrically separates the top and bottom portions of the wafer, and is typically 350-nm thick. The single-crystal-region of silicon above the BOX is typically 250-nm thick. The result is a wafer with a thin surface-silicon-film of high-crystalline-quality, which also has an abrupt interface to a buried layer of stoichiometric-SiO_2.

The SIMOX process has undergone significant evolution since it was first introduced. In early-SIMOX the material was of very-poor quality. That is, the defect-density (dislocations, oxide-precipitates, and polycrystalline-silicon-inclusions) was so high that epitaxial-silicon had to be grown on top of the silicon overlayer in order to obtain a silicon film good enough for device fabrication. In addition, at that time, the implantation of sufficient oxygen to form a buried-oxide-layer in a 2x2 cm^2 area on a silicon-wafer required 24 hours (using a conventional implanter). Thus, most SOI researchers viewed SIMOX as just a laboratory curiosity. Nevertheless, when NTT produced a 1-k SRAM using SIMOX in 1982, the SOI community began to consider SIMOX from a less skeptical perspective.

In 1985, Eaton (now Axcelis) produced the first high-current oxygen-implanter, the NV-200, according to specifications provided by NTT. This machine could implant O^+ ions at a beam-current of 100-mA and at energies up to 200-keV. The availability of such an implanter was a key step toward allowing SIMOX development to go forward.

Before 1985, SIMOX wafers were annealed at 1150°C (which is the highest temperature available in conventional furnaces equipped with quartz tubes). In 1985, it was demonstrated that annealing at 1300°C in a furnace equipped with either a polysilicon or a silicon-carbide tube (and even at higher temperatures - up to 1405°C - in a lamp-annealing system) could dramatically improve the quality of the material, and produce atomically sharp silicon/silicon-dioxide interfaces. Other improvements (to be discussed in more detail later), included the use of multiple-implants, and a lower oxygen-dose. Eventually, by the late 1980s, SIMOX-wafers began to be produced commercially by publicly-traded companies (e.g., Ibis Technology).

As a result of all of this progress, three forms of SIMOX have been developed. The first two forms (*high-dose SIMOX* and *low-dose SIMOX*) use

conventional ion-implantation technology to implant a high-concentration of oxygen. A third version of the SIMOX process, which uses a plasma immersion implanter for the oxygen implant, is called *Separation by Plasma IMplantation of OXygen* (SPIMOX). The latter technique is a potentially low-cost process for fabricating SOI substrates, because the oxygen atoms are implanted by *plasma-immersion ion-implantation* (PIII, see Fig. 11-26). The whole wafer is implanted at once and implantation is thus independent of wafer size. This results in high wafer throughput. A possible drawback of this technique is the lack of ion-beam-selection and the fact that this process may cause more contamination in the SOI wafer than would a conventional ion-implanter.[28] High-dose SIMOX was the first technique to be developed, and it was followed more recently by the low-dose SIMOX-process. The high-dose SIMOX and low-dose SIMOX forms are described in more detail here.

Although SIMOX wafers can be obtained from several commercial suppliers, the high-dose SIMOX wafers are still significantly more costly than bulk-Si or even epi-Si wafers.[29] However, the development of the low-dose SIMOX process appears to significantly cut the premium that must be paid for SIMOX wafers. In fact, it is estimated that SIMOX wafers in the future may not cost any more than epitaxial wafers (whereas in 2000 they cost 2-3x the cost of epi wafers). Future cost-reductions in SIMOX wafers will come from such advances as increased beam-current (e.g., 2x that of present implanter-technology). Note that the SIMOX process has only two major steps: **1**) oxygen implantation; and **2**) annealing. Since SIMOX wafers are commercially available from various suppliers, introducing SIMOX wafers into production does not require any major infrastructure changes in a fab. Furthermore, after years of development work, SIMOX-SOI-wafer and device-fabrication processes have achieved sufficiently high-quality and low-cost to be applied commercially. In August 1998, IBM announced it would apply its leading-edge CMOS technology to commercialize SOI-CMOS products based on low-dose-SIMOX SOI.

11.6.2.1 High-Dose SIMOX

Figure 11-18 shows a schematic of the high-dose SIMOX process. The sequence of steps involved in the high-dose SIMOX process is as follows:

1. The starting material is typically a (100) device-quality Si-wafer. The wafer is first subjected to a high dose ($1\text{-}2\text{x}10^{18}/\text{cm}^2$) oxygen ($O^+$) ion-implantation step at high enough energy (150–300 keV) so that the peak (projected range) of the implant is deep within the silicon (0.3–0.5-μm).

This process is carried out using a dedicated high-current oxygen-ion-implanter, typically with a beam current of 100-mA. For reasons described in the next paragraph, the wafer is maintained at a temperature of ~600°C during the oxygen-implant step.

Note that stoichiometric-silicon-dioxide contains $4.4x10^{22}$ oxygen atoms/cm^3. Therefore, from such considerations alone, an implantation dose of $4.4x10^{17}/cm^2$ should be sufficient to produce a 100-nm-thick buried-oxide-layer. Unfortunately, due to the statistical nature of ion implantation, the oxygen-profile in silicon does not have a box shape, but rather a Gaussian profile. Thus, these implanted-oxygen atoms will spread out over a wider range than 100-nm, and silicon-dioxide-stoichiometry is not reached. If the wafer was annealed after such a $1\text{-}2x10^{18}/cm^2$ oxygen-dose, oxygen precipitates would be formed at a depth equal to the depth of the maximum oxygen-concentration, but no continuous layer of SiO_2 would be formed. Experiments have shown that a dose of $1.4x10^{18}$ oxygen atoms/cm^2 must be implanted in order to create a continuous-buried-oxide-layer. The standard dose (which is the most common) is $1.8x10^{18}/cm^2$, and produces a 400-nm-thick buried-oxide-layer upon annealing. As the dose is increased above $1.8x10^{18}/cm^2$, the overall oxygen-implant-profile no longer shows a higher peak value, but instead becomes broadened. This is due to the fact that at the elevated temperature of the implant, oxygen can readily diffuse to the Si/SiO_2 interface, where oxidation occurs. The dose of oxygen at which the continuous-buried-oxide-layer begins to form ($1.4x10^{18}/cm^2$) is called the "*critical dose*."

2. The wafers are then given a post-implant anneal for sufficient time (3–5 -hours) at an extremely high temperature (1300-1350°C [in furnaces equipped with poly-crystalline-silicon or silicon-carbide tubes], are now considered standard annealing conditions) in N_2 (with about 2% oxygen added, to protect the silicon from the pitting phenomenon which occurs when silicon is annealed in pure-nitrogen at high temperatures), or Ar. It is generally agreed that the higher the annealing temperature, the better the quality of the SIMOX material. An anneal in pure-nitrogen can also be performed if a cap-layer of oxide (usually deposited by CVD) is formed on the wafer after the implant, and prior to the anneal.

This extremely-high-temperature annealing-step has three functions: **1**) It drives the chemical reaction which forms a continuous-buried-amorphous-oxide (BOX) layer of silicon-dioxide near the peak of the

implantation. (The choice of a sufficiently-high O^+-implant-dose and energy provide the required oxygen concentration to enable stoichiometric SiO_2 to be formed); **2**) It removes many of the defects (dislocations) formed during the ion-implantation step; and **3**) It dissolves any SiO_x-precipitates in the silicon device-surface-layer and denudes this layer of oxygen as well (see Chap. 2 for a discussion of denuded-layer formation in CZ-silicon-wafers). To allow integrated-circuit devices to be fabricated on the wafer after the formation of the BOX, a thin layer of crystalline-Si must remain above the BOX-layer. The BOX-layer depth is defined by the depth of the ion-implantation and it's thickness is limited to about 0.5-μm.

The evolution of the SIMOX structure during the high-temperature-anneal (as mentioned above) is quite complicated, but it is useful to summarize the phenomena that occur from a thermodynamic perspective. During annealing, both dissolution of the oxide-precipitates and precipitation of the dissolved-oxide takes place in the oxygen-rich silicon-layers. Oxygen diffuses out of the top silicon-surface, initially denuding the zone near the silicon-surface. Within the bulk of the silicon, small precipitates dissolve into silicon and large precipitates grow from the dissolved oxygen. At any given temperature (and for a given concentration of oxygen in the silicon), there exists a *critical precipitate radius*, r_c, below which a precipitate will disappear, and above which it is stable. The critical precipitate radius increases with temperature, and becomes essentially infinite at very-high temperatures. At temperatures above 1300°C the only stable precipitate is the buried-oxide-layer itself, which has an essentially infinite radius-of-curvature. The complete dissolution of all precipitates at annealing temperatures greater than 1300°C, and the diffusion of dissolved oxygen towards the buried-oxide-layer explain the thickness increase of the BOX when high-temperature annealing is performed.

The fact that a thin-layer of high-quality crystalline-silicon must exist above the BOX-layer implies that the lattice structure of the silicon in this region must maintain a high-degree of crystallinity in the face of the high-dose oxygen-implant step. The high oxygen-implantation-dose does amorphize the layer of silicon that lies above the projected range. However, it would be unacceptable if complete amorphization occurred, and a polycrystalline layer was formed during further annealing. This would happen if the temperature of the wafer was kept too low during the oxygen implant step. On the other hand, by keeping wafers at a temperature greater than 500°C during the implant (by a combination of

beam power and external heating), the amorphization damage anneals out during the implantation process ("self-annealing"), and a high-quality crystal-structure is maintained in the silicon layer at the wafer surface. By maintaining such an elevated temperature during the implant, the implanted oxygen also tends to diffuse away from the vicinity of the projected range to form a broad, flat concentration plateau. Note that the most commonly used implantation temperatures range between 600°C and 650°C. This is because SiO_2 precipitates form in the silicon overlayer at higher temperatures (700-800°C), mostly near the bottom interface. However, large precipitates are also observed near the surface region of the silicon film (and this therefore sets an upper limit of 700°C on the implantation temperature).

3. If the device requires thicker silicon layer 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 the high-dose SIMOX technology, including the following:

1. The formation of SIMOX wafers requires extremely high doses of oxygen, and this necessitates very-long implantation times (which, as we shall see, contributes to the high-cost of SIMOX-wafers). The need to anneal the wafers at the extreme temperatures exceeding 1300°C may also be a limitation for 300-mm wafers.

2. Transition metals may be sputtered from the beam-line or end-station components during the implantation. These metals can deposit onto the SIMOX wafer surface and can then be driven into the silicon by the wafer-heating and subsequent thermal processing. Their presence in the silicon will enhance junction-leakage and reduce the minority-carrier-lifetime of the devices. Metallic contamination is present in all implant processes, but since high-dose SIMOX implants are nearly a thousand times greater than those of conventional high-dose source/drain implants, the metal contamination problem is potentially increased by approximately the same degree.

This issue was especially troublesome during the early days of SIMOX, when implanters were not yet built to limit transition-metal-sputtering. Since that time, however, great efforts were made to reduce the metal contamination sources, and silicon or silicon-dioxide shields have been placed in implanters such that the beam never "sees" any metallic part. The

design of high-temperature furnaces used to anneal the SIMOX material has also been optimized in order to minimize the diffusion of metallic contaminants from the furnace heating-elements into the furnace tubes.

3. Pinholes in the buried-oxide-layer will occur if a particle is present on the wafer during the oxygen-implantation-process. Such particles screen the oxygen implant at that point. The result is a breach in the buried-oxide-layer that represents a flaw in the insulating barrier at that point. Careful wafer cleaning and attention to particulate control in the implant process can control this problem.

4. 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 have been shown to be able to build excellent quality CMOS and bipolar ULSI circuits.[30]

5. SIMOX-wafers are more costly than bulk-Si-wafers, although the cost is expected to sharply decrease as the popularity of this technique expands. The cost reduction will be partly achieved by improvement in the implant equipment used to fabricate SIMOX-wafers. For example, SIMOX-dedicated ion-implanters are now available from IBIS Technology. Such implanters can furnish adequate throughput and acceptably-low levels of metallic-contamination. These production tools have an output of 100-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.

11.6.2.2 Low-Dose SIMOX

A recent trend with SIMOX fabrication is to use a lower oxygen-implant-dose than is used in the high-dose SIMOX approach. With the low-dose technique, improved, lower-cost SOI substrates are obtained. For example, a typical modern SIMOX implanter with a nominal 50-mA of O^+ beam current has a maximum output of only about 25 wafers of 200-mm diameter per day, round the clock operation. The low-dose SIMOX process also drastically increases the crystal quality of the silicon layer above the stoichiometric BOX layer. The oxygen implant dose is reduced to about $4x10^{17}$-$1.0x10^{18}$ ions/cm^2, and at an energy around 120-keV. Note that only a very-narrow process window close to the dose of $4.0x10^{17}$ ions/cm^2 will produce a continuous, precipitate-free BOX layer.

Exhibit 1015 - Page 53 of 120

Table 11-3 compares the process conditions used to produce SIMOX with high-, medium-, and low-doses of oxygen.

A silicon-layer between 150-200-nm deep is produced, but the lower oxygen-implant-dose also reduces the BOX layer to thicknesses of only 90-150-nm (depending on the selected oxygen dose).[31] A high-temperature anneal in Ar (i.e., at 1350°C) is then carried out, just as is done in the high-dose SIMOX process. However, this high-temperature-anneal in an inert ambient is followed by a subsequent high-temperature anneal in a gas mixture of inert gas and 30-60% oxygen (at ~1350°C).

The oxygen in the furnace-ambient at this high-temperature both oxidizes some of the superficial silicon layer, but also diffuses through the surface silicon layer and reacts at the buried-oxide/Si-interface to form new oxide within the Si (Fig. 11-19). This has the effect of increasing the thickness of the BOX-layer.[31] This process of growing the BOX-layer to a greater thickness by annealing the low-dose SIMOX wafers in oxygen is called an *internal thermal oxidation process* (*ITOX*). Note that as long as the oxide grown on the surface of the silicon during ITOX is thinner than 500-nm, there exists a linear-relationship between the thickness of this oxide layer (t_{ox}) and the thickness increase in the buried

TABLE 11-3 Physical and Electrical Properties of Commercial SIMOX

Parameter	Units	Standard Dose SIMOX	Medium Dose SIMOX	Low Dose SIMOX
Silicon Thickness	A°	1500 - 2000	1500 - 2200	1500 - 1800
BOX Thickness	A°	370 - 400	170 - 190	90 - 150
HF-Defects Density	cm^{-2}	< 1	< 1	< 1
Dislocation Density	cm^{-2}	$< 10^7$	$< 10^6$	$< 10^5$
Stacking Fault Tetrahedara	cm^{-2}	$< 10^5$	$10^5 - 10^6$	$10^5 - 10^6$
Surface Roughness (10um x 10um)	A°	7 - 10	5 - 7	3 - 5
Metallics (TXRF)	cm^{-2}	$< 10^{11}$	$< 10^{11}$	$< 10^{11}$
Short Density	cm^{-2}	< 0.1	< 2	< 2
Electron Mobility	$cm^2/V\text{-}s$	850 - 1000	850 - 1000	850 - 1000
Hole Mobility	$cm^2/V\text{-}s$	200 - 325	200 - 325	200 - 325
Oxide Charge	C/cm^{-2}	$2 - 3 \times 10^{11}$	$1.5 - 3 \times 10^{11}$	$1 - 2.5 \times 10^{11}$
Interface State Density	$cm^{-2}eV^{-1}$	$2\text{-}3 \times 10^{11}$	$1 - 3 \times 10^{11}$	$1 - 2.5 \times 10^{11}$
Transconductance (Gmsat)	mS	400 - 440	900 - 1200	1000 - 1300

oxide Δt_{BOX}): $t_{ox} = 0.06\ \Delta t_{BOX}$ (Fig. 11-19c). The internal-oxide also grows at the expense of the bottom of the silicon overlayer.

In the low-dose SIMOX process, the BOX formation consists of oxygen precipitation, growth of the precipitates, and their coalescence during the high-temperature-anneal process. This sequence of events occurs because the oxygen implantation dose is less than the SiO_2 stoichiometry, even at the depth of the peak-oxygen-concentration, and a stoichiometric-SiO_2-buried-layer is required to be formed to produce effective dielectric isolation. By carrying out the ITOX process, the BOX in the low-dose-SIMOX-process is made thicker and exhibits better isolation characteristics. The upper part of the BOX after the ITOX process is essentially a thermal-oxide-layer. Since the upper segment of an ITOX-SIMOX layer is essentially a thermal oxide, such important BOX properties as the breakdown-field, show a marked improvement (i.e., the breakdown-field is increased from 3-4-MV/cm for the case of conventional low-dose SIMOX to more than 5-MV/cm for ITOX-SIMOX, see Fig. 11-19d).[49] This improvement can be attributed to the BOX-thickness incremental-increase due to ITOX. In addition, the ITOX reduces the size and density of silicon pipes in the BOX, which is the main cause of BOX breakdown.[50] In 2001 IBIS Technology licensed a low-dose-SIMOX-technology from IBM called Advantox® MLD.

In 2000, IBM said recent advances in the production of low-dose-SIMOX-wafers will cut incremental processing costs of SOI to only 3% over the costs of producing ICs on bulk-Si-wafers. Low-dose-SIMOX-wafers are commercially available from such vendors as Ibis, Komatsu, and Nippon Steel. For example, the Ibis-1000 model is currently available. The next generation oxygen implanter, the Ibis-2000 (which can accommodate both 200-mm and 300-mm wafers) was under development in 2001, and is forecast to be ready for delivery in the first half of 2002. Figure 11-20 is a photograph of a Power4™ processor made by IBM using a SIMOX-SOI-based process.

11.6.3 Wafer Bonding

Another popular technology for producing SOI wafers is *wafer bonding*.[32] Wafer bonding is a simple concept. Basically, wafer bonding refers to the phenomenon that mirror-polished, flat and clean wafers of almost any material, when brought into contact at room temperature, are locally attracted to each other by van der Waals forces and adhere, or *bond* to each other. The bonding process is started by locally applying a slight pressure to the wafer pair, and the bonded area spreads laterally over the whole wafer area in a few seconds. Because the bonding achieved at room temperature is typically relatively weak, wafer pairs bonded

Exhibit 1015 - Page 55 of 120

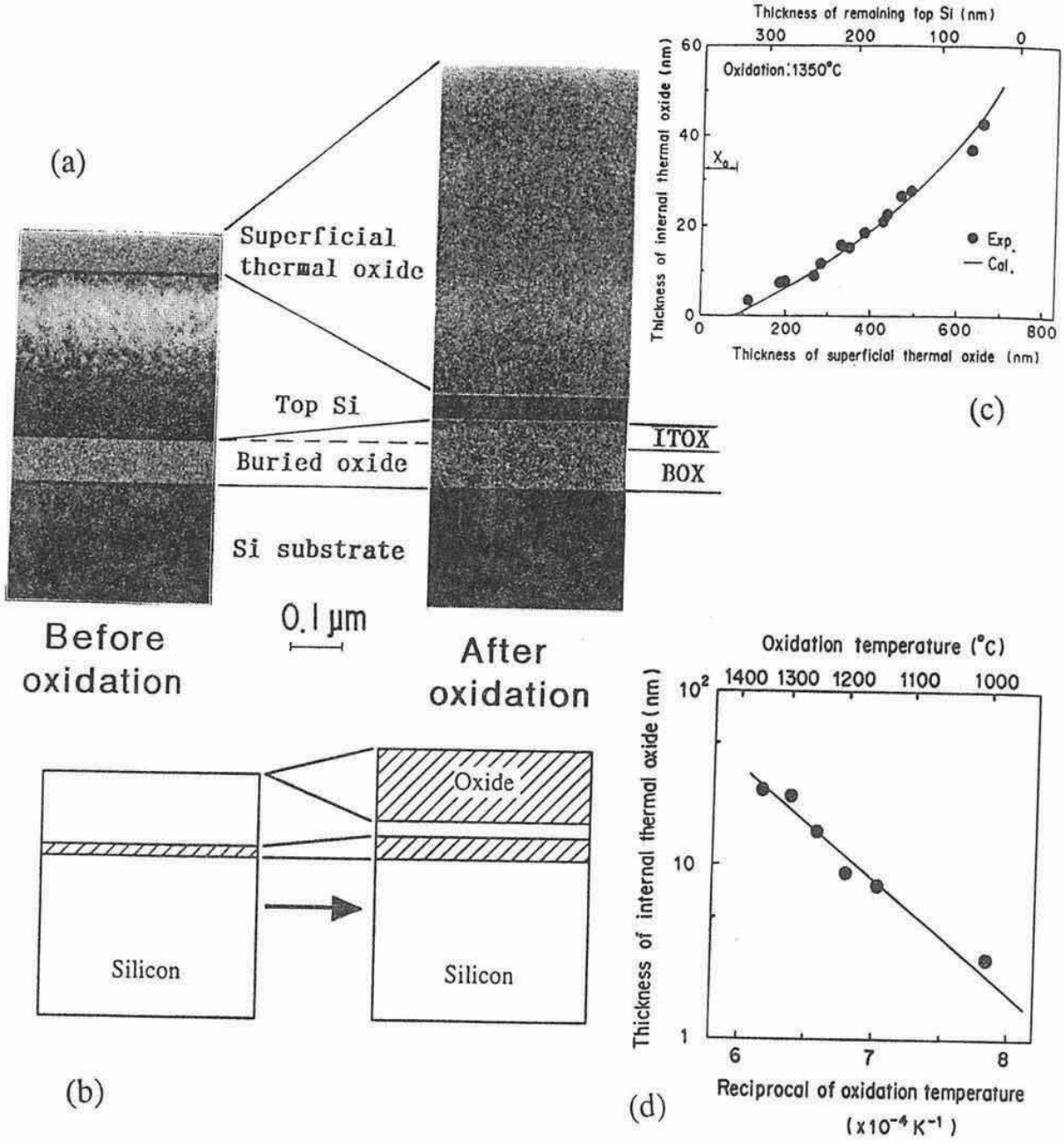


Fig. 11-19 a) An XTEM of SIMOX wafers before and after high-temperature-oxidation at 1350°C; b) A schematic drawing of SIMOX wafers before and after high-temperature oxidation at 1350°C; c) Relationship between the thickness of the ITOX layer and the superficial thermal-oxide-layer formed by high-temperature-oxidation at 1350°C; d) Relationship between the thickness of the ITOX-layer and the reciprocal of the oxidation-temperature.

at room temperature are usually subjected to a heat treatment to strengthen the bonding interface.

In the application of the bonding-process to form SOI-substrates, two thermally-oxidized-Si wafers (or one oxidized-wafer and one bare-Si-wafer) are first

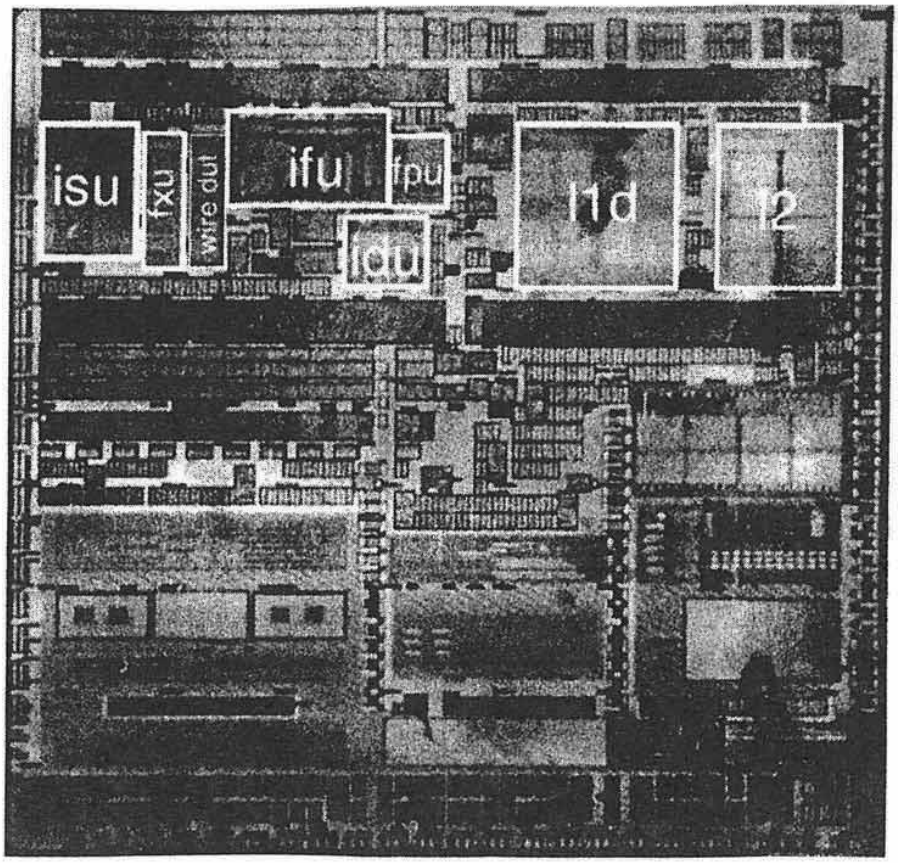


Fig. 11-20 Photograph of an IBM Power4™ processor using a SIMOX-SOI-based process.

first bonded by applying pressure to the wafers at room temperature (Fig. 11-21). A subsequent furnace anneal is used to increase the bonding strength. After such bonding, most of the silicon from one of the wafers is removed, leaving a thinned

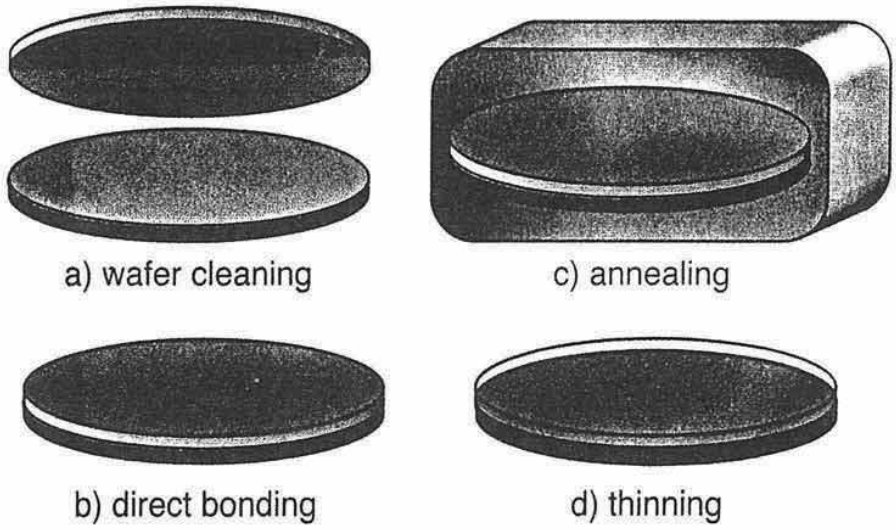


Fig. 11-21 Generic direct-wafer-bonding process flow.

layer of silicon from one of the wafers. The IC devices are built in this remaining thinned-silicon-layer, which now exists over the oxide present at the interface of the bonded wafers (i.e., a buried-oxide). The wide-availability of chemical-mechanical polishing (CMP) in IC processing, as well as a variety of other thinning approaches has led to wide and diverse use of wafer bonding. Unlike SIMOX, the wafer-bonding technique can form BOX-layers up to 50-μm thick. A book has recently become available dealing with the topic of wafer bonding.[33]

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.[34] Bonding is believed to take place by the conversion of oxygen trapped between the wafers, into 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 applies a moderate voltage across two oxidized-wafers placed face-to-face at temperatures in the range of 1100–1200°C.[35] 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 uses 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.[36]

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).[37] 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. It is also possible to achieve full-bonding-energy directly at room-temperature under ultra-high vacuum conditions if the wafer surfaces are free of adsorbates.[38]

It should be mentioned that particles or protrusions on either surface being bonded, do prevent bonding, (at least locally) around these disturbances. These lead to locally-unbonded-areas that are frequently termed *interface bubbles* or *voids*. Such voids are, of course, undesirable, since they can lead to delamination of devices located in imperfectly-bonded-areas during chip-assembly and

packaging. Voids with a diameter of several millimeters are readily created by particles 1-μm or less in size (and these are easily revealed by infrared or ultrasonic imaging). Therefore particles on wafers have to be avoided by proper cleaning and by carrying out all bonding procedures in an ultra-clean environment (Fig. 11-22).

