SILICON PROCESSING

FOR

THE VLSI ERA

VOLUME 4:

DEEP-SUBMICRON PROCESS TECHNOLOGY

STANLEY WOLF Ph.D.

LATTICE PRESS

Sunset Beach, California

TSMC-1034
TSMC Ltd.
IPR2025-01210

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

EMS
Lib.
TK
7874
W64
1986
v. 4

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

SILICON PROCESSING FOR THE VLSI ERA

Vol. 4 – Deep-Submicron Process Technology

DETAILED TABLE OF CONTENTS

PREFACE

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

1.1 THE STRUCTURE OF DEEP-SUBMICRON MOSFETS: (0.25-μm to 0.13-μm) - COMPARED TO THE STRUCTURE OF "CONVENTIONAL" MOSFETS (2.0-μm to 0.5-μm) 2

1.1.1 Evolution of the MOSFET Gate Stack and Contact Structure
1.1.2 Gate Dielectric Materials in Deep-Submicron MOSFETs
1.1.3 Doping-Concentration Profiles of the MOSFET Channel
1.1.4 Evolution of the Drain Structure of MOSFETs

1.2 DEEP-SUBMICRON CMOS STRUCTURES 10

1.2.1 Substrate Issues for Deep-Submicron CMOS
1.2.2 Well Formation in Deep-Submicron CMOS
1.2.3 Dual-Doped Poly in Deep-Submicron CMOS
1.2.4 Shallow Trench Isolation for Deep-Submicron CMOS
1.2.5 Silicon-On-Insulator (SOI) Technology

1.3 LIMITS TO CONVENTIONAL MOSFET SCALING 15

REFERENCES 16

Chap. 2 - 300-MM SILICON WAFERS 17

2.1 300-mm SILICON CRYSTAL GROWTH 17

2.2 GROWN-IN SILICON DEFECTS: 21 CRYSTAL-ORIGINATED-PARTICLES (COPS) & DISLOCATION LOOPS

2.3 DETAILS OF THE FORMATION OF CRYSTAL-ORIGINATED-PARTICLES (COPS) 23

2.3.1 The Radial Distribution of Grown-In Defects on the Wafer Surface

2.4 THE OXYGEN-STACKING-FAULT RING (OSF-RING) 26

2.5 MITIGATING EFFECTS OF COPS BY USE OF POST-CRYSTAL-GROWTH ANNEALING 27

© 2002 LATTICE PRESS Sunset Beach CA All Rights Reserved

2.6 ELIMINATION OF COPS WITH "PERFECT-SILICON" 29
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, & 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

© 2002 LATTICE PRESS Sunset Beach CA All Rights Reserved

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

© 2002 LATTICE PRESS Sunset Beach CA All Rights Reserved 3.10.1

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

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}

© 2002 LATTICE PRESS Sunset Beach CA All Rights Reserved

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
5.3.3 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

© 2002 LATTICE PRESS Sunset Beach CA All Rights Reserved

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)

© 2002 LATTICE PRESS Sunset Beach CA All Rights Reserved

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)
8.4 THE HISTORY OF CMP 333
8.5 MODELING THE MECHANISMS OF CMP 335
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)

© 2002 LATTICE PRESS Sunset Beach CA All Rights Reserved

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

© 2002 LATTICE PRESS Sunset Beach CA All Rights Reserved

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

© 2002 LATTICE PRESS Sunset Beach CA All Rights Reserved

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

© 2002 LATTICE PRESS Sunset Beach CA All Rights Reserved

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

© 2002 LATTICE PRESS Sunset Beach CA All Rights Reserved

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 2^{nd}-Genereation Spin-On-Dielectrics: ($2.5 < k < 2.8$)
14.5.2 2^{nd}-Generation CVD-Dielectrics: ($2.5 < k < 2.8$)
14.5.2.1 2^{nd} Generation Low-*k* CVD-Films: Organo-Silicate Glasses
14.5.2.2 2^{nd} Generation Low-*k* CVD-Films: *FlowFill* CVD Films

© 2002 LATTICE PRESS Sunset Beach CA All Rights Reserved

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

© 2002 LATTICE PRESS Sunset Beach CA All Rights Reserved

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

© 2002 LATTICE PRESS Sunset Beach CA All Rights Reserved

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

© 2002 LATTICE PRESS Sunset Beach CA All Rights Reserved

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

Chapter 8

CHEMICAL MECHANICAL POLISHING (CMP)

As circuit features are scaled to submicron dimensions and additional levels are added to multilevel-interconnection schemes, the required degree of planarization is increased. In fact, the complexity of current microelectronic devices demands global planarity of the substrate after shallow-trench isolation and at all interconnect levels. Thus, such planarization needs to be implemented in both conductor and dielectric materials. In this chapter the technique of *chemical mechanical polishing* (the only known technique that provides global planarization for ULSI applications) is described. The discussion will begin by defining terms through which planarization is characterized, and then a brief recount of the evolution of planarization techniques that preceded CMP will be given. After concluding this introductory matter, the remainder of the chapter is devoted to the CMP process.

8.1 TERMINOLOGY & EVOLUTION OF PLANARIZATION PROCESSES FOR ICS

The term planarization is employed quite frequently (and often loosely) in other technical literature. Thus it is useful to define the term more completely. The example case used for this discussion is that of a dielectric layer deposited after Metal-1 is patterned (i.e., an *intermetal dielectric* [or IMD] film). In the case where no planarization exists (Fig. 8-1a), the step heights on the IMD surface closely approximate the step heights of the Metal-1-layer and the underlying topography. Furthermore, in this case the steps on the IMD surface generally also have steep slopes.

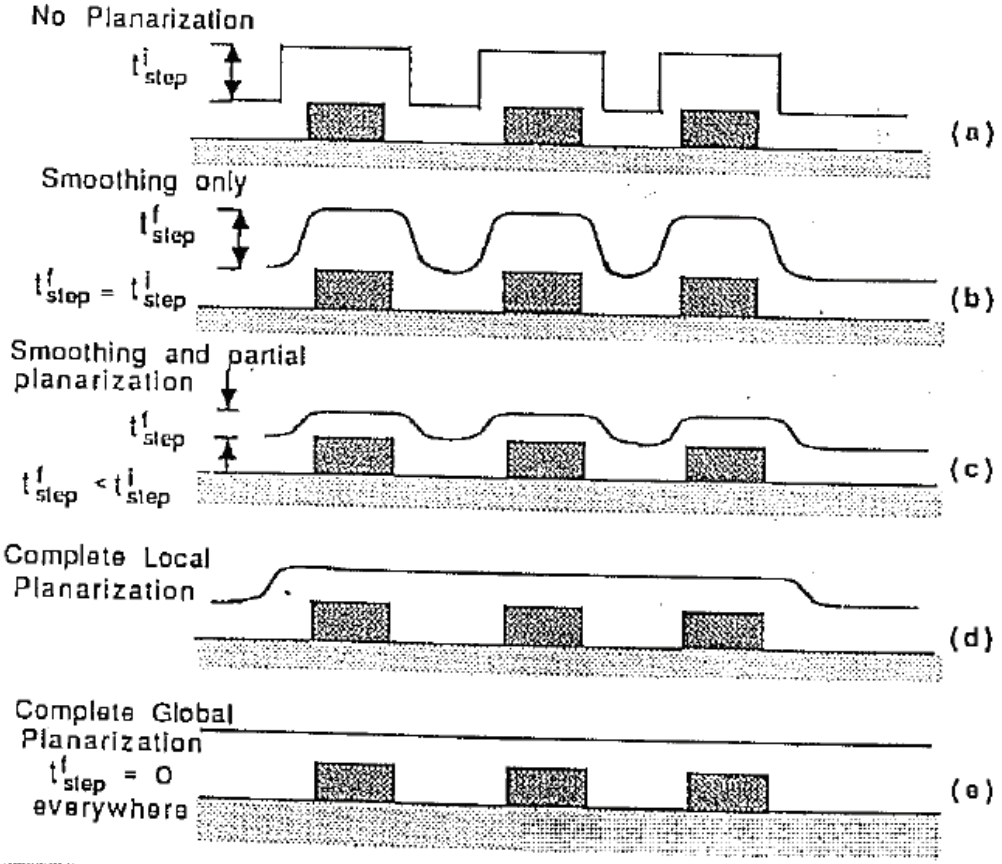
8.1.1 Defining the Degree of Planarization

The steps on the surface of the IMD can be made less severe through various planarization methods. The degree to which this can be successfully accomplished differs according to the technique used. The degree of planarization is classified according to the following qualitative planarization criteria:

1. The first degree of planarization (smoothing) involves a lessening of the step slopes at the IMD surface (Fig. 8-1b). The step heights in this case, however, are not significantly reduced in magnitude.

2. In the second degree of planarization (partial- or semi-planarization) the step heights are reduced (but not eliminated), and the slopes of the steps are also smoothed (Fig. 8-1c).

3. In the third degree of planarization (complete local-planarization), the steps at the surface of the IMD are completely eliminated wherever the spaces in the underlying topography are relatively close together (e.g., < 10-μm



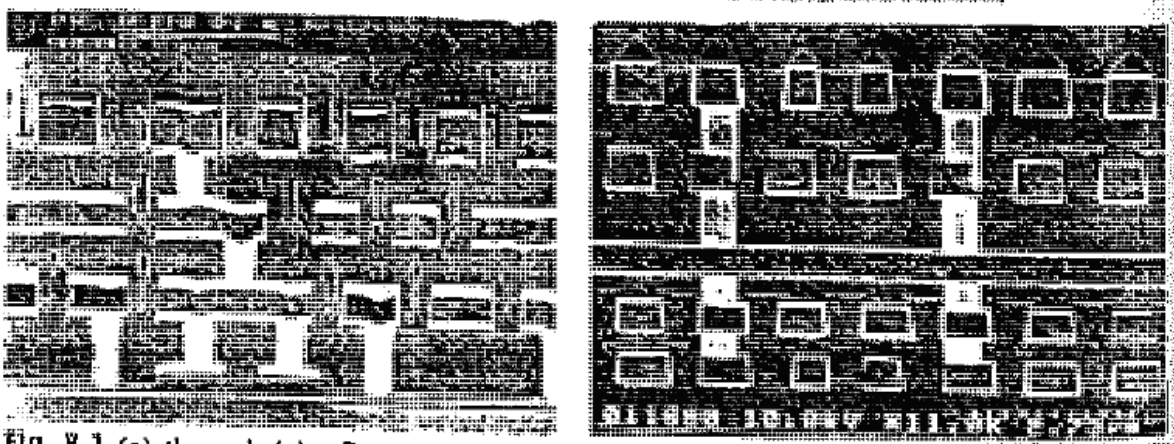


Fig. 8-1 (a) through (e) - Degrees of dielectric planarization; (f) SEM of 6-metal-level interconnect structure, with complete global-planarization realized at each level.

apart), but the steps at isolated, wide features still exhibit a step (Fig. 8-1d).

4. In the fourth degree of planarization (complete global-planarization), the surface is completely planarized over arbitrary topography (Fig. 8-1e).

Figure 8-1f shows a SEM of a 6-level-metal interconnect-system in which complete global-planarization has been realized in each of the metal levels (i.e., using CMP).

A quantitative measure of the step-height-reduction, referred to as the *planarization factor*, β, is given by[1]

$$\beta = 1 - (t^f_{step}/t^i_{step}) \quad (8.1)$$

where t^f_{step} and t^i_{step} are the final- and initial-step-heights, respectively. In complete-planarization cases, $\beta = 1$. If no planarization (or only smoothing) exists, then $\beta = 0$.

8.1.2 The Need for Dielectric Planarization

As the number of levels in an interconnect technology is increased, the stacking of additional layers on top of one another produces a more and more rugged topography. Consider, for example, a two-level-metal, single-poly CMOS technology. Assume the step-height of the semi-recessed field-oxide is 0.3-μm, and the thicknesses of the poly and the first and second metals are 0.4-, 0.5- and 1.0-μm, respectively. The maximum height of the steps on the wafer surface after each of these processes will be 0.3-, 0.7-, 1.2-, and 2.2-μm, respectively (Fig. 8-2). It is apparent that the surface of the wafer must be planarized in some fashion to prevent the topography roughness from growing with each level. Without such planarization, the microscopic canyons that result on the wafer surface from stacking of device features can lead to topography conditions that would eventually reduce the yield of circuits to unacceptably small values.

A nonplanar wafer topography results in three conditions that prevent the fabrication of reliable electrical connections in multilevel interconnect systems. These conditions are as follows:

1. Poor step-coverage of metal lines as they cross over the high and steep steps. A measure of how well a film maintains its nominal thickness is expressed by the ratio of minimum thickness of a film as it crosses a step, t_s, to the nominal thickness of the film on horizontal regions, t_n (Fig. 8-3a). This property is referred to as the *step-coverage* of the film, and is expressed as the percentage of the nominal thickness that occurs at the step:

$$\text{Step coverage (\%)} = (t_s/t_n) \times 100\% \quad (8.2)$$

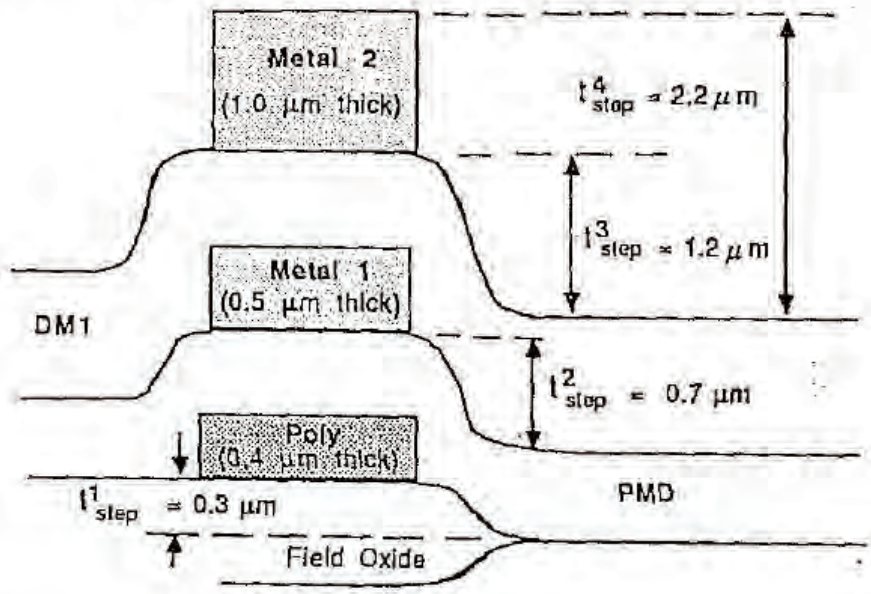


Fig. 8-2 As the number of levels in an interconnect technology is increased, the stacking of additional layers on top of one another produces steps of progressively greater heights.

Step coverage of 100% is ideal for a conductor, but in general the height of the step and the aspect-ratio of the features being covered impact the expected step-coverage. (The *aspect-ratio* is defined as the height-to-spacing ratio of two adjacent steps.) The greater the step-height or the aspect-ratio, the more difficult it is to attain coverage of the step without a corresponding thinning of the film that overlies it. Figure 8-3b shows an example of poor step-coverage. Hence, worse step-coverage is expected under these conditions. In addition to the step-height and aspect-ratio, step-coverage depends on two other topological-factors: the contour and slope of the step. In general, the smoother the contour of a step and the less vertical its slope, the better will be the step-coverage of an overlying film.

2. Metal stringers remain behind at the foot or sides of steep steps when anisotropic-etching is used. This condition is shown in Fig. 8-3c. In addition, the resist is typically eroded by the same gases that etch the metal. The resist must therefore be thick enough in all locations that it will not be completely removed before the etch-process is completed. If the dielectric surface has steps, in certain locations (i.e., over the upper corners of the steps in the dielectric, see Fig. 8-3d) there may not be sufficient resist thickness to survive during the time of overetch that must be used for removal of the stringers. The resist-erosion-rate during the overetch time

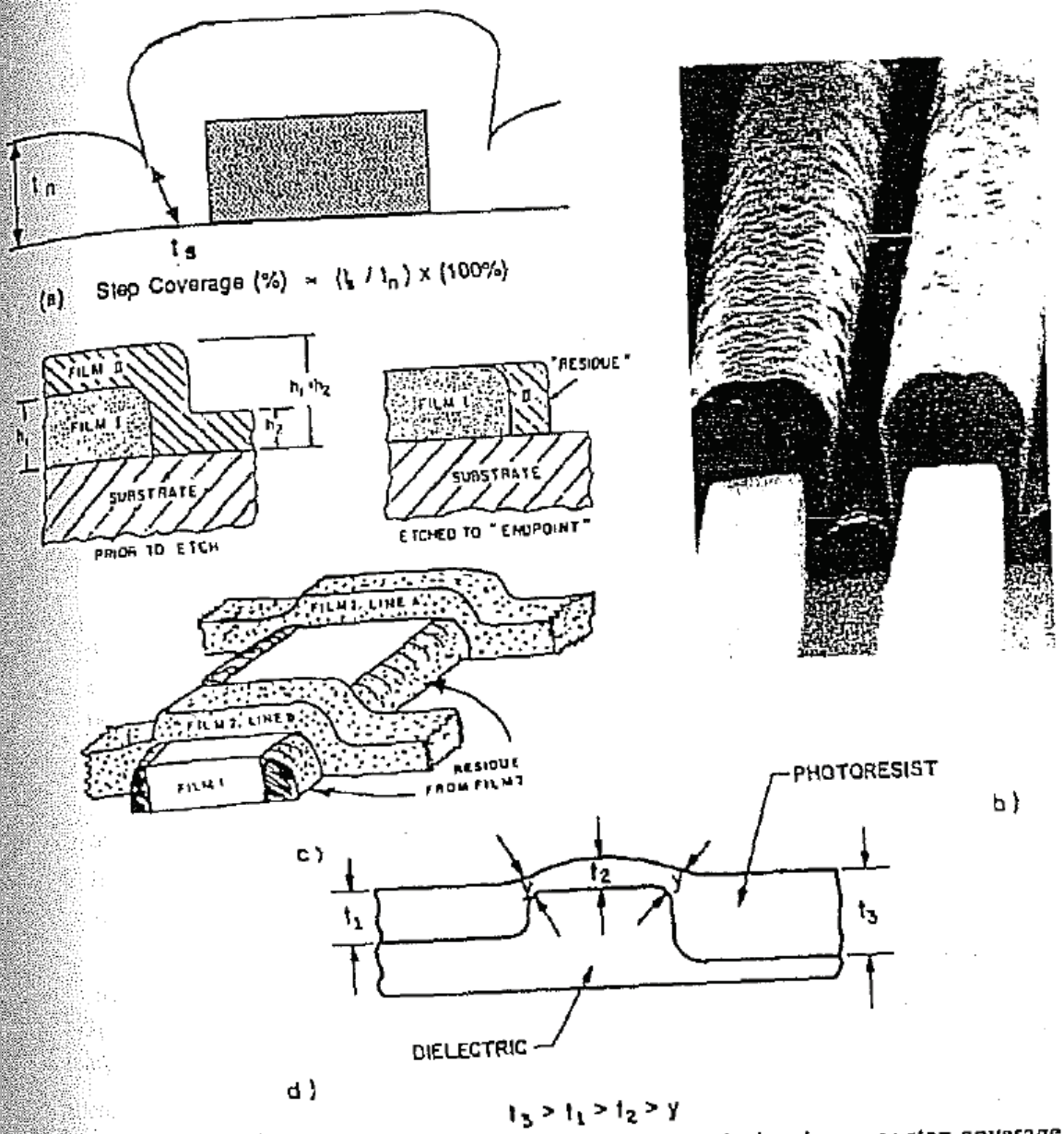


Fig. 8-3 a) Definition of step-coverage. b) SEM micrograph showing poor step-coverage. c) Metal stringers remain behind at the foot or sides of a steep step when anisotropic-etching is used. d) Problem of resist-thinning as film crosses underlying steps.

is also generally enhanced. Because most of the exposed-metal has been etched away, the reactant gases that were being consumed during etching of the metal are now present in greater concentrations. In cases where the resist is completely eroded at thin spots, the metal beneath the resist will be exposed to the reactive species and may be etched away to an unacceptable degree.

Even if the above two limitations could be overcome, the following obstacle would eventually force planarization to be implemented:

3. The depth-of-field limitations of submicron optical-lithography tools will require surfaces to be planar within ±0.5 μm. As a result, if steps larger than 0.5 μm exist on the surface of the IMD (or any other layer on the wafer), it will not be possible to pattern features in Metal-1 to the maximum resolution of the stepper. Thus, planarization will be mandatory if optical lithography is to be usable for fabricating ICs with such submicron feature sizes. (This topic is discussed in more detail in Vol.1, 2nd Ed., Chap. 13.)

Planarization of the intermetal dielectric layers is one of the chief approaches that has been taken to alleviate the problem of rough surface topography.

8.1.3 Design Rules Related to Intermetal Dielectric-Formation and Planarization Processes

The type of intermetal dielectric material, its thickness, and the method of deposition and planarization can all have an impact on the design rules that must be applied when laying out an integrated circuit with a particular multilevel-metal system. Here a two-level-metal system with an underlying poly-silicon level is considered. Assume that the topography beneath the polysilicon is completely planar. In this system, design rules dealing with the following conditions must be specified: 1) minimum polysilicon-to-Metal-1 spacing, assuming no metal overlap of the poly; 2) minimum distance between coincident edges when Metal-1 overlaps polysilicon; and 3) minimum spacing of Metal-1-to-Metal-1.

If polysilicon and Metal-1 come too close together, a gap will exist between them that will lead either to void formation in the CVD oxide PMD layer following deposition (Fig. 8-4a) or to the formation of a deep, narrow crevice that cannot be covered by the next level of metal, even after some planarization procedures have been utilized (Fig. 8-4b).

In the second case (Fig. 8-4c), if polysilicon and metal edges are coincident, the total step height may be too high for adequate coverage by Metal-2 to be achieved. Figure 8-4d illustrates the case in which the problems of coincident edges and a narrow crevice are exacerbated by their simultaneous occurrence.

In the third case, the spacing between adjacent Metal-1 lines must be large enough to prevent void or crevice formation in the deposited dielectric film. The step that exists in the LOCOS field oxide must also be considered in real situations when such design rules are being formulated. One possible way to prevent the formation of voids, crevices, and excessively high steps, is to allow

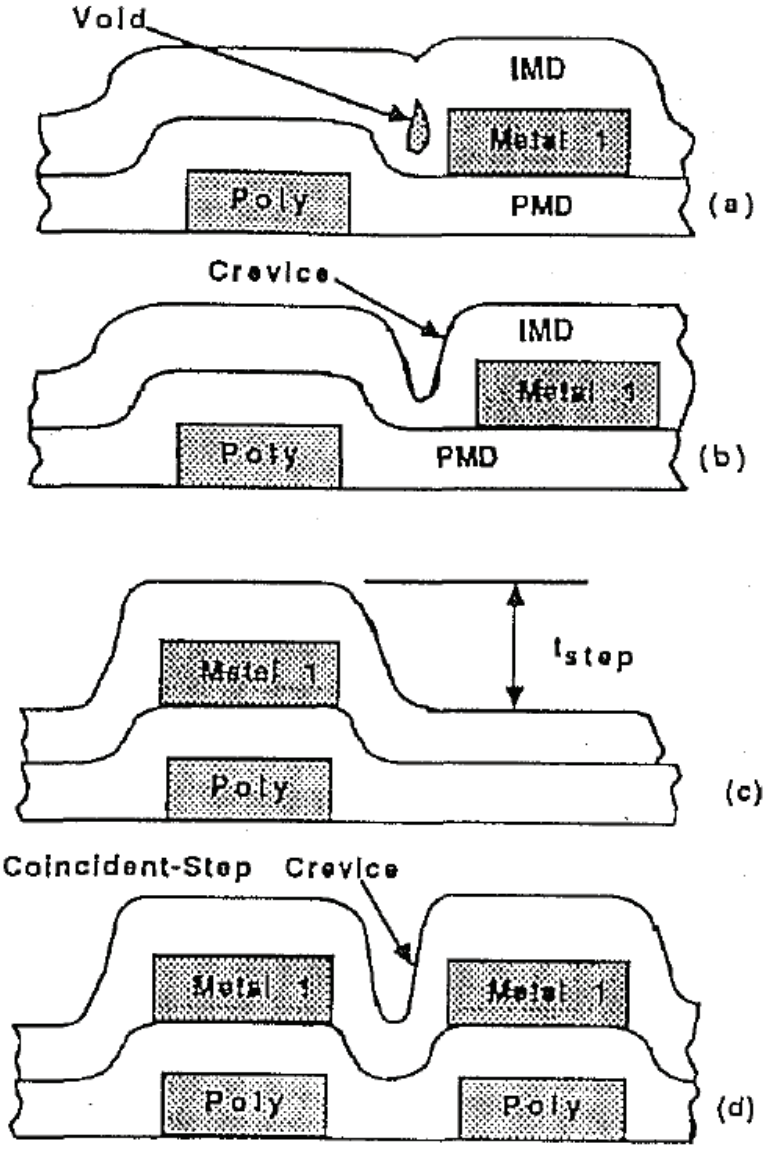


Fig. 8-4 Conditions associated with the relative spacing of polysilicon and Metal-1 that lead to layout design rules of these layers.

the spacings between Metal-1 and underlying features to be sufficiently large so any possible combination of misalignment cannot lead to the problems listed. However, this alternative is impractical, because it increases chip size and causes excessive complexity in layout design.

The trend in advanced IC designs today is toward the elimination of all such restrictive dielectric-layer-related design rules. The development focus is on dielectric planarization processes that allow all spaces between any adjacent or overlapping features (and all possible step heights), to be covered with a dielectric layer whose surface does not exhibit any voids, crevices, or excessively

high steps. Even though this is an actively pursued goal, it has not yet been attained in all processes and technologies. Various planarization approaches that have been developed along the path toward the condition in which no restrictive design rules need be invoked will be described.