11.6.3.1 Bond-and-Etch-back SOI (BESOI): 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 process. The *bond and etch-back SOI* (BESOI) technique is used to produce relatively thick-films of both oxide and silicon. The BESOI wafers typically have thicker, yet better-quality silicon and buried-oxide-layers than do SIMOX wafers.

The silicon-layer on BESOI substrates can be thinned in one of several ways. Mechanical grinding and polishing can be used to thin the Si to a thickness down to ~1-μm, but only to within 10-30% uniformity. Better uniformity can be achieved by using an etchstop. Several etch-stoplayers have been developed, including: a boron-doped-layer, a silicon-germanium-layer, a carbon-implanted-layer, and a porous-silicon-layer.

In the first of these etchstop methods, a lightly-doped B-region can act as an etchstop-layer when it is formed on a heavily boron-doped-substrate (Fig. 11-23). That is, the wafer that will ultimately contain the device consists of a lightly-B-doped epitaxial-layer on a heavily-B-doped-wafer. After oxidation and bonding of the lightly-doped surface of this wafer to another wafer, the heavily doped

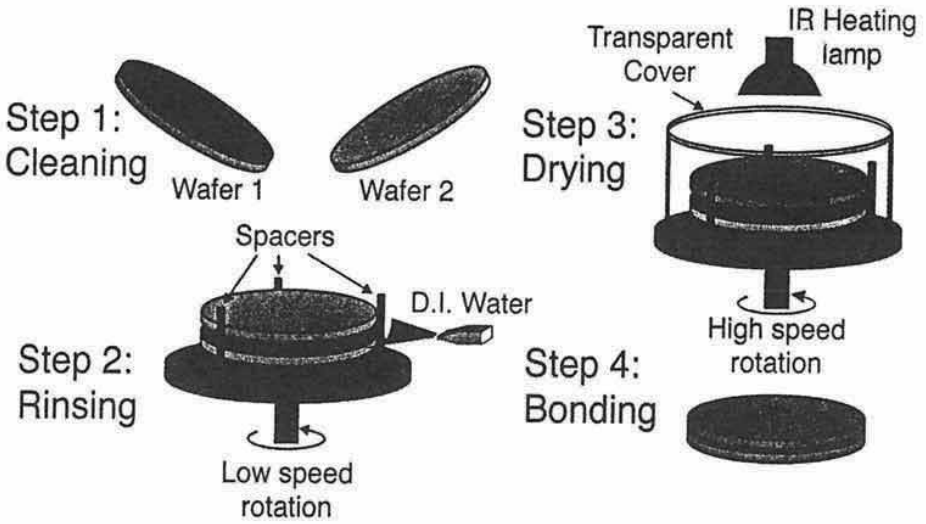


Fig. 11-22 Schematic of the microcleanroom setup and procedure.

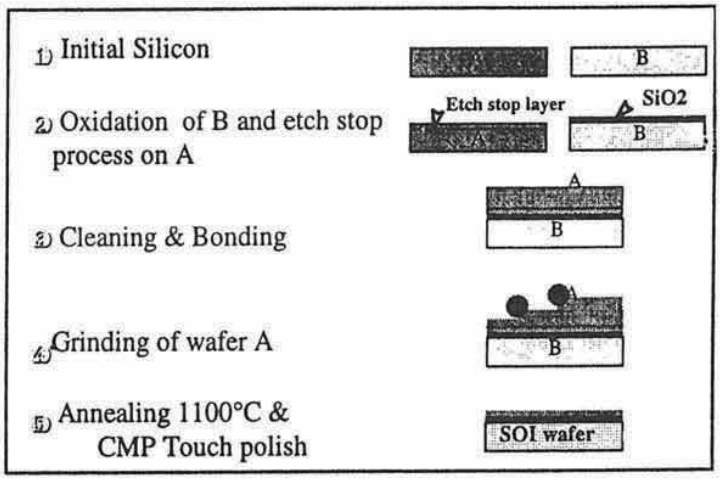


Fig. 11-23 Schematic of the Bonded-and-Etched-Back SOI technology (*BESOI*).

substrate is removed by an etch process that preferentially etches such heavily doped silicon. The epi-layer that remains bonded to the second wafer determines the thickness of the active-device-layer and the sum of the thickness 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. Another is that the uniformity of the thickness of the silicon-layer depends on the etch-stop-selectivity (which can vary, depending on the etch-stop method used, from 10 to 10^5).

Such BESOI methods are widely used for "thick" SOI applications (i.e., > 1-μm), such as power devices (high-voltage and high-temperature devices) and micromechanical machines (MEMS). The power-devices are used for military, industrial, and aerospace applications. The share of the thick-film segment of the SOI market dropped from 80% in 1995 to 53% of the market in 1999 and was taken over by the thin-film segment. Future thick-film SOI applications will be driven by industrial and automotive applications.[54]

11.6.3.2 Hydrogen-Implantation-Induced Layer Splitting to Create SOI Substrates (SMART-CUT®): A recent (and elegant) wafer-bonding-based method has been developed for forming SOI substrates using a technique to obtain very-uniform and thin regions of single-crystal-silicon. This technique employs a process called *hydrogen-implantation-induced layer splitting*.[39] The process is also referred to by its inventor M. Bruel (and by the company that sells such SOI

substrates [SOITEC SA, France]), as *Smart-Cut.*® The Smart-Cut process also allows the sacrificial-wafer to be reused and only one wafer is therefore required to make a bonded SOI substrate. (The SOI material made with the Smart-Cut process is termed UNIBOND.®) There are two reasons why this process may be significantly less-expensive than the process used to fabricate SIMOX-wafers. First, the sacrificial wafer can be re-used. Second, because the implant-dose is an order of magnitude lower than even one in the low-dose SIMOX-process (and the implant is done at room temperature in a conventional implanter), the cost of the implant step is significantly lower, compared to those used in SIMOX.

Figure 11-24 shows how the process works. Two device-quality-wafers (labeled A and B) start the process. Wafer-A is oxidized (to a thickness of about 0.5-μm) to form what will become the buried-oxide of the SOI-structure. Then this wafer is ion-implanted with a high-dose of hydrogen ions ($\approx 5 \times 10^{16}$ H^+ ions/cm^2) through the oxide, using a conventional high-current-implanter. The peak-concentration of this implanted hydrogen-dose is at a projected range R_p of about 0.2-μm below the oxide-surface. At this point, only micro-cavities or micro-bubbles are formed at the R_p of the implantation. For reasons explained next, the depth of the H^+ ion implantation determines the thickness of silicon above the BOX.

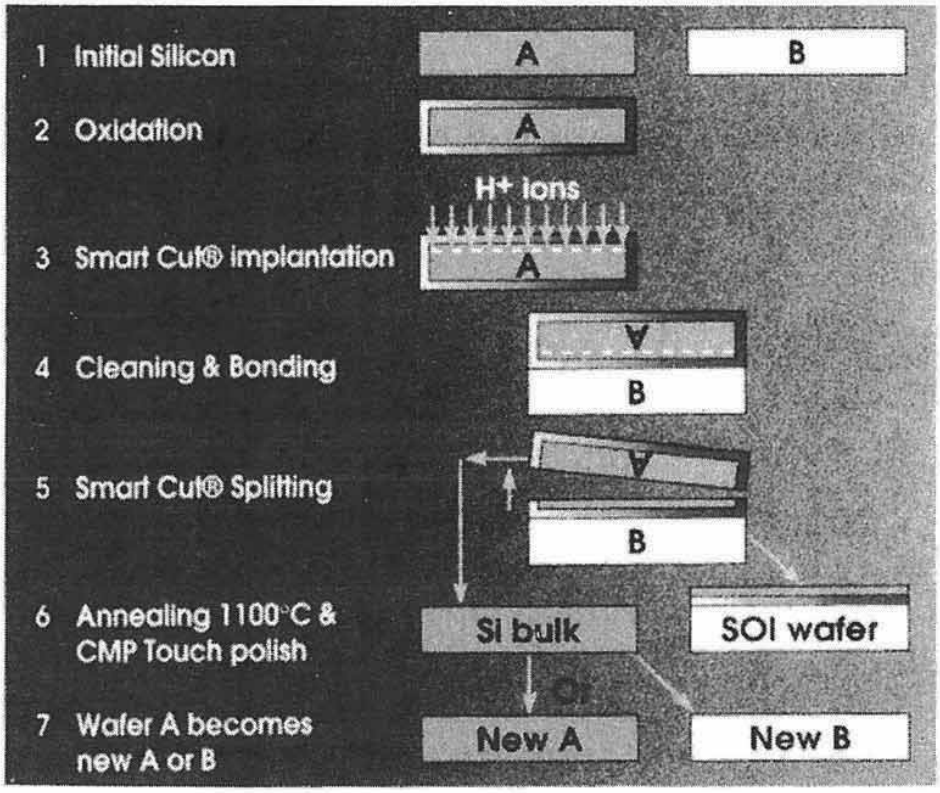


Fig. 11-24 Process flow for fabricating Smart-Cut® wafers.

After implanting the hydrogen into Wafer-A, both Wafers A and B are cleaned and low-temperature bonded face-to-face. The bonded wafers are next annealed (at 400-600°C). During this anneal the implanted-hydrogen precipitates, and this precipitation-process merges the buried network of bubbles and voids at the vicinity of the depth of R_p. As these hydrogen-filled microcavities continue to grow and merge, they eventually form a continuous microcrack along the line corresponding to the peak of the implanted-hydrogen. At that point Wafer-A delaminates from along that line, and Wafer-B (which now has a BOX-layer covered by the silicon layer from Wafer-A) becomes the SOI wafer. Meanwhile, Wafer-A remains available to be used as the sacrificial wafer of another UNIBOND substrate. Both wafers are then annealed at 1100°C to remove defects. This is followed by a "light" CMP (or "touch-polish") to erase the surface-roughness resulting from the layer-splitting process. Wafer-A, which is now only 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 use of ion implantation for layer separation improves the layer-thickness-uniformity over that of bond-and-etch-back SOI-wafers. The method is used to prepare commercially available 200-mm SOI wafers with a silicon thickness above the BOX in the 100-nm range and a uniformity of better than 5%. The dislocation density is $< 10^2/cm^3$, the surface roughness is < 2 Å (rms), and metal contamination is $< 5x10^{10}/cm^3$.[40]

11.6.3.3 Nanocleave® Wafer-Bonding SOI: The fabrication of starting wafers for SOI can also be carried out in other ways. One of these processes, named *Nanocleave®* was introduced by Silicon Genesis Inc. in 1998.[43] The process combines growing a SiGe/Si epitaxial layer, low-temperature wafer-bonding, and a room-temperature cleaving-process to produce sub-100-nm-thick layers of single-crystal-silicon over a buried-oxide substrate. One of the key advantages of the Nanocleave process is that it can produce such thin SOI-layers with a uniformity of several percent over an entire wafer. This type of uniformity is necessary for the successful fabrication of fully-depleted CMOS-SOI-FETs. Other thinning techniques (such as etch-back or CMP have increasing difficulty maintaining such SOI-layer thickness-control when silicon layers of sub-100-nm thickness are being produced. The Nanocleave process uses an atomic-scale cleaving-process that produces cleaved surfaces with an as-cleaved surface roughness less than 5-Å (as measured by AFM). In addition, the Nanocleave process potentially offers a method that is lower in cost than any of the other fabrication processes used to fabricate SOI-substrates. The Nanocleave technique also offers another path toward the capability of fabricating both "ultrathin" Si-

SOI-layers (10-to-50-nm-thick Si) and "nanothin" Si-SOI (5-to-10-nm-thick Si) required for the MOSFETs with gate-lengths smaller than 50-nm that are scheduled to begin production by 2010.

The Nanocleave process begins with the formation of a *cleave-plane*, the growth of a thermal-oxide-layer and Si-device layer on a high-quality "donor" wafer (Fig. 11-25a).[44] This "donor" wafer is then bonded to an inexpensive oxidized "handle" wafer (on which the device layer will ultimately reside).

The cleave-plane is created by first epitaxially growing a layer of Si-Ge on the "handle wafer, and then continuing the growth of this layer with an epitaxial-Si device layer (Fig. 11-26a).[63] The buried Si-Ge-layer is a compressive-strain-layer of a non-dislocated Si-Ge alloy, which contains no porous-regions, inclusions, or voids, as shown in Fig. 11-26a. When device wafer is separated, the resulting separated surface has a surface roughness, as measured by AFM profiling of 0.2-nm, or less.

After the cleave plane has been formed and a thermal-oxide is grown on the surface of the wafer with the cleave-plane (which will become part of the buried-oxide-layer in the finished SOI-device), the "donor" wafer is bonded to the

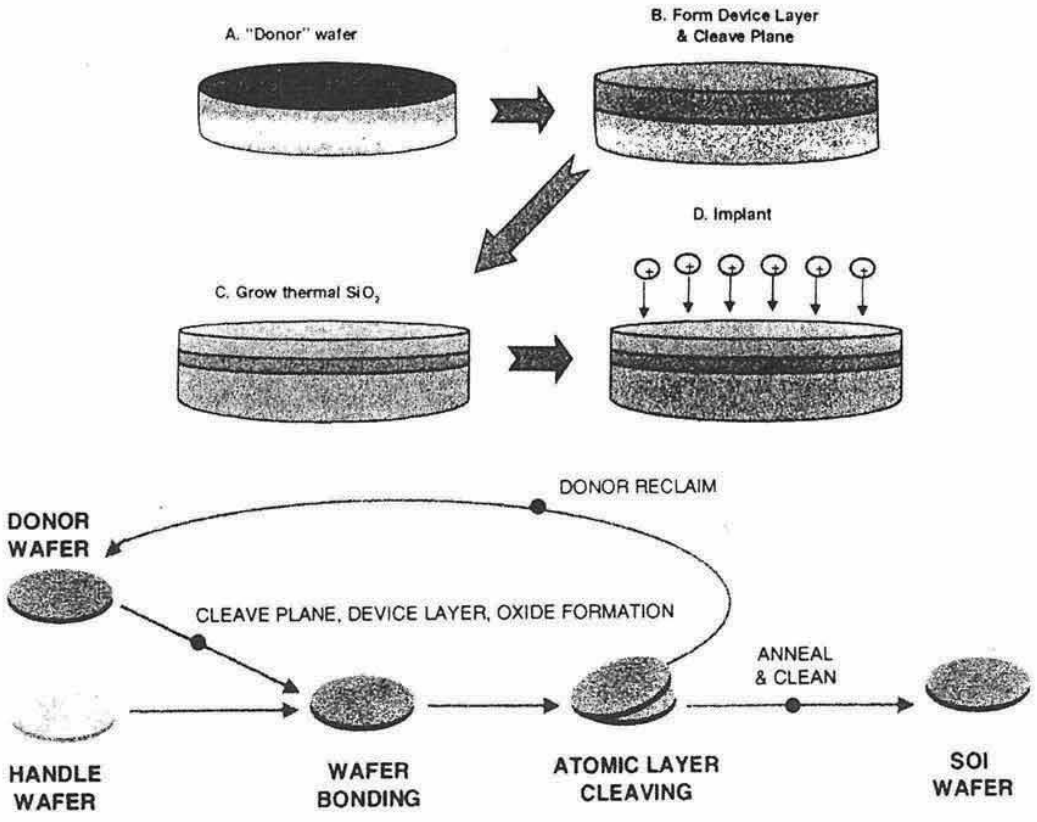


Fig. 11-25 Process flow of the Nanocleave® SOI technology.

"handle" wafer (which also contains a surface-oxide layer). Plasma-bonding techniques are employed to increase the initial bond-strength to sufficient levels for the room-temperature process used to transfer the silicon-layer from the "donor" to the "handle" wafer. This transfer process (by cleaving the device layer from the "donor" wafer from the "handle" wafer) is carried out next (Fig. 11-25b). Since the bond-strength is significantly greater than the interface-strength along the cleave plane, a novel cleaving process is used to separate the Si device layer from the "donor" wafer (which results in the creation of an SOI structure on the "handle" wafer). In the Nanocleave process, the donor wafer cleaving is done at room temperature (unlike in the Smart-Cut-process that uses an elevated temperature anneal for the cleaving step). In the Nanocleave process, the drive force is a static pressure of N_2 gas. A mechanical fixture secures the bonded wafer pair as a high-pressure nitrogen-gas is introduced along the strained Si-Ge layer. That is, cleaving is controlled by fracture along a cleave-front within the Ge-rich Si-layer. A thin-film of material from the "donor" wafer remains bonded to the substrate "handle" wafer, producing an SOI-structure. The result is an SOI-wafer with a silicon-layer whose thickness uniformity is within 1% across a 200-mm wafer. Furthermore, the atomic-layer cleaving-process is claimed to produce an as-cleaved surface roughness of less than 1-nm (typically 2-5 Å).[63] This is an order-of-magnitude smoother than the 80 Å of hydrogen-induced thermal cleaving obtained with the Smart-Cut technique. The smoother surface in the Nanocleave process is acceptable for many IC applications with no additional

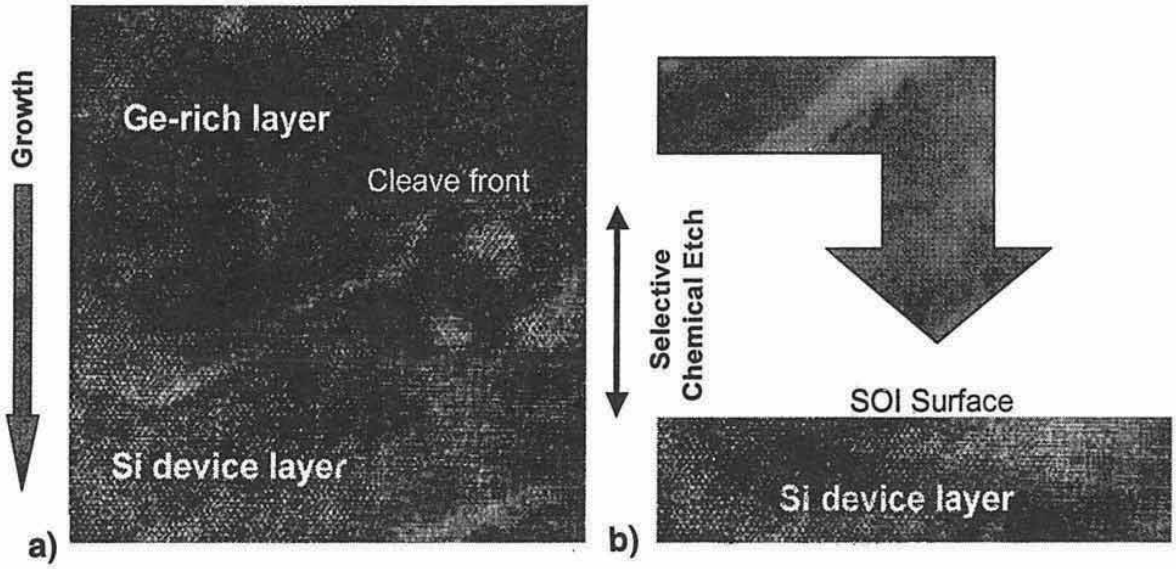


Fig. 11-26 a) HR-TEM image of the boundary area of a Ge-rich, epitaxial strain-layer and a crystalline-Si-device-layer in a state prior to atomic-layer-cleaving; b) The residual Ge-rich layer on the as-cleaved-wafers are subsequently removed through a selective etch step.

surface polishing. The edge of the SOI-layer has a smooth and regular character without the need for edge polishing. The residual Ge-rich layer on the as-cleaved wafers is subsequently removed through a selective etch step (11-26b).

A fully automated tool that separates bonded donor-handle wafer-pairs is available for this purpose. The high-quality wafer can be used to produce many thin-film transfers, reducing the process cost. Another factor that contributes to further cost-reduction arises from the fact that the "handle"-wafer is separated from the device-layer by the buried-oxide-layer. Consequently, a "handle" wafer with considerably relaxed electrical and chemical specifications (and therefore lower cost) can be used in this process.

Post-cleave-processes for layer-thickness-adjustment include: **1**) Si layer thinning using etching processes; and **2**) layer-thickening by growth of additional Si or Si-Ge multilayers. The layer-thickness-control for thickened wafers is better than 0.5% with a within-wafer thickness-uniformity of 2% or better (Fig. 11-27).

11.6.3.4 ELTRAN® Wafer Bonded SOI

A relatively new SOI structure (originated and developed by CANON Inc.), which is also based on wafer bonding, is called *ELTRAN.*® The name ELTRAN comes from *Epitaxial LAyer TRANsfer by BESOI* (Bond and Etch-back SOI method) with selective-etching of porous-silicon (Fig. 11-28).[45,46] ELTRAN wafers became commercially available in 1997 for 200-mm substrates.

The porous-silicon surface on an otherwise bare-silicon-wafer is formed by an anodizing treatment of the silicon surface. The ELTRAN method is characterized as BESOI implemented by growing an epitaxial-layer on a layer of porous-silicon. (It has been demonstrated that high-quality epitaxial-layers can be grown on porous-silicon.) The surface of the epitaxial-layer is then thermally oxidized, and this layer will eventually become the buried-oxide-layer of the SOI-structure. This wafer (with a thermally-oxidized epitaxial-film grown on porous-silicon) is then bonded to a "handle" wafer as the "donor" wafer. In early ELTRAN wafer fabrication process, the "donor" wafer material was removed after bonding by a grinding process (down to the porous-silicon-layer). The exposed porous-silicon that remains after the grinding-process is then removed by a wet-chemical etch. A mixture of HF and H_2O_2 is used for this etching step, since it provides an etch selectivity as high as 105:1 between the porous-silicon and the epitaxial-layer. At the end of the process, only the epi-layer remains on the top of the oxide layer and handle-wafer. The ELTRAN process has been shown to produce 470-nm-thick SOI-layers with a 14-nm thickness uniformity.

In the more recent ELTRAN process (Fig. 11-29), a double porous-silicon

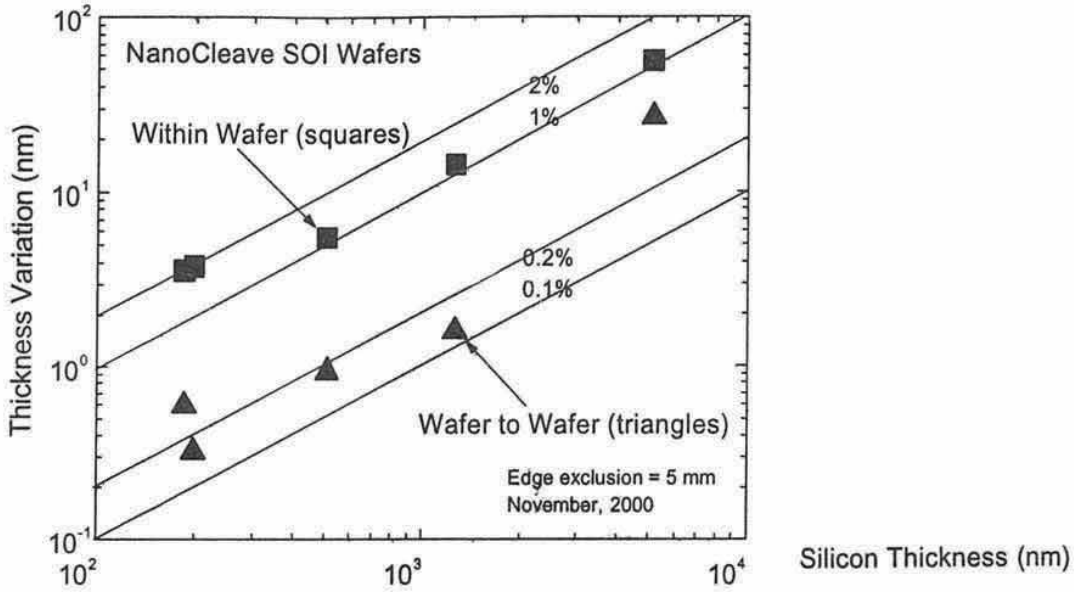


Fig. 11-27 Si-SOI layer-thickness-uniformity and average layer-thickness-control for post-cleave thickening of as-cleaved SOI-wafers with cleaved 200-nm Si-SOI-layers.[63]

layer is grown on the "donor" wafer, with each of these porous-silicon-layers having a different porosity. The porosity of the porous-silicon-films on the wafer surface can be adjusted by changing the value of the anodization current. After bonding it becomes possible to physically separate the "donor"-wafer from the

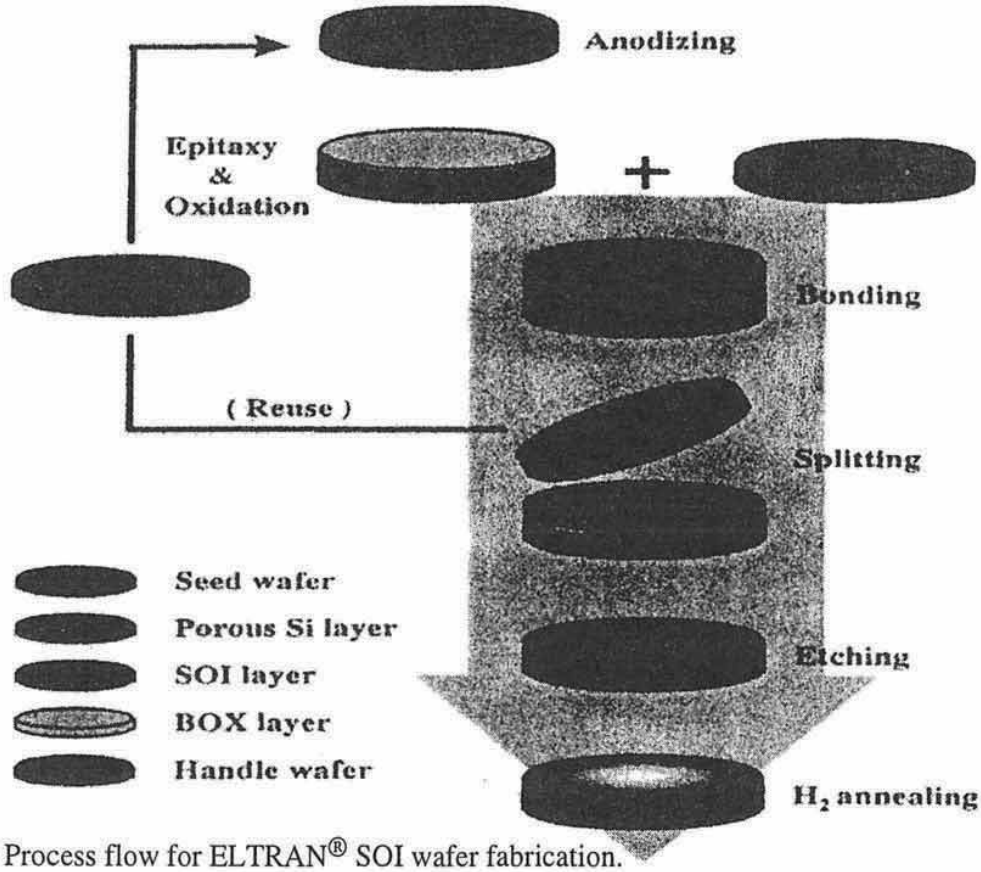


Fig. 11-28 Process flow for ELTRAN® SOI wafer fabrication.

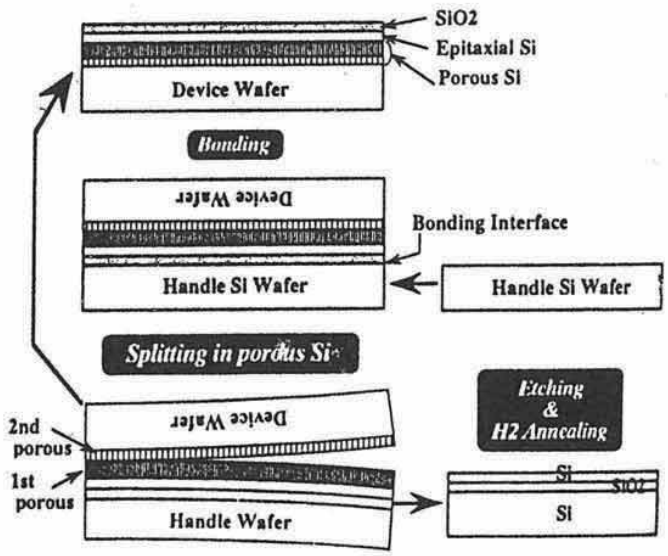


Fig. 11-29 The ELTRAN process flow for device wafer re-usage. Bonded pairs are split within the porous-Si layer. The device wafer sides are used as a starting device wafer. The handle sides result in ELTRAN wafers after etching the porous-Si and H_2-annealing.

"handle" wafer using a physical splitting process (i.e., a water-jet slices the SOI-layer from the seed layer).[62] The splitting occurs along the interface between the two porous-silicon layers. That is, since the porous-layers are a double-stacked structure with different porosities, the strain-energy accumulated at the interfacial area limits splitting to an extremely narrow range (because the interface between the first and second porous layers is mechanically the weakest point in the structure). Therefore, the stress-configuration induced by the double porous-layers is responsible for the splitting mechanism.

For splitting a "water-jet" technology is used. The water jet – a small collimated water stream – is injected at the edge of the bonded-pair and splits the two wafers at the thin porous-silicon layer. The double-layered porous silicon structure contains the stress concentration, with the result that both of the split wafers (SOI and seed wafer) are covered by the porous silicon protective layers. The water-jet drives high-pressure water into narrow gaps by fluid wedging and applies uniform pressure, reducing the chance of damage. The splitting yield is reported to be over 90%, and automatic water-jet splitting-machines are in production use.

The device layer over a buried-oxide is then formed by the etchback of the porous-silicon. After the porous-silicon-film is removed, the surface of the device silicon-layer is flattened by hydrogen-annealing (see Chap. 2, for a discussion of hydrogen-annealing of Si wafers).

There is another economic benefit of the more recent, double-porous-silicon-layer process used to form ELTRAN wafers. In this process (and like in the Smart-Cut and Nanocleave processes discussed in the previous sections), after the splitting process, the "donor"-wafer becomes available to be used as another a "handle"-wafer (instead of being ground away, as in the early ELTRAN process). Canon announced that it has expanded the production of ELTRAN wafers from 2000-per-month to 10,000-per-month in November 2000. This production expansion is being supported by a greater marketing of ELTRAN wafers. To this effect Canon opened a marketing office in the U.S. to achieve this goal.[53]

11.7 LATERAL ISOLATION IN SOI

SOI-CMOS-structures eliminate all inter-device leakage-currents by building *n*- and *p*-type islands on an insulating substrate for *p*- and *n*-channel transistors, respectively. Field-inversion and device-to-device-punchthrough do not exist in SOI-technology, but new parasitic transistors in an SOI-structure can still give rise to leakage-currents. SOI-CMOS-technology depends on high-quality isolation. The methods used to laterally isolate SOI structures will be described here, as well as a description of the leakage paths that arise as a result of each of them. The three techniques for providing lateral isolation in SOI are: **1**) MESA isolation; **2**) LOCOS isolation; and **3**) shallow-trench isolation (STI).

11.7.1 MESA Isolation

MESA isolation was the early method used to create islands of silicon on an insulating substrate (Fig. 11-30). The technique is attractive because of its simplicity. That is, in MESA isolation the active-device-regions (the silicon islands) are formed into a "mesa" by using a mask-step and a silicon-etch step to remove the silicon in the field-areas down to the insulating-substrate. The SOI oxide helps as the etch-stop-layer, while anisotropic-etching allows for minimization of the distance between adjacent mesas. Passivation of the islands (mesas) is performed at the gate-oxidation step, where the gate-dielectric is grown not only on top of the silicon islands, but on their edges as well.

Although the MESA isolation method is simple, it also has some inherent problems. The first involves the etching of the mesas and the thermal-growth of the gate-oxide that is subsequently grown over the entire mesa. That is, the etching of the mesas produces silicon islands with sharp upper-corners. When the gate oxide is grown on these islands, the oxide at the top corner is thinner than on the flat areas of the mesa (typically 30-50% thinner than on the top surface - see

Vols. 1 and 2 for a discussion of oxide-thinning at corners during thermal-oxidation processes.) When the gate-material is run over the mesa and voltage is applied to turn-on the MOSFETs built in the mesa, the thin-oxide at the corner can cause premature gate-oxide-breakdown. The gate-oxide-breakdown-voltage may also be reduced on the sidewalls, since the crystal orientation may be different from that of the top surface of the silicon wafer. The growth-rate of SiO_2 is dependent on the silicon-crystal-orientation of the surface on which it is being grown. Furthermore, crystal-orientations other than (100)-Si generally have larger fixed-oxide-charge-densities and surface-state-charge-densities than (100)-Si, and this can decrease the threshold-voltage of MOSFETs built on such mesas.

The second problem relates to the steep sidewalls produced by an anisotropic-etch of the silicon-mesas. When a polysilicon-layer (used to form the MOSFET gate) is deposited on the mesa (typically to a thickness of 400-nm), an anisotropic-etching of the poly will leave stringers along the mesas, and this may short the gates of devices within the same mesa. Overetching is therefore needed to remove such stringers, but this can lead to poor dimensional-control of the poly gate-lengths.

The third problem involves early inversion along the top of the mesa sidewalls, which behave like lateral (edge) transistors in parallel with the top (main) device. This is due to charge-sharing of the depletion-regions caused by the poly running over the top and sides of the mesa-sidewalls. Thus, the corners show a lower threshold-voltage than the main channel of the MOSFET. This manifests itself as a kink in the subthreshold-characteristic of the drain-current. Such a deterioration of the subthreshold behavior is most pronounced in NMOSFETs in SOI-mesas, and can be alleviated by performing a boron-implant into the field-regions of the silicon wafer before etching the mesas. This boron-implant is then driven-in laterally into the active-area. The mesas are then etched, leaving the sidewalls more highly doped than the active-channel. Figure 11-31 shows a perspective drawing of a laterally-isolated MESA SOI-NMOSFET with a T-gate and a body-substrate contact.

11.7.2 LOCOS-Isolation for SOI

LOCOS-isolation is the second technique that has been used to produce lateral-isolation among active-silicon-islands in SOI. LOCOS is more complicated than MESA isolation, but it has the advantage of already being in wide use in bulk-CMOS technology. LOCOS-isolation in SOI is much the same as LOCOS in bulk-CMOS (see Vols. 1, 2, or 3 of *Silicon Processing for the VLSI Era* for more details on the LOCOS process.) However, there are some issues that do arise

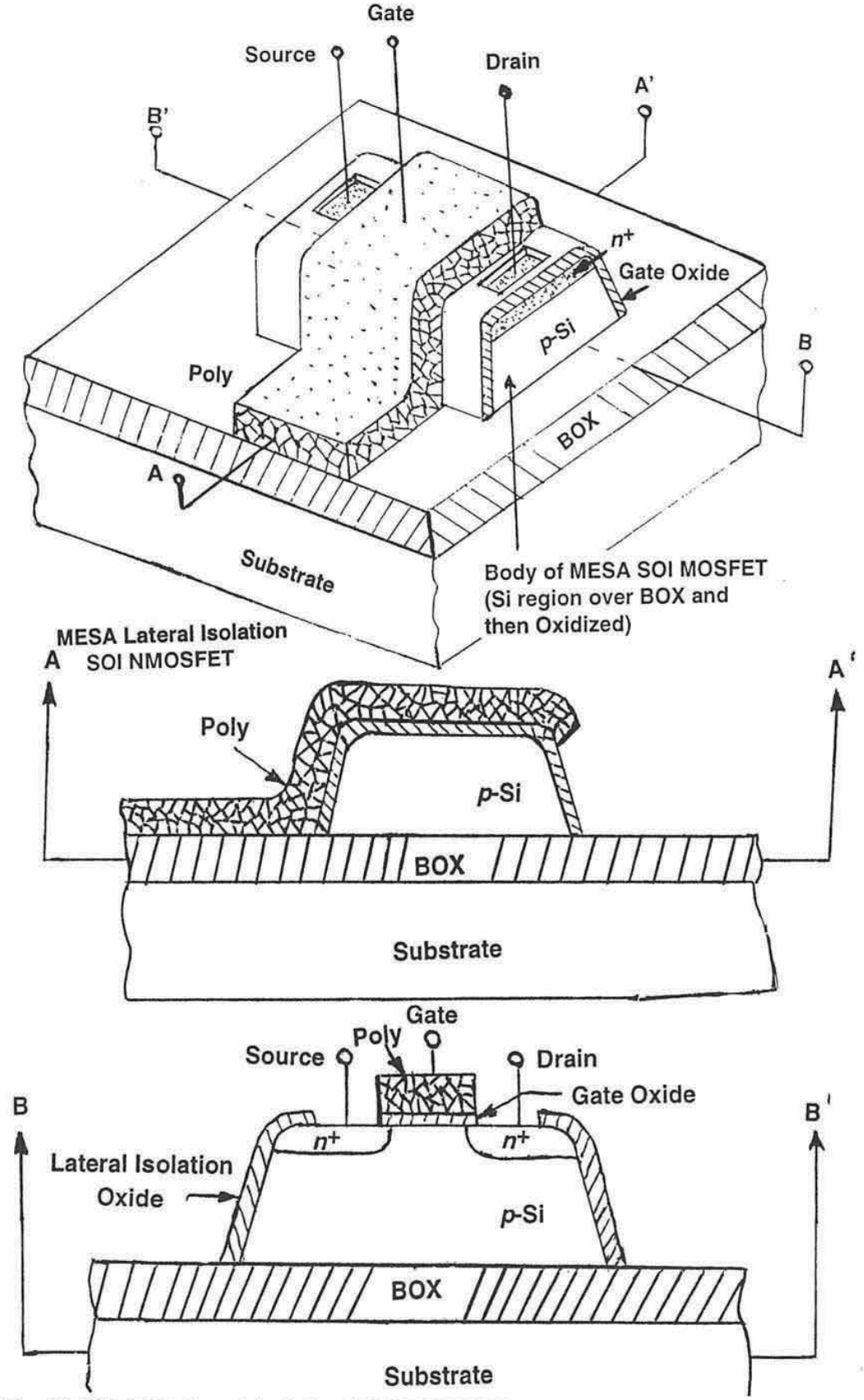
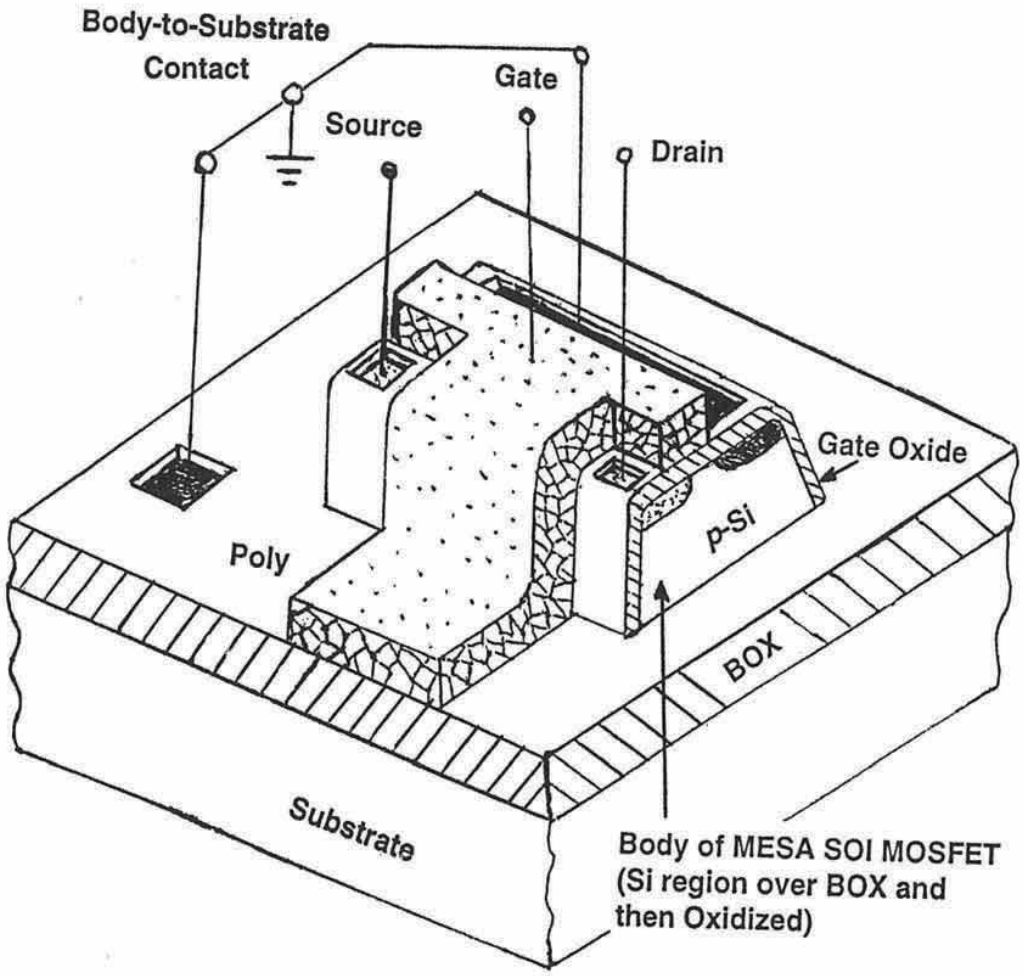


Fig. 11-30 MESA lateral isolation SOI NMOSFET.



11-31 Mesa lateral-isolation SOI NMOSFET with a T-gate and body-to-substrate contact.

when LOCOS is applied as the technique for producing lateral-isolation in SOI. These issues are discussed in this section.

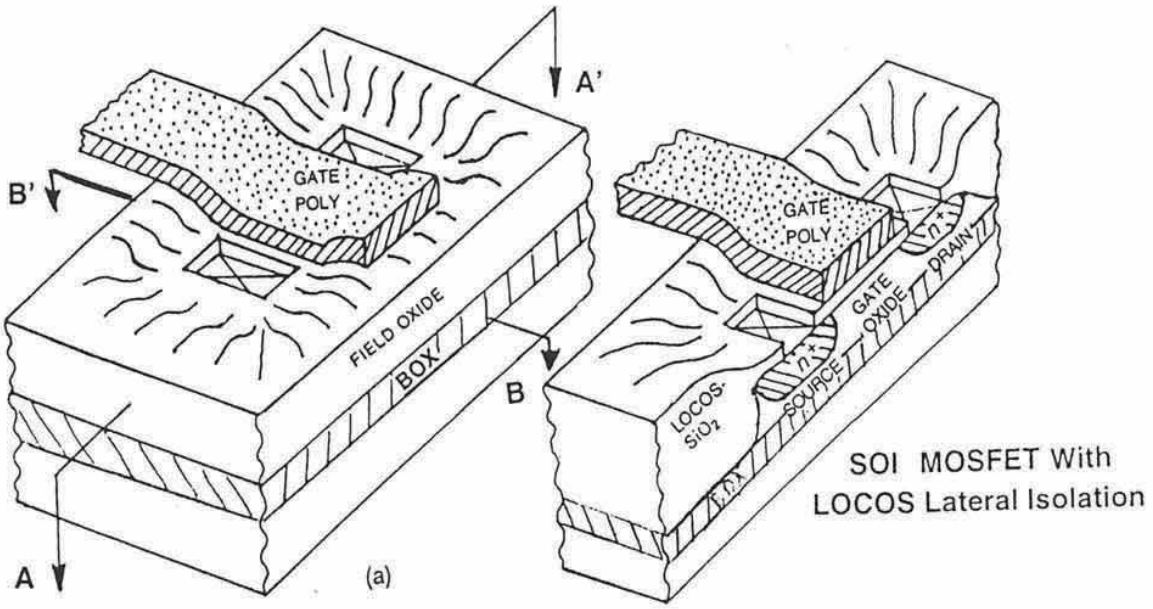
The LOCOS lateral-isolation-technique in SOI consists of the following steps (for producing NMOSFETs): a thin pad-oxide is grown, and a silicon-nitride-layer is then deposited by LPCVD onto the pad-oxide. Using a masking step, the nitride is patterned to define the active-silicon-areas. Boron is then implanted around the islands of *p*-type silicon that will contain the NMOSFETs (i.e., a channel-stop-implant). Next, the field-oxide is thermally grown. No oxide growth occurs over the active-areas which are protected by the silicon-nitride. A nitride and pad-oxide stripping step completes the process. In SOI the field-oxide-growth is carried out so that the thermally-grown field-oxide completely

consumes the silicon in the regions not covered by the nitride, with the thermally-grown field-oxide reaching through to the buried oxide (see Fig. 11-32).

The lateral-encroachment of the field-oxide (called the "bird's beak") is proportional to the thickness of the grown field-oxide. This encroachment has typical values of a fraction of a micron, and this hampers the use of the LOCOS technique for submicron bulk-CMOS processes (where field-oxide-thicknesses of 0.3-05 μm are grown). The problem of "bird's-beak"-encroachment in SOI-CMOS is even worse than in bulk-CMOS processes. This is because the oxidation-kinetics in SOI are somewhat different, particularly when the growing field-oxide reaches the buried-oxide. That is, consumption of silicon during oxide-growth on a bulk-silicon-wafer is 0.45-t_{ox}, where t_{ox} is the oxide-thickness. That means that a thermal-field-oxidation process that produces a silicon-dioxide-layer about twice as thick as the consumed silicon-layer must be used in a bulk-CMOS. For example, a silicon-layer of about 125-nm is consumed when a 250-nm SiO_2 field-oxide-layer is grown.

By comparison, a 125-nm silicon-layer above the buried-oxide of an SOI-wafer is not fully oxidized if only such a 250-nm silicon-oxide growth-process is used. That is, unoxidized-filaments of silicon remain in SiO_2-layers grown by the 250-nm SiO_2-process on an SOI-wafer with a 125-nm silicon-layer above the buried-oxide. To ensure that such silicon-filaments no longer remain in the silicon-dioxide-layer, an oxidation-process equivalent to one that would grow a 350-nm-thick SiO_2-field-oxide on a bulk-wafer must be used. In other words, a process that grows an oxide 3-times-as-thick as the silicon-layer consumed has to be used in SOI to completely laterally-isolate the active-regions with a LOCOS process. The longer field-oxidation-process produces even more "bird's-beak" encroachment (and more lateral-boron-diffusion from the channel-stop-implant into the active-region) than a LOCOS-process in a bulk-CMOS process.

Another problem of the LOCOS method of lateral-isolation is that source-to-drain leakage-current in the subthreshold-region of operation can also be increased by the parasitic-edge-transistor that exists along the sidewalls of the silicon-active-regions in parallel with the main-channel of the SOI-MOSFETs. The problem is most severe in LOCOS-isolated SOI-NMOSFETs for the following reason: A p^+ channel-stop-implant is carried out before the field-oxidation to prevent such edge-transistors from being turned-on prematurely and causing excess subthreshold-leakage-current. However, during the field-oxidation step, such implanted-boron segregates into the growing-oxide on the top of the silicon-surface - and into the buried-oxide as well. As a result, the boron doping at the tip of the bottom corners of the active-device islands ends up being



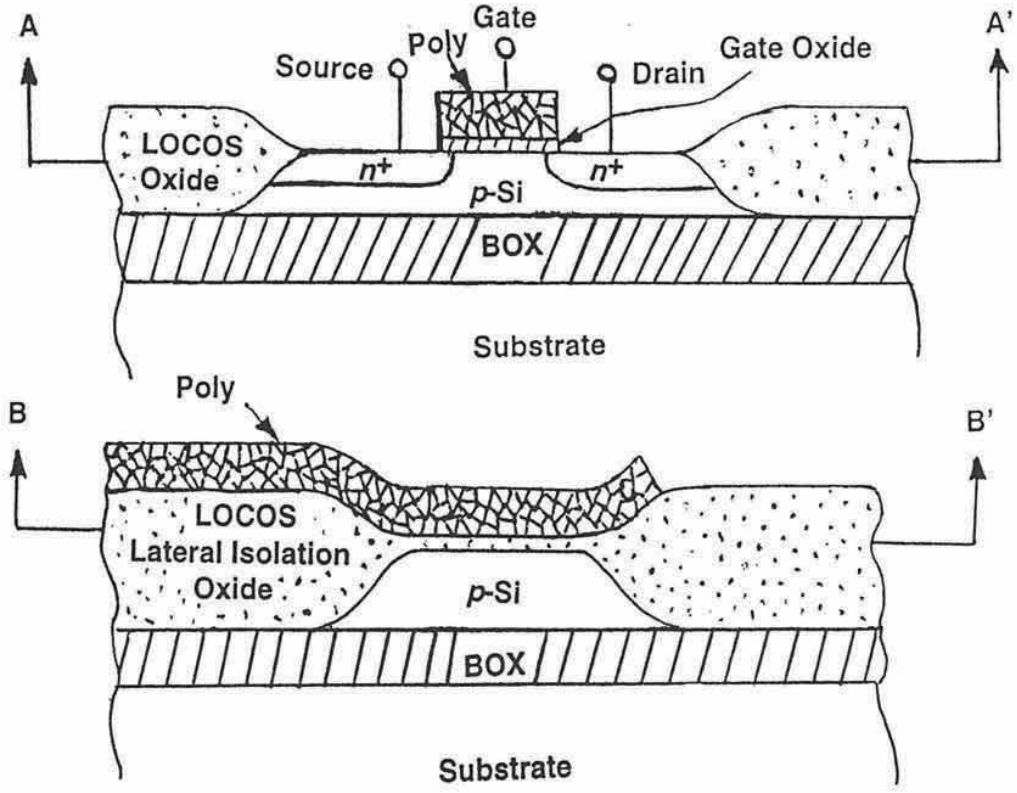


Fig. 11-32 SOI-MOSFET that uses LOCOS lateral isolation.

quite low, allowing edge-transistor-leakage to arise. One method suggested to alleviate this problem is to perform a second boron channel-stop-implant after the LOCOS field-oxide-step has been completed. Since the implant step is carried out after the field-oxide step, such B will be implanted into both active-regions and field-regions of the wafer. The energy of the implant can be adjusted to place the peak of the B-implant-dose close to the meeting point of the grown- and buried-oxides. Thus, this implant step will not impact the doping-concentration of the main MOSFET channels near the surface of the active-regions.