8.1.4 Step-Height Reduction of Underlying Topography Alleviates the Need for Planarization

In two-level-metal processes it may be possible to avoid performing any planarization procedures with respect to the IMD if the topography underlying IMD is planarized by the process steps that precede IMD deposition. Although this approach may appear to be obsolete for ULSI circuits, several of the step-height-reduction techniques turn out to be useful for providing a planar surface under Metal-1. Such a condition will ultimately be required for three- and four-level-metal processes. The survey of planarization techniques of dielectric layers begins by considering this approach. In this section one older step-height reduction technique is described, namely *planarization through sacrificial-layer etchback.*

8.1.5 Planarization through Sacrificial-Layer Etchback

Planarization of CVD-interlevel-dielectric films can also be achieved using the *sacrificial-layer etchback technique* (Fig. 8-5). This method gained the most widespread acceptance in two-level-metal processes down to ~1-μm device tech-

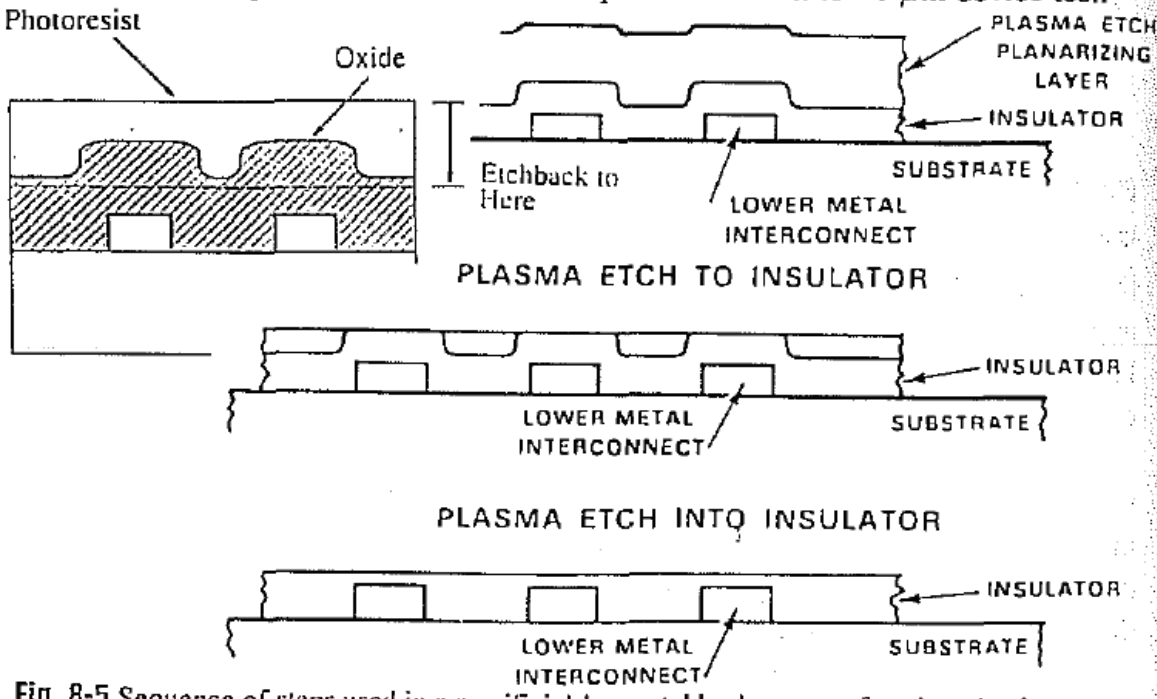


Fig. 8-5 Sequence of steps used in a sacrificial-layer etchback process for planarization.

nologies (and even in some reported three-level-metal bipolar processes). Using this approach, it is possible to achieve a high degree of planarization between steps that are ~2-10-μm apart. With more closely-spaced features, the technique runs into problems. For more widely spaced steps, planarization is less complete. The process is carried out by first depositing the CVD film that will serve as the interlevel dielectric (ILD) or intermetal dielectric (IMD). This layer is then coated with a film that will later be etched off (sacrificed). In most cases, such sacrificial-layers are photoresists, spun-on in liquid form. Upon being baked, these liquids become solid thin films that are thick enough so they provide a film surface that is close to being flat, as shown in Fig. 8-6. As long as the feature-size is smaller than about 10-μm, the degree of planarization can exceed 70% with most photoresists.

In the next step, the sacrificial-layer is first rapidly etched back in a plasma (typically, O_2, or O_2 mixed with CF_4) until the topmost regions of the dielectric layer are just exposed (Fig. 8-5b). The etch chemistry is then modified so the sacrificial-layer material and the dielectric are etched at approximately the same

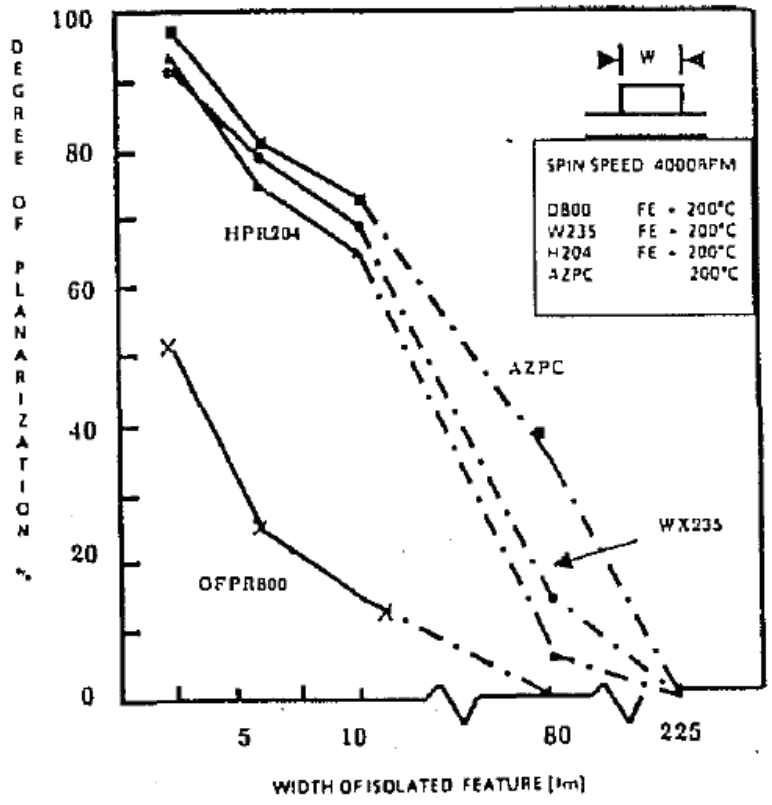


Fig. 8-6 Degree of planarization vs. various feature size for different polymers.[2] (© 1987 IEEE)

rate. The etch is continued under these conditions until all of the sacrificial layer has been etched away. At this point, the surface of the dielectric film is highly planarized since the profile of the sacrificial layer is transferred to the dielectric layer by the etchback procedure. The thickness of the dielectric film over underlying features (such as metal lines), may be thinner than desired after the etchback is completed. In fact, in some processes the etchback step is allowed to proceed until the Metal-1 lines are exposed. In either case, an additional layer of CVD dielectric is generally deposited in order to establish a minimum adequate thickness everywhere on the wafer surface.

8.1.5.1 Sacrificial-Etchback-Process Problems: To achieve the desired degree of planarity and final dielectric-film-thickness, very tight process control is necessary for such parameters as the magnitude of the etch-rate, the etch-rate uniformity across a wafer, end-point-detection, and oxide-thickness-uniformity under the resist layer. The resist etch-rate is sensitive to the cure cycle (especially if a high-temperature bake step [> 150°C] is used to improve surface planarity), and to chamber preconditioning. In batch processes, the etch-rate is also sensitive to the number of wafers present in the chamber. Such tight process-control of a complex process is hard to maintain on a production line.

The process also requires a multitude of steps, which implies a lengthy process, and thus low-throughput. Etchback is a difficult process to successfully implement in a batch-etch mode, resulting in an even lower-throughput. Finally, if the aspect-ratio between adjacent Metal-1 lines with vertical sidewalls exceeds 0.4, voids will form when most CVD-SiO_2 layers are deposited (Fig. 8-7). These voids can trap resist and moisture, which can subsequently be released if the voids are opened during the etchback process. Opened voids will also trap metal during the Metal-2 deposition-step. It will remain in the form of metal stringers following Metal-2 patterning. Thus, this process is not feasible for Metal-1 spacings of less than ~1.25-μm (unless a process is available for deposition of SiO_2 layers at low temperatures without formation of voids in high-aspect ratio spaces [e.g., PECVD TEOS films, O_3:TEOS films, or HDP CVD films]).

8.2 INTRODUCTION TO CHEMICAL MECHANICAL POLISHING (CMP)

As the number of interconnect layers in ICs has increased, the planarization of dielectric and metal layers has become more critical. In general, planarizing techniques such as the thermal flow of oxide layers, sacrificial-resist-etchback, and spin-on-glass, are inadequate for interconnect systems with more than three layers of metal. All of these techniques ultimately provide only a limited degree of smoothing and local planarization (i.e., over distances smaller than 10-μm).

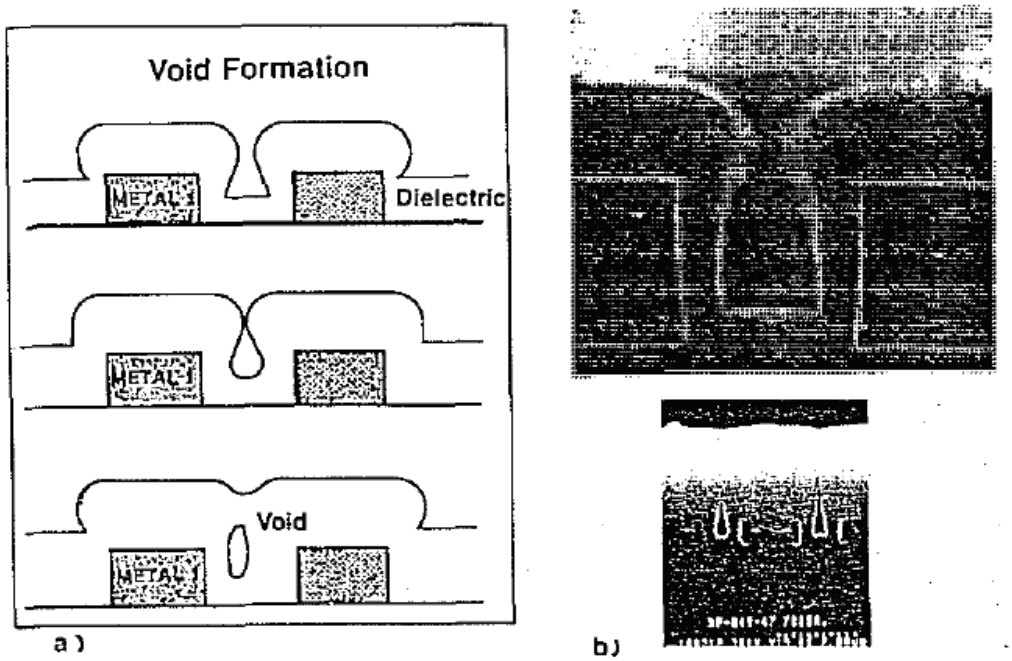


Fig. 8-7 (a) Diagram of void-formation due to cusping of the dielectric. (b) SEM of voids left between 1.5-μm-spaced Metal-1 lines in a 3.0-μm-thick PMD PECVD oxide.

They are not able to provide global planarization.[3] Global planarization of both dielectric and metal layers, however, is possible with chemical-mechanical-polishing (CMP), a process that was pioneered at IBM in the 1980s. In this technique, high elevation features are selectively removed (i.e., material from high elevation features is removed more rapidly than material at lower elevations), resulting in topography with improved planarity (see Fig. 8-10 [in which $h_{final} << h_{initial}$] and Fig. 8-10b). The degree of planarization is expressed by Eqs. 8-1 and 8-3. A step-height reduction of 90–95% is achievable using CMP, yielding values of β in the 0.90–0.95 range. CMP is also capable of planarization of distances up to several millimeters, far in excess of that possible with any other technique thus far developed.[4]

Mechanical polishing, assisted by chemical action, produces such selective material removal. The process is performed by mounting the wafer face down on a carrier. The carrier is then pressed against a moving platen containing a polishing pad. The carrier itself is rotated. An abrasive-containing aqueous slurry is dripped onto the table, and centrifugal force distributes the slurry across the pad. This forms a thin colloidal layer of slurry which saturates the pad (Fig. 8-8). The combination of mechanical effects and chemical reactions results in material removal from the surface of the wafer.

CMP offers three key advantages for fabricating sub-micron ICs. First, for logic devices (such as microprocessors and ASICs), CMP makes it possible to achieve high manufacturing yields and high device speeds with multilevel interconnects having up to eight levels of metal. By planarizing IC topography, CMP avoids the problem of metal thinning over steep topographies in multi-level interconnect structures. Second, memory devices (such as DRAMs) benefit from CMP even though they use interconnect systems with only two levels of metal. For memories, the flat wafer surface after CMP increases the depth-of-focus budget available for lithography (see Chap. 13, Vol. 1, 2nd Ed.). This allows memory-circuit designers to employ smaller critical dimensions, thus reducing chip sizes without decreasing yield. Smaller chip sizes provide a significant cost advantage in the competitive realm of DRAM manufacture. Third, CMP is paradoxically capable of reducing (rather than increasing) defect-density.

Because CMP has become an enabling technology for allowing ICs to be further scaled, the applications of CMP have become pervasive. For example, by 2000, stacked-capacitor DRAMs have upwards of 9 polishing steps, and microprocessors may incorporate as many as 16 polishing steps. As a result, the manufacturability aspects of CMP (namely, the overall cost per CMP step - expressed in terms of $/wafer/polish step) become as important as CMP yield.

In the remainder of this chapter the following topics will be discussed: a) the terminology used to characterize CMP processes; b) the history of CMP; c) the mechanisms involved in CMP of oxides and metals; d) CMP equipment and consumables; and e) the details of CMP for various materials; and f) the challenges facing CMP processing. The latter include: inadequate understanding of the polishing mechanisms, low throughput, inadequate reliability of the polishing tools, degradation of removal-rate with usage-time of a polishing pad,

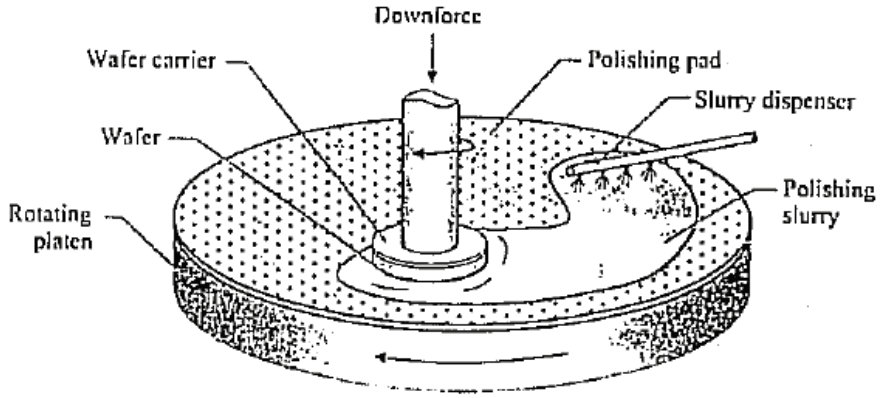


Fig. 8-8 Schematic of a rotary-table chemical-mechanical-polishing system.

non-uniformity of film-thickness-removal across the wafer, high cost of ownership, and high defect density.

8.3 THE TERMINOLOGY USED TO CHARACTERIZE CMP PROCESSES

Before beginning the discussion of the details of the CMP process, it is useful to define the metrics used to characterize the CMP process itself. This discussion is begun by listing the goals desired in a CMP process. They are:

1. The CMP process should have the ability to selectively reduce the step heights of features in the film being polished (i.e., a high degree-of-planarization must be possible);
2. Good planarity across a die (highest spot to lowest spot should be less than 150-nm);
3. Good film thickness uniformity within the wafer (±2–5%, 1σ);
4. High removal rates (100–140-nm/min), for high throughput and low cost-of-ownership;
5. Uniform removal rates from wafer-to-wafer;
6. Good selectivity and/or good endpoint technology to avoid under-polishing or over-polishing;
7. The CMP process should not add excessive defects (e.g., for 0.18-μm CMOS, fewer than 0.125-particles/cm^2 with a size larger than 0.06-μm are allowed at each mask level);
8. The overall cost of a CMP process should be as low as possible (expressed in $/wafer for each CMP process which is conducted).

While it is useful to list these goals in the qualitative manner as given above, the metrics for characterizing a CMP process can also be expressed in quantifiable form. These quantifiable forms are presented next.

8.3.1 CMP Removal-Rate (RR)

The *removal-rate* (RR) describes how rapidly a CMP process removes material from a non-patterned wafer (Fig. 8-9). This quantity is measured by performing a CMP process on a planarized, unpatterned wafer for a specific time, and determining how much material has been polished away. Generally speaking, the RR should be high so that the CMP process can be completed quickly enough to allow a satisfactory throughput. The RR is quantitatively expressed as follows:

$$RR = \Delta T/t \tag{8.3}$$

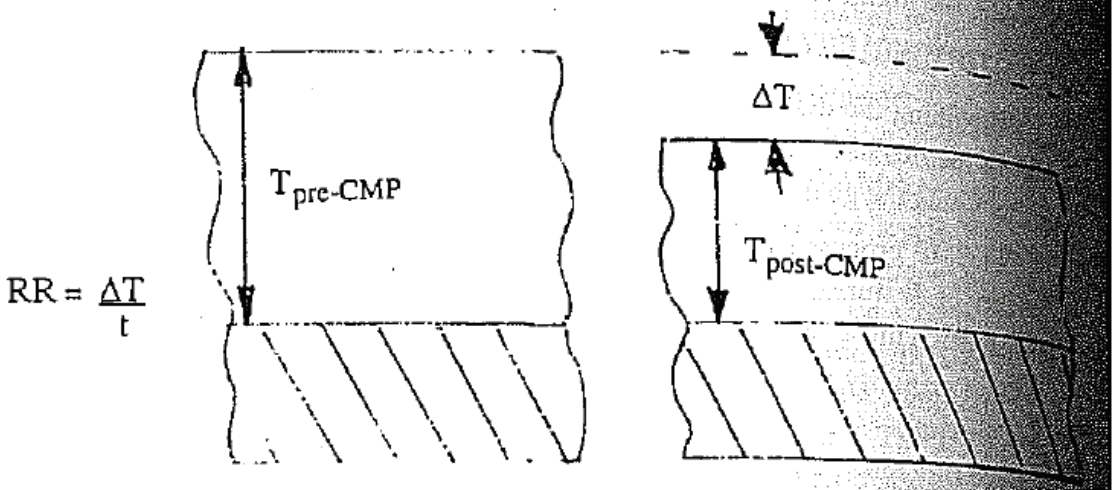


Fig. 8-9 Definition of *removal-rate* (RR).

where ΔT is the thickness of material removed in time t. RR is most frequently expressed in units of Å/min.

8.3.2 The Degree of Planarization (DOP)

In the ideal case, after CMP the surface should be completely planarized (see Fig. 8-1). The quantity that describes how well the CMP process meets this ideal is the *degree of planarization* (%DOP). The %DOP (see Fig. 8-10a) is quantitatively expressed as:

$$\%DOP = [(H_d^i - H_d^f)/H_d^i] \times 100\% \qquad (8\text{-}4)$$

where H_d^i is the height of feature being polished prior to carrying out the CMP step, and H_d^f is the height of the feature after the completion of the polishing step. Note that when the value of the %DOP is 100% this indicates that the feature has been completely planarized. Figure 8-11b shows a wafer surface on which the %DOP is 100% (complete global-planarization). Figure 8-11a shows a wafer surface on which the CMP process has not produced complete planarization (%DOP < 100%).

8.3.3 Within-Die Non-Uniformity (WIDNU)

The rate at which material is removed by CMP within a die may not be uniform. Such *within-die non-uniformity* (WIDNU) is quantitatively defined as the *standard deviation of the film thickness within a die*, σ_{WID}, divided by the *average value* of all "*n*" film-thickness measurements within the die, $\bar{x}_{WID}$:

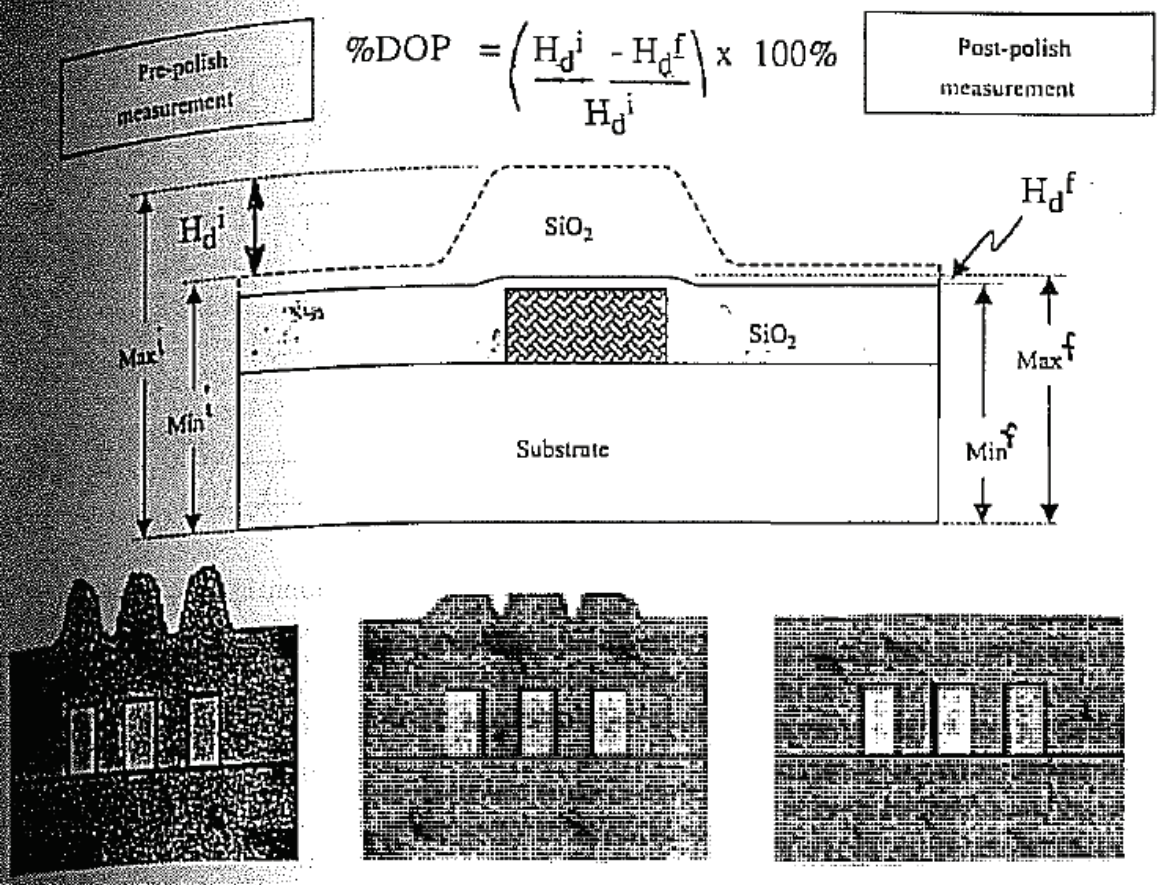


Fig. 8-10 Definition of *degree of planarization* (DOP).

$$\text{WIDNU} = \sigma_{WID} / \bar{x}_{WID} \qquad (8.5)$$

Note that σ_{WID} is the standard deviation of measurements within the die of some specific thickness, depending on the type of thickness non-uniformity being specified (e.g., the thickness variation of the oxide in the field between metal lines in an otherwise planarized oxide surface, see Fig. 8-11), and it is expressed quantitatively by:

$$\sigma_{WID} = \sqrt{[(x_1 - x_{WID})^2 + \ \ldots + (x_n - x_{WID})^2]/(n-1)} \qquad (8.6)$$

where $x_1 \ldots x_n$ are particular measurements on a die. Two examples of effects that cause WIDNU are shown in Figs. 8-12 and 8-13, respectively.

In Fig. 8-12c the step heights after CMP have not been reduced by the same amount within a die. Specifically, narrow, isolated-features tend to be planarized to a greater degree than wide, isolated-features, or narrow, densely-spaced-features. This form of WIDNU is termed *pattern-dependent-planarization*. The causes of such pattern-dependent-planarization are described in Sect. 8.5.3.2.

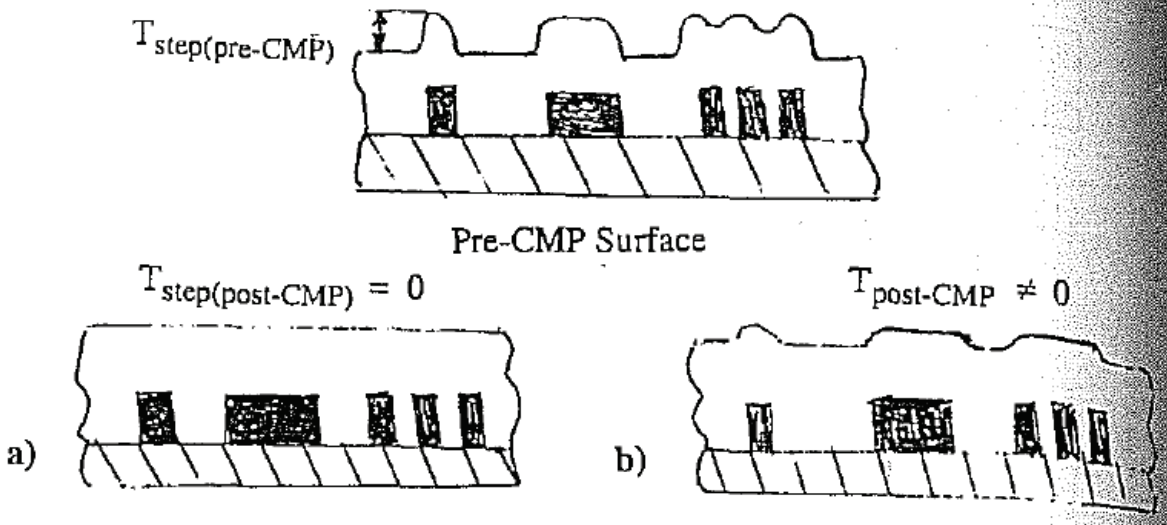


Fig. 8-11 a) Surface on which complete global-planarization has been achieved; b) Surface on which complete global-planarization has not been achieved.

In Fig. 8-13c, although the step heights of the oxide within a die (as shown in Fig. 8-13a) have been completely planarized, some film has been eroded in regions between the steps to a degree that the oxide surface is not planar throughout the die. This causes the distance from the top of the oxide to the planarized substrate surface below the oxide to differ within the die. This form of WIDNU is known as *dishing*, and the effects that cause such dishing are described later.

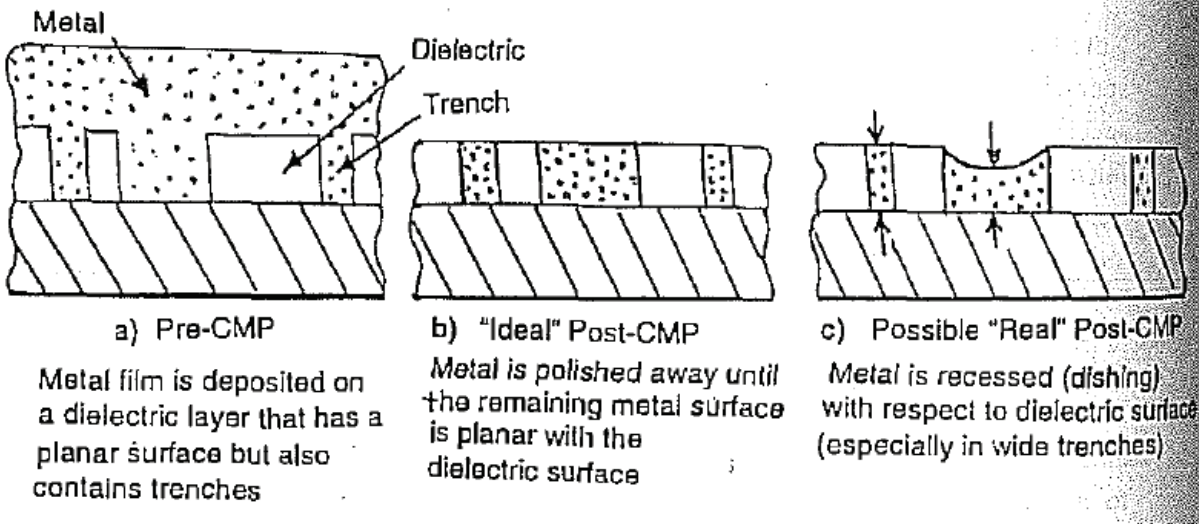


Fig. 8-12 Effect #1 that causes *within-die non-uniformity* (WIDNU).

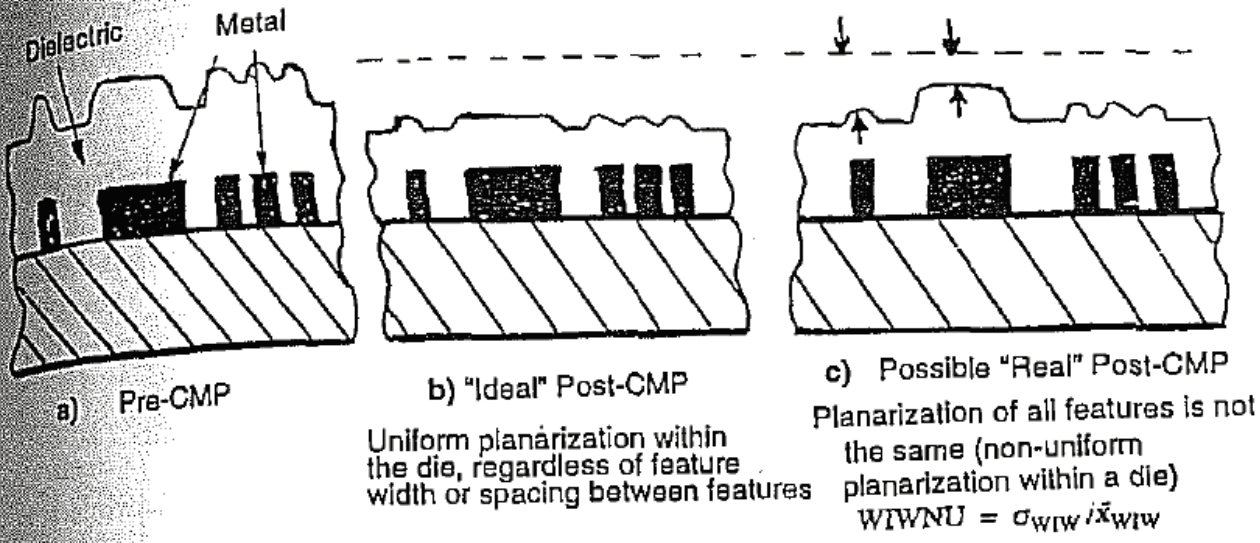


Fig. 8-13 Effect #2 that causes *within-die non-uniformity*, WIDNU (dishing).

8.3.4 Within-Wafer Non-Uniformity (WIWNU)

The *within-wafer non-uniformity* (WIWNU) is defined as the *standard deviation of the film thickness within a wafer*, σ_{WIW}, divided by the *average value* of all "n" film thickness measurements on that same wafer, $\bar{x}_{WIW}$.

$$WIWNU = \sigma_{WIW} / \bar{x}_{WIW} \tag{8.7}$$

Note that σ_{WIW} is the standard deviation of measurements within the wafer of some specific thickness, depending on the type of thickness non-uniformity being specified (e.g., the thickness variation of the distance between the otherwise planarized surface of the oxide and the surface of the planarized substrate below, see Fig. 8-14), and it is expressed quantitatively by:

$$\sigma_{WIW} = \sqrt{[(x_1 - x_{WIW})^2 + \ \ldots + (x_n - x_{WIW})^2]/(n-1)} \tag{8.8}$$

where $x_1 \ \ldots \ x_n$ are particular measurements on a wafer. Two examples of WIWNU are shown in Fig. 8-15.

In Fig. 8-15a the surface of the oxide after CMP is shown to be globally planarized across the entire wafer. However, the distance from the surface of the oxide to the planarized surface of the substrate below is non-uniform across the wafer. An example of two effects that can lead to such WIWNU are: 1) If the thickness of the substrate (wafer) itself is non-uniform, and the CMP process is performed with a completely rigid pad, this will cause WIWNU (see Sect.

$$\bar{x}_{WIW} = (x_1 + x_2 + x_3 + x_4)/4 \qquad WIWNU = \frac{\sigma_{WIW}}{\bar{x}_{WIW}}$$

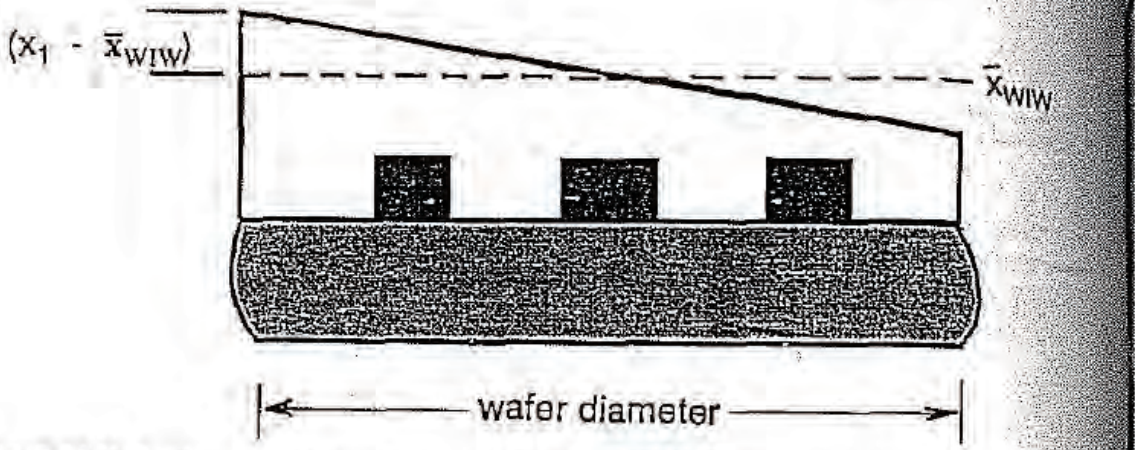


Fig. 8-14 Definition of *within-wafer non-uniformity* (WIWNU).

8.12.2); and 2) In rotary CMP polishing tools, WIWNU can be caused by the relationship between the angular rotation velocities of the wafer carrier and the polishing pad platen. This topic is discussed in Sect. 8.5.

Another effect that gives rise to a special type of WIWNU is shown in Fig. 8-15b. This figure shows that the wafer surface after CMP has an edge bead of material, caused by a smaller removal rate of material by CMP near the edge of the wafer. This topic is described further in Sect. 8.6.2.1.

8.3.5 Wafer-to-Wafer Non-Uniformity (WTWNU)

The *wafer-to-wafer non-uniformity* (WTWNU)) is defined as the *standard deviation of the film thickness from wafer-to-wafer*, σ_{WTW}, divided by the

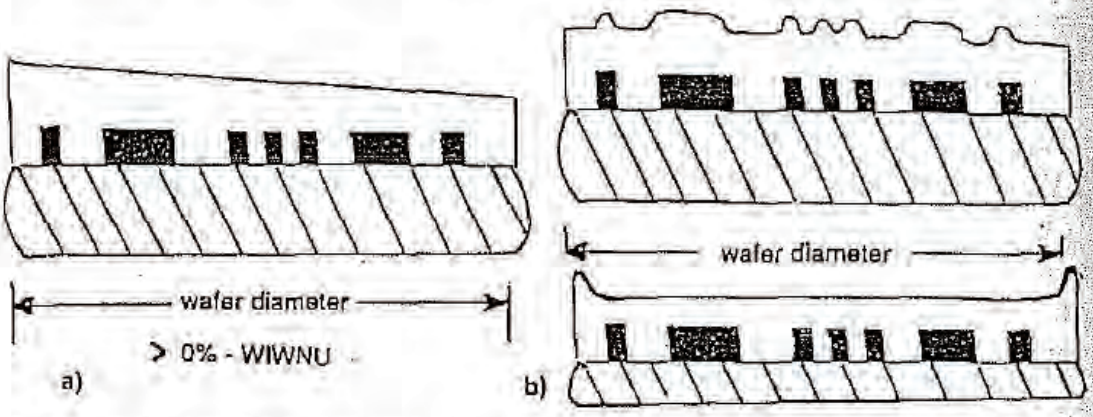


Fig. 8-15 a) Example of one type of WIWNU; b) Example of a second type of WIWNU caused by the *edge-bead-effect*.

average value of all the measurements of the film thickness measurements taken over all n wafers, $\bar{x}_{WTW}$.

$$WTWNU = \sigma_{WTW} / \bar{x}_{WTW} \quad (8.9)$$

Note that σ_{WTW} is the standard deviation of measurements from wafer-to-wafer for "n" wafers of some specific thickness, depending on the type of thickness non-uniformity being specified (e.g., the thickness variation of the distance between the otherwise planarized surface of the oxide and the surface of the planarized substrate below, see Fig. 8-16), and it is expressed quantitatively by:

$$\sigma_{WTW} = \sqrt{[(x_1 - x_{WTW})^2 + \ldots + (x_n - x_{WTW})^2]/(n-1)} \quad (8.10)$$

where $x_1 \ldots x_n$ are particular measurements on a wafer. An example of WTWNU is shown in Fig. 8-16. Two of the many causes of WTWNU include pad-aging (see Sect. 8.6.7), and slurry inconsistency from batch-to-batch (see Sect 8.6.3).

8.3.6 Efficiency of Planarization (EOP)

When an oxide is deposited over a metal line, enough oxide over the metal line must exist after the CMP step to minimize electrical cross-talk and parasitic capacitance between metal layers. This oxide thickness is determined by the IC designer. The CMP process-engineer must set up the CMP process so that after CMP is completed the oxide minimally meets the designers' requirements. The *efficiency of planarization* of a CMP process has an impact on the process integration choices that must be made to ensure that the minimum oxide thick-

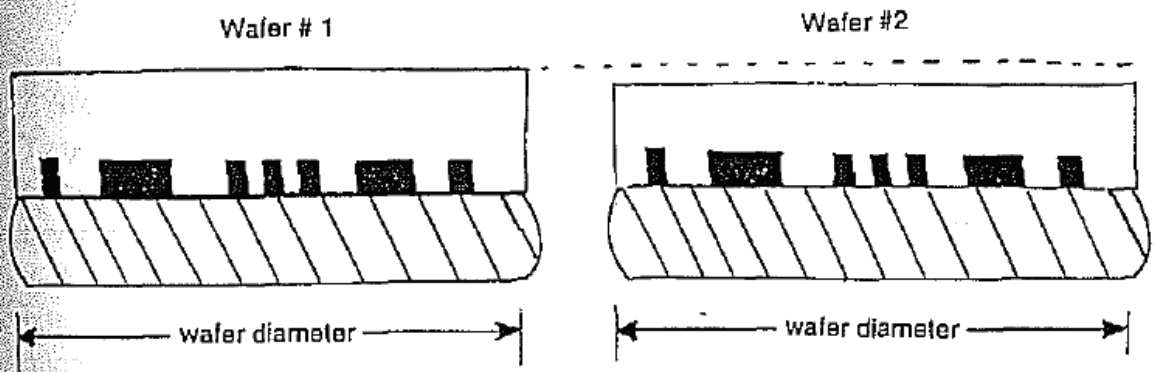


Fig. 8-16 Definition of *wafer-to-wafer non-uniformity* (WTWNU).

ness remains after CMP. The *efficiency of planarization* (%EOP) is expressed as:

$$\%EOP = [(T_r - T_d)/T_r] \times 100\% \quad (8.11a)$$

where T_r is the dielectric thickness removed by the CMP process on top of a feature (e.g., a metal line) and T_d is the oxide-thickness-loss at the center of a wide space between metal lines (see Fig. 8-17). The ideal case would be a CMP process with an %EOP = 100%. If this occurs, that means that complete planarization is achieved without the erosion of any oxide in the center of the wide space between metal lines. However, the %EOP is generally not 100%, and thus some erosion of the oxide between metal lines usually occurs.

It is the task of the CMP process engineer to determine how much total oxide must be deposited prior to the CMP step to ensure that the oxide on the metal lines is at least as thick as that specified by the IC designer. This thickness quantity is normally determined during tool setup, but it is useful to have an expression which allows the CMP engineer a way to estimate the correct value prior to performing the experimental task itself.

To allow the CMP engineer to calculate this thickness value more easily, the %EOP expression given by Eq. 8-11a is rewritten using different parameters. That is, Eq. 8-11a provides an expression for %EOP that makes it easier to visualize the concept of %EOP, but using it makes it cumbersome to calculate the required oxide deposition-thickness. Therefore, another expression is used instead of Eq. 8.11a for the deposited-oxide-thickness calculation. This alternate relationship is given as Eq. 8-11b (see Figs. 8-17c and 8-17d)

$$\%EOP = [H_{metal}/(T_{total} - T_{cov})] \times 100\% \quad (8.11b)$$

where H_{metal} is the height of the metal line being covered by the oxide, T_{total} is the total amount of the oxide film deposited over the metal line prior to CMP, and T_{cov} is the target amount of oxide that should remain over the metal line (as specified by the IC designer). Equation 8-11b is then rearranged to allow T_{total} to be determined:

$$T_{total} = (H_{metal}/EOP) + T_{cov} \quad (8.12)$$

As an example, if the %EOP is 50%, and H_{metal} = 0.5-μm, and T_{cov} is specified as 0.5-μm, then using Eq. 8-12 (with EOP = %EOP/100% = 50%/100% = 0.5), then the amount of oxide that would have to be deposited prior to the CMP process that would result in 0.5-μm of oxide over the metal lines after CMP would be T_{total} = 1.5-μm.

The least-expensive process involves depositing the smallest thickness of oxide (and this would occur if %EOP was 100%, corresponding to a T_{total} = 1.0-

$$\%EOP = [H_{metal}/(T_{total} - T_{cov})] \times 100\%$$

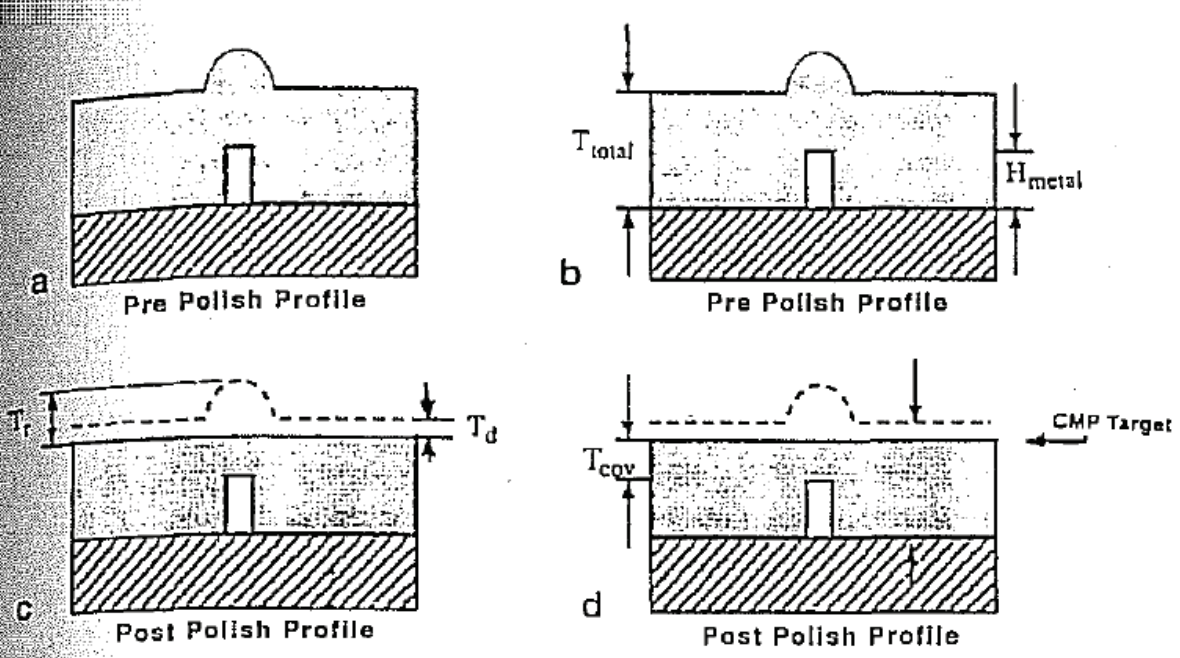


Fig. 8-17 The basic CMP process is one of reducing film-thickness selectivity over the wafer, ending with improved planarity. Here that metric is defined as *efficiency of planarization* (EOP).

μm). Thus, it can be seen that if the %EOP can be increased, the overall IC fabrication cost associated with forming the planarized oxide film over the metal steps will be reduced. For example, if %EOP could be increased to 80% (from 50%), Eq. 8-12 would predict that the $T_{total} = 1.1$-μm.

8.4 THE HISTORY OF CMP

The early history of CMP was chronicled by Fury.[5] He recounts that the process of chemical-mechanical-polishing originated at IBM in 1983. Research and development on CMP was exclusively performed within IBM until 1988. During those five years IBM closely guarded all information about the process. By the late-1980s several factors changed their secretive attitude. First, Intel microprocessors were selected for IBM's PC product line, and IBM decided to share some CMP data to help improve Intel's interconnect technology. Second, at that time the DRAM market was dominated by Japanese manufacturing companies. IBM was wary of having to rely entirely on foreign suppliers for such a critical component of its computer systems. Hence, they elected to help

strengthen a domestic DRAM supplier (Micron Technology, Inc.) by teaching them about the CMP process. Information about CMP was included in the process for the 4-MB DRAM that IBM sold to Micron Technology in the mid-1980s. In 1988, IBM agreed to disclose to SEMATECH its process scheme for multi-level interconnects, although they did not give all the details on CMP. They did, however, provide one of their CMP polishing tools (a Si-wafer polishing-tool sold by Westech, and marketed later as the IPEC Model 372, shown in Fig. 8-18). Motorola was introduced to CMP through the Power-PC microprocessor project, a joint venture between IBM, Motorola, and Apple Computer. In 1989 IBM published the first paper which publicly discussed the CMP of oxide films.[6] These events initiated the rapid proliferation of CMP throughout the IC industry.

The spread of CMP was hastened by other events, chiefly the downsizing of IBM from 1990–94. Many engineers with expertise in CMP left IBM and went to work for other IC companies who were seeking to implement the technology. The effort at SEMATECH involved improving the CMP tool-performance and reliability. Their work with Westech tools led to the development of the early popular CMP-tools, sold later as the IPEC 372 and 472 models. (Westech was acquired by IPEC.)

Although early development of CMP was centered in the U.S., worldwide efforts soon got underway. Polishing tools were initially offered by companies that had been building machines for polishing silicon wafers (Speedfam and Strasbaugh, as well as IPEC [Westech]). Overseas suppliers soon joined them (e.g., Ebara and Cybeq from Japan, and Wolters from Europe). Eventually the

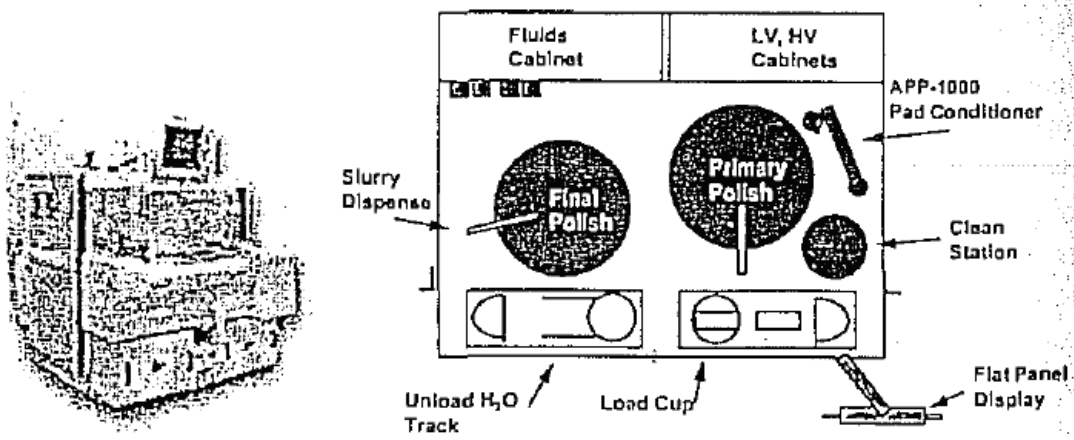


Fig. 8-18 The Speedfam 372 CMP tool.

larger semiconductor equipment vendors entered the market. Applied Materials introduced its tool in 1996, and Lam Research acquired On-Trak Systems, another commercial CMP vendor. The large number of vendors gave rise to some consolidation as well (an example of which was the merger of Speedfam and IPEC in 1999).

CMP was the fastest growing IC process technology of the 1990s, and it was also the most rapidly growing segment of the semiconductor process equipment market in that time frame. From 1995 to 1997 sales tripled, reaching $520 million in 1997. Sales were expected to exceed $1 billion by 2000.[7] The number of CMP polishing steps per year was also estimated to reach 125 million by that date. The first application of CMP was the planarization of dielectric films (mainly CVD oxide), but the number of applications has increased. Currently it is being used to planarize metal films (including tungsten, aluminum, and copper), as well as oxides and low-*k* dielectric films. By the end of the century, CMP had evolved from being a cleanroom-engineers' nightmare, to a critical process technology without which the production of ULSI chips with multilevel-interconnects might not be possible.

8.5 MODELING THE MECHANISMS OF CMP

The mechanisms of CMP are quite complex, and differ for each type of film being polished. Many variables also impact the removal rate, including: film type, pressure (force applied to the polishing pad), velocity, temperature, slurry feed rate, type of polishing motion, slurry particle size, pad hardness, and pad conditioning. The pressing need for production-worthy CMP capabilities forced early researchers to ignore the complexity of the process and go straight to an empirical trial-and-error approach. But, to logically optimize a CMP process, a more fundamental understanding of the mechanical and chemical forces at work is needed. While a complete understanding of CMP mechanisms has not yet been achieved, some insight has been gained. A brief introduction as to what is known about CMP mechanisms is presented here, divided into three categories: 1) the mechanical aspects of material removal by CMP (Preston's Equation); 2) the chemical and mechanical effects that cause planarization of oxides by CMP; and 3) the chemical and mechanical effects that cause planarization of metals by CMP.

8.5.1 The Mechanical Aspects of Silicon Dioxide Removal by CMP: (Preston's Law)

The CMP process for planarizing films in ULSI devices evolved from the technology of chemical-mechanical-polishing used to produce silicon wafers. That is, CMP in the Si wafering process is used to: 1) remove residual sawing and grinding damage that occurs when single-crystal silicon ingots are sliced into wafers; and 2) enhance the flatness (planarity) of the bare, starting-wafer surface.

However, the CMP process used in the Si wafering procedure differs significantly in two ways from the oxide-CMP process used to planarize oxide-layers a ULSI IC:

1. The oxide-CMP process is much more exacting than the Si-wafering CMP process. That is, the thickness of material removed in the CMP step of wafering is on the order of 10-μm $\pm$ 0.1 μm. However, the thickness of material removed in the oxide-CMP processes is only on the order of 0.5-μm $\pm$0.01 μm (i.e., about 5% of the thickness of the material removed in Si wafer polishing). Thus, the degree of control in oxide-CMP processes is much higher than that needed for Si wafer polishing.

2. When a *bare Si wafer is being polished*, the process involves oxidizing the Si surface with a chemical in the slurry solution, and then the slurry particles abrade away this oxide. That is, the CMP of Si wafers is carried out with a slurry containing fine slurry particles dispersed in a solution that *oxidizes the bare-Si wafer.* This is the *chemical component* of the CMP process. The abrasion of this oxide by the slurry-particles is the *mechanical component* of the process. The abrasive process exposes the Si beneath the oxide to the chemical in the slurry, which re-oxidizes the exposed Si, and so on. In fact, the abrasion of the oxide performs two functions: a) The polishing process is continued long enough so that the damaged region near the Si surface from sawing and grinding is eventually consumed by the oxidation and abrasion sequence; and b) Since the abrasion occurs more rapidly on "high" spots on the wafer surface, these "high" spots are eroded more rapidly than the "low" areas. Thus, the wafer-surface becomes flatter (more planar).

However, when an oxide is being polished in a ULSI IC (a process that is carried out in such applications as the CMP of the trench-oxide in shallow-trench-isolation structures (STI), and the CMP-step used to planarize the inter-level dielectric [ILD] and inter-metal dielectric [IMD] layers), the material to be polished *is already an oxide.* Hence, the

mechanism employed to polish the Si wafer (i.e., chemically oxidizing the Si surface, and mechanically abrading it away), is not an operative mechanism in oxide-CMP. A discussion of what are believed to be the phenomena that govern the removal of oxide films in oxide-CMP follows.

In 1997 Tomozawa stated that, like optical-glass polishing, much of the oxide CMP process was still an art rather than a technology. To qualify as a technology, the basic mechanisms of a process must be understood, allowing it to be logically optimized. However. at the time Tomozawa made his comment, the basic mechanisms of oxide CMP were not yet well understood.[8] The descriptions provided in this section briefly summarize what is known about the phenomena associated with oxide-CMP. Since silicon oxides used in IC fabrication are a form of glass, oxide-CMP is a specialized form of the glass polishing technology used to produce optical lenses. Thus, some aspects of oxide-CMP can be understood from the long history of glass polishing (which generally have been derived from trial-and-error, rather than from a fundamental understanding of the process). Optical-glass and SiO_2-films-on-Si-wafers are polished with a slurry consisting of abrasive silica particles (whose hardness is about equal to that of the oxide) suspended in an aqueous alkaline solution (typically KOH or NH_4OH, diluted in DI-water).

An empirical relationship that models the optical-glass-polishing removal-rate, RR, of flat-glass regions (defined as the thickness of glass removed per unit time), was developed by Preston, and is known as *Preston's equation* or *Preston's Law*.[9] This law states that the RR is proportional to the applied-pressure P and linear-velocity v at which a polishing pad moves relative to the optical-glass specimen, or:

$$RR = K_p P v \tag{8.13}$$

where K_p (*Preston's constant*) is a proportionality constant that is independent of pressure and velocity. The same law appears to apply to the rate of removal of oxide by CMP. (Typical values of P and v in CMP processes are 7-psi, and 30-40-ft/min, respectively.)

Note that Preston's equation describes the relationship between the RR of oxide and *mechanical effects*. (That is, mechanical effects are those caused the application of physical forces on a substance, namely in Preston's equation, the pressure exerted on the oxide and the relative velocity between the glass and the polishing-pad.) The Preston equation predicts that oxide-RR is linearly proportional to P and v, but gives no insight into other phenomena (i.e., such physical

and chemical factors as the glass-temperature, the chemical nature of the polishing medium, the hardness of the glass, etc). These effects are all put into the Preston constant, K_p, which contains the influence of such phenomena as glass hardness, Young's modulus, the polishing slurry, and the composition and hardness of the polishing pads. Despite the fact that Preston's equation sheds no light *per se*, on how these other phenomena impact RR, it is still a model that provides useful predictability concerning oxide-CMP. Hence, it is informative to discuss some further issues related to this equation.