11.7.3 Shallow-Trench Isolation (STI) for SOI

Because of the problems associated with the MESA and LOCOS isolation techniques for providing lateral isolation in SOI technology, the more advanced SOI technologies have adopted shallow-trench-isolation as the lateral-isolation method of preference (Figs. 11-1 and 11-33). That is, the active-regions of the silicon are formed as in the MESA isolation method, but in STI the regions between the mesas are filled with a CVD-dielectric-material and then the surface is planarized by CMP to yield a flat wafer surface. Oxide remains in the trenches between the mesas, and its top surface lies in the same plane as that of the silicon-islands. STI eliminates the problems created by the sidewalls of the mesas, and the issues associated with the LOCOS-isolation method for SOI. (That is, the problem of "bird's-beak" encroachment does not occur when STI is used. Also, anomalies in the oxidation kinetics associated with field-oxide growth in SOI do not have to be addressed since the field-regions between the active devices are filled using CVD rather than a thermal-oxidation process). However, the same general issues involved with shallow-trench-isolation methods in bulk-CMOS still have to be addressed when applying an STI approach to SOI. Chapter 9 of this text provides a detailed discussion of STI technology.

11.8 SOI PRODUCT EVOLUTION

Table 11-4 presents a list of the SOI circuits that have been introduced during the International Electron Devices Meeting (IEDM) from 1989 to 1998.[64] It provides an indicator of SOI R&D efforts. The initial publication of the gain in the speed of a relatively small IC fabricated in SOI has been followed by a series of studies of performance-enhancements of larger ICs, mostly using SIMOX wafers. In general, these ICs were fabricated using partially-depleted SOI-MOSFET technologies, and silicon- and BOX-thicknesses were all comparable (and fairly thick). In 1998, IBM announced it would apply its leading-edge SOI-CMOS technology (including copper interconnects), to commercialize SOI-CMOS prod-

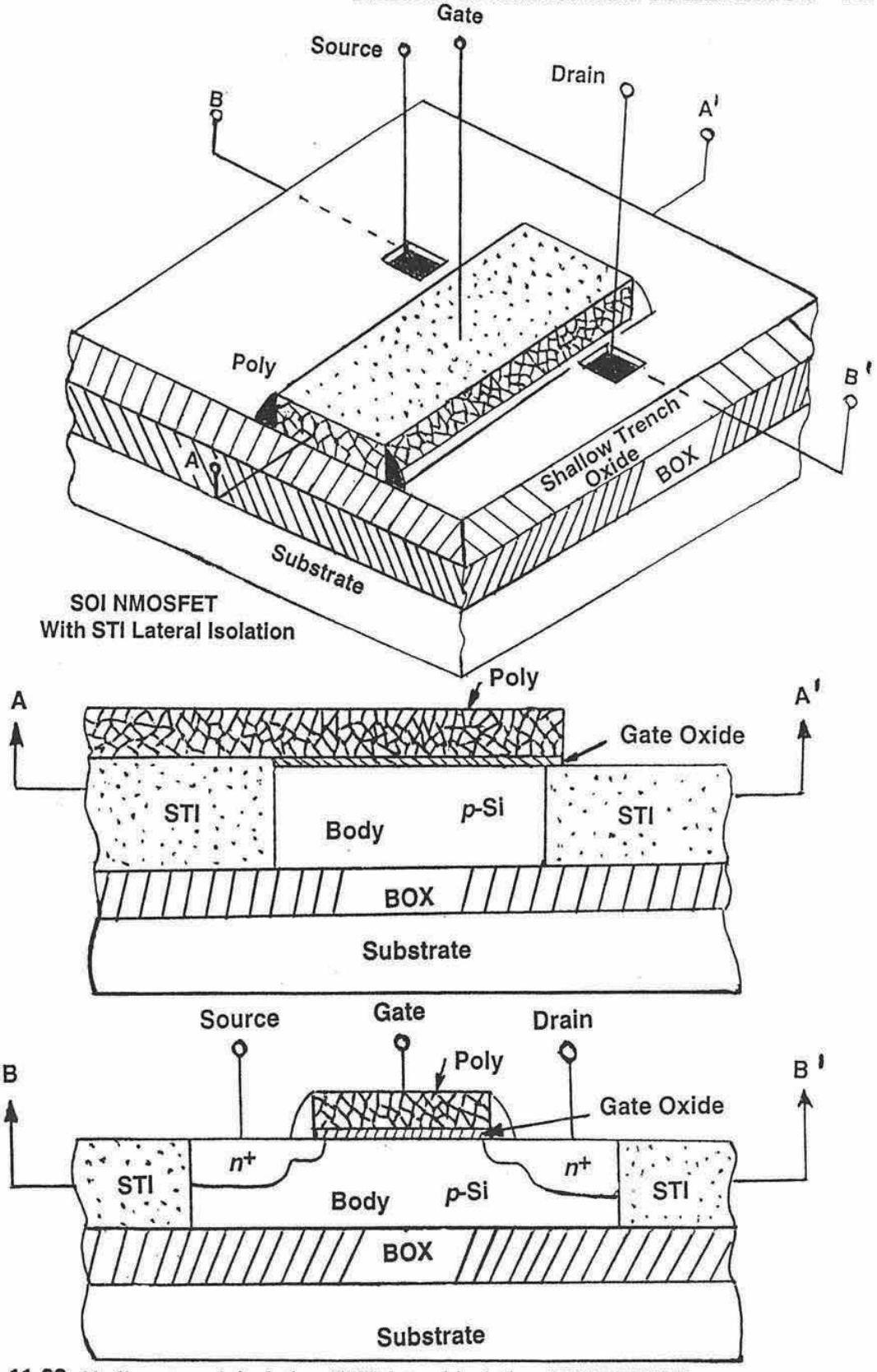


Fig. 11-33 Shallow-trench isolation (STI) lateral isolation SOI MOSFET.

Table 11-4 Summary of SOI Circuits Presented Annually at IEDM[64]

	Bell Labs	IBM	Motorola	Mitsubishi	Hyundai	DEC	IBM	Intel
	1989	1993	1995	1996	1997	1997	1997	1997
Circuit	Prescaler	0.5Mb SRAM	CPU	16M DRAM	1G DRAM	1M SRAM	4 Mb SRAM	4 Mb SRAM
Technology (μm)	0.50	0.50	0.50	0.50	0.17	0.35	0.25	Sub-0.25
V_{DD}		1.0		1.0		2.0	1.8	1.8
Silicon thickness	134		100	100	200	200	200?	150
BOX (nm)	235		400	400	250	200	400	400
Lateral isolation	LOCOS			MESA	LOCOS	STI	STI	STI
Silicon layer characteristics	PD/FD	Non-depleted	PD		Body-cont., thin Si film (75nm) beneath FOX	PD	Non-depleted SOI	PD
Metal levels	1		2			4	6	5
Performance improvement	50%						25–35%	
Power reduction						50%	170–300%	18%
Yield (if known)	Low		Same as bulk			Approaching bulk		Low
Conclusions and claims			Bulk CMOS design, performance 2 × bulk	Both body-tied and floating body MOSFETs	Not clear if circuit is functional!	20% performance improvement	Highest-performance 0.25μm CMOS technology reported. FB effects negligible	No speed advantage over bulk for sub-0.25μm logic applications

ucts based on SIMOX. Figure 11-16 shows a SEM of a 0.18-μm IBM IC structure built using SOI. The company believes SOI processors are 30% faster than bulk-CMOS versions, and that they provide comparable circuit yields and add less than 10% to wafer-processing costs. Other companies have also announced the transition to SOI for high-speed microprocessors or low-power logic circuits. This is a positive development that supports IBM's claim that CMOS could move more towards SOI in the next few years (see Fig. 11-5).

In 1999, the presentation of new SOI devices occurred primarily at the *International Solid-State Circuits Conference* (instead of at the IEDM). Announcements were made concerning the successful fabrication of a number of SOI circuits in this ISSCC 1999 Conference. These circuits included 32- and 64-bit microprocessors, and 16-Mbit DRAMS on SOI. Depending on the technology and circuit layout performance, 10-40% enhancement was reported. A 0.18-μm SOI-CMOS technology for high-performance logic applications, however, was described by IBM in the 1999 IEDM. This SOI technology used a low-dose SIMOX substrate, shallow-trench-isolation, dual gate-oxide-thickness, dual threshold-voltage (V_T), dual-doped-polysilicon, cobalt-silicide, low-k dielectric

in their interconnect scheme, and seven layers of copper metallization. The SOI circuit devices were partially-depleted MOSFETs, and the test chip built using this technology contained a microprocessor-circuit that demonstrated performance above 1-GHz.[58,59] These reports indicate a new-degree of maturity in SOI technology.

The 0.18-μm SOI process of IBM was qualified by that firm for production in the summer of 2000, and work was continuing to qualify a 0.13-μm SOI process in 2001. The 0.13-μm SOI technology would allow production of high-performance devices with 44-million transistors. IBM also said that it had begun shipping its AS400 servers with 0.22-μm SOI devices early in 2000, and that their newer "Turbo" Unix machines would employ 0.18-μm SOI-processors running at well over 800-MHz. The 0.18-μm designs utilize shallow-trench-isolation, dual-work-function-poly, local-interconnect, and borderless tungsten-interlayer interconnect.

At the end of 2001, IBM reported on their 2nd generation 0.13-μm SOI process. The power supply voltage was 1.2 V, and a thinner SOI-layer and a gate-dielectric material with a higher concentration of nitrogen were used. The interconnect structure can have as many as 8 levels of metal. A completely low-k dielectric stack (i.e., with SiLK for the main dielectric material and SiC as the hardmask material and the post-CMP cap-dielectric). The Cu wires are formed using a dual-damascene process.

11.9 PROBLEMS ASSOCIATED WITH SOI

11.9.1 Floating Body Effects in Partially-Depleted SOI MOSFETs

In partially-depleted SOI MOSFETs, the depletion-region under the gate when the MOSFET is operated in inversion does not extend all the way to the buried-oxide interface. Instead, a neutral-region still exists. When the body is grounded (say, with an independent body-contact), a partially-depleted SOI-MOSFET behaves very much like a bulk-MOSFET, and most standard $I_D(V_{GS},V_{DS})$ equations and design concepts still apply. However, if a body-contact (connected by a conducting-line to ground) is not made to the SOI-MOSFET, so called *floating-body-effects* arise, some of which have a detrimental impact on the SOI MOSFET behavior. These are discussed in more detail in the following sections.

11.9.1.1 Kink Effect in Partially Depleted SOI NMOSFETs: The *kink effect* manifests itself as a "kink" in the output characteristics of partially-depleted SOI MOSFETs (Fig. 11-34). The kink appears in the graph of the drain current of

Exhibit 1015 - Page 77 of 120

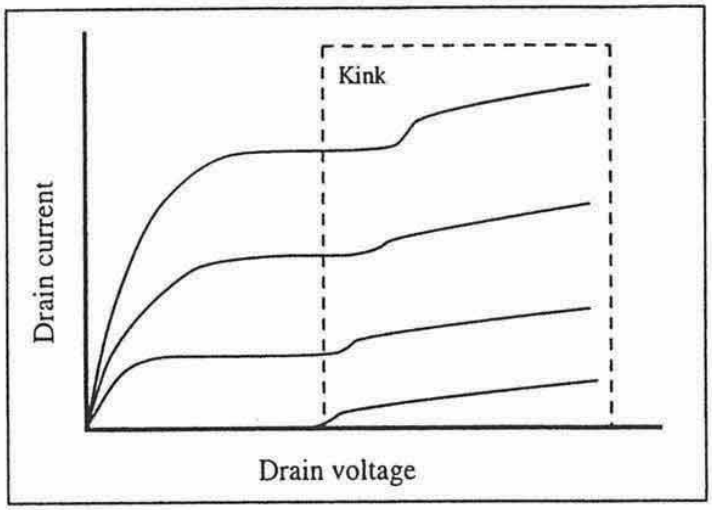


Fig. 11-34 Illustration of the kink-effect in the output characteristics of a partially-depleted SOI-MOSFET.

such devices above some value of the drain voltage. It is often quite pronounced in SOI-NMOSFETs, but is usually absent in SOI-PMOSFETs. The kink-effect is also not observed in any bulk-MOSFETs operated at room temperature if a substrate or well contact exists. It may be observed in bulk-NMOSFETs if they are operated a low temperatures or in NMOSFETs fabricated in a floating *p*-well.

The kink-effect occurs for the following reason. When the drain voltage is large enough in an NMOSFET, the electrons near the drain are accelerated across the channel-drain depletion-region by an electric-field that is strong enough to cause electron-hole pair generation by impact-ionization (in PMOSFETs the coefficient of electron-hole pair generation by energetic-holes is much lower than that of energetic-electrons - which are the carriers in NMOSFETs). Drain voltages typically get this high when the MOSFET is operating in the saturation region (Fig. 11-34).

In a bulk-NMOSFET, electrons generated by this effect rapidly move into the drain, adding to the drain-current. The generated holes exit the device through the body-contact where they become substrate-current. In partially-depleted SOI-NMOSFETs, the generated electrons still move out of the drain and add to the drain-current. The generated holes, however, collect in the NMOSFET body, where they raise the potential of the body, with respect to the source (which is grounded). The amount by which the potential of the body is raised can be calculated, but this requires iterative solving of a set of complex equations, and

will not be described here. However, when the body-potential rises about 700-800 mV, it starts to significantly impact the output-characteristics of the NMOSFET.

That is, an increase in the body-potential has the same effect as biasing the substrate (body-effect). In other words, when the body of an NMOSFET is positively biased, it will give rise to a decrease in the threshold-voltage, V_T, of the NMOSFET (this is termed the *body-effect*). The decrease in V_T causes an increase in the drain-current, and this increase shows up as a "kink" in the drain-current curve of the SOI-NMOSFET. The kink represents an annoyance when analog circuits are designed with SOI NMOSFETs, but can actually lead to the fabrication of faster digital-circuits.

The situation regarding floating-body-effects in fully-depleted SOI-MOSFETs is somewhat different than in partially-depleted devices for several reasons. First, in the fully-depleted SOI-MOSFET, the body of the MOSFET at threshold has the same potential as the surface of the silicon at the top Si/SiO_2 interface. Hence, any holes generated by impact ionization at the drain will migrate out of the source and not collect in the body (which, in a partially-depleted SOI MOSFET, will have a lower potential than the source, and will thus attract the generated holes). Hence, there is no collection of holes in the body region, and the body-effect is also insignificant. In other words, the generation of holes by impact-ionization in fully-depleted SOI-MOSFETs does not decrease V_T, and hence no kink-effect is observed in fully-depleted SOI-MOSFETs, even in the absence of a body contact (i.e., body contacts are thus no longer needed). Second, since the potential of the body-region is higher, the electric-field at the drain will be smaller, reducing the number of electron-hole-pairs generated by impact-ionization.

There is, however, some effect on the subthreshold-slope of both partially-depleted and fully-depleted SOI-MOSFETs that is produced by the electron-hole pairs generated by impact-ionization at the drain. That is, as the MOSFET is being turned-on (but is still in the subthreshold-regime), the generated-holes will collect in the body (which has a lower potential than the source in both partially- and fully-depleted SOI-NMOSFETs if it is being operated in the subthreshold-regime). These injected-holes charge the body and produce a forward-bias on the body-source junction that reduces the threshold-voltage. As a result, the subthreshold-current increases, and the subthreshold-slope of the $I_D(V_{GS})$ curve is not as steep (i.e., the turn-off behavior of the device is degraded).

11.9.1.2 Transient I_D Effects: The floating-body may also produce transient-effects in the drain-current of a partially-depleted SOI-MOSFET. For example, an *overshoot* in the drain-current occurs when the MOSFET is turned-on (e.g., by increasing the gate-voltage from 0 to V_{DD}). What is happening is that majority-carriers are being expelled from the volume of the silicon of the floating-body under the gate that is becoming a depletion-region, and these carriers collect in the remaining neutral-region of the floating-body. This causes the potential of the floating-body to rise, and (as is described in the section on the kink-effect), this causes an increase in the drain- current. However, these excess-majority-carriers recombine and eventually equilibrium is re-established. At that point, the potential of the floating-body is returned to the value it would have without the presence of the excess injected- majority-carriers. Thus, the drain-current gradually subsides with time to a lower value. A reciprocal transient *undershoot* in the drain-current is observed when the gate voltage is switched off (e.g., if the gate-voltage decreases from V_{DD} to 0).

Such short-duration transient spikes in the drain-current can impact such SOI-circuits as SRAMs and DRAMs, which use pass-gate configurations. The amount of charge that is passed by the transient floating-body-effects must be accounted for in the design of such circuits.

11.9.1.3 Effect of Building Partially Depleted SOI-MOSFETs Without Body-Contacts: As discussed in the previous sections, the floating-body of partially-depleted SOI- MOSFETs has an impact on various MOSFET performance characteristics, most of which appear to be undesirable. However, Shahidi made a radical proposal in 1994. He suggested that in many important commercial circuit applications, partially-depleted SOI-MOSFETs could be used without contacts to the floating-body, and that this could be accomplished by making some modifications to the circuit design. Furthermore, by leaving the body floating, not only would the density be increased, but circuit-speed would be improved as well.

Early applications of SOI were limited to space and military programs where radiation-hardness was more important than performance and cost. These products were built with partially-depleted SOI, and fixed potentials to the body were provided by well- and substrate-contacts, just as is done in bulk-CMOS technology. In commercial applications, the considerations of performance and cost become paramount, and radiation-hardness is almost never an issue. Thus, Shahidi suggested that the floating-body-effects would not significantly degrade the performance of many SOI-ICs tailored for specific applications. If that was the case, the floating-body-contacts could be left out. In fact, several bulk-SRAM

circuits and microprocessors were implemented in a partially-depleted SOI-technology that did not employ floating-body-contacts, and these SOI-ICs did indeed exhibit high-performance. In addition, they met all other commercial product specifications (such as stress, reliability, burn-in, and yield). Although some circuit-modifications to the bulk-CMOS circuit design had to be made when implementing them in SOI, these modifications resulted in a higher performance gain than would have been obtained by merely re-mapping the bulk-CMOS design into SOI.[51] The floating-body-effects on the change in drain-current were compensated for in the circuit designs by a number of solutions, and these allowed the SOI-ICs to be successfully designed and fabricated using partially-depleted SOI-MOSFETs without contacts to the floating-body.[52]

Another floating-body issue arises when designing 0.13-μm SOI-CMOS circuits (and beyond), as a result of the very thin gate-oxide-thicknesses needed for these generations of devices. That is, such thin-gate-oxides give rise to a significant amount of gate-to-body tunneling-current. The charge introduced into the floating-body modulates the body-voltage and therefore causes a so-called *history-effect*, which can become excessive on 0.1-μm SOI-CMOS circuits.[60]

This history-effect can be explained as follows: The gate-current in these circuits can change the device "ON" V_{bs}, but does not usually change the "OFF" V_{bs}. Therefore, the gate current will not impact the delay when the input signal switches from "0" to "1" for the first time, but it will impact the delay when the input signal then falls from "1" to "0" (1^{st} switch V_{bs}) - and the impact of this delay will persist when the input signal switches the second time from "0" to "1" (2^{nd} switch V_{bs} – Fig. 11-35). This history-effect must be minimized in 0.1-μm SOI-CMOS by appropriate device-design.

Minimizing the history-effect can accomplished by making two modifications to the SOI-MOSFET device-design: **1**) the silicon-dioxide gate-dielectric is replaced with a dielectric having a higher *k*-value (which reduces the gate-current for an equivalent-oxide-thickness); and **2**) the value of gate-side junction capacitance C_j is adjusted.[60] (That is, C_j controls the drain-capacitance coupling to the body, and hence determines the 2^{nd} switch V_{bs}.) By adjusting C_j at a given level of gate leakage-current, the 1^{st} and 2^{nd} switch delays can be balanced (thereby eliminating the history effect in the switching delay). From a performance point of view, a low-value of C_j is beneficial. C_j can be reduced effectively by scaling the SOI silicon-layer thickness.

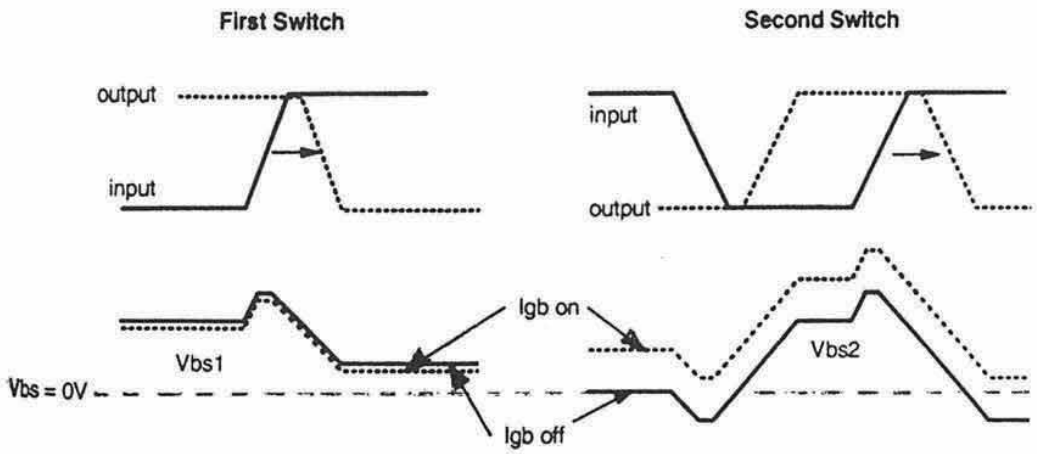


Fig. 11-35 Definition of 1st and 2nd switch rise-to-fall delay for the NMOSFET in an SOI-CMOS inverter. (© IEEE 2000).[60]

11.9.2 Short-Channel-Effects in SOI-MOSFETs

SOI-MOSFETs offer good short-channel behavior, especially for devices with channel lengths smaller than 0.1-μm. Fully-depleted SOI-devices are being studied for low-power/low-voltage applications because they appear to exhibit the best short-channel behavior. However, the silicon-thickness and the buried-oxide-thickness have to be made ultrathin in single-gate fully-depleted SOI-MOSFETs to achieve the best short-channel behavior (i.e., the silicon-thickness must be considerably smaller than the junction-depth of a comparable bulk technology). Even in such ultra-thin SOI-MOSFETs, back-channel-potential control-issues can exacerbate such short-channel-effects as threshold-voltage-rolloff and this issue must be addressed.[54,55]

The problem of threshold-voltage-rolloff arises in deep-submicron single-gated fully-depleted SOI-devices because the depletion-region-charge of the source/drain-regions cannot terminate in the buried-oxide region, and thus must terminate in the substrate below the buried-oxide or in the channel-region (Fig. 11-36).[55] Because the buried-oxide is thick, a portion of the depletion-charge will terminate in the channel. Thus, the source and drain behave like fringing back-gates. If the channel is short, the fringing-fields from these parasitic back-gates can induce back-channel-conduction, which enhances the problem of threshold-voltage-rolloff with decreasing channel-length. This phenomenon has been

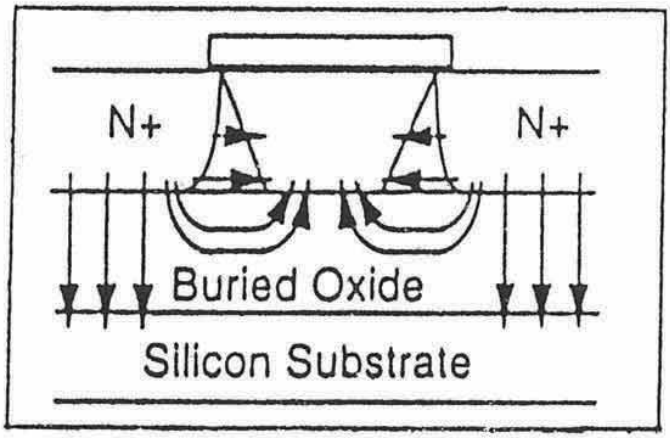


Fig. 11-36 Electric-field lines in the source/drain and body *pn*-junctions in a short-channel SOI-MOSFET. Some of these lines terminate in the substrate beneath the buried-oxide (BOX), while others terminate in the body, producing fringing-electric-fields.[55]

studied by simulation and empirically modeled.[55] Reducing the buried-oxide thickness can also help alleviate the 2-D charge-sharing through the buried-oxide, because the bottom substrate is brought closer to the source and drain regions, provided there is no significant depletion in the substrate. Simulation studies have been presented that predict that single-gate SOI-MOSFETs, even with ultra-thin silicon channels, do not provide adequate immunity against short-channel-effects in 25-nm channel-length devices.[57]

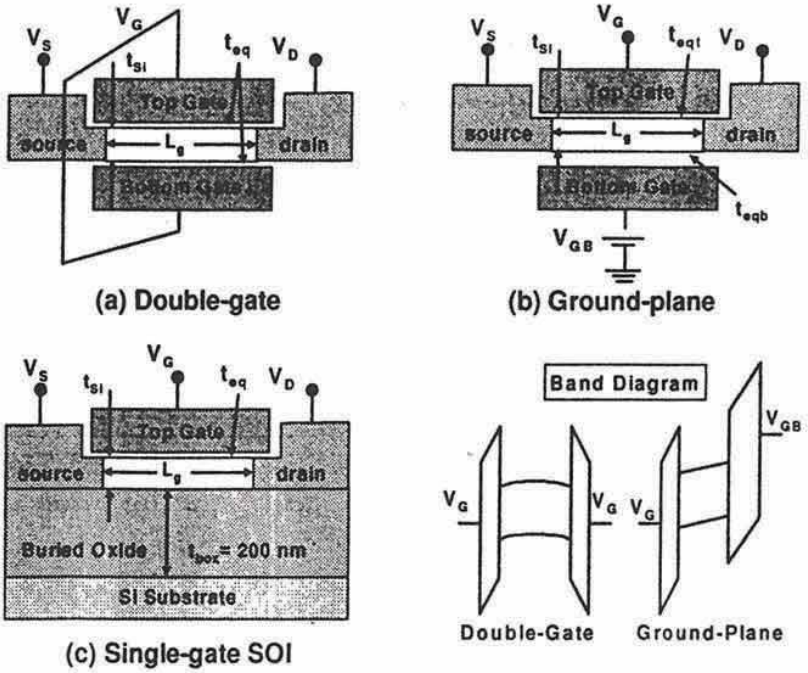


Fig. 11-37 a) Double-gate; b) Ground-plane; c) Single-gated SOI-MOSFET. Gate material work-function is set at the mid-gap of the silicon band gap. On-chip biasing of the ground plane is assumed. The source/drain junction is abrupt. The channel length (L_g) is the metallurgical channel length. (© 1998 IEEE).[57]

As a result, two alternative approaches have been proposed to reduce the threshold-voltage-rolloff problem described above. That is, it has been suggested that either a *double-gate SOI-device*, or a *ground-plane SOI-device*, be used instead of a single-gate SOI-device (Fig. 11-37). Simulation studies show that double-gate SOI-MOSFETs give the best control over the threshold voltage rolloff problem, but the difficulties in fabricating double-gate SOI MOSFETs are daunting. Instead, a compromise structure (namely an SOI device with a low-resistance ground plane) has been proposed.[56,57] In the ground-plane SOI MOSFET, the low resistance ground plane can be biased with a power supply separate from that used to control the gate. This offers the flexibility of tuning the threshold voltage by means of the bottom gate bias.[57] Both double-gate and ground-plane SOI-MOSFETs are shown to able to meet the electrostatic design requirements.