The most important criticism of the Preston equation is that it was formulated empirically rather from first principles (i.e., it was not mathematically derived from a physical model of the glass-polishing process). Several authors have proposed equations from first principles that predict the dependency of oxide RR on pressure and velocity, including Brown and Cook (1984),[10] Liu *et al.* (1995),[11] Runnels and Eyman (1994),[12] Tseng (1997),[13] and Zhao and Shi (1999).[14] A critical review of the evolution of these models is found in Shinn *et al.*[15]

Many of the models (and supporting experimental evidence) indicate that the dependence of oxide-RR in CMP on pressure and velocity is not a simple linear relationship. Recent work has proposed multiple variations of Preston's equation of the form:

$$RR = K_p \, v^{\beta} \, P^{\alpha} \tag{8.14}$$

Where α and β are adjustable parameters dependent on the system to be polished. The values of these coefficients must be experimentally determined.

In the model of oxide RR of Zhao and Shi (which is the first model that also includes the effect of pad and wafer hardness), the RR demonstrates a sub-linear dependence on pressure given by:

$$RR = K_p \, v \, P^{2/3} \tag{8.15}$$

Figure 8-19 shows experimental data demonstrating the dependence of RR for oxide films.[16] Zhao and Shi also found that the pressure dependence of RR significantly influences planarization efficiency and pattern-density dependence. For instance, not all slurry particles in contact with a wafer necessarily remove material efficiently. A particle rolling against the surface removes little material, yet a particle held firmly by the pad slides against the wafer surface, changing the surface structurally and chemically. Sliding and removing requires a minimum local pressure, giving a revised polishing rate of:

$$RR = K_p \, v \, (P^{2/3} - P_{th}^{2/3}) \quad \text{when } (P \geq P_{th}) \tag{8.16}$$

and

$$RR = 0 \quad \text{when } (P \leq P_{th}) \tag{8.17}$$

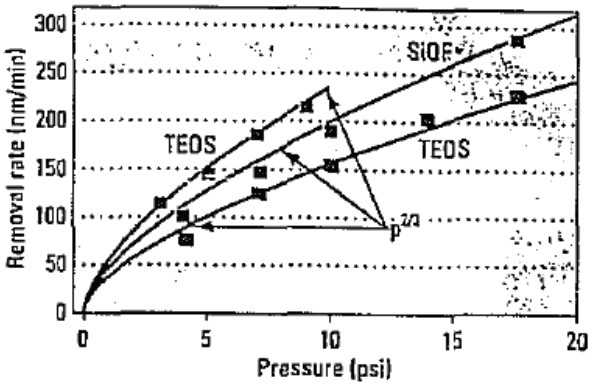


Fig. 8-19 Experimental data correlates well with the $P^{2/3}$ dependence of removal-rate on applied pressure.[16]

where P_{th} is the threshold pressure. In addition, they found that planarization efficiency also improves with P_{th} (Fig. 8-20), requiring less material removal to achieve a given planarization result.[16]

8.5.1.1 Model that Treats WIWNU Effects In Rotary-CMP Tools Based on Preston's Law: One of the limitations of CMP is that within-wafer-non-uniformities (WIWNU) occur, as defined in Sect. 8.3.4. The Preston equation states that the oxide-RR is linearly proportional to the *linear velocity* of the pad with respect to the oxide-surface being polished. Thus, one way to reduce such WIWNU is to ensure that the same linear velocity is experienced by all areas of the wafer. In rotary-CMP tools, the wafer carrier rotates, as does the platen that holds the polishing pad. If the pad did not rotate, and only the carrier rotated about its axis, the oxide-RR at center of the wafer would drop to zero, since there would be no relative motion between the wafer and the pad at that point. In addition, the RR would increase radially from the center to the edge of the wafer.

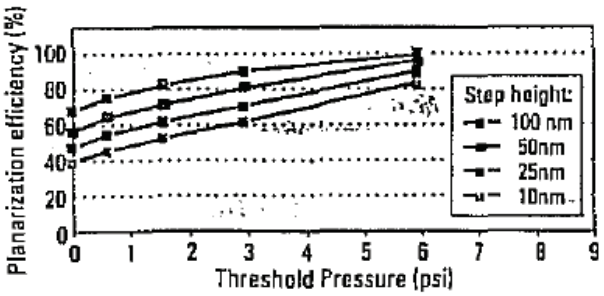


Fig. 8-20 Increasing the threshold pressure improves planarization efficiency for various final step heights (500-nm initial step height).[16]

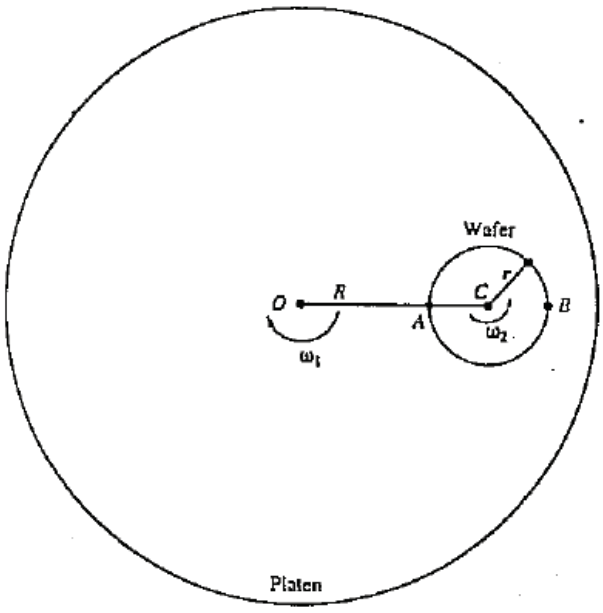


Fig. 8-21 Plan view of a single-carrier single-platen rotary-polisher, showing relative rotary motions of both the carrier and the platen.

Now assume that the platen and the wafer carrier both rotate. In this case (referring to Fig. 8-21), it can be noted that the centers-of-rotation of the wafer-carrier and the platen are offset from one another. If the wafer-carrier and the platen have the same rotational frequency, and are rotating in the same direction, point B of the outside edge of the wafer in that figure experiences zero relative velocity, and the point A on the inside edge of the wafer experiences the greatest relative speed of the surface of the wafer to the pad. Thus, the instantaneous removal rate at the edge of the wafer is greatest when the edge is directly between the two centers (point A), and zero when in the extreme position (point B). However, since all points on the edge of the wafer pass through points A and B at a constant frequency, the time-average RR at the edges will be the same as at the center of the wafer. Thus, it is the average-radius from the center of the platen, r_{av}, that defines the distance the wafer travels and the average speed of travel, S_t, for purposes of the total film removed and RR calculations:

$$S_t = 2\pi r_{av} f(\text{rpm}) \qquad (8.18)$$

In practice, within-wafer-non-uniformity of RR is observed in rotary-CMP tools (e.g., WIWNU). One form of such WIWNU is a spoke-like pattern due to vibration harmonics. Driving the carrier in rotational synchronization with the platen usually causes the worst vibrations. Hence, the wafer-carrier is usually driven at a constant difference in frequency to the platen. Since the correct wafer-carrier rotational-frequency for perfect circular symmetry is not known, it must be experimentally determined.

8.5.2 Models Which Describe Factors that Impact Oxide-RR in CMP Through the Preston Constant

There are many factors that impact the oxide-RR in CMP besides the applied pressure and velocity. These are expressed in Preston's Law by changes in the Preston constant. Here two models are given that have been proposed to explain the dependence of oxide-RR on other phenomena besides pressure and velocity.

8.5.2.1 Impact on the Oxide-RR from the Electrochemical Interactions Between the Slurry Particles and the Oxide Surface: A model was proposed by Cook in which he postulated that a mechanism for oxide-removal by CMP involved contact-mediated reactions with the solvent (water).[17] That is, local events in CMP may be considered to be the approach to contact of two opposing asperities: 1) pad plus adhering particle; and 2) the oxide surface. Asperity-contact causes reactions in the vicinity of the contact, and these result in mass removal from the oxide. The depth and width of the swath removed is determined by the size of the slurry particle and the chemical-reaction-rate, which is largely controlled by the interaction force between the contacting surfaces. The length of the swath gives an indication of the duration of the asperity contact.

For the case where the slurry particle and the surface are of the same material (as occurs in oxide-CMP with a silica-based slurry), the forces due the Zeta-potential act to repel the silica particles from the oxide surface, while the Van-der-Waals forces cause the oxide surface to attract the slurry particles (see Sect. 8.7.1 for a discussion of the Zeta-potential and the Van-der-Waals forces). The curve that graphs the force-versus-distance for such a case of opposing forces between the particles and the wafer-surface exhibits a maximum repulsion at some characteristic separation distance. This may be perceived as an activation-energy-barrier for closest approach. The particle must thus have a kinetic-energy greater than this activation-energy-barrier in order for direct contact to occur between the slurry particle and the oxide-surface. A reduction in this activation-energy-barrier should result in a higher probability of contact, and therefore cause an increase in oxide-RR.

Based on this model, Cook suggested that an increase in RR could be achieved by using a slurry whose particles were different than that of the oxide (e.g., use CeO_2 slurry particles instead of silica), and in which the Zeta-potential had a value that caused the slurry particle and the oxide to attract one another. This would strongly increase the probability of contact between the slurry particle and the oxide-surface and thus increase RR.

8.5.2.2 Dependence of Oxide-RR in CMP on Dielectric Hardness: Dielectric materials vary in hardness from the very-hard and inelastic silicon-nitride, to softer dielectrics such as doped-oxides and spin-on-dielectrics (SODs). The RR of dielectrics has been found to increase with a decrease in their hardness. Figure 8-22 shows that as the hardness of fluorinated-silicate-glass (FSG) is decreased, the RR increases.[18] (The hardness of the FSG in this study was decreased by increasing the concentration of fluorine that is incorporated in the oxide.) In another work, the RR of PSG and BPSG was studied as a function of the P-concentration in a P-doped oxide-film, and as a function of the B- and P-doping in a BPSG oxide film (Fig. 8-23).[19] Higher doping is seen to cause the RR to increase. This is correlated with oxide hardness, since the incorporation of P_2O_3 and B_2O_5 into an SiO_2-network causes a softening of these oxide-films.

8.5.3 Models Which Describe Chemical-Mechanical Phenomena that Give Rise to Planarization Phenomena Associated With CMP of Oxides

Various models have been proposed to explain how chemical phenomena act in concert with the mechanical forces affecting oxide-RR (as expressed by Preston's Law). These models also provide insight as to why oxide-CMP can selectively remove the "up-regions" of an oxide film on wafers faster than the "down-

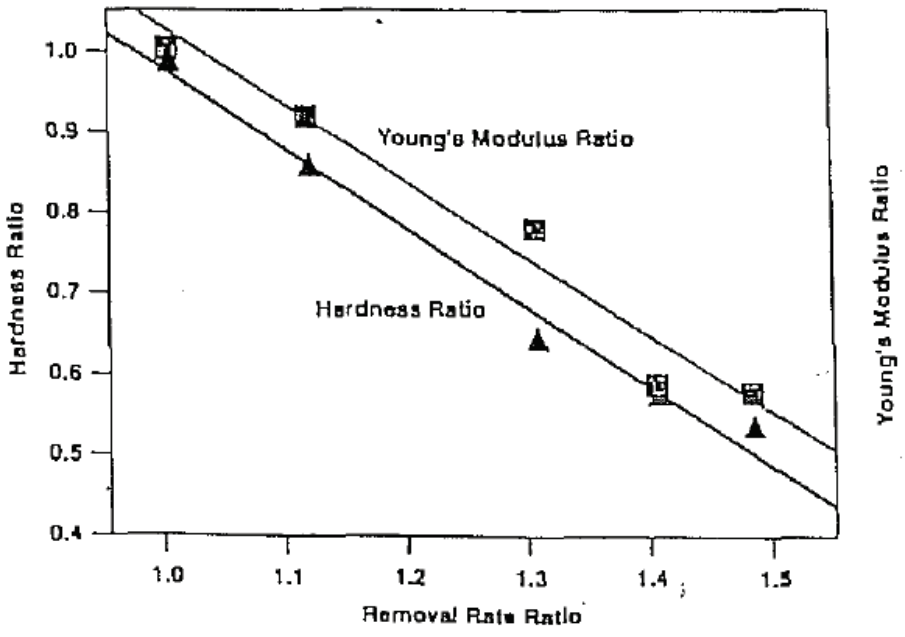


Fig. 8-22 A linear relationship exists between the CMP-removal-rate and the hardness and elastic modulus of F-SiO_2. Removal-rate, hardness, and elastic modulus are all normalized with respect to those parameters of undoped TEOS-oxides.[18]

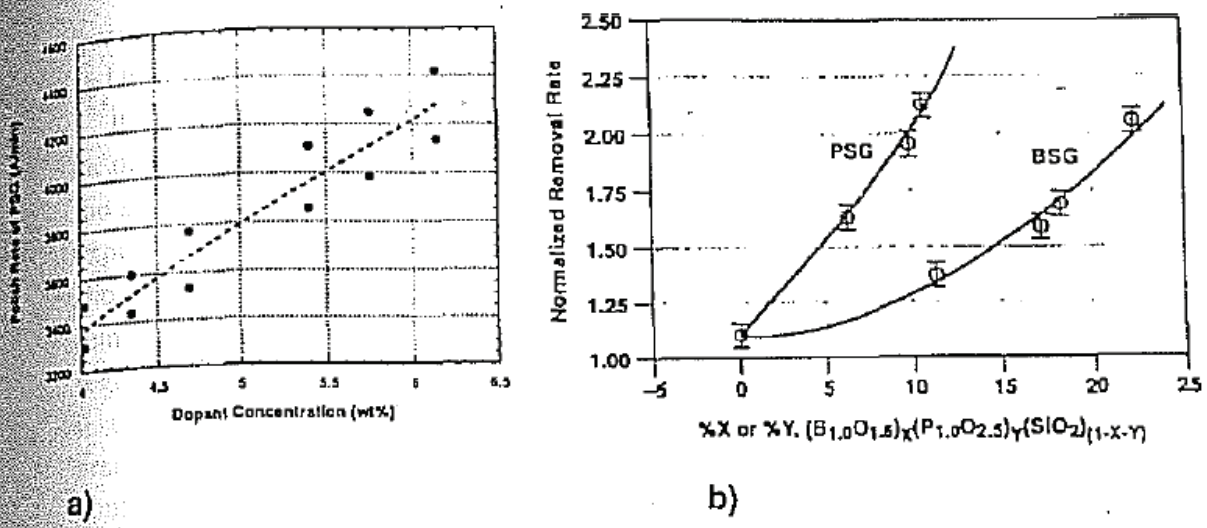


Fig. 8-23 a) Removal-rate of PSG as a function of phosphorus-concentration: pressure 4-psi, table speed: 40-rpm, carrier speed: 643-rpm; b) Removal rates of P-doped oxides and B-doped oxides as a function of P- and B-doping-concentration in the oxide.[19]

regions" (thereby allowing planarization of the film to occur). Some ideas from these models are discussed in this section.

Izumitani proposed that optical-glass-polishing "proceeds by the formation of a hydrated-layer by means of a chemical-reaction between the glass-surface and water and then the removal of this hydrated-layer by the abrasive particles."[20] This proposed mechanism is consistent with the observation that the polishing-rate is extremely low without the presence of water to cause the hydration. (Polishing with a slurry of silica-particles in a non-aqueous solution [e.g., in oil] exhibits substantially lower polishing rates.) The hydrated-layer that thus forms on the oxide-surface generally has a much lower hardness and mechanical strength than the SiO_2 beneath it. Therefore, plastic-deformation results in mechanical removal of the hydrated-layer (by a plowing effect) when it is rubbed by the abrasive particles.

The same type of mechanism is expected for oxide-CMP. Removal of the oxide takes place by a combination of mechanical and chemical action, using an abrasive slurry dispersed in alkaline solution (Fig. 8-24). Since silica-particles are used in the slurry of oxide CMP, they have about the same hardness as the SiO_2. Thus, the unreacted SiO_2 is not appreciably removed (nor damaged) by the plowing effect when it is rubbed by the silica particles in the slurry. However, the silica-slurry-particles still cause the oxide to plastically deform. This deformation

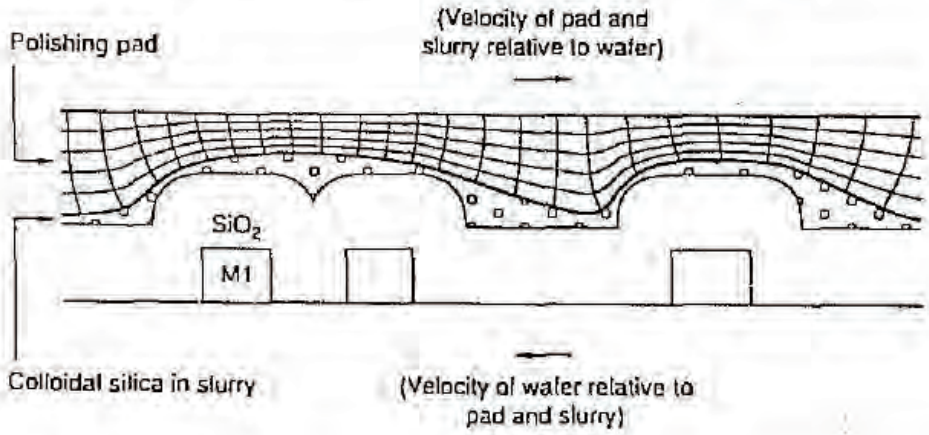


Fig. 8-24 Depiction of polishing-pad compression and localized pressures applied to topographic features. Pressures are greatest at the top-leading-edges of steps, and are lowest in the wakes of the steps.

is likely to be assisted by frictional heating. It has been determined that oxide readily undergoes hydration (water entry) if it is plastically deformed by a diamond indenter. When entered by water, the hardness and mechanical strength of the oxide is generally decreased. Thus, during CMP, it seems that hydration of the oxide film surface is caused by the plastic deformation of the oxide by the abrasive particles. The degree of deformation is increased as the pressure of these particles on the surface is increased, and by frictional heating. The softer hydrated-layer is then removed by the plowing action of the abrasive particles, which increases as the linear-velocity of the polishing pad is increased. These conclusions are substantiated by Preston's Law. The properties of the pad also appear to play an important role in CMP. It is probable that the polishing pad has to be soft enough to grip the abrasive particles, but hard enough to transmit the pressure to the oxide. It is generally agreed that some sort of chemical mechanism (such as SiO_2-dissolution) is a cofactor in causing oxide-CMP. Various chemical processes have been proposed.[21] The larger polishing rate in higher-pH slurry is often cited as evidence for silica-dissolution.

8.5.3.1 Model of the Mechanism that Produces Planarization of Steps in Oxide CMP: While the above model gives some insight into how oxide is polished with a silica-based slurry, it does not address the issue of how planarization of the oxide by such polishing is achieved. To obtain planarization of a film with topography, the surfaces of elevated features ("up-areas") must be eroded faster than those at lower elevations ("down-areas"). The selective removal of material in "up-areas" is empirically observed to depend on many factors, including the

feature dimensions, the structure (and surface condition) of the polishing pad, and the height of the step being removed

It is important to note that the case being discussed here involves the CMP of an oxide layer that proceeds only until the oxide surface is planarized, and that the surface of the wafer after the CMP step is *homogeneous* (i.e., only oxide is exposed over the entire wafer surface at the conclusion of the CMP step, as shown in Fig. 8-25a). This differs from the case when a film is polished until it is removed from the surface in some regions, but is allowed to remain in recessed regions of the wafer surface after the CMP step. In that case (which is described in a later paragraph), the wafer surface is *heterogeneous*, insofar as some exposed areas of the wafer contain the polished film residing in recesses, and others contain material that underlaid the film that was polished (Fig. 8-25b). Note that in the second case, the CMP process is essentially a *blanket-removal-process of a planar film* because the film being removed is already roughly planar when the CMP process is started.

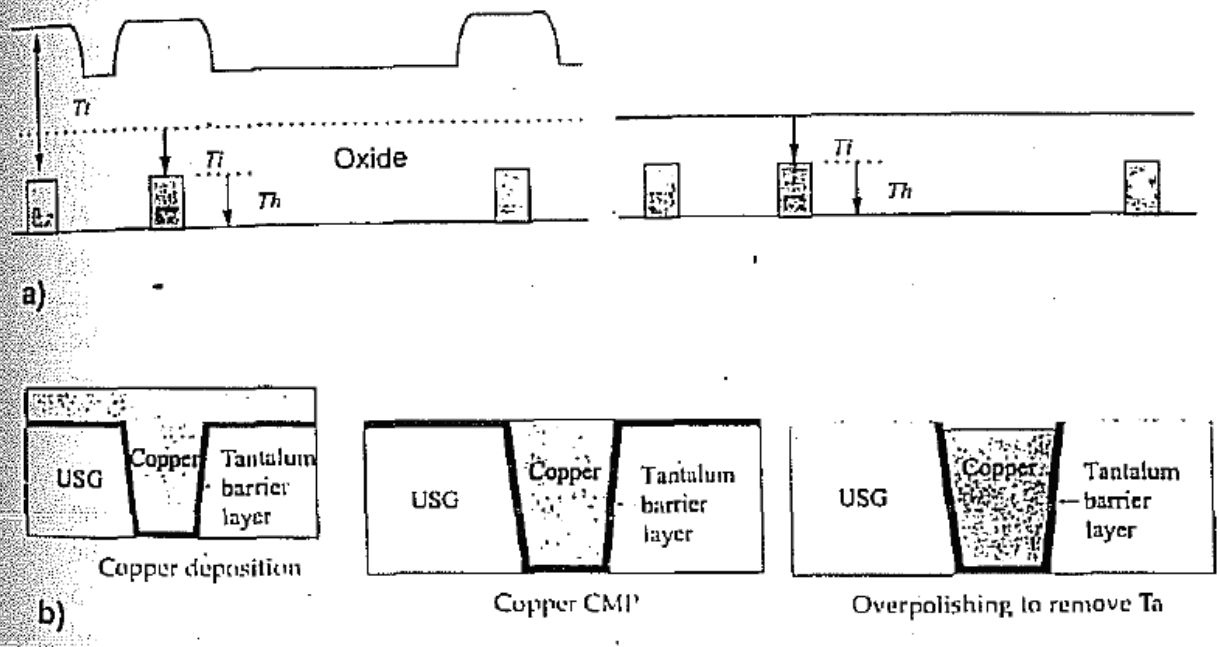


Fig. 8-25 The process of removing material by CMP falls into two categories: a) Removing homogeneous material and stopping while the removal of the surface film is incomplete (e.g., planarization of an ILD oxide); b) Removing deposited material, but continuing until all material is entirely removed on some regions of the wafer surface, while in other regions it remains recessed below the surface (e.g., Cu-damascene feature formation by CMP).

Selective removal of material that protrudes above the surrounding topography (as shown for the case of Fig. 8-25a) is thought to arise in part from the compliant nature of the polishing pads.[22] Figure 8-24 depicts the localized compression occurring within a pad as features on the wafer surface are driven across it. The pad applies pressure to the wafer surface through the slurry, with the largest pressures being exerted where the pad is most compressed. Higher pressure is applied to elevated regions ("up-areas" - Fig. 8-26) and the leading edges of the features that are moving against the compressed pad, compared to lower-lying regions ("down-areas") and trailing edges.[23] The difference in local applied pressure causes the material removal rate to be fastest at the upper corner of the leading edge of a step. Figure 8-27 shows how a feature would be planarized as the leading edge is eroded away as a function of time during a CMP process. (In the next section it will demonstrated how this phenomenon - in concert with wafer rotation during polishing - causes smaller features to be rounded-off and polished-aster than wider features.)

As the steps erode, the "down-areas" are exposed to increasing pressure, and the removal-rate there increases (approaching that of the "up-areas"). As this happens, the rate of step-height-reduction decreases toward zero. In the latter stages of the polishing process the selectivity is so low that small residual steps remain, and in practice cannot be fully removed by continued polishing.

In the second case described above (Fig. 8-25b), the material of the blanket-deposited layer is removed non-selectively by CMP. This non-selective removal continues until all of the film is polished away from the areas of the substrate, and only film material remains in the trenches. However, if the CMP process is continued beyond that point, material within the trenches will continue to be removed. This removal will be greater in wide trench areas than in narrower trenches. The result of such removal from wide trenches gives rise to a CMP problem called *dishing*. The phenomenon of dishing and methods developed to mitigate it are also discussed in other parts of the book. In Sect. 8.12.1, the general aspects of dishing are covered, and the impact of dishing effects on the fabrication shallow-trench-isolation structures is found in Chap. 9. Chapters 15 and 16 provide additional information on the impact of dishing on Cu-CMP and on the CMP processes used to polish inlaid metal lines in damascene processes.

8.5.3.2 Pattern-Dependence of the Planarization Rate in CMP Processes: The degree of planarization (DOP) of features for the CMP processes depicted by Fig. 8-17a has been found to depend on the feature sizes and the density of features within a region of a die. This DOP dependency on feature size and pattern density is one cause of WIDNU. The reasons for such WIDNU are described here.

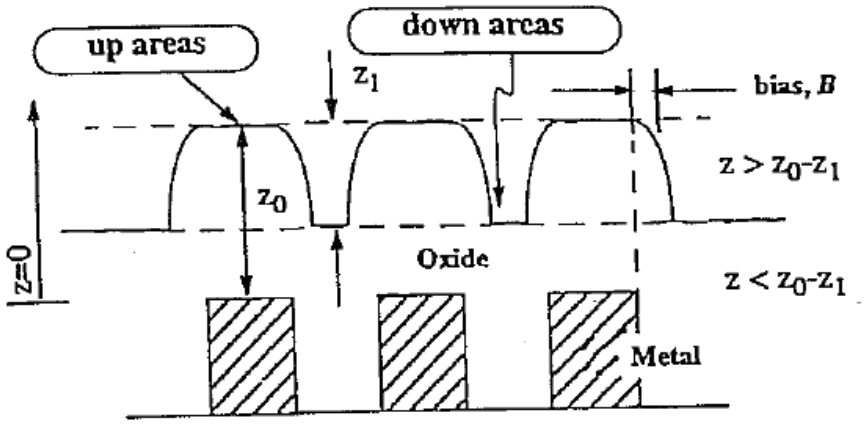


Fig. 8-26 Cross section of conformal dielectric on patterned metal, showing the "up-regions" and "down-regions" of the dielectric film.

The first question to answer is why narrow, isolated-features tend to planarize faster than wide isolated features. From Fig. 8-27 we see that the leading edge of a step will be removed as a function of time in an oxide CMP process. However, in a process in which the wafer carrier is rotating, all top edges of the feature being polished become the "leading edge" as the feature is rotated by the carrier. As a result, the entire feature is planarized according to the sequence shown in Fig. 8-28a. If the feature is wide, the leading edges will be polished away, as shown in Fig. 8-28b, but the material in the center of the wide feature will not be polished as quickly as material at the feature edges. As a result, wider, isolated-features will planarize more slowly than narrow, isolated features. This will cause such WIDNU as shown in Fig. 8-13.

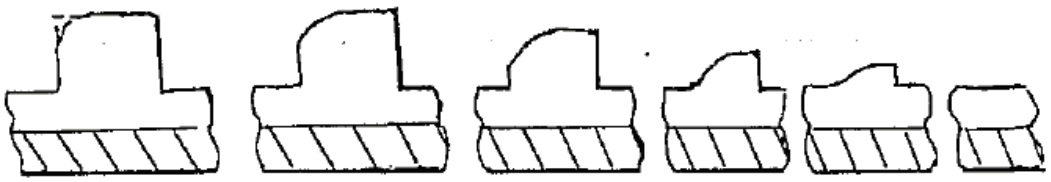


Fig. 8-27 A depiction of how the leading edge of a feature of an oxide step (i.e., the edge that is pressed against the pad by the combined movements of the carrier and platen), is eroded by CMP.

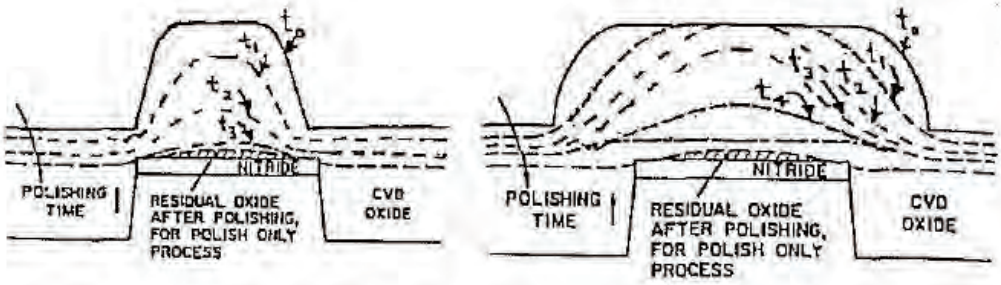


Fig. 8-28 Schematic cross-section of a narrow and wide feature of oxide being polished by CMP, as a function of time. Since the leading edges are polished most rapidly (and the feature continually rotates, so that all edges become the leading edge for the same amount of time), the narrow feature is planarized more rapidly than the wide feature.

The next question is: Why are densely-spaced narrow features planarized more slowly than isolated, narrow features? As first discussed by Burke,[24] the time required to planarize an "up-area" structure as shown in Fig. 8-26 (or to reduce its height to some fraction of its initial value) is dependent on the "down-area" adjacent to it. Specifically, the time to planarize a step depends on the width of the "up-area," the width of the "down-areas" between "up-areas," and the ability of the polishing-pad to rigidly span that distance between "up-areas." The stiffer the pad, the more rigid the span - and the slower the rate of material removal from "down-areas." This gives good planarization. The softer the pad, the less able it is to bridge larger distances. Thus, more material is removed by softer pads in "down-areas" between adjacent "up-areas." This gives poorer planarization. When features to be polished are separated by distances greater than the leveling-distance of the pad, there is no selectivity between "up-areas" and "down-areas," and planarization does not occur. In practice, some selectivity in RR occurs between "up-" and "down-areas," but this selectivity is not complete. Ideally, the leveling-distance of a pad should be large enough to planarize the widest "down-areas" on a chip. On the other hand, the greater the leveling-distance, the more a pad will respond to wafer-scale non-flatness parameters, such as TLV and bow. This results in non-uniformity of the polished film thickness across the wafer (a subject discussed in more detail in Sect. 8.12.2).

In summary, planarization will occur most slowly for step heights in an oxide over the largest or densest interconnect structures adjacent to wide unpatterned

field areas, scribe lines, and wafer edges. It will occur most rapidly for step heights in the oxide over smaller arrays and narrow, isolated features. It is also necessary to specify an acceptable degree of planarization, because in the latter stages of the CMP process, material is removed nearly as fast from "down-areas" as from "up areas." Step heights no longer decrease at appreciable rates.

Like metal-CMP processes, oxide-CMP in a production environment is often a multi-step process. When in-line endpoint-detection is employed (as will be discussed in Sect. 8.6.9), two or more polishing steps may be needed to obtain the correct oxide thickness. In addition, a separate final-polish (or buff) procedure is frequently used to provide light polishing and to remove slurry particles and small wafer-defects. A small amount of oxide can also be buffed away with this buffing procedure (typically 30–40-nm), which aids in removing defects (especially residual slurry particles and the small scratches created by the primary polish). Use of a buff procedure is one of the reasons why CMP can actually *reduce* defect-densities (including eliminating some types of defects created in other processes), rather than increase them.[25]

8.5.4 Metal-CMP Mechanisms

The mechanisms involved in the CMP of metals which cause planarization of the metal film surface, follow a relatively simple three-step sequence proposed by Kaufman.[26] This three-step sequence is as follows (Fig. 8-29):

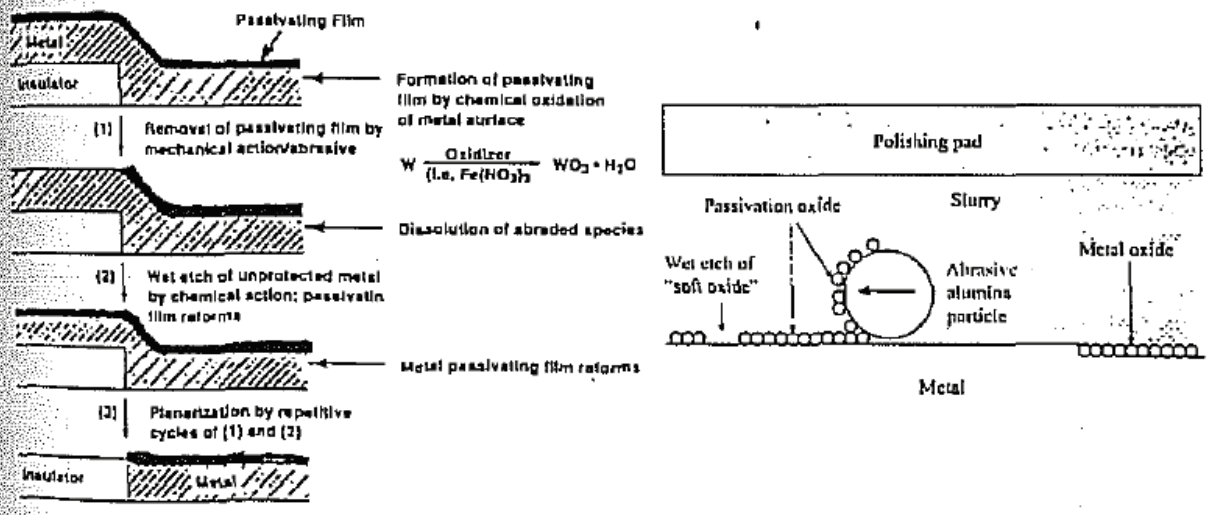


Fig. 8-29 Schematic representation of the model of CMP of metal films.[26]

1. A passivation layer (e.g., an oxide of the metal) is chemically formed on both the "up-areas" and "down-areas" of the metal film, due to the oxidation potential of the solution. (For effective polishing, it must be possible to both oxidize the metal and closely match the hardness of the slurry abrasive to the hardness of the material being removed.)

2. This freshly formed oxide film is softer than the metal itself and is mechanically removed from the "up-areas" by the pad/abrasive. (While the "down-areas" are not contacted due to the partial flexibility of the pad.)

3. The corrosive nature of the slurry re-grows the passive film on the "up-ares," while the corrosive elements are blocked from the metal in the "down-areas" by the passivation film.

The "up-areas" are continually corroded and polished until the pad flexibility allows it to contact the "down-areas." At that point, the planarization process again depends on the hardness of the pad and the down-force applied to the wafer. Thus, the corrosion properties of the slurry are chosen to convert the metal surface into a film that is soft enough to be abraded by the pad-plus-abrasive-material, yet hard and impermeable enough to prevent outright etching of the metal.

The goal is to find slurries with just the right combination of chemical and abrasive aggressiveness to produce high removal-rates, good-planarity, local-uniformity, and high-selectivity. Although the metal-CMP-model is simple to understand, successfully carrying out the above process for specific metals can be challenging. The slurry's chemical-composition tends to play a more important role in metal-CMP than in dielectric-CMP. As a result, it is possible to achieve fairly high metal-to-oxide selectivity (e.g., 120:1 for tungsten to silicon dioxide). However, the slurry solution may also etch the metal-oxide that is formed. A careful balance between etching and oxidation must therefore be maintained. If etching dominates, then the metal will etch isotropically and planarization will not be obtained. The metal-deposition-step may also have to be modified to prevent metal deposition near the wafer edge. Otherwise, metal residue may re-main on the wafer's beveled edge after CMP, causing yield and reliability problems.

Pourbaix diagrams are often used to identify which pH-range is most ideal for CMP of metals.[27] *Pourbaix diagrams* are graphical representations of the domains of stability for different oxidation states of metals in aqueous solutions with respect to pH and electrochemical potential (i.e., contour maps in a solution potential versus pH). For example, the Pourbaix diagram for W (shown in Fig. 8-30) indicates that in order to form an un-dissolved protective WO_3 layer over the W, the pH of the solution should be below 4, and have a slightly positive oxida-

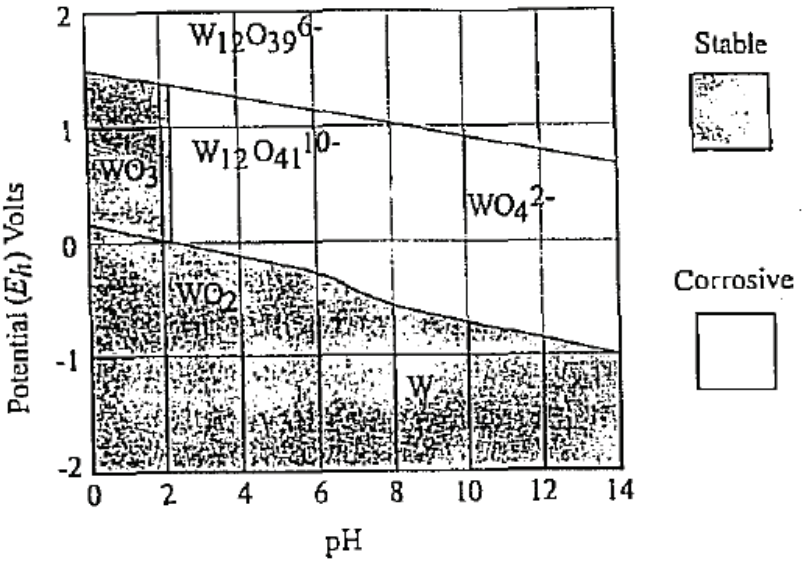


Fig. 8-30 Tungsten solubility versus solution-pH (Pourbaix) diagram.[27]

tion potential. For this reason, most tungsten slurries have a low pH, typically with pH-values between 2 and 3.

Unlike oxide CMP processes (whose purpose is to remove a specific amount of the dielectric layer), the metal-CMP processes require some degree of overpolish to ensure that metal is completely removed from field areas. The challenge here is to optimize such overpolish so that process time and oxide loss are minimized (i.e., to maximize throughput), while compensating for any non-uniformities from prior steps (and without causing any new uniformity problems from dishing or erosion).

It should also be mentioned that one aspect which makes metal-CMP processes easier to control than oxide-CMP processes, is that commercially available endpoint methods are much more sensitive for metal CMP. That is, metal-CMP processes are carried out by polishing-back blanket-metal-layers until all of the metal on the wafer surface is removed (but the metal in the via holes or within the trenches remains). Once the metal on the top surface is removed, the underlying oxide in the field is exposed. Consequently, this endpoint is easier to detect by monitoring a variety of phenomena associated with the clearing of the metal over the oxide, such as: a) a change in surface reflectivity; b) a change in surface temperature; or c) a change in the torque of the platens or wafer carrier (due to the different frictional coefficients of the materials being polished). In any case, the optimum metal-CMP process would have a clear endpoint-signal -followed by a short overpolish step (to remove any

metal remaining in dished areas of the wafer, or in surface scratches). The result should be a surface in which the metal and the barrier-layer is completely removed, with a minimum of the underlying dielectric being eroded, and where a minimum amount of dishing is produced.

8.6 CMP EQUIPMENT

There are number of different tools and consumables used in a complete CMP process, including:

1. CMP polishing tools;
2. Slurries;
3. Pads;
4. Slurry distribution systems;
5. Endpoint detection systems;
6. Cleaning tools;
7. Metrology tools.

These items are described in the following sections.

8.6.1 CMP Polishing Tools

The basic polishing tool used for CMP evolved from the machines used to polish bare-Si-substrate-surfaces. When the idea for CMP was first conceived at IBM, they purchased silicon polishing tools (used by wafer manufacturers) from Westech (now SpeedFam-IPEC) and Strasbaugh. The basic configuration of the wafer-polisher was retained, but the tools were modified to make them suitable for developing CMP. As mentioned earlier, in such polishers the wafer is mounted upside down on a wafer-carrier, and the carrier is pressed downward against a polishing pad (Figs. 8-31 and 8-32). Both the wafer carrier and polishing pad are rotated. Slurry is dripped onto the polishing pad, and it becomes saturated with slurry. The pad thus drags slurry beneath the rotating wafer carrier, where it can polish the wafer surface (as described earlier in the section on mechanisms of CMP). The first polishers adapted for CMP had a single wafer-carrier.

The CMP-polisher differs from the wafer-polishing tool in a number of ways. In polishing silicon-wafers, several tens-of-microns of material are removed, while the CMP-process removes only about 0.5–1.0-μm of material. The uniformity of material removal across a wafer in CMP is also much more stringent than in silicon-wafer polishing (±0.5–0.1-nm for CMP). Furthermore, CMP-tools require a higher-level of automation, throughput, reliability, and critical process-parameter-control. Thus, despite more than 30 years of experience in silicon-wafer polishing, new processes and equipment had to be developed by the IC industry for carrying out CMP.

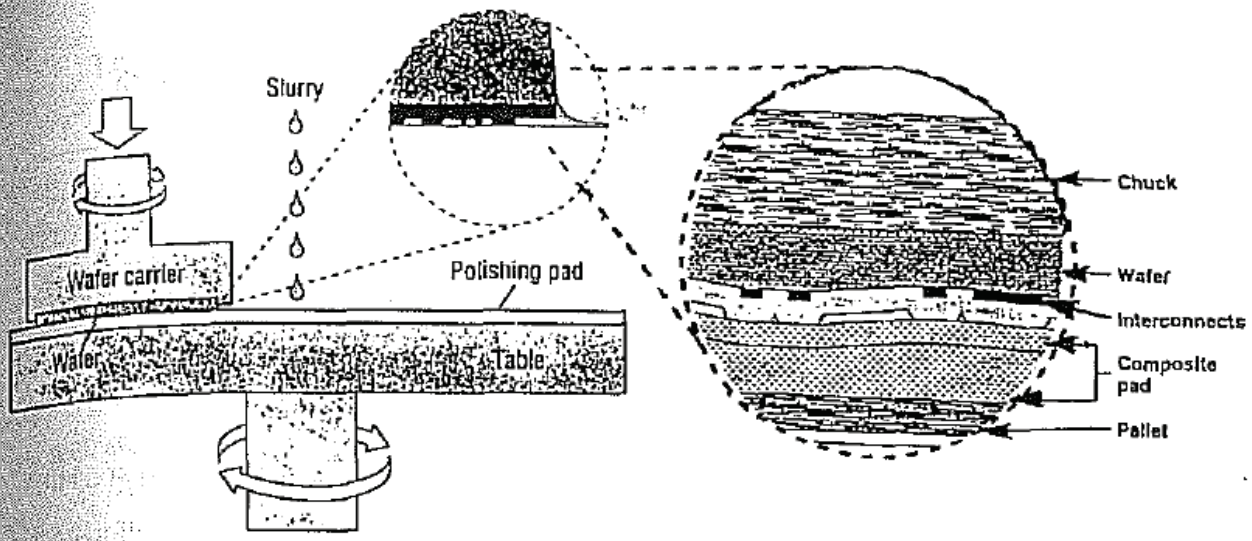


Fig. 8-31 Schematic diagram of a wafer being polished by CMP.

The Westech and Strasbaugh polishers first used by IBM were also the tools that found widest use in the early CMP processes. IBM provided to SEMATECH one of these silicon-wafer polishing tools from Westech. This tool was used by SEMATECH to launch a program that focused on improving the tool's ability to do CMP on the thin films encountered in VLSI structures, and to increase the tool reliability. The results of this program were used in designing the IPEC model 472®-CMP-tool, which achieved wide adoption for CMP applications during the mid-1990's. Figure 8-33 shows a plan view of an IPEC Avanti 776® polisher, a modified version of the 472 tool.

8.6.1.1 CMP Polishing-Tool Configurations: While early CMP tools used a single wafer-carrier, more recent designs contain multiple-carriers (and multiple-platens). This increases the throughput of the polisher and allows multi-step CMP processes to be carried out on a single tool. The classes of CMP polishing tools can be categorized according to the manner in which motion between the wafer-carrier and polishing table is produced. These categories are as follows:

1. *Rotary CMP Tools* (in which both the table [or platen] and the carrier rotate);
2. *Orbital CMP Tools* (in which the carrier rotates, but the table moves in an orbital path);
3. *Linear CMP Tools* (in which the carrier rotates and the table moves in linear path);
4. *Fixed Table CMP tools* (in which the table is stationary, but the carrier moves in an orbital path).

Table 8-1 lists some commercial CMP polishing tools and categories into which they fall.

8.6.1.2 Rotary CMP Tools: Figure 8-32a shows a schematic of a single-carrier/single-platen *rotary CMP tool.* The motion between the wafers and the pad is created by the rotation of the platen together with the rotation of the wafer-carrier. (The carrier also holds the wafer face down, and applies a downward force against the surface of the pad mounted on the rotating platen.) The pad is afixed to the flat table by a waterproof-adhesive. The following tools are examples of rotary-motion CMP tools with multiple wafer-carriers: Strasbaugh Symphony,® which has four wafer-carriers and three platens (Fig. 8-34); SpeedFam-IPEC Auriga® which has five wafer-carriers and two platens (for primary and secondary polishing, Fig. 8-35); Cybeq Nano-Tech-Isoplanar-8000® which has six-wafer carriers; and Applied Materials Mirra® with four wafer-carriers and three platens (each with an *in-situ* pad-conditioning arm, Fig. 8-36). In the Mirra,® a central spindle transports the wafer-carriers from one platen to the next. A three-platen-configuration (such as is available on the Symphony® and the Mirra®) permits three-step CMP-processes to be performed on a single tool.

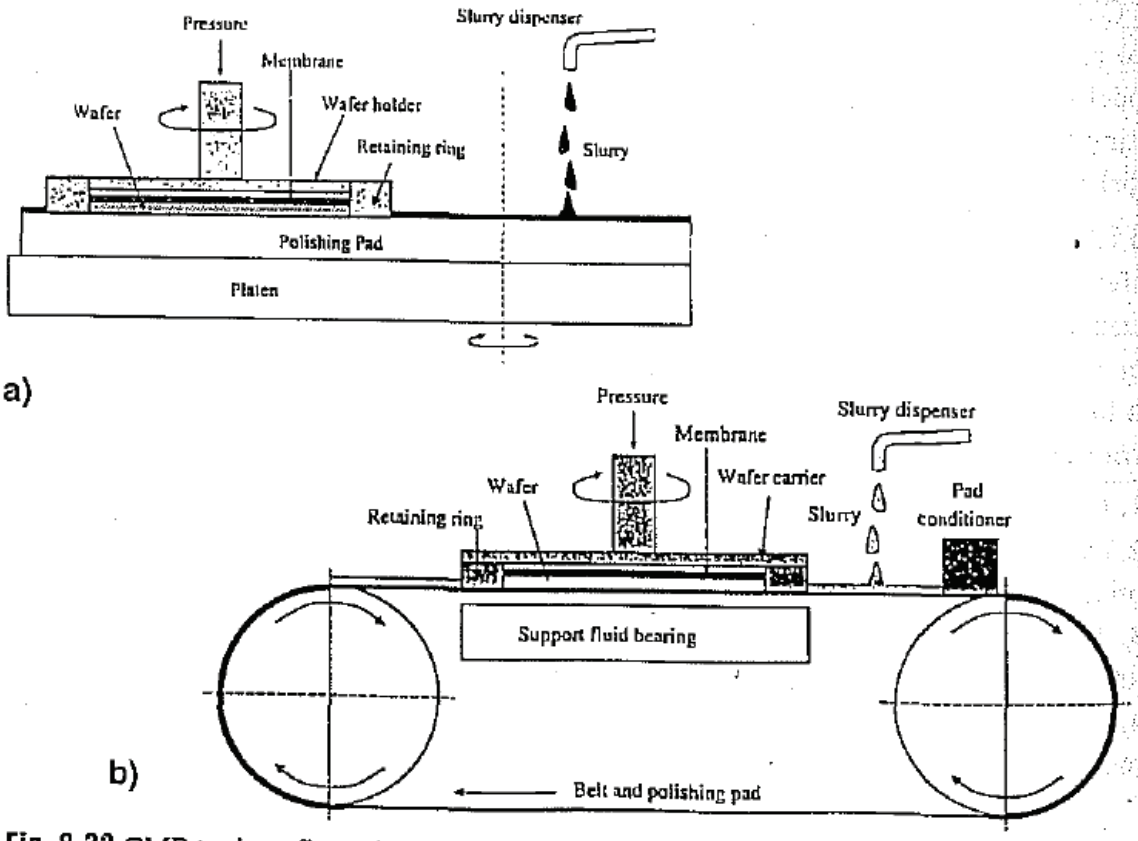


Fig. 8-32 CMP tool configurations: a) Rotary tool; b) Linear tool.

TABLE 8-1 Classification of Some CMP Polishers

Vendor & Model	Table motion–number	Carrier type–number	Built-in endpoint methods!	Special features
	Established	Polishers		
AMAT Mirra	Rotary–3	BB-CRRC–4	Red laser reflectometer	Sequential use of 3 tables
Cybeq 8000	Rotary–1	BB-CRRC–6		
IPEC 372/472	Rotary–2	FBC/NRRC–1	Absolute temperature	Sequential use of 2 tables
IPEC 676	Orbital–4	BB-CRRC–4		Parallel tool
IPEC 776	Orbital–4	BB-CRRC–4		Ditto, Integrated Ontrak scrubber
Speedfam/ Auriga	Rotary–1	FB-NRRC-5	Visible light reflectometer	
Strasbaugh/6DS-SP	Rotary–1 or 2	FB-NRRC-2		2 carrier size buffing platens, optional 2nd table
Strasbaugh/ Symphony	Rotary–3	FB-CRRC–4		
Lam/Teres-	Developing	Polishers		
Ontrak/Synergy	Linear-2	?–2		Fluid bearing platen
Obsidian	Integrated Cleaner Fixed–1	Orbital, BBC?–1		Disposable 3-M slurryless pad

8.6.1.3 Orbital CMP Tools: In *orbital CMP tools*, the wafer-carrier rotates in the same manner as in the rotary CMP tools. The platen, however, executes an orbital path about the center of the rotating carrier (instead of rotating about its own central axis (Fig. 8-37). In this approach the removal-rate-non-uniformity problem encountered in rotary CMP tools is circumvented. This occurs because the travel distance of all loci on the wafer surface relative to loci on the platen surface are equivalent. Another advantage of this tool configuration is that it uses a platen with a smaller radius than that needed in a rotary polisher. As a result, the footprint of the polisher can be significantly reduced. Orbital CMP tools are sold by IPEC-Speedfam in their AvantGaard 676,® AvantGaard776,® and Momentum® models. Figure 8-38 shows the SpeedFam-IPEC Momentum,® which has four wafer carriers and four platens.

8.6.1.4 Linear CMP Tools: CMP tools that have a pad in the form of a continuous belt (that moves in the same fashion as a treadmill) are also available. The pad is thus moved past the rotating wafer carrier in a linear (rather than rotary) motion (*linear CMP tools*). These tools offer the potential benefits of higher linear

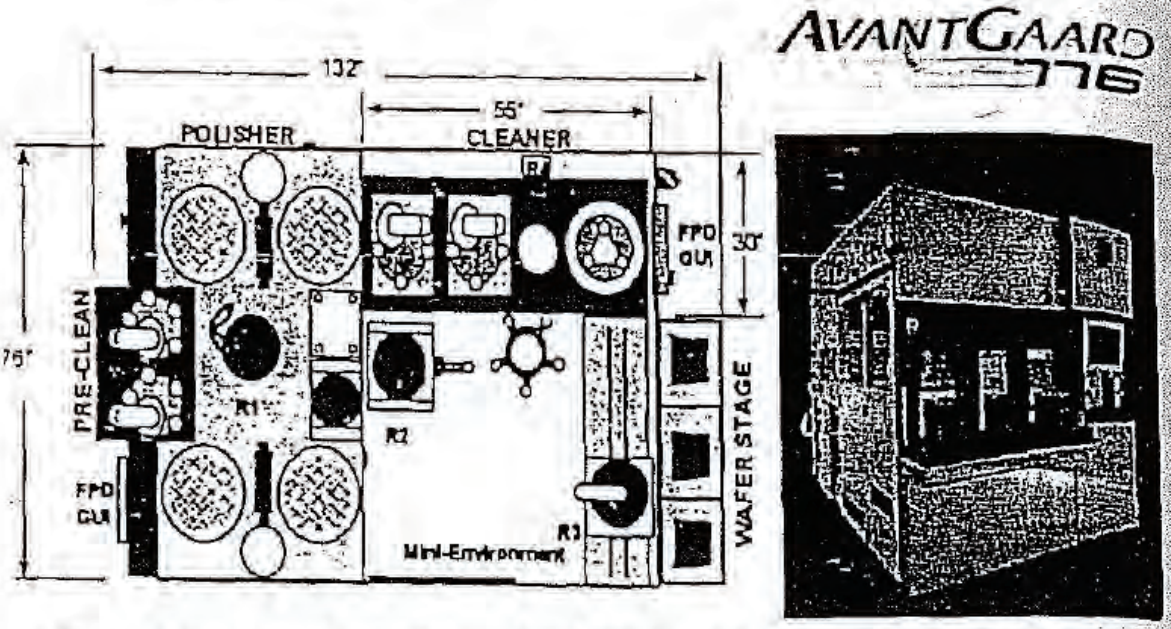


Fig. 8-33 Plan view of an IPEC *Avanti* 776® CMP tool. Courtesy of Speedfam/IPEC.

linear speeds 400–500-ft/min (versus 34-ft/min in rotary CMP tools), harder pad materials, and reduced polishing-pressure (i.e., because the polish rate can be maintained at a high value with reduced applied pressure, as a result of increased pad velocity). This type of tool may be able to produce better planarization of oxides in a shorter time and reduce the problem of dishing in CMP (see Sect. 8.12). Linear CMP tools are offered by Lam Research (Teres® Fig. 8-39), Aplex, and Doosan. The Teres® has two linear polishing belts, and one station with a rotating platen (for carrying out buffing or cleaning steps). The wafer carriers move independently from one station to another. The Teres® also employs an *air bearing* beneath the polishing belt. This allows the pressure distribution across the wafers to be profiled. At a given pressure and speed, adjustments to the pressure-distribution profile can be made to significantly reduce film-thickness non-uniformities (Sect. 8.12). The removal-rate non-uniformity problem encountered in rotary CMP tools is also circumvented in linear CMP tools. Improved planarization efficiencies (i.e., less oxide needs to be deposited to achieve complete planarity) have also been claimed for this configuration.