11.9.3 Self-Heating Effects in SOI MOSFETs

SOI MOSFETs are thermally insulated from the substrate by the poorly heat-conducting buried-oxide layer. Consequently, removal of excess heat generated by Joule heating within the transistors is less efficient than in bulk-CMOS technology. As a result, the device-temperature in SOI technologies can rise substantially above what it would be in a comparable bulk technology. Although the heat transfer rate is lower in an SOI device, heat is nevertheless still transported out through a number of paths, namely: **1**) through the buried oxide layer; and **2**) laterally through the silicon-island into the contacts and metallization. As we shall see below, however, such heating effects have not proved to be a significant problem in SOI CMOS technology.

The increased operating temperature within an SOI-MOSFET can have an impact on the device-performance. The most significant effect is that the mobility of the carriers in the channel is reduced. That is, if a MOSFET is being operated in the saturation-region, and the drain voltage is increased, there will be a larger voltage drop at the drain-junction. This will increase the power dissipated in the device and its temperature will rise. Once the device temperature is increased by this effect by 50°C (to 150°C), a significant reduction in carrier-mobility will occur. This effect will manifest itself in the output characteristics of the device in the form of a drop in the drain-current with increasing drain voltage (Fig. 11-38).

However, in CMOS circuits, all of the MOSFETs normally operate in a standby-mode. Large drain-currents in any device occur only during switching, and these cease after a very brief period of time (< 1-ns). Since the thermal time constants of SOI MOSFETs are on the order of several tens of nanoseconds, it

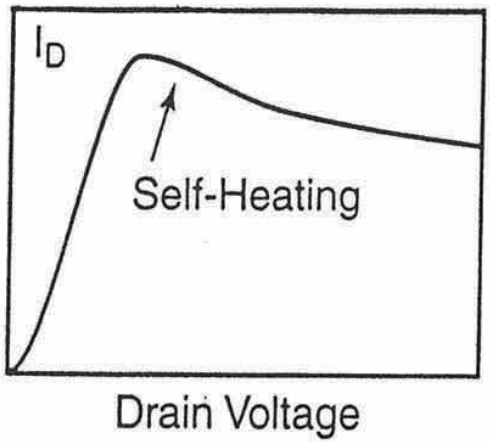


Fig. 11-38 The effect of self-heating in SOI-MOSFETs on output characteristics of the device.

seems that this self- heating effect generally will cause no problems in CMOS digital circuits, with a few exceptions. That is, in logic-circuits that exhibit short transition times, temperature rises of only 2-3 degrees are observed compared to bulk-CMOS. However, in such circuits as off-chip drivers (that can be ON for a significant portion of the switching cycle), the temperature rise is observed to be higher (e.g., ~10 degrees).

11.9.4 Reliability Issues for SOI

There are also some reliability issues unique to SOI-technology that should be addressed when considering SOI for ULSI applications.[61] Such an analysis indicates that SOI-circuits do not exhibit significantly worse reliability behavior than comparable bulk-CMOS-circuits. Some of the reliability issues associated with silicon-on-insulator ICs are reviewed here.

First, bulk- and SOI-CMOS circuits appear to react comparably with respect to *off-state current degradation* when subjected to bias-temperature stress-testing. Hence, this apparently does not represent a special reliability-problem for SOI. Second, channel hot-electron-degradation (CHE) in SOI-devices is more complex than in bulk-MOSFETs because charge-trapping and interface-state-generation can occur at both the gate-oxide-interface and at the buried-oxide-interface. Third, the defects responsible for gate-oxide yield and reliability have been minimized by 2000 in state–of-the-art SOI-materials to acceptable levels. This claim is supported by TDDB tests on gate-oxides on both bulk- and SOI-materials (i.e., the oxides grown on both the bulk- and SOI- materials show

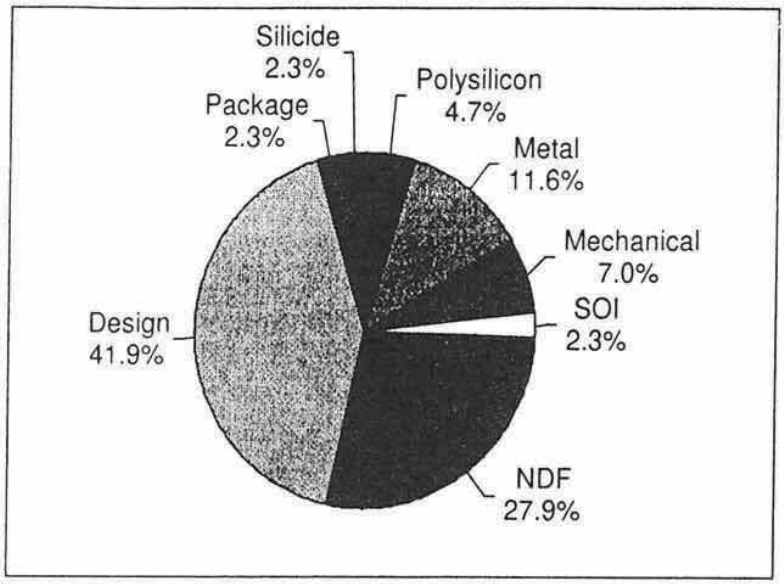


Fig. 11-39 SOI unique defect fails (induced by burn-in stress), are negligible part of the overall failure population.[61] (© 2000 IEEE)

similar failure distributions). Fourth, SOI-devices appear to be less susceptible to charging damage than bulk-CMOS devices. Finally, when failure mechanisms in SOI-microprocessor-circuits are catalogued, the percentage of fails due to SOI-unique defects represent a negligible portion of the failure distribution (Fig. 11-39). In summary, these reports indicate that by using state-of-the-art SOI material, highly reliable products can be produced.

REFERENCES

1. International Technology Roadmap for Semiconductors, 1999 Edition, Semiconductor Industry Association, San Jose, CA.

2. "Proc. 5th Inter. Workshop on Measurement, Characterization, and Modeling of Ultrashallow Doping Profiles in Semiconductors," *J. Vac. Sci. Technol.*, **B-18**, 2000, p. 337-604.

3. G. Shahidi, *et al.*, "Mainstreaming of the SOI Technology," *Proc. IEEE Inter. SOI Conf.*, p. 1-4, 1999.

4. A. Agarwal, "Ultra-Shallow Junction Formation by Spike Annealing in a Lamp-Based or Hot-Walled Rapid Thermal Annealing System: Effect of Ramp-Up Rate," *Materials Sci. in Semiconductor Processing*, 1, p. 237, 1998.

5. K.K. Ng and J. Lynch, "The Impact of Intrinsic Series Resistance on MOSFET Scaling," *IEEE Trans. Electron Dev.*, **ED-34**, p. 503, 1987.

6. M.I. Current *et. al.,* "Process Integration Issues for Doping of Ultrashallow Junctions," *J. Vac. Sci. Technol.* **B-18**, p. 468, 2000.
7. C.-T. Chaung *et al.*, "SOI for Digital CMOS VLSI: Design Considerations and Advances," *Proc. IEEE,* **86**, April 1998, p. 689.
8. S.T. Liu *et al.*, "Radiation Response of SOI Materials," *Mtg. Abs. Electrochem. Soc,*. Spring 1999, Abs. 414.
9. J.-P. Colinge, "Subthreshold Slope of Thin-Film SOI MOSFETs," *IEEE Electron Dev. Letts.,* **EDL-7**, pp. 244, 1986.
10. C.L Chen *et al.*, "High-Frequency Characterization of Sub-0.25-μm Fully-Depleted Silicon-on-Insulator MOSFETs," *IEEE Electron Dev. Letts.,* **EDL-21,** October 2000, p. 497.
11. *ibid.* Ref. 7.
12. *ibid. Ref. 9.*
13. J.-P. Colinge, "Fully-Depleted SOI CMOS for Analog Applications," *IEEE Trans. Electron Dev.*, **45**, May 1998, p. 1010.
14. G.G. Shahidi *et al.*, "A Room-Temperature 0.1-μm CMOS on SOI," *IEEE Trans. Electron Dev.,* **41,** December 1994, p. 2405.
15. N. Kistler *et al.*, "Sub-Quarter Micron CMOS on Ultrathin SOI, *IEEE Electron Dev. Lett.*, **EDL-12**, May 1992, p. 235.
16. M. Chan *et al.*, "Recessed-Channel Structure for Fabricating Ultrathin SOI MOSFET with Low Series Resistance," *IEEE Electron Dev. Lett,* **EDL-14**, January 1994, p. 22.
17. A. Toriumi *et al.*, "Performance and Reliability Concerns of Ultra-Thin SOI and Ultra-Thin Gate Oxide MOSFETs," *Tech. Dig. IEDM,* 1995, p. 847.
18. G.W. Cullen and C.C. Wang, Eds., *Heteroepitaxial Semiconductors for Electronic Devices,* Springer-Verlag, 1978.
19. G.W. Cullen *et al.*, *J. Crystal Growth*, **56**, 287 (1982).
20. M.S. Abrahams and C.J. Buiocchi, "Cross-Sectional Electron Microscopy of Silicon on Sapphire", *Appl. Phys. Lett.*, **27**, 325 (1975).
21. K.E. Violette, 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).
22. A. Gupta and P.K. Vasudev, "Recent Advances in Hetero-Epitaxial Silicon-on-Insulator Technology, Part II," *Solid State Technology,* **26**, 129 (1983).
23. 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).
24. 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).

Exhibit 1015 - Page 87 of 120

25. G.W. Cullen, M.T. Duffy, and R.K. Smeltzer, "Recent Advances in Heteroepitaxial Si on Sapphire Technology," *VLSI Science and Technology/1984*, Electrochemical Society, N.J., 1984, p. 230.
26. D. Munteanu *et al.*, "Silicon-on-Sapphire Material and Device Properties at High and Low Temperature," *Mtg. Abs. Electrochem. Soc.,* Spring 1999, Abs. 435.
27. K. Izumi *et al.*, *Electron Lett.*, **14**, (1978) 593.
28. J. Liu *et al.*, "Formation of Buried Oxide in Silicon Using Separation by Plasma Implantation of Oxygen," *Appl. Phys. Letts.*, **67**, p. 2361, 1995.
29. One supplier of SIMOX wafers is Ibis Technology Corp, Danvers, MA.
30. C.E. Chen *et al.*, *1984 IEDM Tech. Digest*, 701 (1984).
31. S. Nakashima *et al.,* "Investigations on High-Temperature Thermal Oxidation Process at Top and Bottom Interfaces of Top Si of SIMOX Wafers," *J. Electrochem. Soc.,* **143**, p. 244, January 1996.
32. *Wafer Bonding and SOI Technologies*, Abstracts of the Fall 1997 Meeting of the Electrochemical Society, Paris, France.
33. Q.-Y. Tong and U. Gosele, *Semiconductor Wafer Bonding*, VCH-Wiley, New York (1999).
34. J.B. Lasky *et al.,* "SOI by Bonding and Etchback," *Tech. Dig. IEDM*, 1985 p. 684.
35. R.C. Frye *et al*, "A Field-Assisted Bonding Process for Silicon Dielectric Isolation," *J. Electrochem. Soc*, **133**, 1673 (1986).
36. H. Ohashi, "Improved Dielectrically Isolated Device Integration by Silicon-Wafer Direct Bonding Technique," *Tech. Dig. IEDM*, 1986 p. 210.
37. Q.-Y. Tong, W.J. Kim, T.-H. Lee, and U. Gosele, "Low-Vacuum Wafer Bonding," *Electrochemical and Solid State Letter*, **1**, 52 (1998).
38. A. Plossl and G. Krauter, *Mater. Sci. and Engr Reports,* **R25**, 1 (1999).
39. A.J. Auberton-Herve, "SOI: Materials to systems," *Tech. Dig. IEDM* 1996, p. 3.
40. Unibond Wafer Specification Sheet. SOITEC USA Inc. Peabody, MA, (1998).
41. IBM Microelectronics
42. 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).
43. S. Farrens *et al.,* "High-Quality SOI Structures by the Genesis Process," *Mtg. Abs. Electrochem. Soc.,* Spring 1999, Abs 396.
44. M.I. Current *et al.*, "Ultrashallow Junctions or Ultrathin SOI?" *Solid State Technology*, September 2000, p. 66.
45. K. Sakaguchi *et al.,* "ELTRAN by Splitting Porous Si Wafers," *Mtg. Abs. Electrochem. Soc* , Spring 1999, Abs. 395.

46. T. Yonehara *et al.*, "ELTRAN: Epitaxial Layer Transfer," *Mtg. Abs. Electrochem. Soc.*, Spring 1999, Abs. 394.
47. K. Suzuki *et al.*, "Scaling Theory for Double-Gate SOI MOSFETs," *IEEE Trans. Electron Dev.*, **40**, 2326 (1993).
48. C.H. Wann *et al.*, "A Comparative Study of Advanced MOSFET Concepts," *IEEE Trans. Electron Dev.*," **43**, 1742 (1996).
49. M. Tachimori *et al., Proc. Electroche,m. Soc. on SOI Tech & Dev.* Vol. 96-3, p. 53 (1996).
50. A. Matsamura *et al.*, "Recent Progress in Low-dose SIMOX Wafers Fabricated with Internal-Thermal-Oxidation (ITOX) Process," *Mtg. Abs. Electrochem. Soc.*, Spring 1999, Abs. 391.
51. D.H. Allen *et al.*, ISSCC, 1999.
52. F. Assaderaghi and G. Shahidi, "Advanced SOI Circuits for the Microelectronics Market," *Mtg. Abs. Electrochem. Soc.*, Spring 1999, Abs. 383.
53. "Canon Brings Big Company Muscle to the SOI Market With Yet Another Technology," *Solid State Technology*, June 2001, p. 50 (and 167, an Ad which announces the expanded production of ELTRAN wafers).
54. T. Ogawa, "At Long Last, SOI Market is On the Move," *Solid State Technology*, February 2001, p. 62.
55. L. Su *et al.*, "Deep-Submicrometer Channel Design in Silicon-On-Insulator (SOI) MOSFETs," *IEEE Electron Dev. Letts.*, Vol. 15, September 1994, p. 366.
56. T. Ernst and S. Cristoloveanu, "Ground Plane Concept for Reduction of Short-Channel Effects in Fully-Depleted SOI Devices," *Mtg. Abs. Electrochem. Soc.*, Spring 1999, Abs. 431.
57. H.-S. P. Wong, D.J. Frank, and P.M. Solomon, "Device Design Considerations for Double-Gate, Ground-Plane, and Single-Gated Ultra-Thin SOI MOSFETs at the 25 nm Channel Length Generation," *Tech. Dig. IEDM* 1998, p. 407.
58. E. Leobandung, *et al.*, "Scalability of SOI Technology into 0.13 μm 1.2 V CMOS Generation," *Tech. Dig. IEDM* 1998, p. 403.
59. E. Leobandung *et al.*, "High-Performance 0.18 μm SOI CMOS Technology," *Tech. Dig. IEDM* 1999, p. 679.
60. S.K.H. Fung *et al.*, "Controlling Floating-Body Effects for 0.13 μm and 0.1 μm SOI CMOS," *Tech. Dig. IEDM* 2000, p. 231.
61. R. Bolam *et al.*, "Reliability Issues for Silicon-on-Insulator," *Tech. Dig. IEDM* 2000, p. 131.
62. K. Sakaguchi and T. Yonehara, "SOI Wafers Based on Epitaxial Technology," *Solid State Technology,* June 2000, p. 88.
63. M. Current *et al.*, "Atomic-Layer Cleaving and Non-Contact Thinning and Thickening for Fabrication of Laminated Electronic and Photonic Materials," *Proceedings 2001 Materials Research Society Spring Meeting*, Paper 18.3.

64. A. Kamgar and J.T. Clemens, "Isolation's Path to SOI Technology," *Solid State Technology*, October 1999, p. 109.
65. J.W. Sleight *et al.,* "A High Performance 0.13 μm SOI CMOS Technology with a 70 nm Silicon film and with a Second Generation Low-*k* Cu BEOL," *Tech. Dig. IEDM*, 2001, p. 621.
66. R. Chau *et al.,* "A 50-nm Depleted Substrate CMOS Transistor (DST)," *Tech. Dig. IEDM,* 2001, p. 621.

Chapter 12

MULTILEVEL INTERCONNECTS FOR ULSI

The semiconductor industry continues to pursue the development of ever more powerful ICs. This effort basically involves a two-pronged approach. The first approach seeks to increase the speed of individual transistors through a continual reduction in the minimum size of device features. Such scaling also provides an increase in the device density on the chip. The second approach entails developing increasingly complex interconnect systems, which now use multilayer structures of metal wiring separated by interlayer dielectrics (ILDs). This enables the higher-speed devices to be interconnected (resulting in ICs with enhanced performance and system functionality). Efforts in both of these development arenas are supported by continually more sophisticated device and circuit designs.

The tradeoffs between individual device and interconnect performance depend on the details of the circuit architecture. When the IC industry was in its early stages (i.e., from the mid-1970s to about 1990), circuit design rules were in the 1-5-μm range. During that time, interconnect delays could generally be ignored because they were typically much smaller than the device switching times (which were ~50-Mz). However, by the mid-1990s, microprocessors and other large area ULSI devices with millions of circuit elements and speeds greater than 1-GHz were envisioned. Designers then began to realize that interconnect delays would become comparable to (and could eventually exceed) the intrinsic delays of the transistors. It was generally agreed that beyond the 0.25-μm technology node, the overall circuit performance would be dominated by the efficiency with which devices are connected, rather than by the speed of individual devices (Fig. 12-1).[1]

In addition, by the mid-1990s, two different technological domains had also started to collide – namely circuits designed for purely for radio-frequency (rf) applications and others designed merely for digital applications. That is, rf circuits were typically low-density and high-frequency (GHz) designs used in avionics and wireless communication, whereas digital circuits strove for high-density and covered a wide performance spectrum (1-100 MHz). The merger of performance requirements (or *mixed-signal products*), reflects a growing trend,

Exhibit 1015 - Page 91 of 120

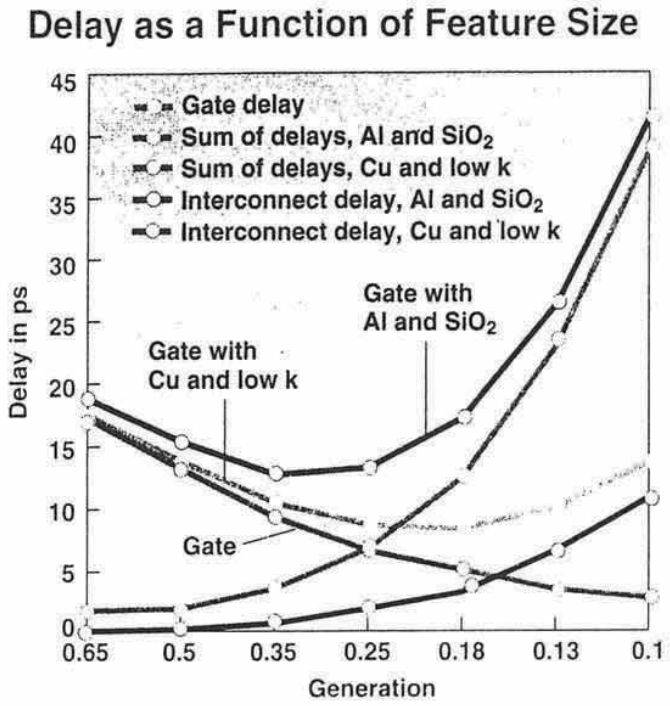


Fig. 12-1 RC-delay as a function of feature size.[1] (© IEEE 1997)

and places additional demands on interconnect systems as performance requirements continue to escalate.

In this chapter, some of the challenges related to the performance and density issues of ULSI interconnects will be discussed. The discussion will be divided into two parts:

1. The performance requirements of ULSI interconnect systems;
2. The evolution of the materials and architecture of interconnects for ICs.

The exact feature size at which the interconnect RC delay equals and then surpasses the intrinsic transistor delay depends on the details of the circuit design. Furthermore, the total overall interconnect delay t_{int} is composed of four components and can be approximated by:[2]

$$t_{int} = R_0 C + 0.4\,[(RC)^{1.6} + (tof)^{1.6}]^{1/1.6} + 0.7\,RC_{in} \qquad \textbf{(12.1)}$$

where R_0 is the output resistance of the driver device, tof is the time of flight of the electromagnetic wave in the interconnect, and C_{in} is the input capacitance of the device in the next gate. These four terms contribute differently to the overall delay, depending on the gate-length and gate-width, the interconnect line-length, and other geometrical parameters. However, most signs point to the feature size of the order of 0.5-μm at which the interconnect RC-delay starts to become a significant part of the overall signal delay.

12.1 EVOLUTION OF THE STRUCTURE AND MATERIALS USED IN IC INTERCONNECTS

Interconnects for ICs have evolved from the single-level-metal system on the original IC patented by Noyce, to the multi-level-metal interconnect systems used in today's ULSI. In this section we will describe the evolution of these interconnect systems. This will point out the limitations that existed for previous generations of IC interconnect technologies, and how they were overcome by learning how to build more advanced interconnect systems. By describing the state-of-the-art of current interconnects, we will also be able to see the limitations they face, and what kinds of techniques are being explored to overcome them.

12.1.1 Single-Level-Metal Interconnects in Early ICs

In 1959 Robert N. Noyce invented the concept of the monolithic-integrated-circuit (IC) while at Fairchild Semiconductor. His patent application described how planar silicon-bipolar-transistors and resistors could be interconnected with thin and narrow aluminum lines adherent to an underlying silicon-dioxide film (SiO_2) that covered the surface of the silicon chip. Contact holes were etched in the oxide film, and these allowed the aluminum lines to make contact to the circuit devices fabricated within the silicon substrate. After the aluminum lines were patterned, a passivation layer (e.g., a CVD P-doped SiO_2 film) would be deposited over the entire wafer. Figure 12-2 shows a schematic of how such a single-level-metal interconnect system was fabricated. The materials used in these early ICs (Al and SiO_2) were selected because they had many unique advantages. These advantages proved to be so compelling that this same combination of materials was used for many generations of IC interconnect

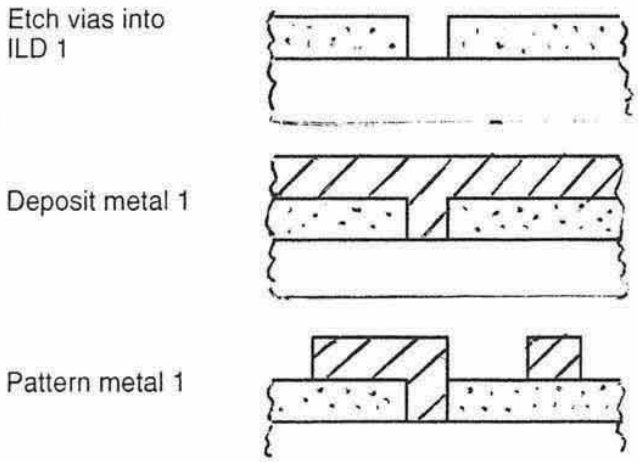


Fig. 12-2 Schematic of process steps used to produce a single-level-metal interconnect.

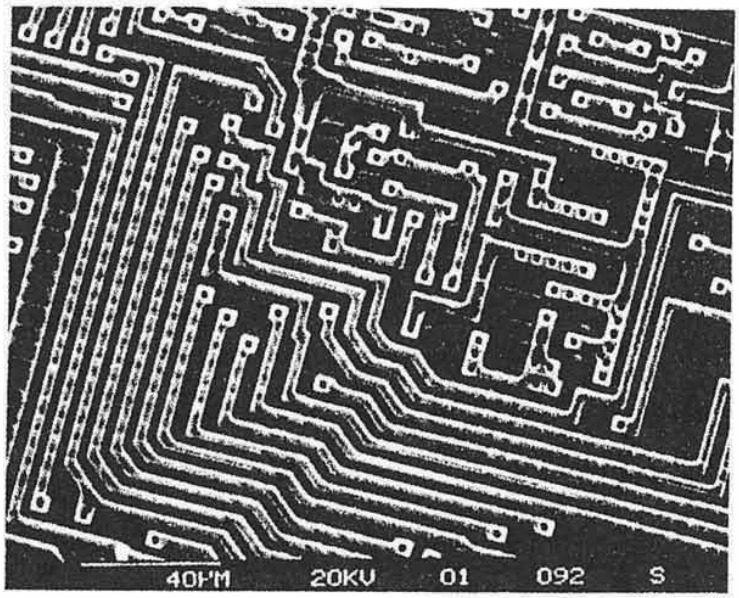


Fig. 12-3 SEM showing the patterned metal layer on a circuit with only one layer of metal.

technology. As a point of reference for materials parameters, the resistivity of pure Al is ρ_{Al} = 2.7-μΩ-cm, and the dielectric constant of SiO_2 is k_{SiO_2} = 4.2 (for the type of SiO_2 used in dielectric films under the aluminum). Later versions of this interconnect system used Al films alloyed with Si to improve contact stability. Figure 12-3 shows an SEM of a circuit using only one-level of metal.