8.6.1.5 Fixed-Table CMP Tools: A CMP tool that employs a stationary platen and a wafer-carrier that moves in an orbital path over the platen (Fig. 8-40) was announced by Obsidian Corp. in 1999. (Another possibility would be to rotate the carrier while at the same time moving it linearly back and forth across the table,

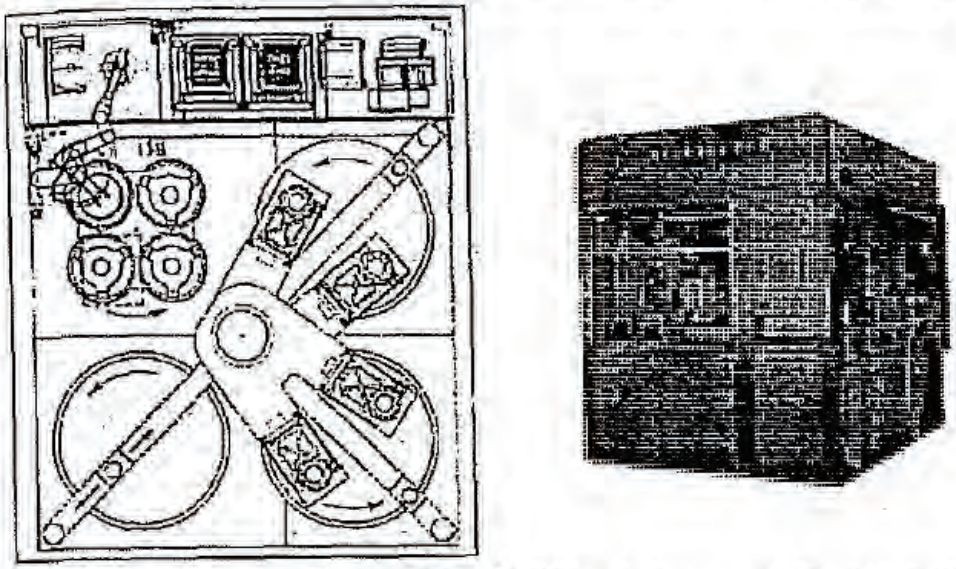


Fig. 8-34 Schematic drawing of the Symphony® CMP tool. Courtesy of Strasbaugh Corp.

see Fig. 8-41). This tool also employs a slurry-less pad supplied by the 3M Corp. The pad material is mounted on a roller (akin to a roll of paper towels). After each wafer is polished, the roll is advanced to remove the used pad-surface. This makes a new segment of pad material available to polish the next wafer (roll-to-roll transport of the slurry-free pad). Applied Materials acquired Obsidian in 2000, but by the end of 2000, the tool had not yet been commercially released.

8.6.1.6 300-mm CMP Tools: Many semiconductor equipment manufacturers are readying CMP polishing tools for 300-mm wafers. By mid-2000 more than a dozen CMP equipment manufacturers have exhibited or announced intentions to participate in the 300-mm market.[27] It is reported that most 300-mm CMP tools are capable of reaching the performance capabilities of the 200-mm tools. However, challenges remain with respect to the issues of consumables costs, integration of metrology, and full automation. In some cases a reduction in consumables cost is being achieved. For example, in the Speedfam-IPEC 300-mm polishing tool, slurry use is increased by only 20% over that needed in the 200-mm tool, whereas the area of the wafer being polished is increased by a factor of 2.25. This has a net positive effect on the CoO. Edge exclusion zones of 3-mm or less are also reportedly being achieved on these tools.

8.6.2 Wafer Carriers in CMP Polishing Tools (and the CMP Edge Effect)

In Fig. 8-42 the wafer is shown mounted on the carrier. The traditional wafer carrier consists of a circular plate (chuck) within a pocket (which holds the

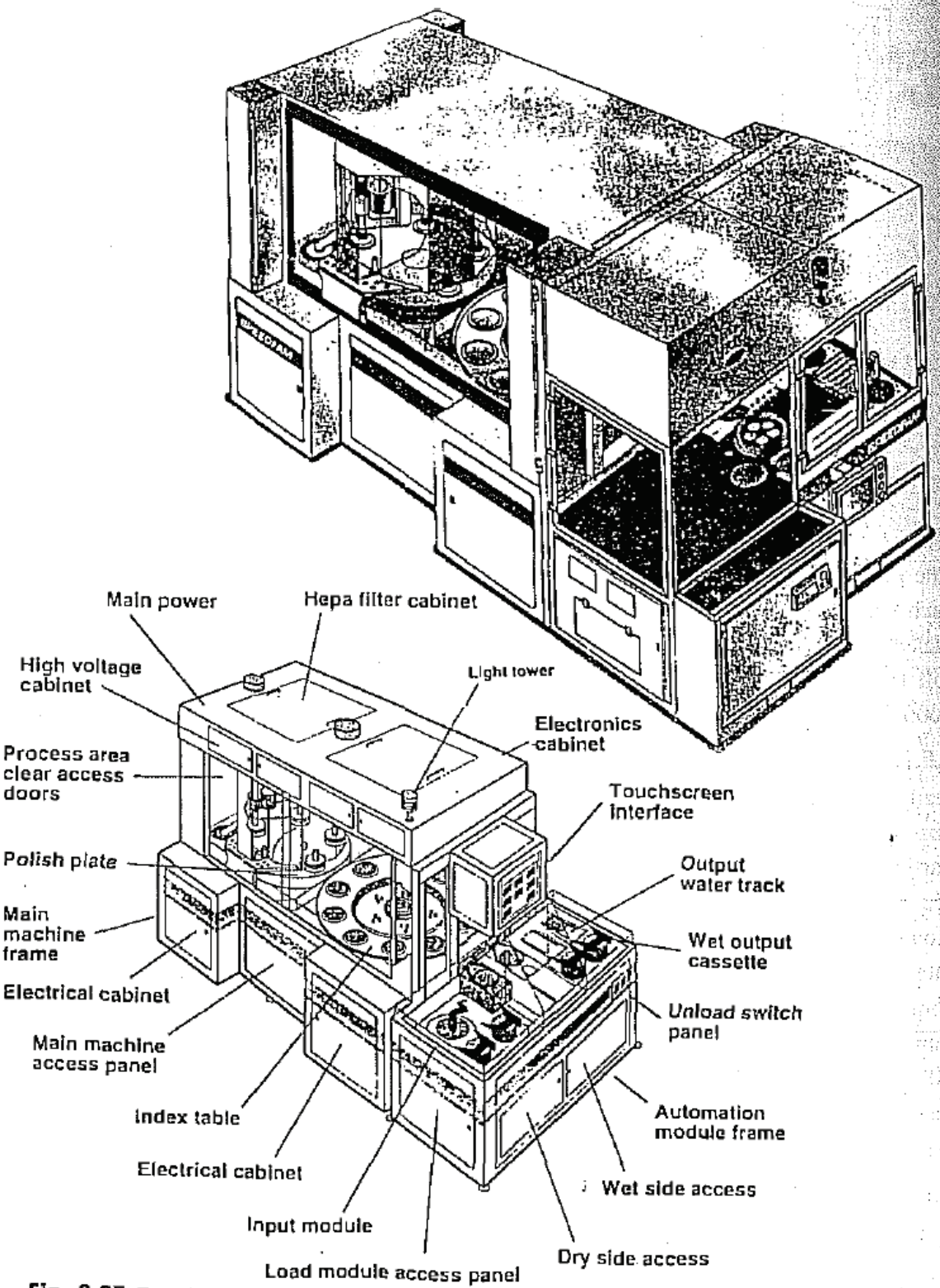


Fig. 8-35 Detailed view of the SpeedFam-IPEC Auriga® *five-carrier polishing tool.* Courtesy of SpeedFam-IPEC, Inc.

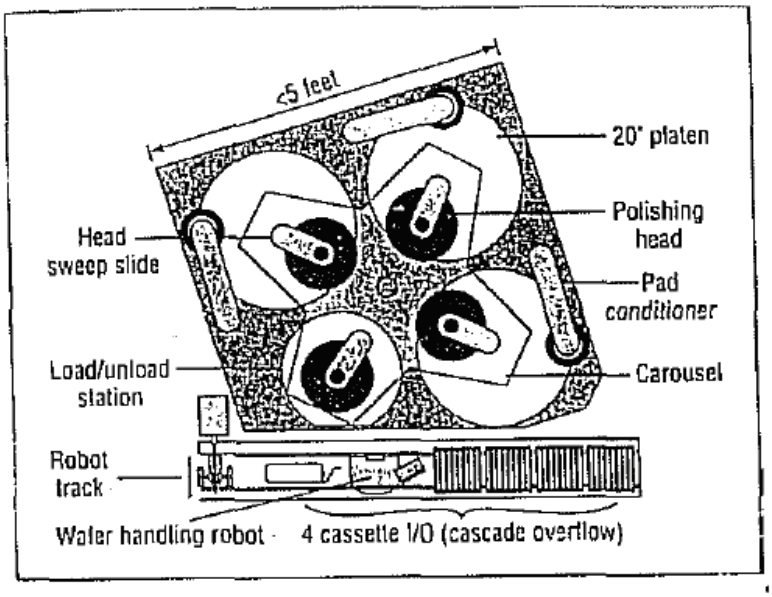


Fig. 8-36 Architecture of the Mirra® three-platen CMP tool. Courtesy of Applied Materials, Inc.

wafer), attached to a rod by a gimbal (Fig. 5-43)). Downward force is applied to the plate by the rod during polishing. The depth of the pocket in which the wafer resides is chosen to allow the top-surface of the wafer to extend beyond the pocket, permitting its surface to be polished. Instead of the wafer being directly in contact with the steel chuck, a backing film is sandwiched between the wafer and the chuck. This film provides elasticity. Without a flexible backing film, any

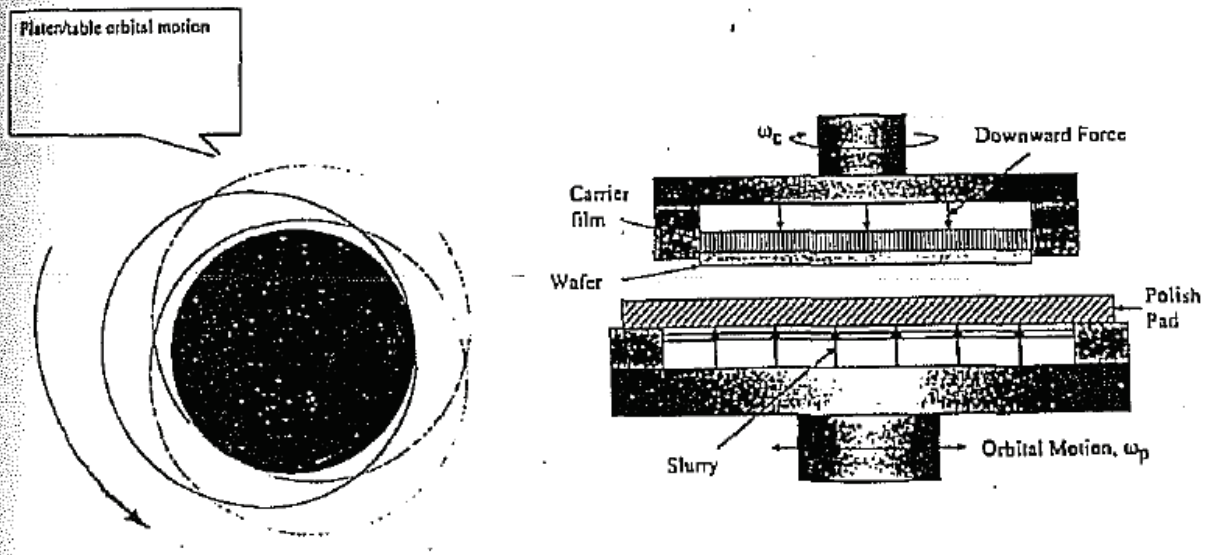


Fig. 8-37 Schematic drawing of orbital platen motion polisher.

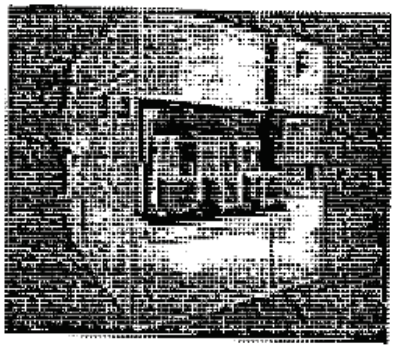


Fig. 8-38 Photograph of the Momentum,[10] an orbital platen motion polisher. Courtesy of SpeedFam-IPEC, Inc.

defect or particle on the chuck (or the back of the wafer) will cause a thin spot in the film being polished. In addition, the possibility of wafer breakage is

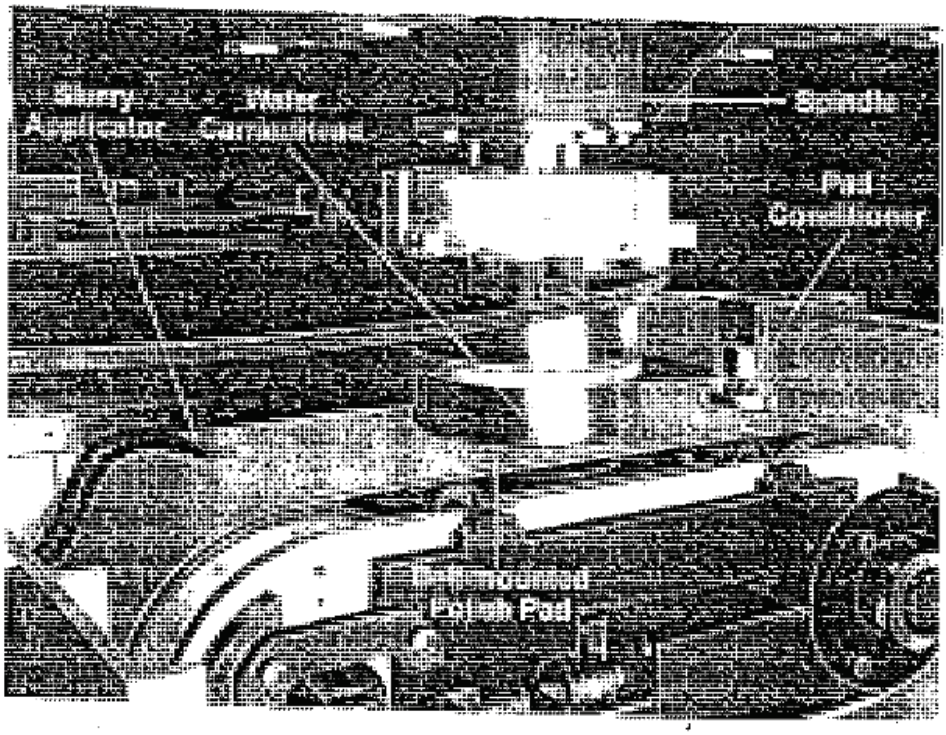


Fig. 8-39 Schematic drawing of the Teres® linear CMP tool. Courtesy Lam Research, Inc.

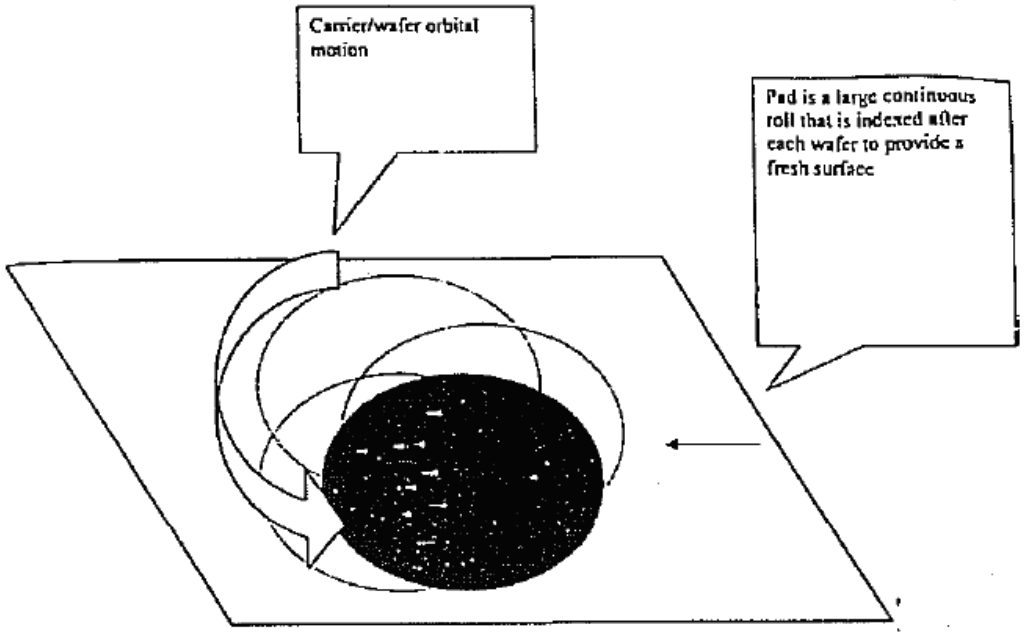


Fig. 8-40 Schematic drawing of orbital carrier-motion with indexing pad.

increased. Early backing films were merely thin foam layers. They did little to lessen the polish-performance-dependency on the variations of wafer-thickness or flatness. In order to more fully compensate for wafer-to-wafer variations in flatness and film-thickness, much development work has been done to improve the wafer-carrier design.

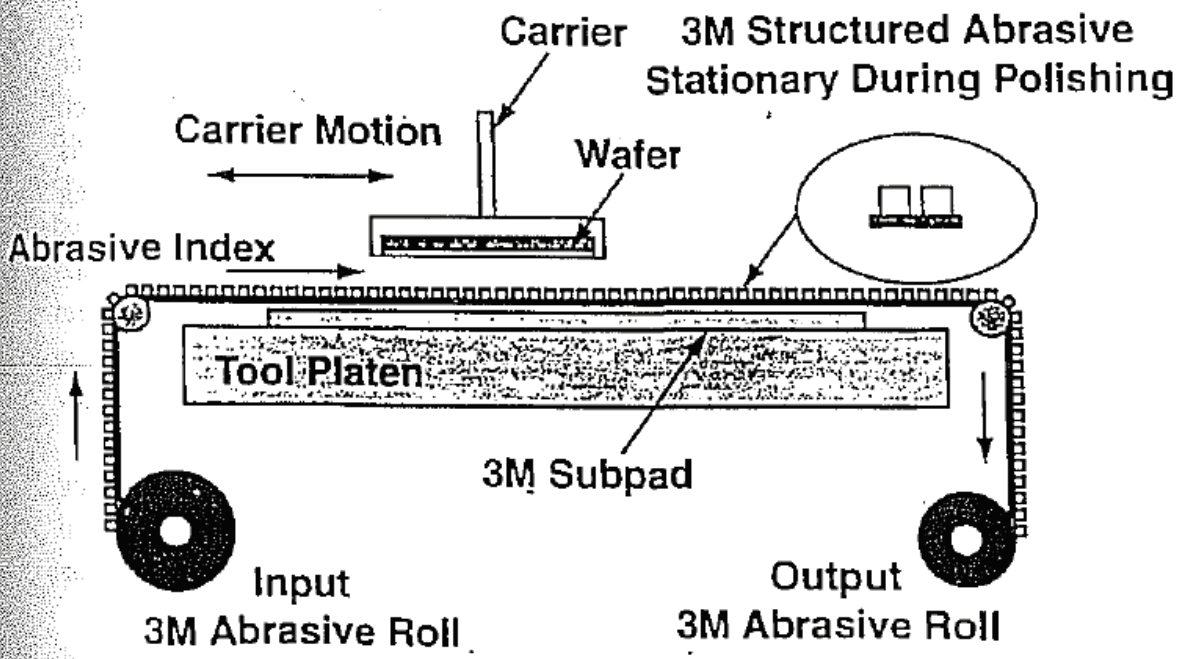


Fig. 8-41 Fixed-table with roll-to-roll pad. Carrier rotates and also exccutes linear motion.

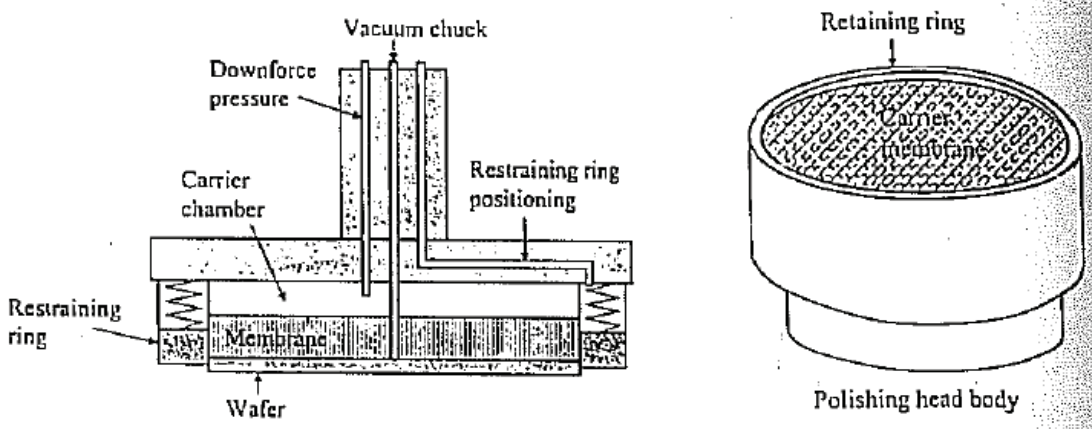


Fig. 8-42 Schematic drawing of the components of a CMP wafer-carrier.

Two key design elements in the improved wafer-carriers are:

1. The retaining-ring;
2. The wafer-leveling mechanism.

The primary function of the *retaining-ring* is to prevent the wafer from slipping out of the pocket when high down-force and high-speeds are used. Wafer retention can be maintained even with an average down-force of about 2-psi if the retaining-ring extends past the mid-thickness of the wafer. To keep the retaining-ring surface from being polished, some wafer-carriers may keep the retaining-ring about 100-microns off the pad surface. Wafer-carriers with such rings are termed *non-contact retaining-ring carriers* (NRRC). In NRRCs there is no pressure applied to the pad by this ring. The other type of retaining-ring

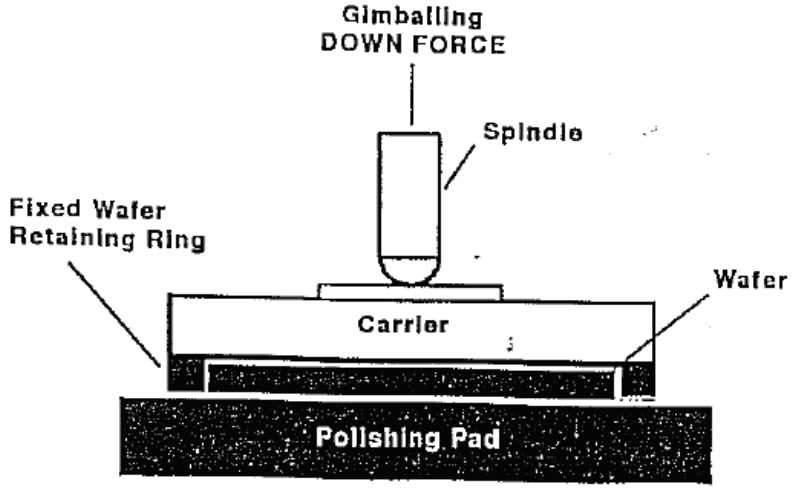


Fig. 8-43 Carrier being subjected to gimballing down force.

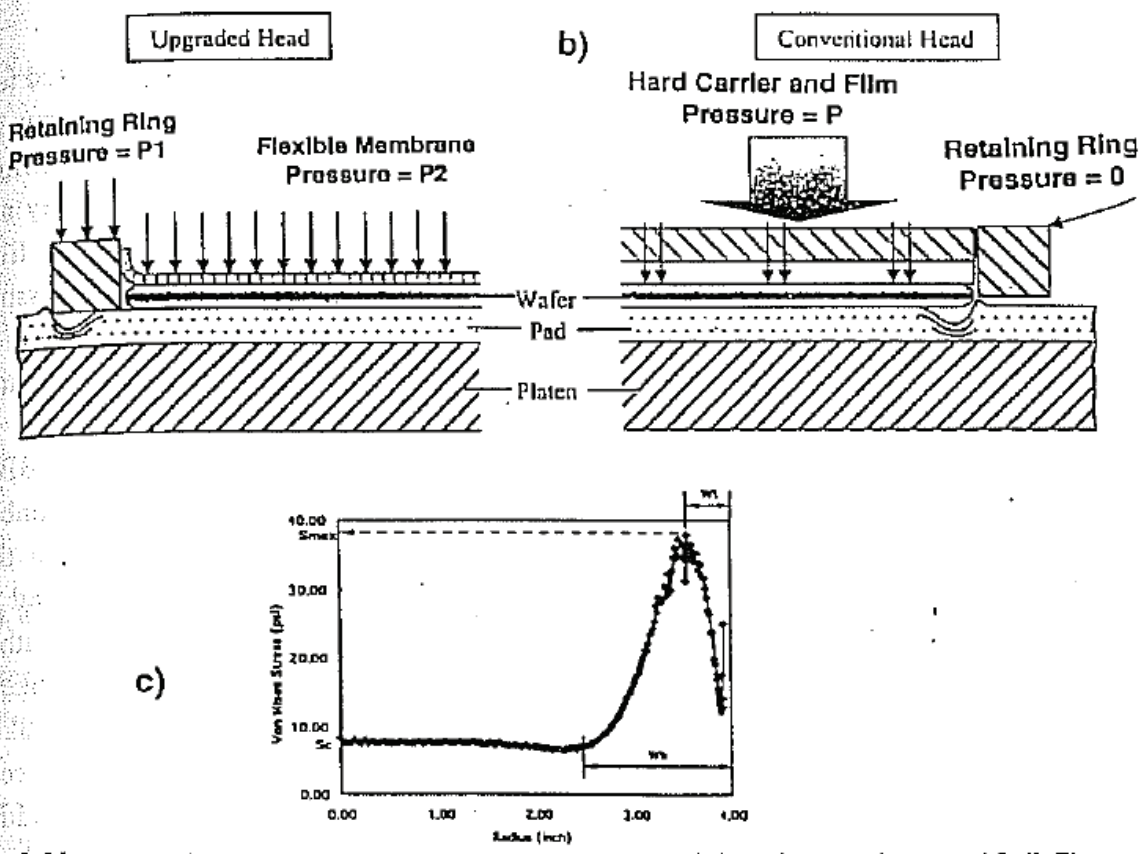


Fig. 8-44 Comparison of: a) film-backed/non-contact retaining-ring carrier (FB-NRRC); and b) bladder-backed/contact retaining-ring carrier (BB-CRRC); c) distribution of pressure profile at wafer edge for FB-NRRC.

design causes the retaining-ring to contact the pad. Hence, carriers with these rings are called *contact retaining-ring carriers* (CRRC). The CRRC also has the capability of varying the down-force on the wide (typically 10-25 mm-wide) bearing surface of the ring. Wafer-carriers used in some polishing tools (e.g., Applied Materials' Mirra® and Strasbaugh's Symphony®) are designed with a retaining-ring that can apply variable pressure to the pad.[28] This allows the pad profile at the wafers edge to be controlled. A smaller edge-exclusion zone on the wafer can therefore be used (2-3 mm).

Wafer leveling by the wafer-carrier is accomplished by backing the wafer with: 1) a compressible film (i.e., *film back carrier*, FBC); 2) a pneumatic bladder (*bladder-backed carrier*, BBC); or 3) a lip-sealed air bearing *(air-backed carrier*, ABC).

The earliest wafer carriers were FBC-NRRC (Fig. 8-44a). Perforations in the backing film and the flat surface plate are used to vacuum-attach the wafer to the carrier during transport. The backing film acts as a cushion to distribute the down

force evenly to the wafer. Early backing films were chemically-impregnated felt materials that had to be applied by hand, but they were only useful for about 25 wafers. Newer backing films (such as the DF200, the WB, and the R200 films) last about 200-800 wafers. (The reason backing films wear out is that they are compressed and dc-compressed each time they are used. At some point, the film will lose its ability to recover and takes a "set.") In FBC-NRRCs the pad pressure at the edge of the wafer is not well controlled, and the polishing rate of the film is decreased at the wafer's edge. Hence, a wider edge-exclusion ring of 5-7-mm must be enforced for wafers with such carriers.

The second type of wafer-carrier is the BBC-design (Fig. 8-44b). In the BBC a rubber inner-tube-style bladder is substituted for the backing film. The chamber created by the bladder can be alternately vacuumed or pressurized. When the carrier brings the wafer into contact with the platen, first pressure is applied to the retaining-ring bladder (with a sufficient down force to clamp the wafer within the retaining-ring). Next, the wafer-bladder may be vented or pressurized. Usually the platen is stationary when the wafer-carrier retaining-ring makes contact and the wafer-bladder is vented until the wafer-carrier comes to speed. Then the wafer-bladder is pressurized. Since the wafer "floats" on the bladder bubble, global and local planarity of the wafer surface is accomplished without gimbaling the wafer-carrier to the head of the tool.

The air-backed wafer-carrier (ABC) functions much like the BBC. However, instead of a bladder to contain the air pressure (or vacuum) behind the wafer, it has a compliant lip-seal around the backside edge of the wafer. However, the seal is not easy to maintain, and slurry particles can enter and contaminate the backside of the wafer. As a result, ABC types have not yet been commercialized.

8.6.2.1 CMP Edge-Effect

The removal-rate of material by CMP should ideally be uniform across the wafer (i.e., the WIWNU should = 0). However, for various reasons, non-uniformity of material removal within a wafer is generally observed. One of the most common forms of WIWNU is the *CMP edge-effect*, in which the removal-rate is much smaller at the edge of the wafer, leaving an edge-bead of material at the conclusion of the CMP process. Figure 8-45 shows two typical thickness profiles of wafers polished by two different polishers, along a wafer-diameter. In both cases, an edge-bead, 100-200-nm high, exists after the polishing step. This edge-bead will generally render devices fabricated in that area inoperable. As a result, devices fabricated within a region near the edge of the wafer are considered to be excluded as candidates for working die, and this ring near the edge of the wafer

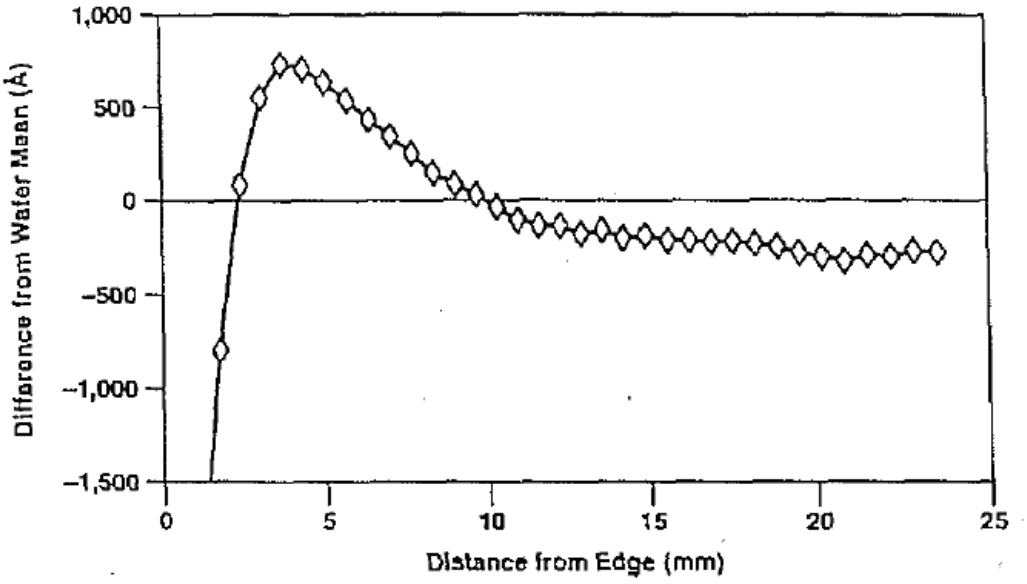


Fig. 8-45 Edge effect observed in CMP.

is termed the *exclusion ring* (Fig. 8-46). In current polishers, the exclusion ring is 3-mm wide. However, as wafer sizes are increased, the number of die at the wafer's edges keeps increasing (Table 8-2).[29] In order to obtain more yielding die per wafer, methods to further reduce the edge exclusion ring are being pursued.

The edge-bead-effect is attributed to the pad-deformation-profile and the pressure-distribution near the edge of the wafer. As a result, this effect can be significantly reduced by optimizing the pressure distribution at the edge of the wafer, primarily through the use of contact-retaining-ring-carriers (CRRC). The bulk physical properties of the polishing pad (such as its hardness, flexibility, compressibility, energy-absorption, and rebound) also all influence the width of

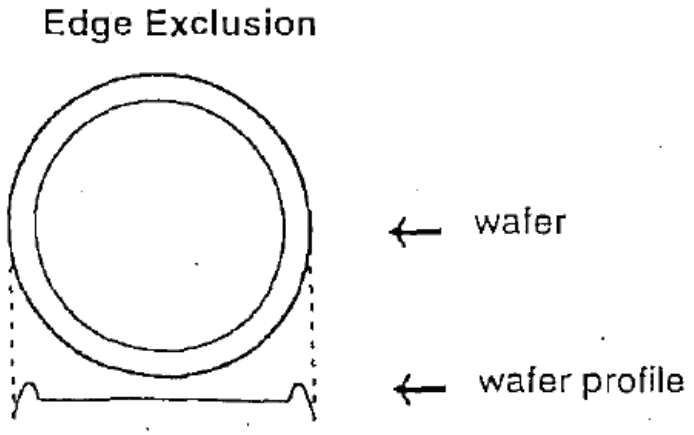


Fig. 8-46 Exclusion ring.

TABLE 8-2 Edge Die Count versus Wafer Diameter

Wafer diameter	Total wafer die count	Edge die count	Edge die percentage	Max edge revenue ($B)
300 mm	515	107	21%	2.76
200 mm	217	69	32%	1.78
150 mm	115	48	42%	1.24
127 mm	71	38	53%	1.00
100 mm	43	26	60%	0.68
76 mm	21	16	76%	0.43

the exclusion ring. Pad materials that provide the hardness necessary for planarization (but remain compliant enough to transfer dynamic contact-pressure more evenly), will deliver improved edge-exclusion. New materials for polishing pads and sub-pads have been designed to improve the uniformity of contact.[30]

8.6.3 CMP Consumables (Slurries)

Slurries consist of small, abrasive particles (generally inorganic oxides) of specific shape and size (typically 10–100-nm) suspended in an aqueous solution. The abrasive particles have roughly the same hardness as the film being polished. Reagent chemicals, such as acids, bases, surfactants, buffers, corrosion inhibitors, and complexing agents, are also added to the solution, depending upon the material to be polished. The slurry parameters that impact the polishing-rate include the chemical composition of the slurry solution (and its pH) and the concentration (in wt %) of the solid-particles in the slurry. Also important are the shape, size, and size distribution of the slurry-particles. An automatic slurry feeding system is used to ensure uniform wetting of the polishing pad and proper delivery of the slurry. Slurries are marketed by a number of suppliers, the largest of which is Cabot Corp. More slurry vendors are entering the market, including Rodel Inc., Fujimi Inc., Olin Microelectronic Materials, Praxair Surface Technologies, and EKC Technology, Inc.

The goal for each CMP process is to find a slurry that produces high removal-rates, good-planarity, local-film-uniformity, and high-selectivity. Also important are the ease with which the slurry can be cleaned from the wafer, the aggressiveness with which it attacks the polishing equipment, its storage stability, and the cost of disposing of the slurry waste. From a manufacturability perspective, consistency is a critical issue. That is, slurries have tended to exhibit quite a lot of variability from batch-to-batch. In the future, suppliers will be

expected to deliver slurry that has the same performance consistently. If this can be done, the need for monitor wafers will be significantly reduced, allowing the cost of the CMP process to be decreased.

The shelf-life of slurry is also an important characteristic. This determines how long a slurry can be stored before use after manufacture. Some slurries need to be kept within a specific temperature-range during shipping and storage to maintain their performance. The shorter the shelf-life of the slurry, the narrower the window in which it can be used, and inventory-control becomes important. An ideal shelf life for slurries is at least one year.

As CMP technology has evolved, slurry technology has therefore also had to progress to keep pace with the demanding requirements of the industry. Today's CMP slurries are increasingly tailored for specific process applications (like STI or Cu interconnects) as well as for specific polishing tools or other consumables (i.e., pads). Some of these issues have been discussed in Sect. 8.5.2, and others will be addressed in Sect. 8.11. The remaining slurry topics are considered here.

One other aspect of slurry management should also be pointed out. Slurries tend to dry rapidly on a wafer after leaving the wet environment of the polishing table. Once a slurry dries, it is very difficult (if not impossible), to remove all the slurry particles from a wafer surface. To prevent this from occurring, wafers are

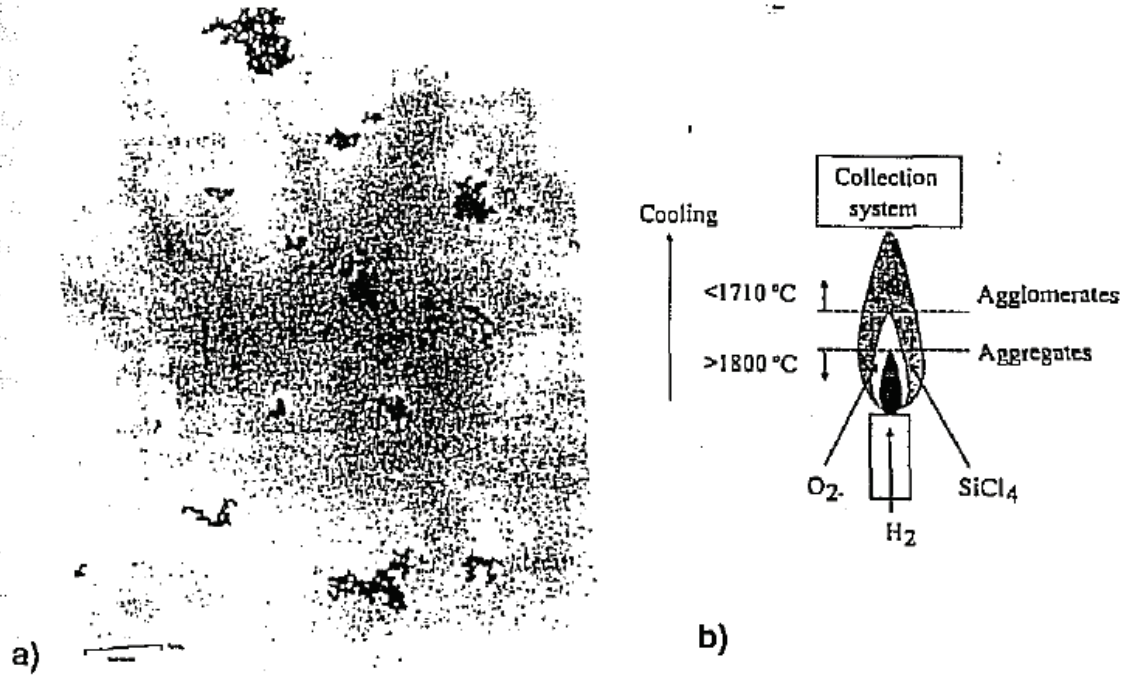


Fig. 8-47 a) SEM of fumed-silica slurry-particles. Courtesy Rodel; b) Fumed-silica formation.

rinsed while still on the pad, and then quickly transported into wet-cleaning tanks to remove the slurry. In fabs that do not use polishing tools with integrated cleaning systems, logistical difficulties may delay the transport of wafers from the polisher to the cleaning system, inviting the above problem.

8.6.3.1 Slurries for Oxide-CMP: Slurries for oxide-CMP traditionally consisted of colloidal-silica or fumed-silica suspended in an alkaline solution (typically potassium-hydroxide [KOH] or ammonium-hydroxide [NH_4OH], diluted in DI-water). The first generation of slurries for oxide-CMP contained fumed-silica abrasives, and they still retain a large fraction of the market share. Fumed-silica is manufactured as a dry powder by the oxidation of silicon tetrachloride ($SiCl_4$) at high temperature in a flame (see Fig. 8-47). The powder must be dispersed and blended into the final slurry. The fact that the fumed-silica must be mixed is inconvenient, and adds cost. The average selling price of fumed-silica slurries in 1998 was \$10-\$15 per gallon. Figure 8-47 also shows a SEM of fumed-silica slurry particles.[31]

Colloidal-silica slurries have entered the market as an alternative to fumed-silica. Colloidal-silica is manufactured by the hydrolysis of sodium-silicate (or other silicates). This is followed by an ion-exchange to reduce the sodium levels for IC applications. (Early on, colloidal-silica was avoided due to the possibility of sodium contamination, but improvements in this aspect of colloidal-silica processing have opened the door for its adoption.) Colloidal-silica is prepared directly in solution, thus avoiding most of the handling problems of fumed-silica. Its microstructure is a distribution of spherical particles (see Fig. 8-48).[32] Colloidal-silica slurries also tend to be more stable and less prone to agglomeration than fumed-silica slurries. In addition, their more uniform size

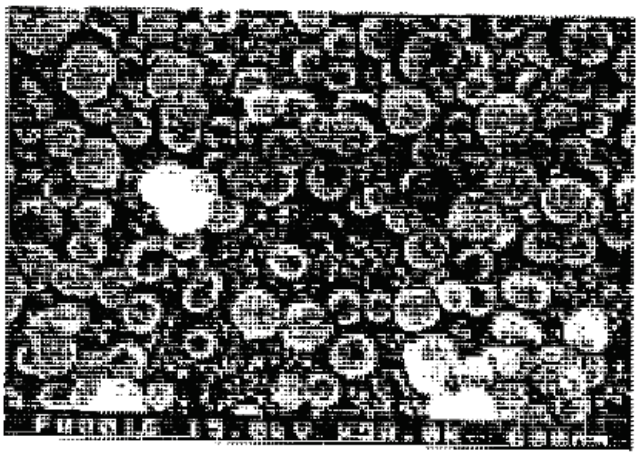


Fig. 8-48 SEM of colloidal-silica slurry particles. Courtesy of Fujimi Corporation

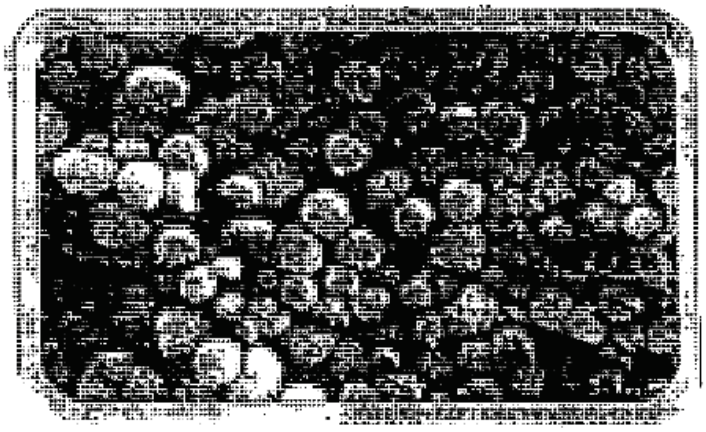


Fig. 8-49 SEM of cerium oxide slurry particles. Courtesy of Fujimi Corporation.

distribution makes it less likely that they will cause microscratches. The cost of colloidal-silica slurries is about 20% higher than that of fumed-silica slurries ($18 versus $15 per gallon). The use of colloidal-silica slurries has rapidly increased (replacing fumed-silica slurries), and they find use in STI, ILD, and polysilicon CMP processes.

The solids content of oxide-CMP slurries is 5-35-wt%, and the pH is between 9.0–11.0. By controlling pH within this range, agglomeration is minimized. CMP slurries generally contain few chemical additives. As mentioned above, these few typically consist of: 1) an alkaline chemical to provide colloidal stability (i.e., to prevent slurry-particle-agglomeration) and to increase oxide removal rates; and 2) a buffering agent (used to keep the pH constant throughout the slurry volume and over time). The polishing rate of oxide increases with: 1) increasing pH,; 2) increasing slurry particle concentration; and 3) increasing particle size.

Slurries for oxide-CMP based on other abrasives than silica have also been introduced in recent years. Ceria-slurries (CeO_2-based) have been formulated for STI and ILD applications, although ceria is difficult to manufacture at small particle sizes. A self-stopping, dishing-free process using a ceria-slurry was reported by Nojo.[33] Ceria-slurries exhibit a higher removal rate at a much lower pH than silica-slurries (e.g., at a pH of 6 to 7). Ceria-slurries have been adopted for STI-CMP because of their high-selectivity to oxide over nitride (as high as 250:1). Ceria particles can also be blended with silica particles in a slurry. The addition of silica helps keep ceria particles in suspension. In 1998, the average selling price of ceria-slurries was $18 to $20 per gallon. Figure 8-49 shows cerium oxide (CeO_2) particles.[34]

Zirconia-based slurries (ZrO_2) have been investigated for use in low-k-dielectric-CMP, especially the CMP of non-fluorinated polyarylene ethers (e.g.,

FLARE™, from Honeywell, see Chap. 14). Zirconia is harder than silica but softer than alumina. Manganese-oxide slurries (MnO_2) have also been studied for the direct polishing of low-*k* dielectrics (and for the CMP of W). It was reported that the deep scratches formed with CMP using a conventional silica-based slurry were reduced from a 20-nm depth to a 2-nm depth by the use of a MnO_2-based slurry. Residual MnO_2 slurry particles can also be removed using H_2O_2 with a pH of 6, which does not attack the fluoro-carbon low-*k* material being polished (e.g., a poly-AE layer, FLARE®).[35]

8.6.3.2 Slurries for Metal-CMP: Slurries for metal-CMP applications typically have more constituents than do oxide slurries. Besides the *oxidizer* (e.g., ferric-nitrate and hydrogen-peroxide), and a buffer agent, they may contain: 1) a *surfactant* (to assist in keeping the slurry particles dispersed, see Sect. 8.7.1.4); 2) a *complexing agent* (used to facilitate the removal of metal from surfaces, especially in the case of copper); 3) *corrosion inhibitors* (e.g., benzotriazole [BTA] in Cu slurries to permit the use of acidic slurries); and 4) *additives to change the selectivity to other films*. The oxidizer is often separated from the abrasive for reasons of colloidal stability or shelf life. In the year 2000 time-frame, choosing and balancing all of these constituents still makes the formulation of metal-CMP slurries as much an art as a science.

Metal-CMP slurries typically use alumina particles suspended in various liquids (depending on the specific metal being polished). Alumina slurries are preferred because they exhibit higher metal-removal-rates and higher selectivity to the oxide ILD-layers (> 20:1, metal-to-oxide). The alumina particles may be fabricated either by a fumed process similar to silica, or by precipitation from aluminum hydroxide.

The pH range of metal-CMP slurries is low (e.g., pH = 1.0-5.0 for W-CMP alumina slurries, and pH = 3.0- 9.0 for Cu-CMP alumina slurries). As noted in the previous section, silica particles in oxide-slurries can be kept well dispersed for relatively long periods of time by controlling the pH of the slurry (giving them a long shelf-life). However, alumina-based slurries have traditionally exhibited the drawback of poor colloidal stability. This causes slurry particle agglomerates to form. Such clumps produce unacceptable microscratches in the films being polished. Recently, improvements have been made in the colloidal stability of alumina based slurries.[36] Nevertheless, to overcome this problem from another direction development of slurries based on other particle types (e.g., cerium oxide and zirconium oxide particles) for metal-CMP has been pursued. In 1998, the cost of alumina slurries was between $25 and $30 per gallon.

8.6.3.3 Multi-Step Slurries: When carrying out a metal-CMP process (e.g., W or Cu) two materials must be polished (i.e., the main metal, and the underlying barrier [or glue] layer). This CMP-step can be carried out with a process that uses one slurry, or two slurries. While the one-slurry process is simpler, it has not been possible to restrict the problems of dishing and erosion to the necessary degree with such a process. Hence, the two-slurry approach has been adopted. That is, for Cu-CMP, the first slurry removes the bulk of the Cu in the field areas, while the second removes the barrier (e.g., Ta or TaN) from the field areas. These two-step CMP processes significantly reduce the copper dishing and erosion effects. (Generally, the formulations of these slurries still remains proprietary). Use of a CMP-tool with multiple-platens also facilitates the manufacturability of such two-step CMP processes.

8.6.4 Slurry Distribution Systems

Slurry management (i.e., mixing, storage, distribution, and dispensing) is another key issue of a CMP process. Reliable *slurry distribution systems* (*SDS*) are an increasingly important component of the overall CMP-equipment-set. It is critical that the user of a CMP system view the SDS as an integral part of the CMP manufacturing process, and treat it with as much care as a polisher is given. This will greatly minimize process variations, ensure smooth functioning of manufacturing, and reduce unscheduled downtime.

Slurry management is challenging for a number of reasons.[37] Precise mixing and consistent batch blends are necessary to achieve the repeatability demanded by process engineers. The slurry consists of particles in a liquid, and they must be kept suspended in it in a homogeneous state. However, these particles will agglomerate into clumps unless steps are taken to suppress this tendency. (The main source of surface deformations in CMP is generally believed to be the presence of undetected coarser-size particles in the CMP slurry.)[38] The slurry-pH plays a role in the dispersion of silica particles. As is described in more detail in the section *CMP Cleaning*, at a pH above 7.5 the silica particles acquire sufficient surface charge to generate electrostatic repulsion, and this effectively keeps the slurry well dispersed.[39] (Above a pH of 10.7, however, the particles dissolve and form a silicate.) In addition, from the time the raw slurry is introduced into the SDS it is also critical that a blanket of humidified air be provided over the slurry to prevent agglomeration in the dispense system. Finally, slurries for Cu- and W-CMP require constant re-circulation to maintain particle suspension. (It is typically recommended that a minimum flow-rate of 1-m/sec be used for re-circ-

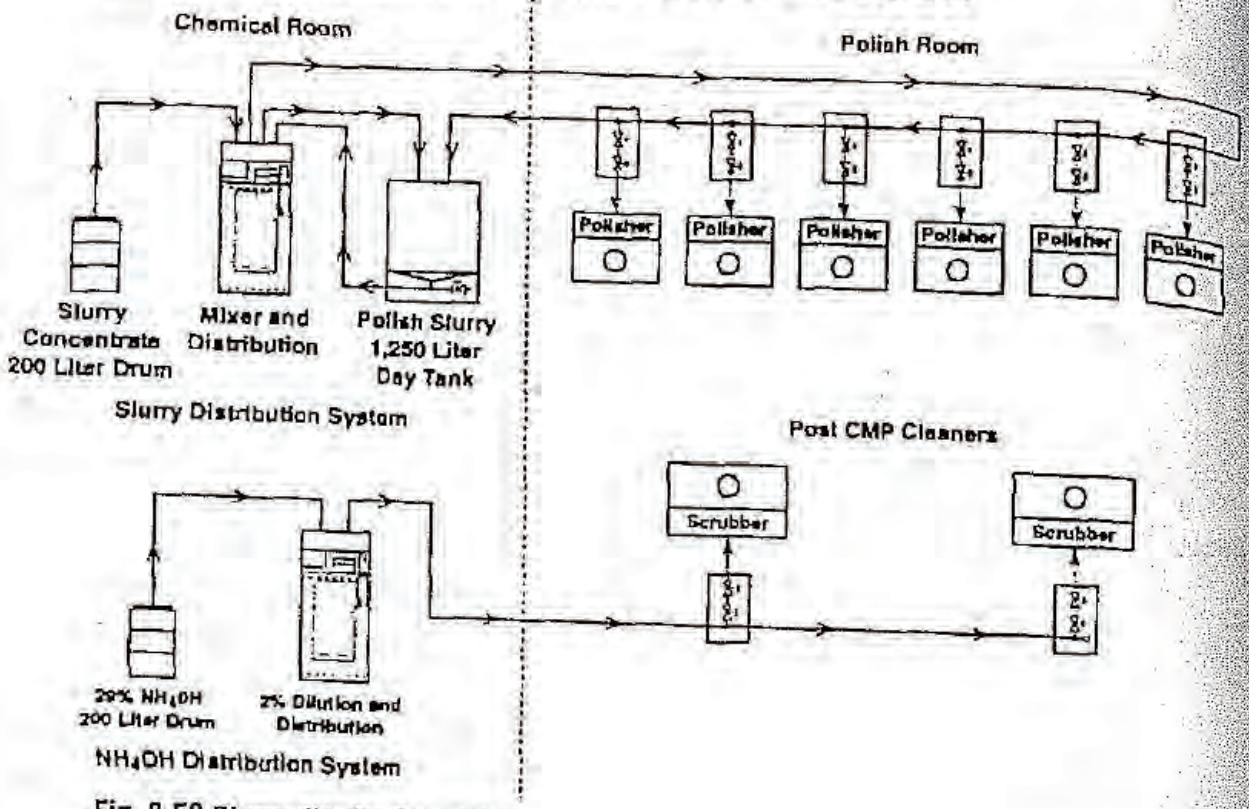


Fig. 8-50 Slurry distribution system.

ulation). Slurries have a finite shelf-life (after which the agglomerate-level becomes excessive). This shelf-life ranges from less than one week to about one month. Filtering is therefore necessary to prevent any agglomerates (which might have formed during that time), from reaching the wafer being polished.

Slurry handling in an SDS is a delicate task and care must be exercised to prevent "damaging" the slurry. That is, the slurry must not be subjected to excessive shear (which could cause the slurry to agglomerate), or to insufficient shear (which would permit slurry settling). The SDS must also be regularly maintained. Such maintenance involves periodic cleaning out and flushing of the system (as well as the polisher).

The SDS must continuously provide slurry to a number of polishers in a sustained manufacturing production environment (i.e., 24 hours a day, seven days a week, and in high volumes). The pH and solids-content of the slurry must be monitored in-line to ensure consistency. Furthermore, since slurries are abrasive particles contained in chemically aggressive liquids, they present a harsh environment to the components of the distribution system. That is, pumps, piping, valves, and measurement apparatus in contact with slurry, can be stressed to the point of failure in a short period of time. Nevertheless, bulk slurry

distribution systems have been designed to handle these conditions. These systems are sophisticated pieces of equipment, far more advanced than other bulk-delivery systems. Figure 8-50 shows the schematic layout of an SDS, and Fig. 8-51 shows the actual pieces of equipment in an SDS. Such SDS systems are available from vendors including: Mega Systems & Chemicals and FSI International. These vendors supply detailed recommendations regarding design and maintenance of SDS, piping, mixing, and dilution procedures, filtration of the slurry, filter changeouts, and re-circulation of the slurry.

8.6.5 Environment, Safety, and Handling Issues of CMP Slurries

As CMP further moves into large scale IC production, the disposal of spent slurry, CMP process by-products, and metal contaminants (i.e., material removed from the wafer surface) become important issues.