Note that the metal wires made in this traditional way of creating the IC interconnect structure are formed by: **1**) depositing a blanket layer of the metal over the entire wafer; and **2**) selectively removing the metal from regions on the wafer where the metal is not needed (i.e., with a *subtractive* etching process).

12.1.2 Double-Level-Metal Interconnects in Early Bipolar ICs

As the density of ICs increased with decreasing feature size, it became necessary to have more than one level of interconnect. In NMOS ICs, the polysilicon-gate material could serve as a local-interconnect level (eventually covered with a silicide to form a polycide), and this level could be covered with a single-level of metal (Al-Si) to form a two-level interconnect system.

In bipolar technology there is no MOSFET poly-gate structure, and thus to get two-levels of interconnect, it is necessary to use two-levels of metal (Fig. 12-4). Thus, the first two-level-metal interconnect systems for ICs were used in bipolar technology. (Figure 12-5 shows the steps used to fabricate such a 2-level-metal

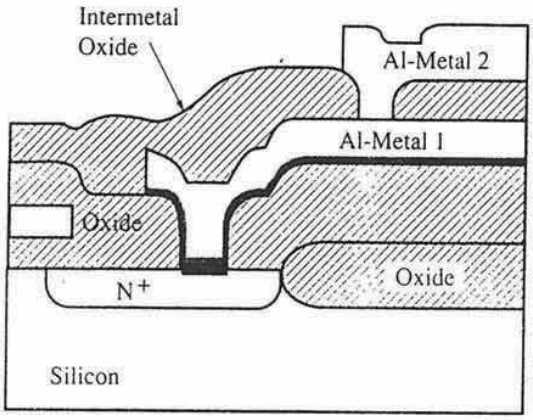


Fig. 12-4 Schematic of a double-level-metal structure used in early bipolar-ICs.

interconnect structure.) In these 2-level-metal bipolar-ICs there was no attempt made to planarize the dielectric layer that separated the 2-levels of metal (i.e., the *interlevel dielectric*, or *ILD*). The holes opened in the ILD to allow Metal-2 to contact Metal-1 were termed *vias*. While the two-level-metal process could be successfully implemented on bipolar-ICs, the step coverage of Metal-2 as it ran over Metal-1 lines was marginally acceptable. Thus, it would not have been possible to continue to use this process to form a three-level-metal interconnect, even in bipolar-ICs.

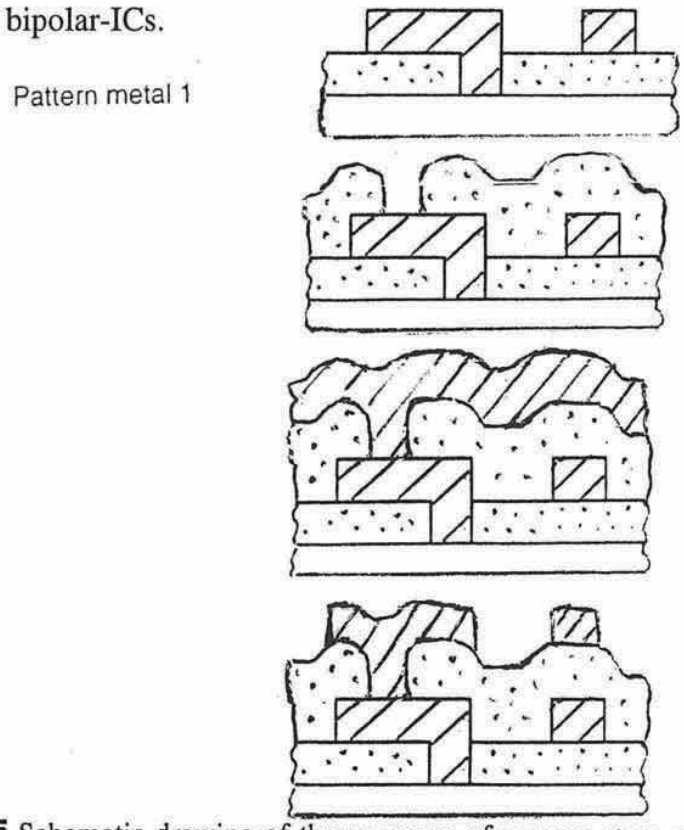


Fig. 12-5 Schematic drawing of the sequence of process steps used to produce a double-

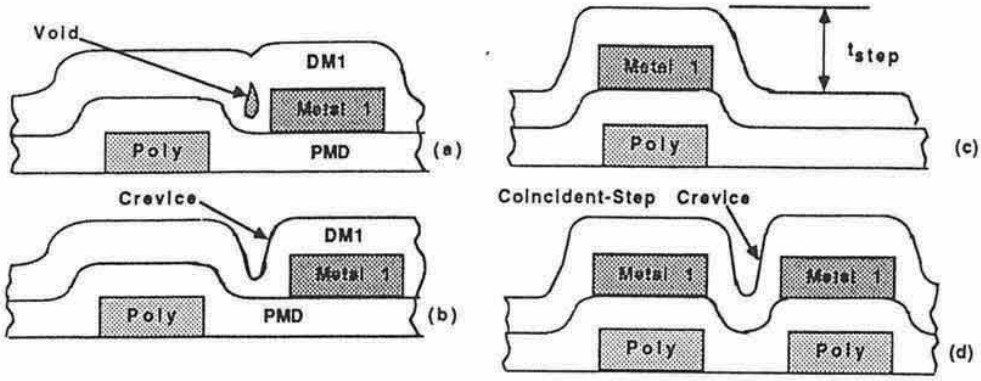


Fig. 12-6 Topography conditions that arise in MOSFETs that make it difficult to fabricate a double-level-metal structure without some kind of planarization method.

Likewise, when CMOS ICs came to replace NMOS ICs (in the mid-1980's), it was no longer possible to use the polysilicon as a local interconnect layer in the same manner that it had been used in NMOS. This was because polysilicon in early CMOS technologies was *n*-doped, and therefore it could not be used to make contact to PMOS source/drain regions. Instead, it also became necessary to use a double-level-metal interconnect structure in these CMOS ICs. Unfortunately, the same approach to fabricating a double-level-metal structure used in the early bipolar ICs could not be used in CMOS ICs. That is, the steps created by the polysilicon layer (together with the steps caused by Metal-1) were too high to allow adequate step coverage when Metal-2 lines had to run over such (poly + Metal-1 hills), as is shown in Fig. 12-6.

12.1.3 Double-Level-Metal Interconnects in CMOS ICs (Partial Planarization of the ILD)

To enable double-level-metal interconnect structures to be fabricated in CMOS ICs, the fabrication sequence used for making double-level-metal interconnects in bipolar ICs had to be modified. That is, it was necessary to be able to partially planarize the ILD prior to depositing Metal-2. The partial-planarization technologies used in the CMOS ICs of the late 1980's included *resist etchback* and the use of *spin-on glass* (SOG). Details concerning these partial-planarization methods can be found in Vol. 2, Chap. 4. Furthermore, the contact holes and vias were opened with a two-step etch process. The first half (which opened the top half of the via) used an *isotropic* etch process, while the second etch step (that opened the remainder of the via) used an *anisotropic* process. This

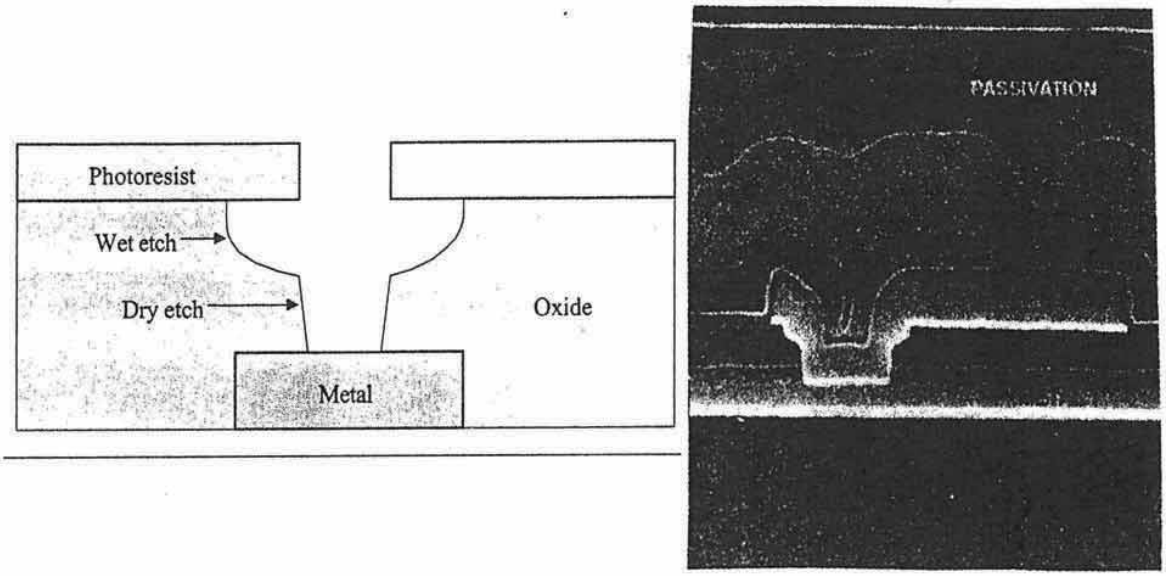


Fig. 12-7 a) Schematic of a via with a "champagne-glass" shape; b) SEM of actual via.

two-step via-opening process produced via profiles that had a "champagne-glass" shape (Fig. 12-7). Such vias were easier to adequately fill with Metal-2. Note that Al-alloy films were typically used (e.g., Al:2%Cu) because they exhibited better electromigration characteristics than did pure-Al films. Figure 12-8 depicts a schematic of a double-level-metal structure for CMOS technology, and Fig. 12-9 shows a SEM of an actual CMOS IC with such a two-level-metal interconnect-structure. Three-level-metal structures in bipolar-ICs could also be fabricated with this approach.

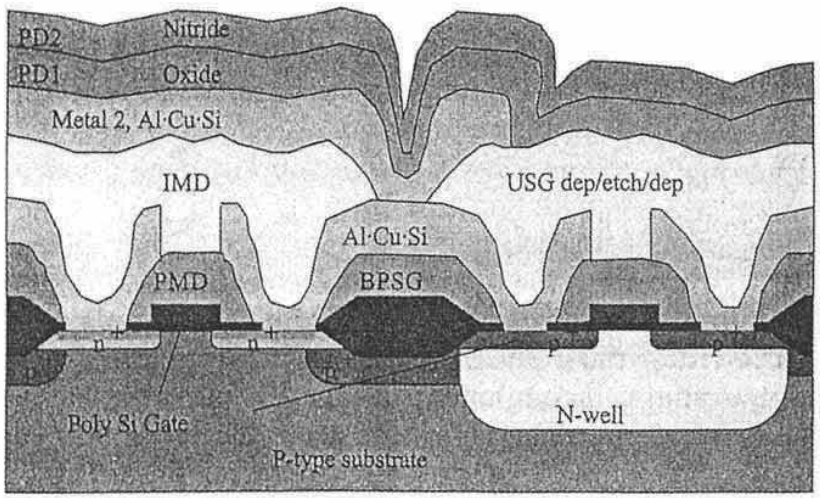


Fig. 12-8 Schematic of a double-level-metal structure in CMOS with partial-planarization.

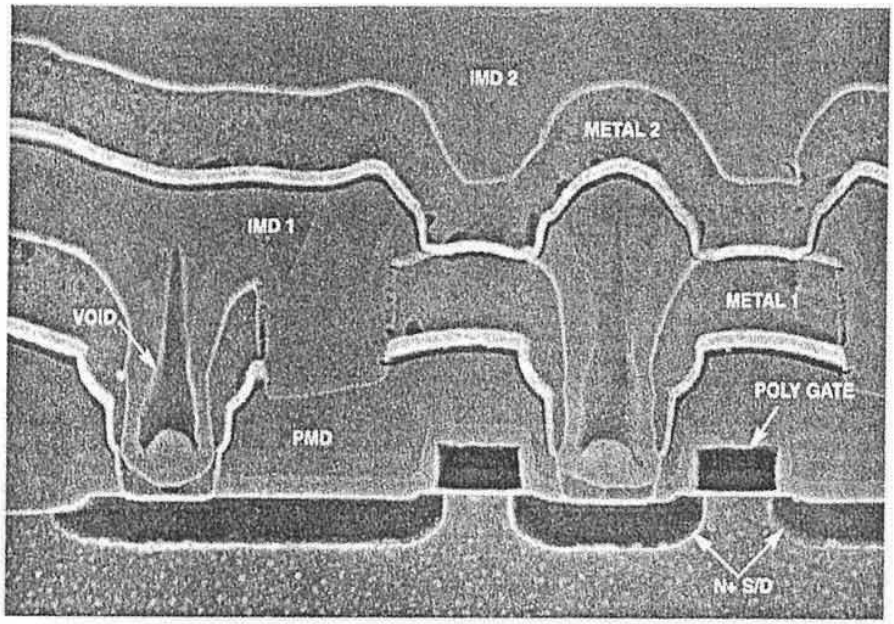


Fig. 12-9 SEM of an actual double-level-metal structure in CMOS with partial planarization.

The partial-planarization of the ILD allowed the step-coverage of Metal-2 to be sufficiently improved, and the "champagne-glass" via profile allowed Metal-2 to adequately fill the vias. Nevertheless, the topography of the IC remained too rough to allow this technology to be extended for producing three-level-metal interconnects in CMOS-ICs.

12.1.4 Three-Level Metal Interconnects in CMOS ICs: (CMP of the ILD & W-Plugs)

To enable three-level-metal interconnects (or more) to be fabricated in CMOS-ICs, a technique for globally-planarizing the ILD after it was deposited was needed. The technology developed for this purpose was *chemical-mechanical-polishing (CMP)* of the CVD-SiO_2 ILD. With this technique, the next level of metal would not have to cover any steps, but could be deposited on a completely planar surface (Fig. 12-10).

While CMP of the ILD solved the problem of metal lines having to cross high steps, it created a new difficulty. That is, as shown in Fig. 12-10, the vias that have to be opened in the planarized ILD are of varying depths (with some being quite deep). Such deep vias could no longer be adequately filled by the traditional PVD process used to deposit metal layers (sputtering), even if a two-step via-etch process was used. To address this issue, a CVD (instead of PVD) process was

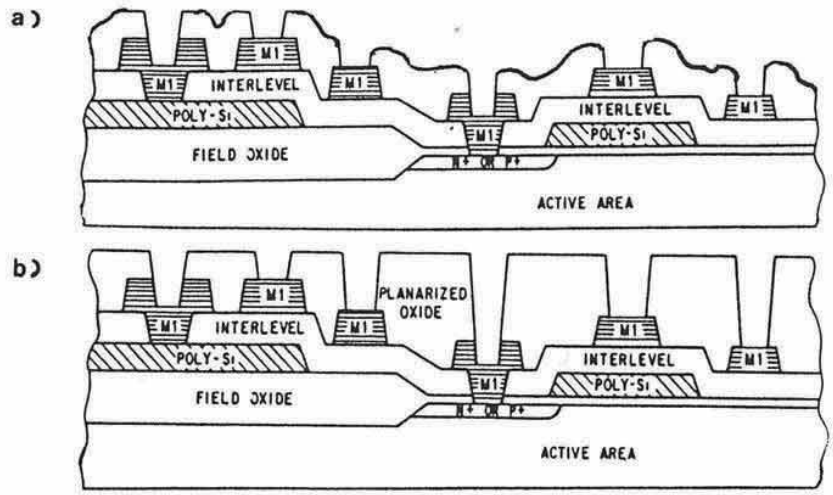


Fig. 12-10 a) If no planarization is used, the via and contact holes between adjacent conductor levels all have approximately the same depth; b) If planarization of the intermetal dielectric layer is achieved, the vias will vary widely in depth.

used to deposit the metal to fill the vias. The metal chosen to fill the vias in this manner was tungsten (CVD-W). A blanket CVD-W process was developed that could deposit a layer of W over the entire wafer covered with a planarized ILD (Fig. 12-11), such that variable-depth-vias across the wafer could also be completely filled. This blanket-W film would then removed from the planar wafer surface (by etching or CMP), leaving W only in the vias (i.e., *W-plugs* were created). The process of filling recesses in a dielectric layer with a metal film and removing the overburden with etchback or CMP (rather than depositing, patterning, and etching the metal on top of a dielectric layer) is called a

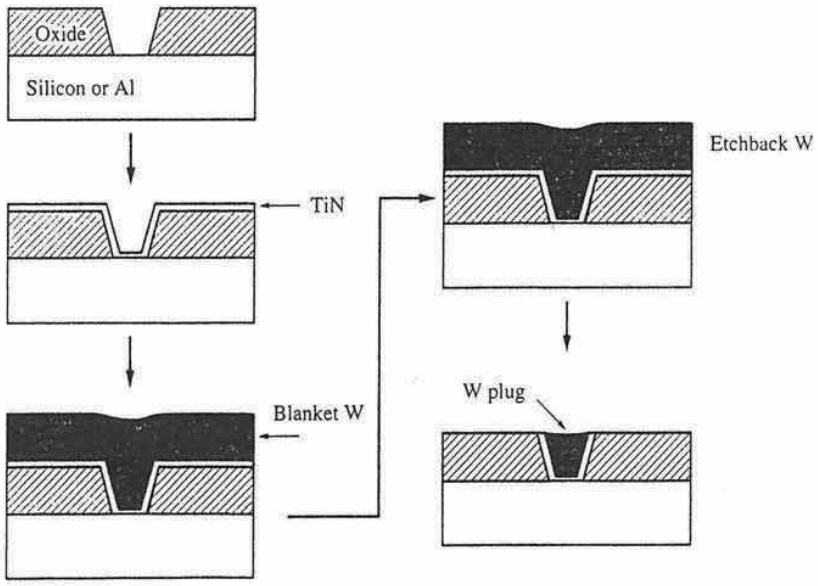


Fig. 12-11 Schematic of the process used to form W-plugs with CVD-W and etchback.

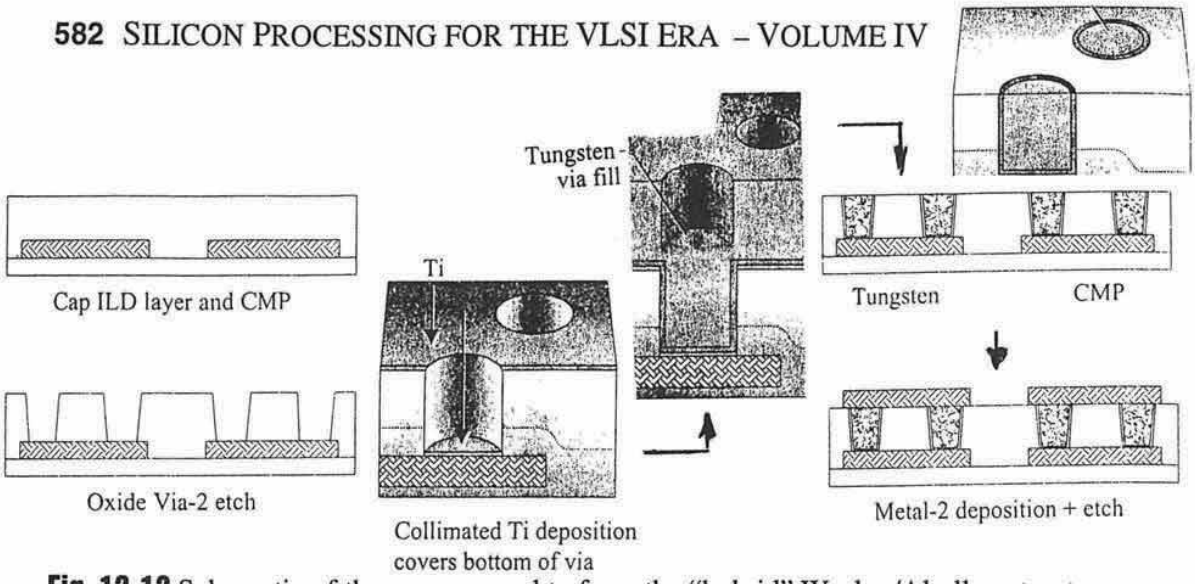


Fig. 12-12 Schematic of the process used to form the "hybrid" W-plug/Al-alloy structure.

"*damascene*" method. Its name comes from the ancient decorative practice of inlaying metal in wood or ceramics that originated in Damascus.

The layer of aluminum alloy (e.g., Al:2%Cu) from which the interconnect lines are patterned is then deposited and etched by the usual subtractive process. Fig. 12-12 shows a schematic of the steps used to form one-level of this type of interconnect structure. Figure 12-13 shows a schematic drawing and a SEM of such an interconnect structure with stacked vias.

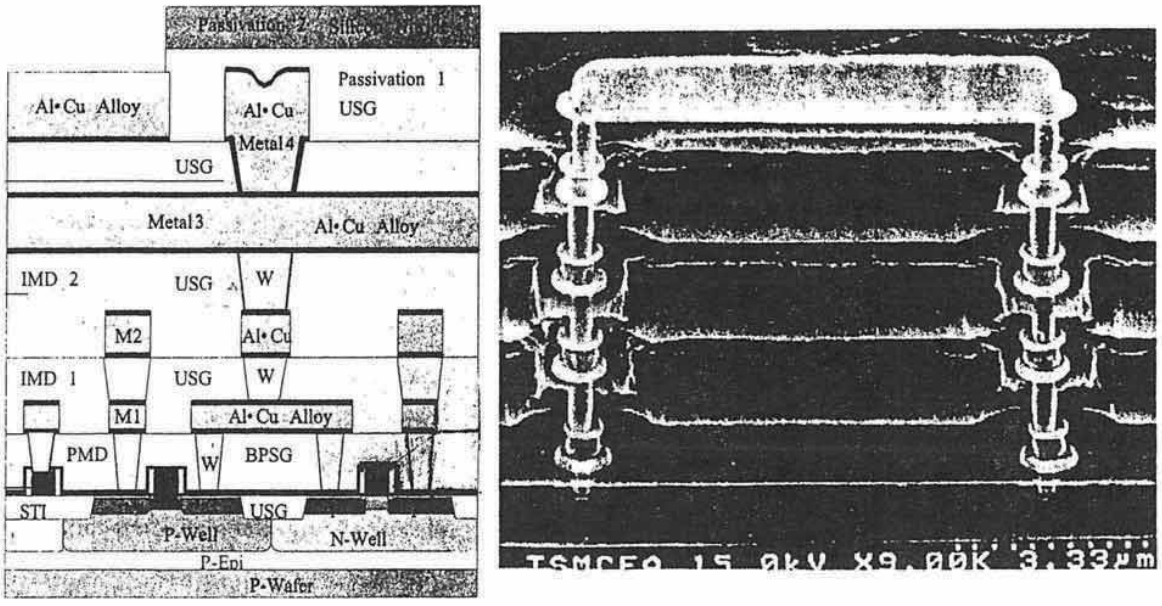


Fig. 12-13 a) Schematic of a W-plug/Al-alloy interconnect structure with stacked vias; b) SEM of an interconnect structure with stacked vias and five-levels of metal.

Note that in several aspects this sequence involves a hybrid process-technology approach. First, two-types of metal are used, one for the plug that fills the via (W), and another to form the metal lines (Al:Cu). Second, a single-damascene process is used to remove unwanted W from atop the entire surface of the chip, while a subtractive etch-process is used to selectively remove Al:Cu from some regions of the chip surface, and thus form the interconnect lines.

The ILD-CMP/W-plug/Al:Cu-line interconnect-fabrication-approach solved some significant problems, allowing it to be used to form interconnect structures with an indefinite number of levels of metal. These solutions are: **1**) the metal lines on every level have no steps to cover; **2**) vias of varying depth can be completely filled using the CVD-W step; and **3**) the vias can be stacked on top of one another if a CMP process is used to planarize the surface of the W-plugs (which allows the density of the interconnect structure to be increased).

12.1.4.1 Problems of W-plug/Al-line Interconnect Structures: While this approach to fabricating multi-level interconnects represented a significant advance, it also introduced some drawbacks. First, it is a complex process, in that it uses two-types of metal, and two-different metal-deposition-processes (each with its own set of process tools). This fabrication complexity is compounded by the fact that two-CMP steps, one ILD-etch-step, and one metal-etch-step are required. Second, the overall resistance of the Al-line/W-plug-structure can grow to be too large for small sized interconnect paths (whereas this resistance value can be reduced by using a lower-resistivity conductor-material - e.g., Cu). Finally, there are some problems associated with W-contacts and vias. That is, electromigration problems in Al often occur in the vicinity W-contacts and vias. This may be partly due to mechanical stresses that can develop at W/Al interfaces (which may enhance electromigration, or promote void and hillock formation). But it is also because the different material in the contact or via serves as a point for flux-divergence. Al also has a lower electrical resistivity than W. This can be detrimental with respect to electromigration effects, especially in circuits where current flow between different materials is significant.

12.1.5 Dual Damascene Cu/Low-*k* Interconnects

An alternative approach to forming interconnects on ULSI ICs is the *dual-damascene approach.* The Cu/low-*k* dual-damascene interconnect scheme has been pursued because it has several advantages over the subtractive W-plug/Al-alloy interconnect approach. First, an all-Cu interconnect structure has a lower resistivity and better electromigration resistance characteristics than the W-

plug/Al-alloy system. Second, by incorporating low-*k* dielectrics, the parasitic capacitances of the interconnect structure are reduced, also improving the RC-delay. Third, only one CMP step is needed (i.e., CMP of the Cu/diffusion barrier film stack), compared to the two-CMP steps needed for the subtractive interconnect (i.e., CMP of the W-plug and the ILD dielectric layer). Fourth, the damascene approach avoids having to etch Cu lines, a difficult process to implement in manufacturing. Fifth, only a single interface exists for the via (i.e., the barrier metal in contact with the Cu film in the metal level below, versus the two interfaces at the top and bottom of the W-plug).

The Cu/low-*k* dual-damascene interconnect also has some difficult challenges, some of which are listed here. First, Cu is a fast-diffusing and device-failure-inducing contaminant, even when present in trace concentrations in the silicon. Therefore, it must be surrounded by a diffusion barrier on all sides to prevent it from entering the silicon substrate. Deposition of diffusion-barrier-materials into high-aspect-ratio trenches and vias with sufficient thickness everywhere becomes harder as the via and trench dimensions are scaled. Second, CMP of Cu (and the underlying diffusion-barrier-metal-layer) is also a complex process. Third, the deposition of Cu-layers by electroplating introduces another new, and potentially contaminant-producing process technology into the wafer fab. Fourth, developing low-*k* materials to replace CVD-SiO_2 as the ILD layer for the dual-damascene structure is not simple, and new materials will need to be found for several generations of low-*k* dielectric applications. These issues are discussed in detail in other chapters of this book, namely: low-*k* dielectrics (Chap. 14); issues of dual-damascene interconnects (Chap. 15); and copper deposition and CMP (Chap. 16).

The following sequence of steps is used to create a dual-damascene interconnect structure (see Fig. 15-3 in Chap. 15):

1. A low-*k* interlevel dielectric (ILD) film is first deposited;

2. Vias and trenches are next patterned and etched in the ILD (two masking and etch steps are needed to accomplish this result);

3. A multi-layer film (i.e., barrier-metal/Cu) is deposited into the vias and trenches of the ILD (the barrier-layer and Cu-seed-layers are deposited by PVD, and the main Cu-layer by electroplating);

4. The barrier-layer/Cu film is removed from regions above the surface of the ILD by CMP, leaving the barrier-layer/Cu recessed in trenches and vias;

5. The entire wafer surface is covered with a non-conductive diffusion-barrier-layer material (e.g., SiN) to complete the encapsulation of the Cu.

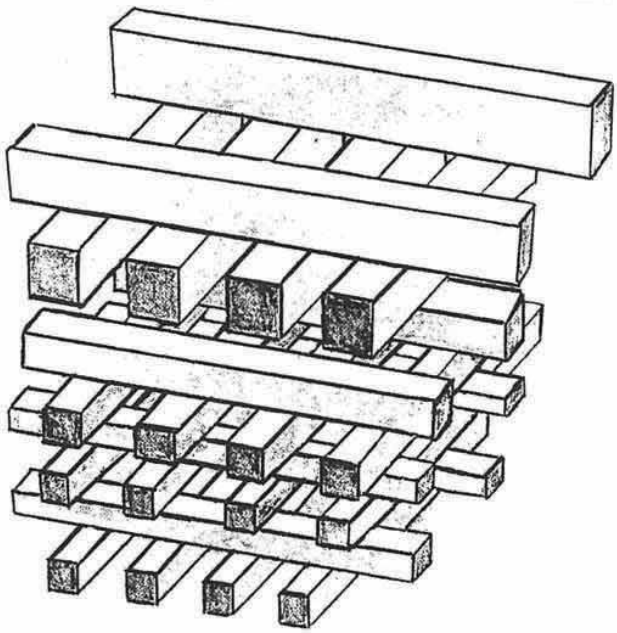


Fig. 12-14 Schematic showing the orthogonal architecture of multilevel interconnect lines.

12.1.6 Other Geometrical Aspects of Multilevel ULSI Interconnects

There are two other geometrical aspects of ULSI multi-level interconnects that should be discussed: **1**) orthogonal directions of alternating metal levels; and **2**) tiers of metal lines containing the same pitch.

In a multi-level-metal interconnect structure, alternating levels are run orthogonal to the levels above and below (Fig. 12-14). This serves three purposes. First, maximum flexibility of routing signals is made possible by using such an orthogonal multilevel structure. Second, the intralevel-capacitance between levels is minimized. Third, crosstalk can be reduced. Figures 12-15 and 12-16 show examples of such an orthogonal structure in a W-plug/Al-line interconnect structure, and in a Cu dual-damascene interconnect structure, respectively. Figure 12-17 shows a cross-sectional SEM of a six-level-metal interconnect structure that reveals the use of orthogonal metal-level layout as well.

Multilevel interconnect structures are also typically arranged in *tiers*. Tiers are pairs of orthogonal metal levels with equal pitch. In some interconnect systems, only two tiers are used (i.e., a *local interconnect-tier* and a *global interconnect-tier,* Fig. 12-18a). In a *local interconnect-tier,* the wires are only used within a cell or a block. In such local tiers the finest metal-pitch allowed by the design rules is employed. In a *global interconnect-tier,* the wires are used for inter-block communication. Thus, the lines may be quite long, and are thicker and wider than those in the local interconnect tier. In interconnect structures that contain six or more levels of metal, there may be a more optimized three-tier network (i.e., one

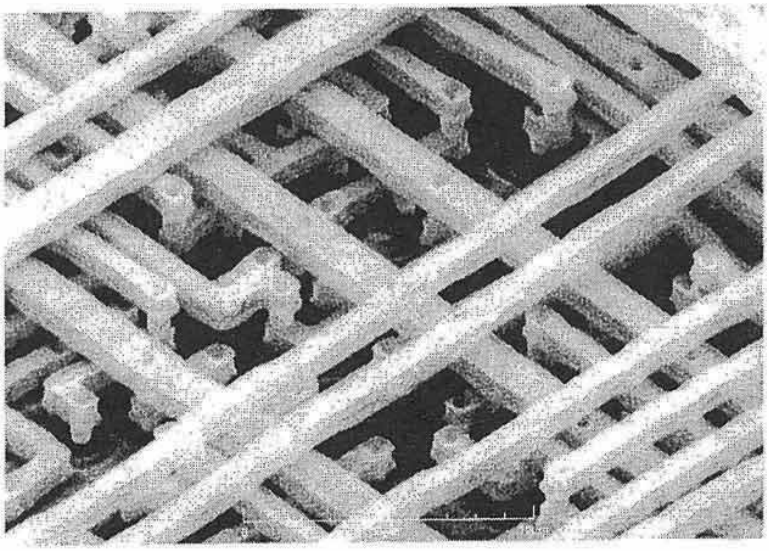


Fig. 12-15 SEM of orthogonal architecture of metal lines in a W-plug/Al interconnect.

that contains a third, *semi-global tier*, as shown in Fig. 12-18b).

12.2 PERFORMANCE REQUIREMENTS OF ULSI INTERCONNECT SYSTEMS

By 2001, microprocessors for consumer PCs were operating at speeds greater than 1-GHz (and even faster in the most advanced workstations). Data in Table

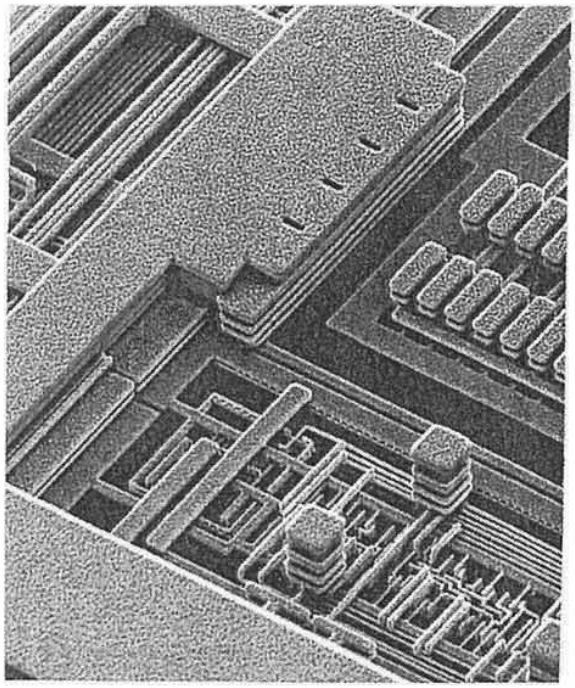


Fig. 12-16 SEM of orthogonal architecture of metal lines in a Cu interconnect.

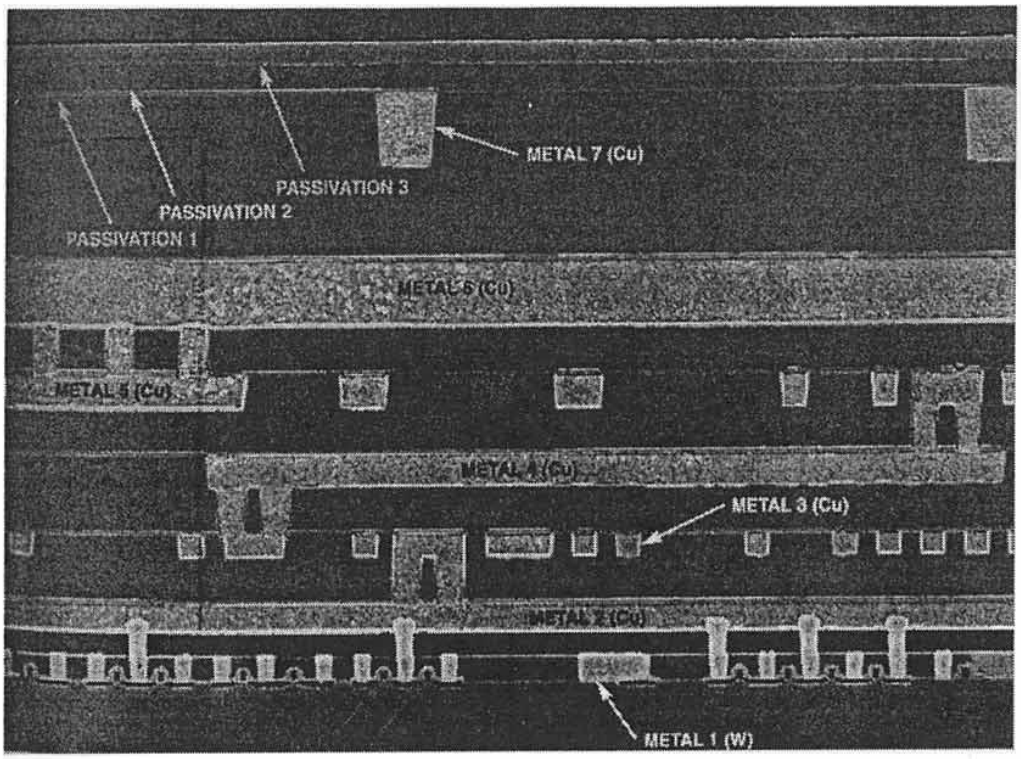


Fig. 12-17 SEM of a cross-section of a 6-metal-level interconnect.

12-1 taken from the 1999 ITRS,[3] summarizes the projected interconnect needs of microprocessors out to 2014, with respect to performance, density, feature size, and other critical technology parameters. Here we discuss those parameters and issues involved in allowing multilevel interconnect structures to achieve the

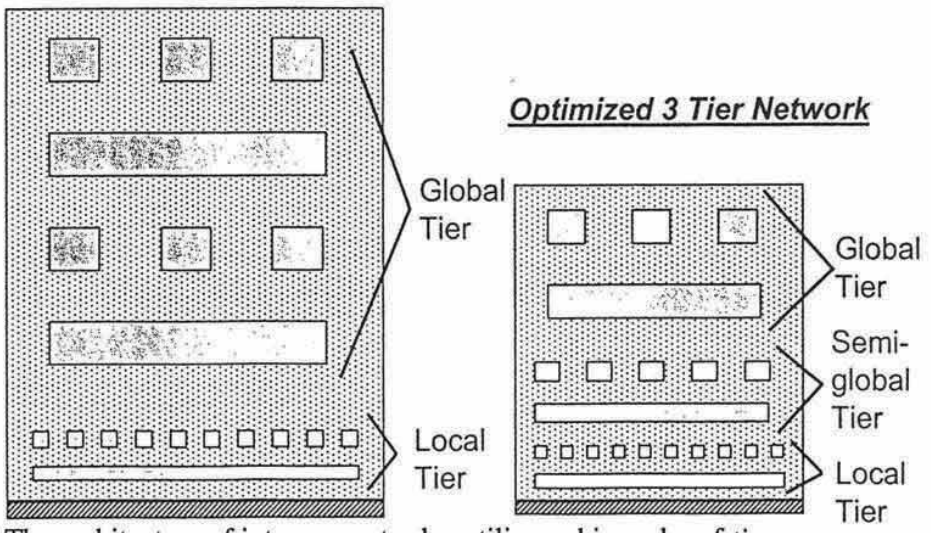


Fig. 12-18 The architecture of interconnects also utilizes a hierarchy of *tiers*.

TABLE 12-1 Selected Overall Technology Characteristics based on the 1999 International Technology Roadmap for Semiconductors.[3]

Year	1999	2002	2005	2008	2011	2014
Technology node	180 nm	130 nm	100 nm	70 nm	50 nm	35 nm
MPU $\frac{1}{2}$ pitch	230	160	115	80	55	40
MPU gate length (nm)	140	85–90	65	45	30–32	20–22
DRAM $\frac{1}{2}$ Pitch (nm)	180	130	100	70	50	35
MPU chip frequency on-chip across chip clock–MHz (Cost-performance)	600	800	1100	1400	1800	2200
MPU chip frequency on-chip local clock–MHz (High performance)	1250	2100	3500	6000	10000	13000
High volume-MPU (Cost-performance) chip size at ramp (mm^2)	170	214	235	269	308	354
High volume-MPU (Cost-performance) Mtransistors/cm^2 at ramp (including SRAM)	7	18	41	100	247	609
Memory chip size at production (mm^2)	132	152	174	199	229	262
Memory chip Gbits/cm^2*	0.20	0.52	1.23	3.05	7.51	18.5
Maximum power w/high performance heat sink (Watts)	90	130	160	170	174	183
Maximum power (Watts) –Battery (hand-held)	1.4	2.4	2.0	2.0	2.2	2.4
Minimum logic Vdd–(Maximum performance) Volts	1.8	1.5	1.2	0.9	0.60	0.60
Minimum logic Vdd–(Lowest power) Volts (battery power)	1.5	1.2	0.90	0.60	0.50	0.30

performance goals listed in Table 12-1. The performance parameters that will be addressed are the following:

1. Propagation-delays (*RC*-delays) in digital circuits;
2. Power dissipation issues in ULSI interconnects;
3. Crosstalk in ULSI interconnects;
4. Contact resistance;
5. ESD protection.

12.2.1 Propagation-Delays in ULSI Multilevel Interconnects

As noted in the introduction of this chapter, for technology nodes beyond 0.25-μm the overall performance of ICs will generally become limited by the performance of the interconnect system on the chip. The most widely used figure of merit of an interconnect system is its $R_{int}C_{int}$-*delay*, which represents a measure of the time-delay for signal propagation. R_{int} denotes the resistance of the interconnect metal-line and C_{int} denotes the effective capacitance between the line and its surroundings. For the multi-level interconnect systems found in ULSI chips at the 0.25-μm technology nodes and beyond, the estimated effect of various geometrical and materials parameters on the $R_{int}C_{int}$ delay can be modeled using the test-structure shown in Fig. 12-19.[11] This test-structure consists of a set of metal lines between two grounded metal planes, all embedded in the interlayer-dielectric medium (ILD). Let us assign the following symbols to those aspects of the interconnect system depicted by this test-structure.

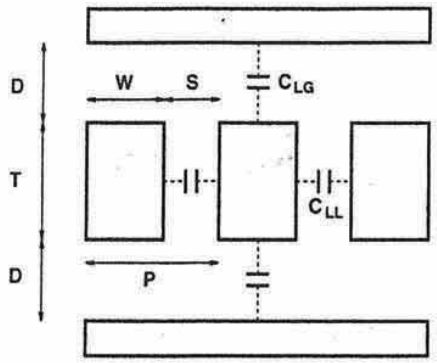


Fig. 12-19 Simplified *RC*-delay test-structure.[11] (© IEEE 1995)

That is, the metal-line *width* is w, the metal-line *thickness* is t, the *spacing* between the lines is s, the metal-line *pitch* is p ($p = w + s$), and the *thickness of the ILD between the top of the metal lines and the grounded planes* is d. Also let the *length* of the metal lines be l, and the *resistivity* of the metal be ρ Some equations can now be derived that allow the contributions made to R_{int} and C_{int} from: **1**) the materials properties; and **2**) the interconnect architecture, to be expressed. In these derivations the pitch p and the metal thickness t will serve serve as the parameters in the equations. In that way it will be possible to set $w = ap$, and $d = bt$, where a and b are constants appropriate for the given geometry.

Using this approach, the resistance of a metal line is calculated from:

$$R_{int} = \rho l/wt = \rho l/apt \qquad \textbf{(12.2)}$$

Next, the capacitance between the metal lines and the grounded planes C_{LG} (separated by an ILD whose relative dielectric constant is k), is calculated from:

$$C_{LG} = k\varepsilon_o wl/d = k\varepsilon_o apl/bt \qquad \textbf{(12.3)}$$

and the intralevel capacitance C_{LL} (assuming the same ILD) is found from:

$$C_{LL} = k\varepsilon_o tl/s = k\varepsilon_o tl/(p - w) = k\varepsilon_o tl/(1 - a)p \qquad \textbf{(12.4)}$$

where ε_o is the permittivity of free space. In Eqs. 12-3 and 12-4, C_{LG} represents the "vertical" capacitance between a metal line and a ground plane above (or below), and C_{LL} represents the "horizontal" capacitance between two neighboring lines. The total capacitance C_{int} of a metal line relative to its surroundings is thus written as:

$$C_{int} = 2(C_{LG} + C_{LL}) = 2k\varepsilon_o l\{[ap/bt] + [t/(1 - a)p]\} \qquad \textbf{(12.5)}$$

For the $R_{int}C_{int}$ time delay constant, it follows that:

$$R_{int}C_{int} = 2\,R_{int}\,(C_{LG} + C_{LL}) \qquad \textbf{(12.6)}$$

By combining Eqs. 12-2, 12-3, 12-4, and 12-6, we get:

$$R_{int}C_{int} = 2\rho k\varepsilon_o l^2\{[1/bt^2] + [1/a(1 - a)p^2]\} \quad \textbf{(12.7)}$$

This equation can also be rewritten as:

$$R_{int}C_{int} = 2\rho k\varepsilon_o (l^2/p^2)\{[p^2/bt^2] + [1/a(1 - a)]\} \quad \textbf{(12.8)}$$

The right-hand side of Eq. 12-7 has a natural interpretation, in that the factor $2\rho k\varepsilon_o$ represents the materials contributions, and the factor $l^2\{[1/bt^2] + [1/a(1 - a)p^2]\}$ represents the geometrical contributions, to the interconnect time delay.

Equation 12-8 can also be examined from a slightly different perspective if one defines the aspect ratio A of the metal lines as the ratio of line-thickness to line-width; i.e., $A = t/w$, or equivalently $A = t/ap$. In terms of A, Eq. 12-8 can be expressed as:

$$R_{int}C_{int} = 2\rho k\varepsilon_o (l^2/p^2)\{[1/a^2bA^2] + [1/a(1 - a)]\} \quad \textbf{(12.9)}$$

It can be observed that the term $1/a^2bA^2$ in Eq. 12-9 represents the "vertical" contribution to the total capacitance, and $1/a(1 - a)$ represents the "horizontal" contribution to the total capacitance.

Several conclusions about the performance of a multilevel interconnect system can be drawn by further examination of Eqs. 12-7 to 12-9. Two separate cases can be considered: **1**) where the same ILD is used in the interlevel and intralevel spaces (i.e., the value of k is constant throughout); and; **2**) where the spaces between metal lines are assumed to be filled with a material of a lower k.

12.2.1.1 Case When the *k*-Value of the ILD is Constant Throughout the Interconnect System: In this case, from Eqs. 12-8 and 12-9, it can be seen that for a fixed length l and a fixed aspect-ratio A, the RC-delay increases quadratically with decreasing feature size (i.e., decreasing pitch p). These two equations also indicate that for $A << 1$, the interlayer "vertical" capacitance is more important to the RC-delay than the intralayer (or "horizontal") capacitance., whereas the opposite is true for $A >> 1$. Also note that for the simplest geometry (where $a = 0.5$ and $b = 1$ (i.e., $d = t$), Eq. 12-9 reduces to:

$$R_{int}C_{int} = 8\rho k\varepsilon_o (l^2/p^2)\{[1/A^2] + [1]\} \quad \textbf{(12.10)}$$

12.2.1.2 Case When the *k*-Values of the Interlevel-ILD and the Intralevel-ILD are Different: In this case, one ILD is used between the metal lines and ground planes (whose "vertical" [between levels] k-value is given by k_{LG}), and another ILD is used between the metal lines themselves (whose "horizontal" [intralevel] k-value is given by k_{LL}). In that case, Eq. 12-9 becomes:

$$R_{int}C_{int} = 2\rho\varepsilon_o (l^2/p^2)\{[\,k_{LG}/a^2bA^2] + k_{LL}[1/a(1 - a)]\} \quad \textbf{(12.11)}$$

Again, for the simplest geometry where $a = 0.5$ and $b = 1$ (i.e., $d = t$), Eq. 12-11 reduces to:

$$R_{int}C_{int} = 8\rho\varepsilon_o(l^2/p^2)\{[k_{LG}/A^2] + [k_{LL}]\} \qquad \textbf{(12.12)}$$

Equation 12-12 demonstrates several important conclusions. First, if a high aspect ratio interconnect structure is used on a given level (i.e., A is large), the large value of A will mitigate the influence of k_{LG} (i.e., it doesn't matter as much if an ILD with a high or low value of k_{LG} is used between interconnect levels). However, the value of k_{LL} plays significant in determining $R_{int}C_{int}$ for this case. Second, Eqs. 12-11 and 12-12 indicate that the contribution of the intralayer capacitance to the total capacitance (namely, that due to the term 1/{a[1-a]}), is minimized when $a = 0.5$; i.e., when w is one half the pitch p.

While Eqs. 12-11 to 12-12 offer some general insight for minimizing the $R_{int}C_{int}$-delay, the interconnect structure can also be treated as consisting of two tiers, termed: **1**) the *local interconnect tier*; and **2**) the *global interconnect tier* (see Sect. 12.1.6). There are some differences in the ways in which the minimum *RC*-delay can be achieved in each of these tiers. The issues of are considered next.

12.2.1.3 Minimizing the *RC*-Delay in Local-Interconnect Tiers: The local interconnect tiers are those used to route interconnect lines *within* a cell or block, and in which the finest metal pitch allowed by the design rules are employed. In these tiers, the routing lines are so short that it can be assumed that $R_{int} << R_{trans}$ and $C_{int} >> C_{trans}$ (i.e., the wire resistance of lines shorter than about 1-mm is much less than the transistor on-resistance R_{trans}). In addition, the lines are so close that the largest component of C_{int} is intralevel capacitance, C_{LL}. In this case, the *RC*-delay in local interconnects is not found from Eqs. 12-2 through 12-4 (which assume the RC is controlled by R_{int} and C_{int}) but instead is described by the following equation:

$$R_{trans}C_{int} \sim R_{trans}\, k\varepsilon_o tl/s \qquad \textbf{(12.13)}$$

Thus, unlike in the global interconnect tiers where the *RC*-delay is controlled by $R_{int}C_{int}$ (and where the delay is thus dependent on the *square* of the interconnect length l), in the local interconnect tiers the delay is only *linearly-dependent* on the interconnect length. In addition, in this region C_{int} is dominated by the intralevel coupling between adjacent lines. Furthermore, in the local interconnect regions it is also clear that changes in the metal resistance will not play a big role in altering the RC delay, since $R_{trans} >> R_{int}$. This means that in this line-length regime, the resistance and capacitance are decoupled. Consequently, factors that reduce C_{int} (such as lowering the value of k between metal lines, and reducing

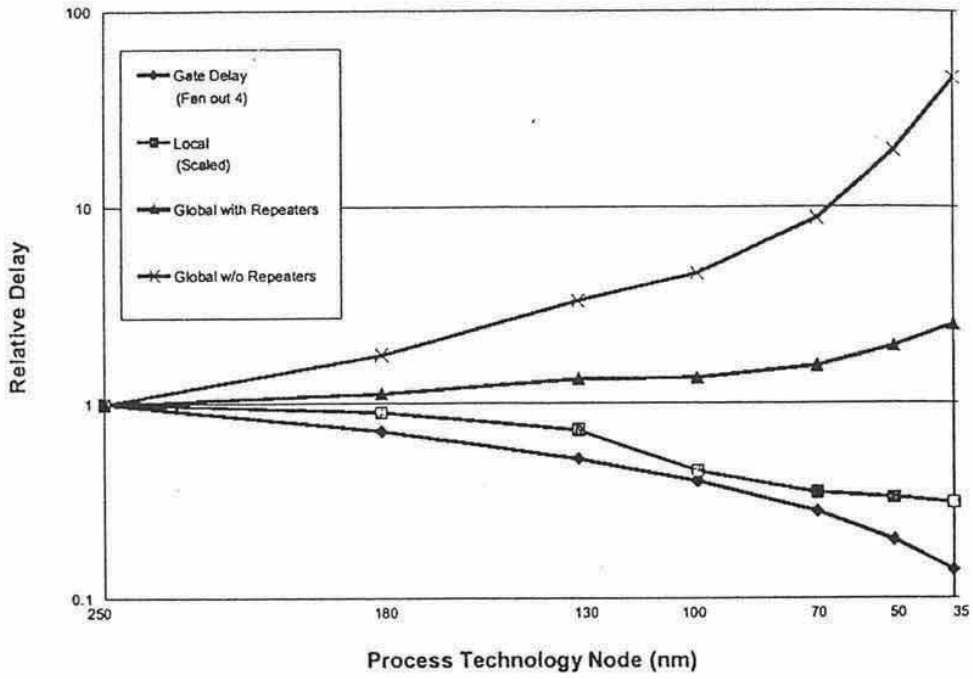
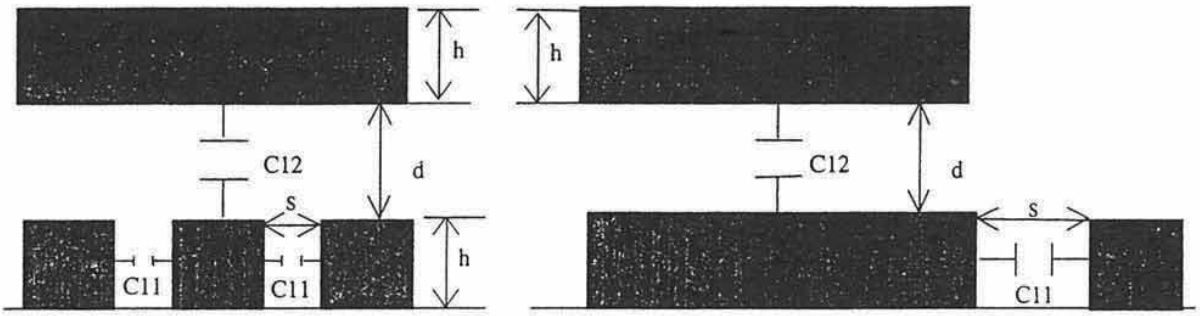


Fig. 12-20 Interconnect versus delay, from ITRS99.[3]

the height of the metal lines), will help decrease the *RC*-delay (provided that the interconnect resistance of the metal line does not exceed that of R_{trans} due to the reduction in the height of the metal lines). Figure 12-20 shows it is predicted that the relative delay of local interconnect tiers will remain roughly comparable to the gate delay as devices are scaled down toward 50-nm gate-lengths.

12.2.1.4 Minimizing the *RC*-Delay in the Global-Interconnect Tiers: The factors that minimize the *RC*-delay in the global tiers of the interconnect structure are somewhat different than the factors that govern minimizing *RC*-delay in the local interconnect tiers. (The *global interconnect tiers* are those used for inter-block communication employing the upper metal levels.) First, there are two coupling regimes that are encountered in global interconnect tiers – the *intralevel regime* and the *interlevel regime*. Figure 12-21a depicts the conditions that exist in the intralevel regime (which is typical of the denser interconnect lines found in the lower levels of global interconnect tiers). In this regime the intralevel capacitance dominates ($C_{LL} >> C_{LG}$). In the interlevel regime (which tends to exist in the upper levels of the global interconnect region, where wide wires are used to lower the resistance of long runs – Fig. 12-21b), the interlevel capacitance dominates ($C_{LG} >> C_{LL}$).

In the *intralevel global interconnect-regime*, $R_{int}C_{int}$ is calculated by combining Eq. 12-2 and 12-4, to get:



(a) Dense metal case where C11>C12
Major contribution in metal level

(b) Wide metal case where C11< C12
Major contribution between metal levels

Fig. 12-21 Capacitance coupling contributions for interconnect systems: a) Dense interconnect case where major coupling is between minimum-spaced lines; and b) Wide interconnect lines where the major coupling is between metal-levels.

$$R_{int}C_{int} = 2R_{int}C_{LL} = 2\,k\varepsilon_o\rho l^2/ws \qquad \textbf{(12.14)}$$

In the *interlevel global interconnect-regime* the interlevel-capacitance dominates, and hence $R_{int}C_{int}$ is calculated by combining Eqs. 12-2 and 12-3:

$$R_{int}C_{int} = 2\,k\varepsilon_o\rho l^2/td \qquad \textbf{(12.15)}$$

We see that in both cases the *RC* delay is strongly impacted by the length of the interconnect line (i.e., *RC* increases as the square of the line-length).*

In many circuits it is undesirable to operate under these conditions (i.e., in which the interconnect limits the propagation delay of the circuit). To overcome this limitation, some circuit designs include *repeaters* or *buffers* along the interconnect lines that exceed a critical length. The use of repeaters (while not desirable from a circuit-density standpoint), is nevertheless an essential technique for optimizing signal propagation. Implementing repeaters can even extend the usefulness of a given process technology. Their use will optimally halve the delay of a long line, with a small penalty of the repeater-delay and some device area.[13] Figure 12-22 shows how the addition of repeaters to an interconnect system can reduce the number of levels of metal as the functional density of ICs is increased.[16] In another example, Figure 12-23 depicts the interconnect delay contribution vs. feature size for varying repeater insertion lengths of 3000-μm

* Note that the intermediate situation (i.e., where both intralevel- and interlevel-capacitance effects are comparable in magnitude) is treated by Eqs. 12-8 and 12-9, since this situation lies between the two extremes and includes capacitive contributions both in- and between-metals-levels.

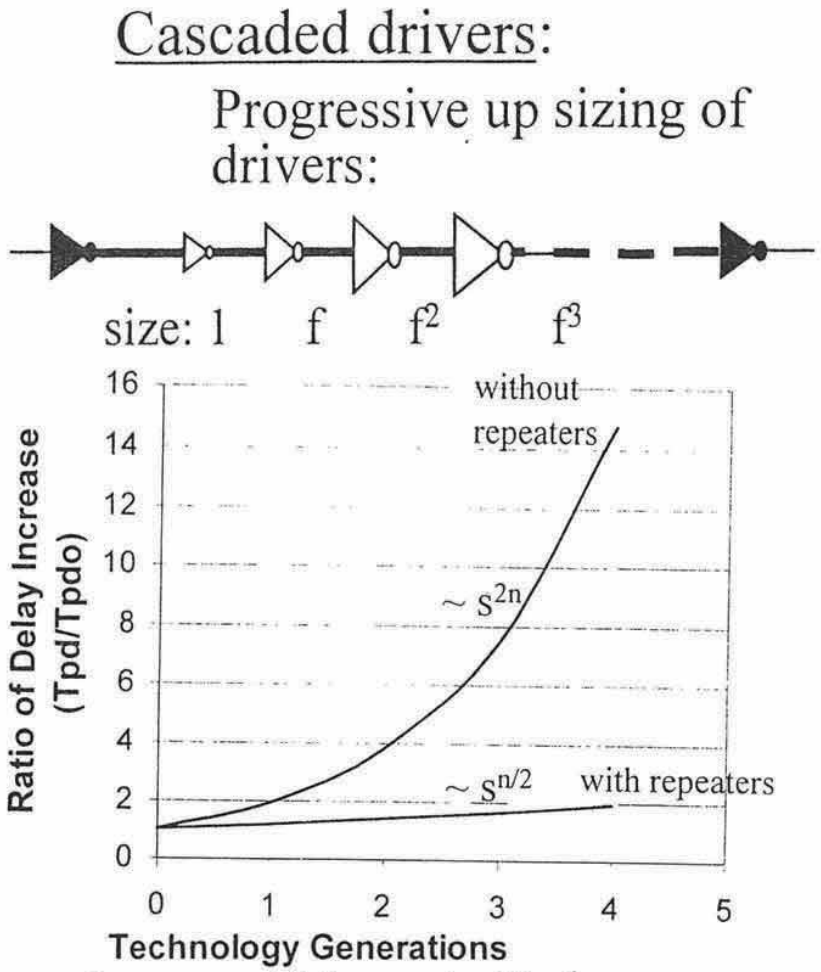


Fig. 12-22 Impact of repeaters on global propagation delay.[5]

and 5000-μm.[14] Note that reducing the spacing at which repeaters are placed can compensate for delays due to RC parasitics and circumvent the need for new materials (e.g., Cu and low-k dielectrics).

On the other hand, where long interconnect lines must be used, large pitches p must be employed, and the lines should be made wider and thicker. This implies that long interconnect lines should only be implemented in the upper interconnect metal layers where there is this room for increased pitch (wider lines and spaces), and where the lines can be made thicker. This is the approach that is commonly used today.

Thus, a multi-level interconnect system requires a hierarchy of scales, with the smallest lines in the first metal-level, and progressively wider (and perhaps thicker) and longer lines in the upper levels. One disadvantage of using thicker and wider lines in the upper levels is that fewer metal lines can be fabricated on any level. Consequently, to achieve the same functionality in the face of continued feature size scaling, the number of metal levels used in interconnect systems will also continue to grow.

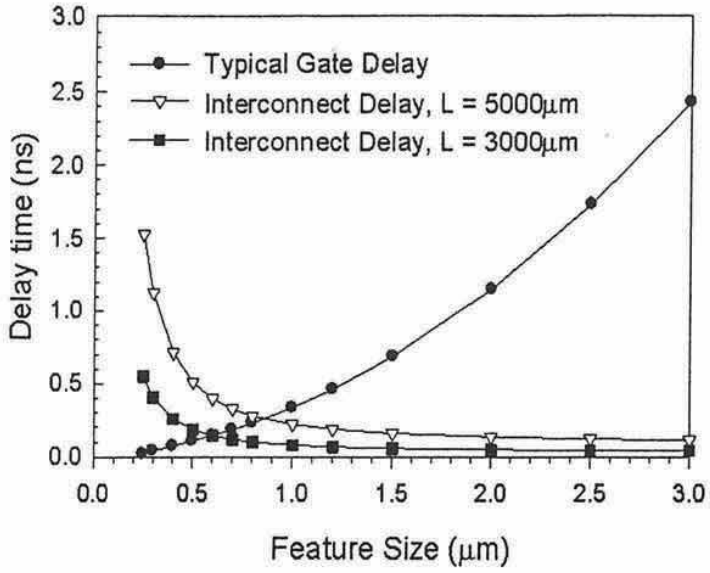


Fig. 12-23 Gate and interconnect delay versus feature size. Interconnect delay is shown for repeater spacings (L) of 3000-μm and 5000-μm.[14]

On the materials side of this issue, it is evident that the lowest $R_{int}C_{in}$-delay in global interconnect-layers will be achieved if a metal with the lowest possible ρ is used together with an ILD having the lowest possible k-value. With respect to the metal, this means using copper. (The details of fabricating metal interconnects with copper are described in Chap. 16). With regard to the ILD, it is especially beneficial to selectively reduce k_{LL} between the metal lines. We will explain how this can be accomplished in Chap. 14.

12.2.2 Power-Dissipation Issues in ULSI Interconnects

As the requirements for performance in ULSI continue to increase, the power demands will also arise unless the overall system capacitance, resistance and/or operating voltages can be reduced. The power-dissipation associated with a functioning IC can be calculated using either of the following two relationships:

$$P \sim f\, C_{total} V^2 \qquad \textbf{(12.16)}$$

or

$$P \sim I_{total}^2\, R_{total} \qquad \textbf{(12.17)}$$

where: P = power dissipated (watts)
f = operating frequency (Hz)
C_{total} = total IC capacitance
= $C_{trans} + C_{int}$ (farads)
R_{total} = total IC resistance
= $R_{trans} + R_{int}$ (ohms)

V = IC operating voltage (volts)
I_{total} = IC operating current (amperes)
C_{trans}, C_{int} = transistor & interconnect capacitance contributions
R_{trans}, R_{int} = transistor & interconnect resistance contributions

Although Eqs. 12-16 and 12-17 indicate that the IC operating voltage plays the greatest role in determining the power dissipation, its selection is based on circuit and system design issues. On the other hand, the system resistance and capacitance are the chief parameters that can be controlled by the fabrication technology. Currently, the transistor resistive losses are responsible for most of the chip power dissipation (and these are discussed in Chaps. 1 and 5). However, as the ICs are being operated at higher and higher frequencies, an appreciable fraction of the power dissipation and voltage drops are associated with the resistive component of the interconnect. Due to the decreased size of interconnects on ULSI (~200-nm minimum feature sizes in 2000), interconnect resistance is increasing and losses can become significant if the wiring is not designed properly. The overall resistance of the interconnect includes both the line resistance and the resistance of one or more contacts and/or vias along the wiring path. Joule-heating caused by current flowing through the resistive interconnect enhances electromigration and represents a growing reliability concern. The power dissipated at the higher-resistance transition-regions of the contact or via interface is of special concern, especially for large pulsed currents. Since power-dissipation in interconnects is a growing problem, it is necessary to ensure interconnects are properly optimized to reduce such power-dissipation.

12.2.3 Crosstalk in ULSI Interconnects

Another issue that impacts the performance of ULSI interconnects is *crosstalk*. Interconnect lines of substantial length act as "antennae" which experience unwanted voltage pulses induced by the rapid voltage swings in neighboring wires of close proximity (which arise from the propagation of digital pulse waveforms along the driven lines). Induced voltages due to intralevel-coupling (the largest crosstalk component between dense wires in the same level) can easily achieve 20% of the supply voltage for very rapidly changing clock waveforms.[4] Figure 12-24 shows a simulation of the amount of crosstalk experienced by an undriven wire that is 1-mm long, separated by 100-nm from an adjacent interconnect line of the same length and driven with a very fast clock edge.[5] V1 is the driving waveform and V2 is the crosstalk voltage induced in the undriven line. The rate of rise of the clock edge imposes a large induced voltage

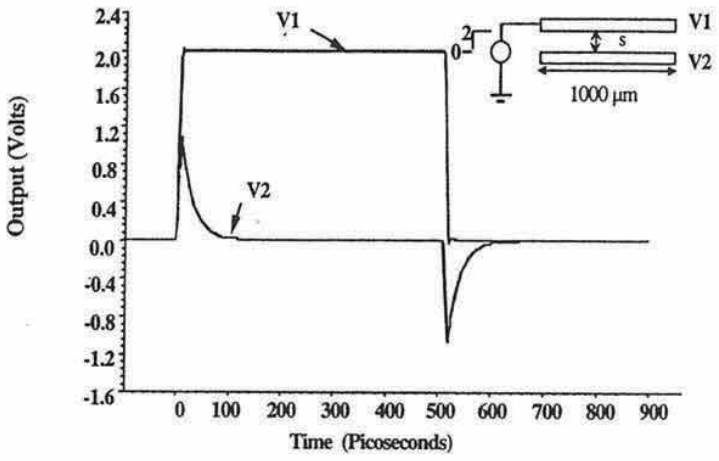


Fig. 12-24 Electrical modeling result showing the crosstalk voltage induced from a driven interconnect-line (V1) into an undriven line (V2) seeing a 20-psec waveform-edge. The lines were spaced at s = 1000-nm with a common length of 1-mm.[10]

which is related to the parasitics of the interconnect. In this example, where the clock-edge rise time is 20-psec, the magnitude of the imposed crosstalk voltage can be as much as 60% of the voltage waveform V1 (2 x 0.6 = 1.2V). As the intralevel-spacing between interconnects decreases, the magnitude of the crosstalk noise can become substantial, as is evident from the simulation data shown in Fig. 12-25.

It is more difficult to establish the level of crosstalk in dynamic systems than in the example given above. More sophisticated circuit simulators are required to

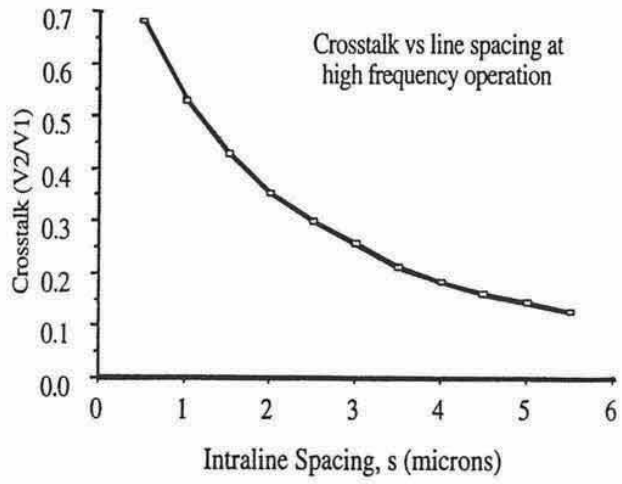


Fig. 12-25 Simulation results showing that the magnitude of the crosstalk between 2 interconnects separated by different line spacings, s.[10] Conditions described in Fig. 12-24.

estimate the impact of such crosstalk on circuit device performance. The potential of surrounding wires, wire resistance, wire length, and the capacitive load all affect the amplitude of crosstalk signals. Nevertheless, it is possible to reduce intralevel crosstalk by changing the relative dielectric environment. That is, by inserting an embedded low-k dielectric between dense intralevel wires, electric-field lines can be diverted to the short orthogonal lengths of underlying interlevel wiring (lowering capacitive-coupling between intralevel interconnects).

Crosstalk will become a significantly greater problem as ULSI mixed-signal devices begin to emerge in the marketplace. It is predicted that noisy digital wires and circuit elements will cause a number of problems for analog elements that are in close-proximity to such sources of noise. New techniques involving such novel electrical-isolation-methods as shielding and signal conditioning will be needed. This is because, in some circuits, analog devices must be immune to such "noise" down to levels that are three to six orders of magnitude smaller than the power-supply voltage-levels.

12.2.4 Contact Resistance in ULSI Interconnects

The magnitude and uniformity of contact and via resistance is another forthcoming challenge (see Vol. 2, Chap. 3 for more background information on contact and via resistance). Contact and via resistances can be substantial and have a significant impact on the overall wiring resistance of the interconnect system. These resistance contributions act in the same way as adding length to an interconnect line, and thus can factor heavily into voltage drops and delays in the wiring. For 0.25-μm contacts with an interfacial-contact-resistivity ρ_c to heavily doped silicon of $5x10^{-8}$-$1x10^{-7}$-Ω-cm^2, such contact-resistance values are approximately 100-200-ohms/contact.[6] Vias for systems using Al alloys with W-plugs have practical metal-to-metal contact resistivities of approximately $2x10^{-9}$ Ω-cm^2 and a resistance of about 5-Ω/via at the same geometry.[7] Vias which use copper metallization, however, are expected to show substantially lower contact-resistance because they have only one critical barrier interface per via (instead of two interfaces when W-plugs and Al alloy interconnect lines are used), and because the interfacial-contact-resistivity should be lower than $1x10^{-9}$-Ω-cm^2.[8]

Cleaning prior to the deposition of the contact- and via-filling conductor-films and control of dimensional tolerances are critical to the reproducibility of via or contact resistance. It is predicted that production variability will increase with the migration to smaller feature sizes and have a negative impact on resistance. This

resistance must be accounted for in the delay analysis when multiple levels with a large number of contacts and vias are used.

12.2.5 Electrostatic Discharge Protection Issues of Interconnects

Electrostatic discharge events (ESD) can cause large transient currents to flow in an IC. This can cause "fusing" in upper wiring levels, thus jeopardizing the long-term reliability of the IC. ESD robustness has been shown to be a severe reliability concern by several research groups, for interconnect generations with metal features smaller than 0.25-μm.[9,10] Since short term discharges can exceed 2000 volts for durations of hundreds of nanoseconds, it has been shown interconnects are becoming more prone to failure due to this type of electrical stress. Copper interconnects exhibit a substantially better resistance to this type of failure, when compared to aluminum interconnects, due to the higher melting point and thermal conductivity of copper.

12.3 SUMMARY OF USLI INTERCONNECT ISSUES AND FUTURE TRENDS

Based on the discussion of this chapter, the following trends are likely to be encountered in ULSI interconnect systems:

1. If the materials of an interconnect system are not changed (i.e., the metal resistivity ρ and ILD-k values remain the same), the RC-delay will continue to increase as device dimensions (and therefore interconnect pitch p) are reduced. There are two main reasons why this increase in RC occurs:

a. Increasing the aspect ratio A of metal lines (to keep R constant) beyond about 2 produces diminished improvements in RC - even for a constant pitch p (because C_{LL} rises while R remains constant). However, if the pitch is reduced, this makes C_{LL} grow even larger, which, in turn causes an additional rise in RC.

b. The metal line lengths tend to scale less strongly than does the metal pitch p. Hence, even as devices keep shrinking, the ICs will continue to contain some rather long lines (which will have larger C_{total}/unit length, for lines with the same length).

We noted earlier, however, that in short- and intermediate-length local interconnects, changes in the metal resistance will typically not play a major role in the RC-delay (since in this regime $R_{trans} >> R_{int}$). For such cases, the resistance and capacitance are decoupled. That is, factors that reduce interconnect capacitance (such as a reduction in the dielectric

constant of the ILD, and/or a reduction in the height of the metal lines), will help improve *RC* delay (as long as the interconnect resistance increase due to reducing the metal film thickness does not approach R_{trans}).

2. If ρ and/or k can be decreased (especially a reduction in the value of k of the ILD used in intralevel metal spaces k_{LL}) this will cause a decrease in *RC*. Hence, there will be continual effort expended to integrate Cu and low-k dielectrics into ULSI interconnect systems. The issues involved with incorporating low-k ILDs are covered in Chap. 14, and the challenges encountered when attempting to implement Cu-metal-interconnects are described in Chap. 16. Figure 12-26 shows a possible migration path for materials use in IC interconnect structures.[15]

3. More levels of metal will need to be used as the feature sizes are decreased.

4. Optimizing interconnect length is a powerful method for reducing *RC*-delay. This can be influenced by design philosophy and device architecture. Use of repeaters, while not desirable from a density standpoint, is another key approach for optimizing signal-propagation, and can extend the use of a given process technology. Since incorporation of repeaters into an interconnect system is a circuit-design issue, this approach not considered further in this text.

5. Increasing the metal line aspect-ratios and adding more levels of interconnect are two methods for improving interconnect performance and density. However, these approaches will reach practical limits in a few

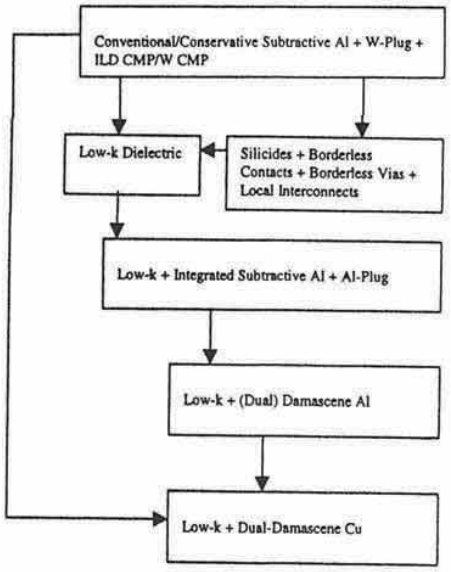


Fig. 12-26 Schematic illustrating an interconnect migration.[15]

generations.[11] (For example, continuing to increase the aspect ratio A above $A = 2$ will bring only diminished improvements in RC because the reduction in the metal resistance will be offset by an increase in intralevel capacitance.) New interconnect materials, improved circuit-design techniques, and possibly novel interconnect schemes will be needed to address the growing limitation that interconnects pose to ULSI performance. Circuit design needs to more accurately model interconnect performance to avoid overdesign or unexpected performance limitations, and automatic layout tools need to make more efficient use of existing interconnects. The novel interconnect approaches include RF-transmission and optical-interconnects. More information about these novel techniques can be found in Ref. 12.

REFERENCES

1. S.C. Sun, "Process Technologies for Advanced Metallization and Interconnect Systems," *Tech. Dig. IEDM*, 1997, p. 765.

2. R. Liu *et al., Proc. Int. Interconnect Technol. Conf.* (IEEE), 1999, p. 21.

3. International Technology Roadmap for Semiconductors, 1999 Edition, Semiconductor Industry Association, San Jose, CA.

4. C. Cregut *et al.,* "Low-*k* Dielectrics Influence on Crosstalk: EM Analysis and Characterization," *Proc. of the International Interconnect Technol. Conf.,* San Francisco CA, June 1-3, 1998, p. 59.

5. H.P. Bakoglu, *Circuits, Interconnections, and Packaging for VLSI*, Addison Wesley Publ., Reading MA, 1990.

6. I.S. Park, *et al.,* "Fabrication of 256 Mbit DRAM with a Fully Planarized Double Level Interconnection," *Proc. 12th VLSI Multilevel Interconnect Conf. (VMIC)*, Santa Clara CA, 1995, p. 45.

7. K. Kikuta *et al.,* "Comparitive Study of W-plug, Al-plug, and Al-Dual-Damascene for 0.18 μm ULSI Multilevel Interconnect Technologies," *Proc. of the Internat. Interconnect Technol. Conf.,* San Francisco CA, June 1-3, 1998, p. 140.

8. J. Zhang *et al.,* "CVD Cu Process Integration for sub-0.25 μm Technologies," *Proc. of the Internat. Interconnect Technol. Conf.,* San Francisco CA, June 1-3, 1998, p. 163.

9. S.H. Voldman, "ESD Robustness and Scaling Implications of Aluminum and Copper Interconnects in Advanced Semiconductor Technology," *IEEE EOS/ESD Symposium , 1997, p. 316.*

10. K. Banerjee *et al.*, "Characterization of VLSI Circuit Interconnect Heating and Failure under ESD Conditions," *1996 IEEE 34th International Reliability Physics Proceedings*, April 30-May 2, 1996, Dallas, TX, p. 237.
11. M.T. Bohr, "Interconnect Scaling – The Real Limiter to High Performance ULSI," *Tech. Dig. IEDM*, 1995, p. 241.
12. M.E. Thomas and R.H. Havemann, "Overview of Interconnect," Ch.12, in *Handbook of Semiconductor Manufacturing Technology*, Y. Nishi and R. Doering, Eds. Marcel Dekker, New York, 2000, pp. 303-4.
13. H.P. Bakoglu and J.D. Meindl, " Optimal Interconnection Circuits for VLSI," *IEEE Trans. Electron Dev.,* May 1985, p. 903.
14. J.C. Eble *et al*, "Minimum Repeater Count, Size, and Energy Dissipation for Gigascale Integration (GSI) Interconnects," *Proceedings of the International Interconnect Conf.*, San Francisco CA, 1998, p. 56.
15. J. Givens, "Engineering CMP for Manufacturability and Yield," *Abs. Mtg. Electrochem. Soc.,* Fall 1999, Abs. 1275.
16. S. Wills, "Interconnect Architecture," in *IEDM* 2000 *Short Course*: *Advanced Interconnects: Design, Process, Integration*, p. 27.