Slurry also always contains aggressive chemicals. In some cases these chemicals are highly toxic. Furthermore, the material removed by a CMP process ends up in the expended slurry and the rinse water. Some IC fabs estimate that CMP processes may come to consume 30-40% of the total water utilized in IC manufacturing. The issue of purification and disposal of this waste water (or

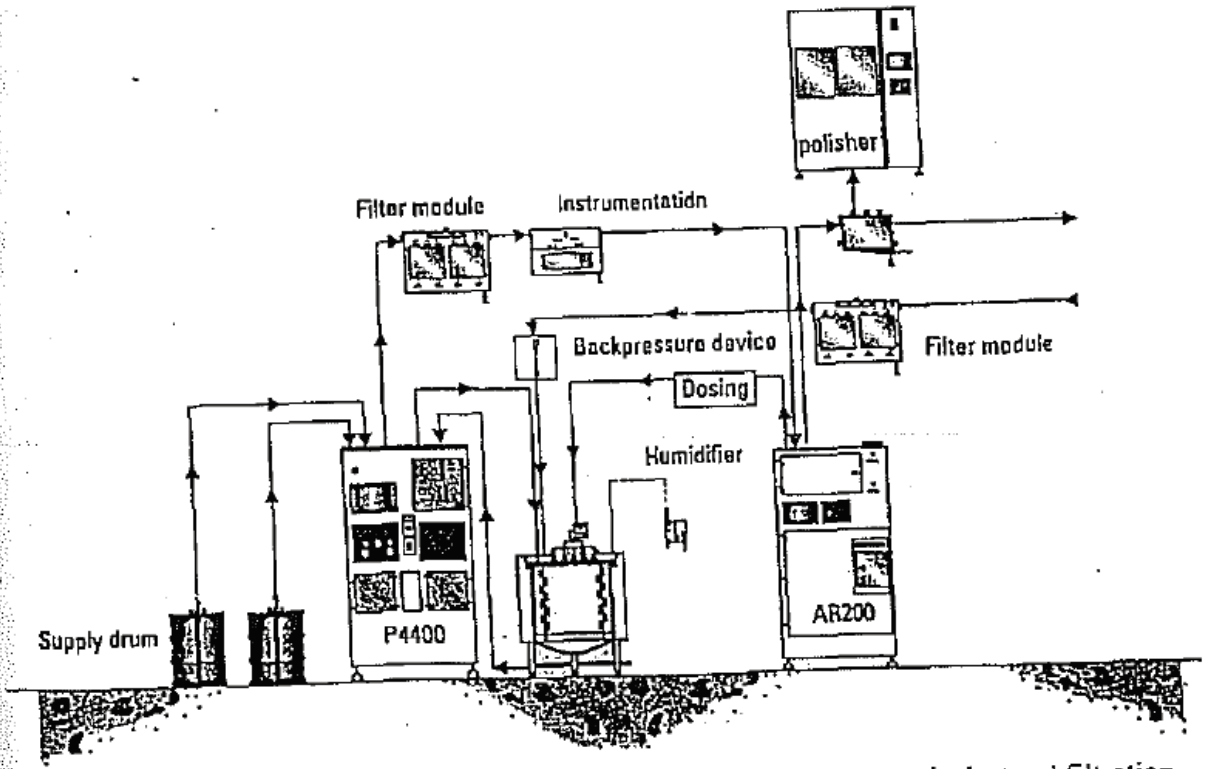


Fig. 8-51 Schematic of a slurry management-system with metrology, auto-dosing and filtration.

recycling), becomes more critical as new materials such as copper, gold, and tantalum become adopted in IC manufacturing. For example, copper is a known toxin in water. Therefore, fabs require waste-water treatment systems to recycle the drain water of CMP processes. The most common limits for waste streams are those for Total Suspended Solids (TSS), heavy metals (like copper, chromium, zinc, and nickel), pH, and total organic content (TOC).[40]

The slurry particles are suspended particles in a solution, and hence they can be removed from the waste water by a process of *flocculation* (i.e., a process by which the dispersed slurry particles are caused to clump together to form larger clusters of matter [i.e., to flock]) followed by filtration. Flocculation involves adding a flocculent and adjusting the pH to a level where the flocculent works best. The slurry particles then agglomerate and settle out of suspension. After the solids are separated, the next step involves ion removal. Either ion-exchange or reverse-osmosis may be used to do this. Note that as slurry volumes go up, they will generate huge amounts of solid waste, estimated at 100,000-400,000 pounds per year, per typical fab, and this solid waste must also be disposed of safely. A decrease in slurry consumption by the CMP-processes will be needed to reduce this waste-disposal burden.

Some companies are also examining the issue of reusing the water from a polisher. That is, while a polisher is in an idle (or pad-wet) mode, there is a continuous stream of DI-water that is used to keep all parts of the polisher wet and reduce the possibility of formation of dry particles. This water generally retains a high purity level and offers the potential of recycle and reuse, thus reducing the waste-treatment-costs for CMP.

One suggested method of reducing slurry consumption is to re-cycle the slurry. In one study of re-using slurry, KOH-based fumed-silica slurry was used to polish a PECVD-TEOS film. The slurry was then captured after being flushed from the CMP tool. By adding new slurry and adjusting the pH to compensate for the decreased solids content and pH, the slurry appeared to be capable of re-use up to six times, while maintaining roughly equal removal rates.[41] In a more comprehensive slurry-recycling process, the slurry-waste would be treated by the following processes: filtration, oxidizer-replenishment, removal of metals, and replenishment of buffers and other necessary additives. The task of slurry re-cycling (and eventual disposal) could be carried out by the slurry manufacturer. On the other hand, it has also been discussed that the slurry could be leased by the end user, and that it would be sent back to the slurry manufacturer after use for regeneration. The concept of slurry re-use is fraught with a number of

difficult challenges, and is today still in its infancy. The issues of co-operation and development between the end-users and suppliers still need to be addressed before this process becomes an accepted procedure.

8.6.6 CMP Consumables (Polishing Pads)

Pads and slurries are the consumables of a CMP process. Historically, the polishing tools were considered to be the technology's major contributors. But consumables are now being recognized as being equally important. With a new group of consumables being made available, some customers select the slurries first, then review their compatibility with existing tools, and based on the results, select a polishing tool. The importance of CMP consumables is expected to increase with the advent of Cu-CMP and low-*k*-CMP. The general characteristics of polishing pads are described here.

Polishing pads are made of either: 1) polyurethane-impregnated felts; or 2) closed-cell cast and sliced polyurethane (with filler material to control hardness). Urethane is utilized because urethane chemistry allows the pad characteristics to be tailored to meet specific mechanical property needs. Table 8-3 lists the important properties of five polishing pads (all manufactured by Rodel). A thorough discussion of the mechanical properties and the surface roughness of CMP pads is found in pp. 68-83 of Ref. 42.

The pads made from impregnated felts are more porous and can carry a larger volume of water or slurry, due to the basic fiber structure of felt. The Rodel Suba

TABLE 8-3 Rodel Polyurethane Pad Properties

Pad	Specific Gravity	Compressibility	Hardness	Type
Politex Supreme	*Not available*	High	Soft	Micro porous felt
Suba IV	0.3	15%	61 (Shore A)	Urethane impregnated polyester felt
Suba 500	0.34	7%	72 (Shore A)	Urethane impregnated polyester felt
IC-1000	0.63–0.85	2.25%	57 (Shore D)	Rigid micro porous cast polyurethane
IC-1400	Top layer IC-1000, bottom layer foam	0.7%–7%	Flexible	Equal to stacked IC1000/Suba IV

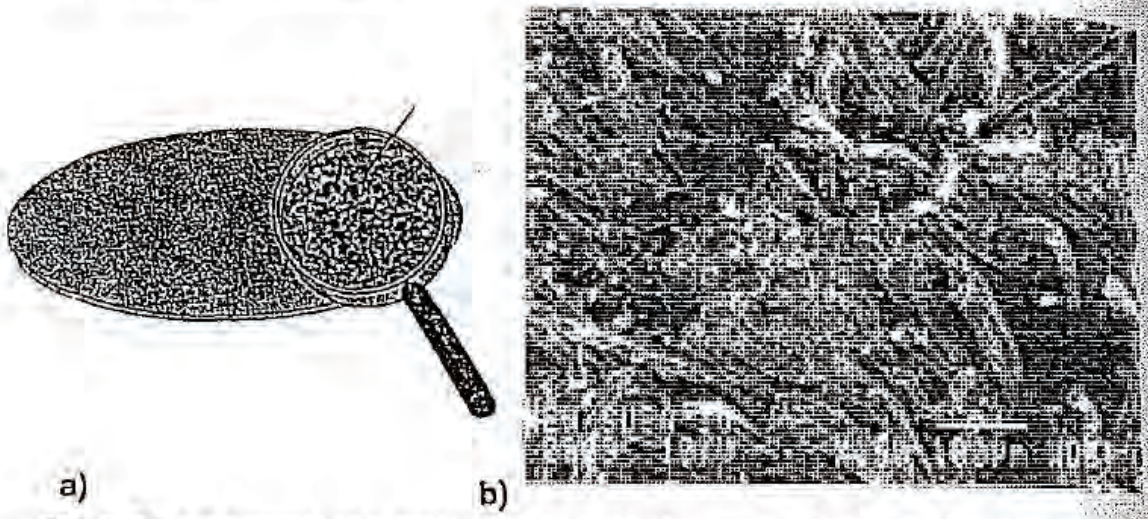


Fig. 8-52 a) Schematic of polishing pad; b) SEM of a new Rodel Suba IV® Pad. Courtesy of Rodel Inc.

IV pads are made of polyurethane impregnated polyethylene felt (Fig. 8-52), which shows a SEM of the surface of a new Suba IV pad.[43] They are classified as *soft-polish pads*, and exhibit greater compressibility than the closed-cell polyurethane types. The Rodel Politex pad is also made from a flexible felt-like material, and it is classified as a *finishing pad*, used to buff wafers with water and other chemicals to remove particles and reduce scratch defects.

The Rodel IC-1000 pads are micro-porous, filled, cast pads that have low compressibility (Fig. 8-53).[44] Pores make up about 30% of the volume of the pad, and the average diameter of a pore is about 30-nm. Such micro-porous, cast pads are classified as *hard-polish pads*. The IC-1000 pad has been the most widely used pad for oxide-CMP processes. The Rodel IC-1400 is a hybrid pad. While there is no consensus, it is generally agreed that the hardness and porosity

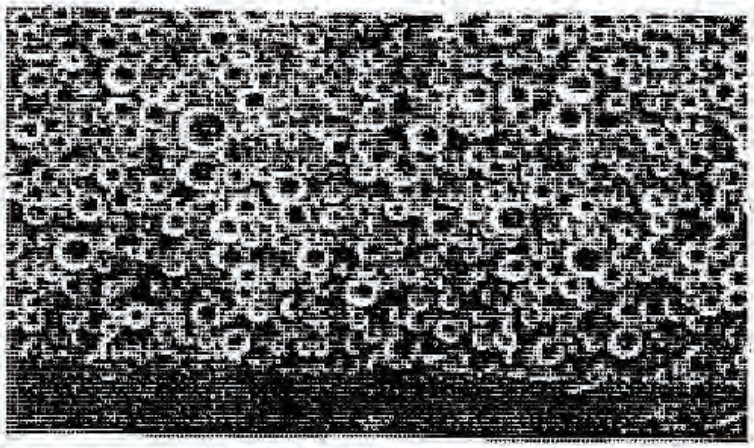


Fig. 8-52 SEM of a new Rodel 1000 IC Pad. Courtesy of Rodel Inc.

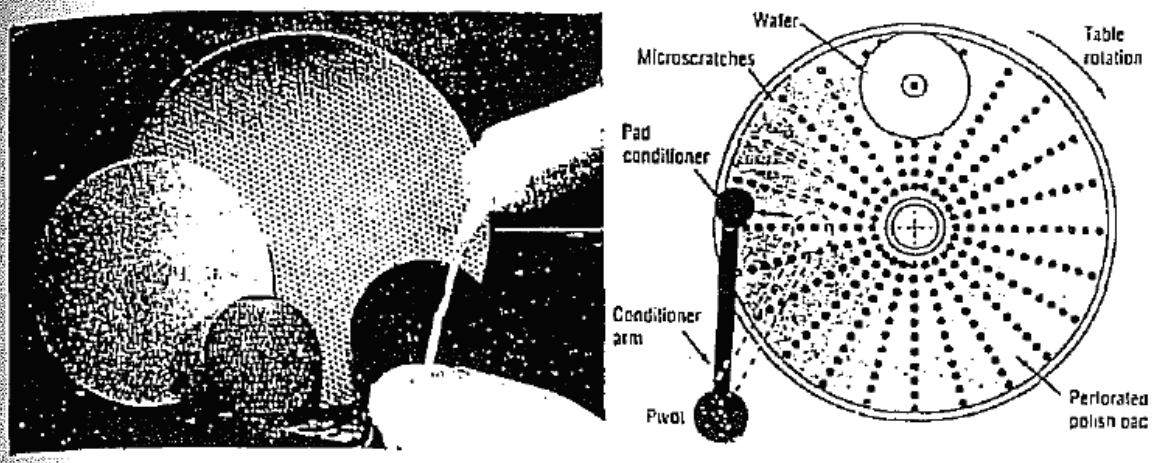


Fig. 8-54 Photograph of polishing pads, showing pad perforations. Courtesy of Rodel Inc.

of the pad are important parameters and that the pad plays a crucial (but still poorly understood) role in determining the final polish results. However, there is no single physical property of a polish pad that is known to lead to good quality for all CMP tools and consumable sets. Polishing pads are supplied by a number of vendors, with Rodel Corp. having the largest market share. Other suppliers of pads include Thomas West, Freudenberg, Fujimi, and Teijin.

In many CMP applications, two pads are simultaneously used, because a hard pad gives better local (within die) planarity, but a softer pad gives better uniformity of material removal across the entire wafer. Using a combination of two pads provides a compromise between these extremes. For example a hard IC-1000 pad is mounted onto a softer Suba-IV pad to form a stacked pad. On the other hand, the IC-1400 is a single molded pad. That is, it is an IC-1000 pad molded to a more porous backing pad that simulates the compressibility of the Suba-IV pad.

As shown in Figures 8-54 and 8-55, the surface of the pad is either cut with concentric-grooves or contains perforations 1-mm-wide and 250-μm-deep punched into the pad. These grooves (or perforations) form channels to help transport the slurry between the pad and wafer. Pores present on the pad surface also aid in slurry transport.

The desired goal of cost-of-ownership for CMP pads has been identified as less than \$0.05 per wafer. The cost of closed-cell polyurethane pads is high (\$0.06-\$0.10/cm^2). Thus, for a 200-mm CMP rotary tool (which requires a 300-mm diameter pad) the initial cost for the pad is about \$50. For a 300-mm CMP tool, the pad must have a diameter of 510-550-mm, and the cost of the pad is about \$125. To achieve a CoO for the pad of less than \$0.05 per wafer, the



Fig. 8-55 Polishing pad cross-section, showing grooves in pad surface.

smaller pad (i.e., for 200-mm wafers) must polish 800 wafers, and the pad for the 300-mm CMP tool must polish 2500 wafers before being discarded.[45]

8.6.7 Pad Conditioners

One crucial aspect of CMP is the state of the surface of the pad. That is, surface roughness and porosity determine: 1) slurry transport to the wafer surface; 2) material transport away from the wafer surface; and 3) the contact area of the pad to the wafer surface. As a result, unless the surface of a pad is continuously maintained in the optimal state of roughness and porosity, removal-rates can vary with time. In fact, a decrease in removal-rate has been most commonly observed (approximately logarithmic with time).[46] For example, starting with a new pad, the removal-rate can drop from 210-nm/min to 75-nm/min after only 50 wafers have been polished. This effect occurs because during the polishing process the surface of the pad rapidly undergoes plastic deformation so that the surface becomes smoother and the pores become filled with pad material (Fig. 8-56). This phenomenon has been termed *glazing*. Such glazing is undesirable, since as the pores of the pad become closed, slurry delivery is decreased to the wafer surface, and reduced polishing rates result. In addition, large waste-particles can become embedded in the pad and cause scratches.



Fig. 8-56 Surface of a glazed pad (Rodel's Suba IV® pad) after polishing.

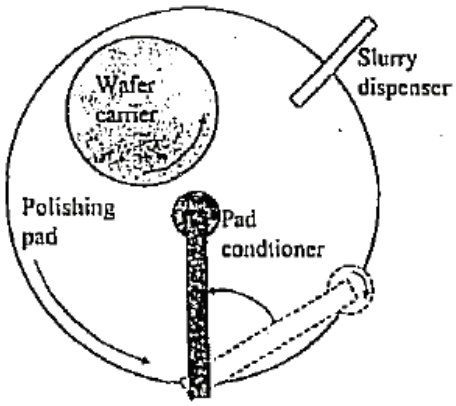


Fig. 8-57 Schematic drawing of pad conditioning. Reprinted with permission of Solid State Technology. Published by PennWell.

A consistent (and rapid) removal rate is necessary for production-worthy CMP processes. This can be achieved by a process that refreshes the surface of the pad, and this technique is called *pad conditioning*. Figures 8-57 and 8-58 show a typical conditioning process in which the pad-conditioner (a disk whose surface is embedded with a diamond grit) sweeps back-and-forth across the rotating pad to restore the pad to its required surface planarity and surface roughness. The pad-conditioning process also opens the pores of the pad by forming micro-scratches on the pad surface. An example of what the pad-conditioning process accomplishes is shown in Fig. 8-59, which contains SEM micrographs of the

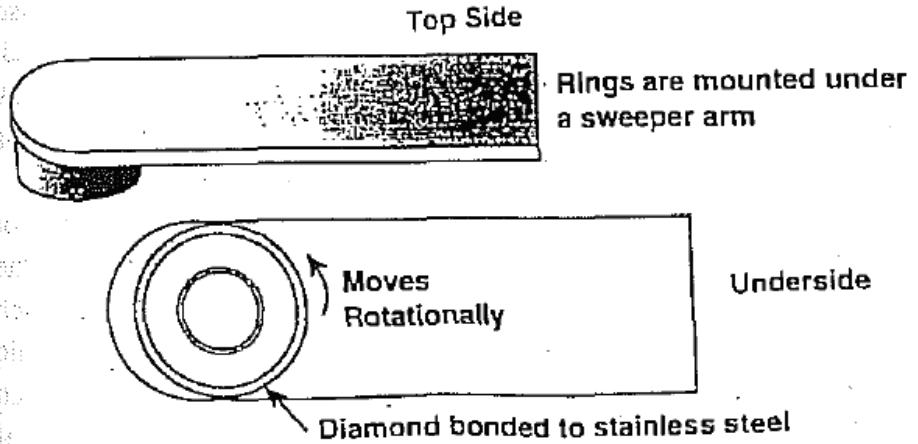


Fig. 8-58 Diamond embedded ring pad conditioner.

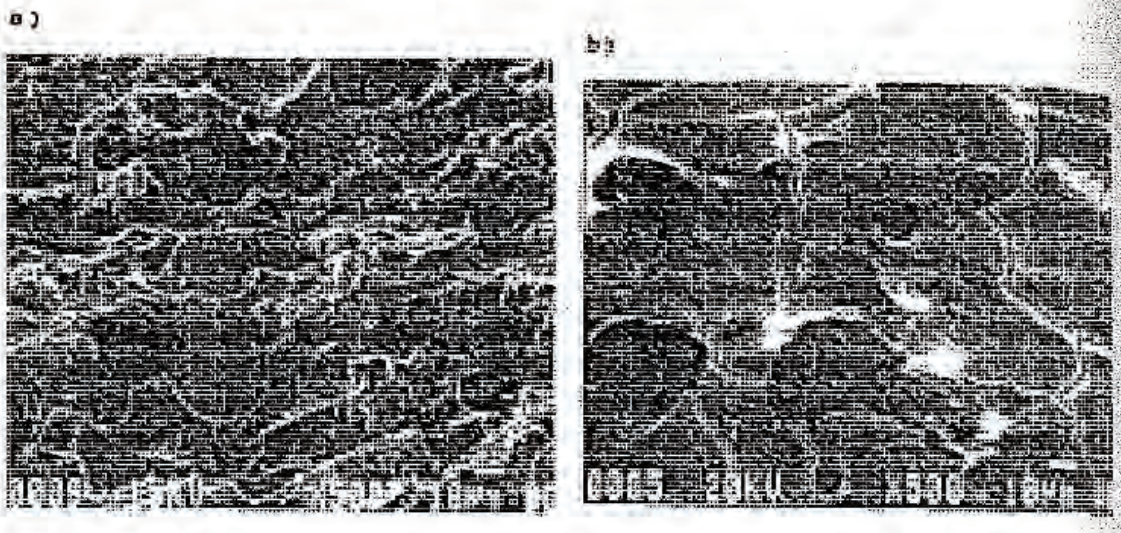


Fig. 8-59 (a) Polishing pad cross section. (b) Effect of pad conditioning.[47] Reprinted with permission of Solid State Technology. Published by PennWell.

pad surface of a Rodel IC-1000/Suba IV stacked pad, with and without conditioning. The SEM of the non-conditioned pad shows that its pores are completely clogged.[47]

The simplest form of pad-conditioning is done with a nylon-toothed brush. This is adequate for large particle removal or for conditioning soft, woven (or high-nap) pads. However, for the harder, molded closed-cell polyurethane pads (i.e., the Rodel IC-1000 or IC-1400 series), pad-conditioning is carried out by sweeping the pad surface with a disc containing 80-120-grit size diamonds embedded in a nickel plate (Fig. 8-60). Diamond, the hardest material available, is well suited for conditioning the hard-polyurethane-pad materials, but it is also brittle and can fracture. Alternative conditioning-abrasives have been pursued, since diamond particles may fracture and break away from the conditioning disk and become embedded in the pad. Since 80-grit diamonds are 100-200-microns in diameter, this can lead to severe scratching of the surfaces being polished.

Several approaches are used to mitigate the scratching problem. First, superior types of bonding mechanisms have been developed, which chemically bond the diamonds to the base of the conditioning-pad. Diamonds adhere better to such pads than to those that employ a plating technique for bonding. Second, periodic cleaning of the conditioning wheel is also carried out. Finally, some pads (such as the Rodel IC-1000) are fabricated with grooves created in the top surface of the pad. These grooves not only provide the pad with more slurry-carrying capacity, but dislodged 80-grit diamonds tend to be collected in the deep surface

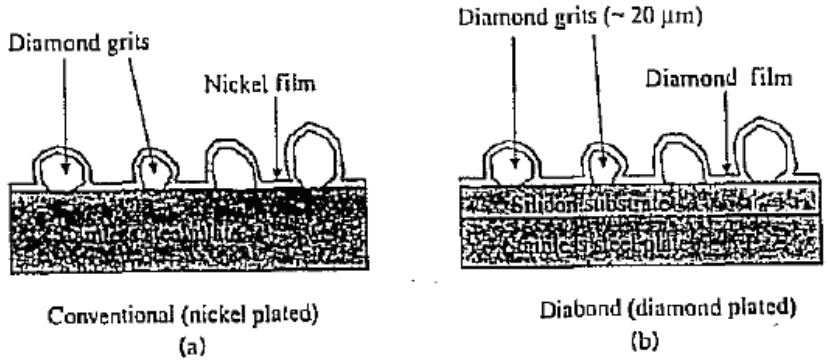


Fig. 8-60 Surface of pad conditioner showing bonded diamond grit.

grooves. Since the ridge tops of the pad between the grooves perform the polishing work (and not the pad surface within the grooves), the collection of the diamonds in the grooves minimizes scratches on the wafer surface. The pad is also cleaned with a high-pressure spray between wafers to dislodge and wash away the diamonds embedded in the grooves. (This high-pressure spray is more critical near the end of a pad's life, because at that point the grooves have become shallow. At that point, the diameter of the diamonds collected in the shallow grooves can exceed the groove depth and scratch the wafers). By using all of these techniques, diamond loss (and resultant microscratching), now represents much less of a problem.

Many polishers incorporate a continuous *in-situ* pad conditioning capability (a unit that moves across the pad during polishing, see Fig. 8-61 and Fig. 8-62).[48] This allows each wafer on the platen exposure to a pad with the same amount of conditioning, resulting in a more uniform removal-rate from wafer-to-wafer.

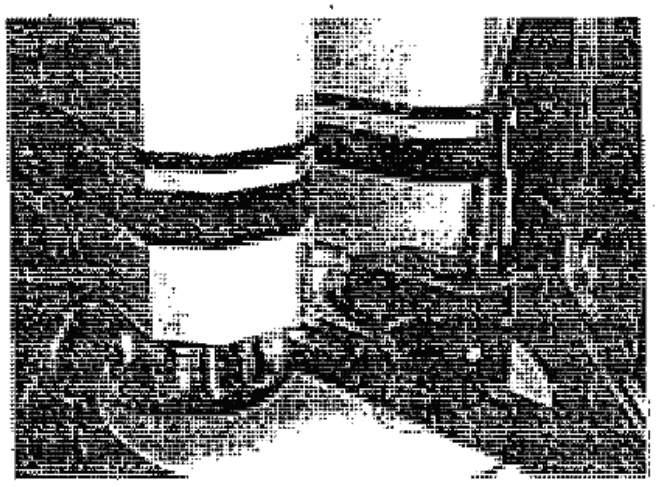


Fig. 8-61 In situ pad conditioner. Courtesy of Strasbaugh Inc.

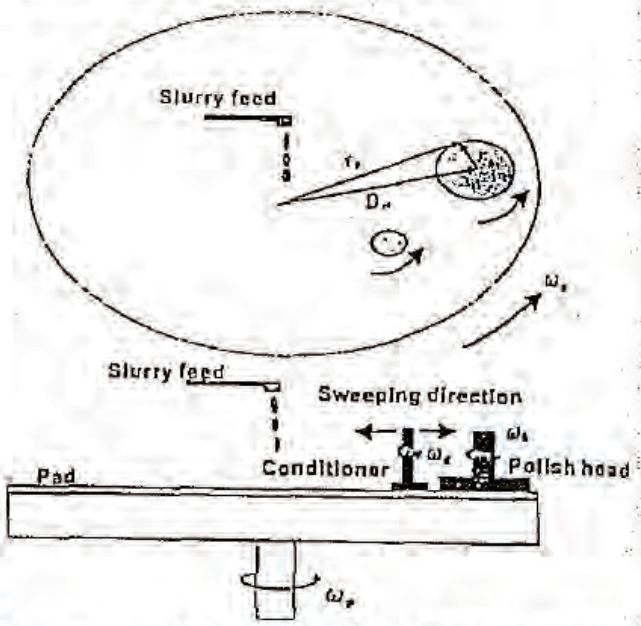


Fig. 8-62 *In situ* pad conditioner.

New pads also require a break-in conditioning-cycle to establish a constant removal-rate. In this procedure a set number of dummy wafers are polished for the purpose of stabilizing the pad condition and performance. Break-in conditioning-cycles can consume a significant fraction of the pad's useful life, and the polishing tool is not available for productive use while this is in-progress.

As a final note, it should be pointed out that conditioning of the pad is used to reduce the variation in removal-rate that occurs over a pad's lifetime. Yet the act of pad conditioning becomes yet another variable. As a result, elimination of pad conditioning would be an approach toward improving process stability. Use of a fixed-abrasive, one-use-pad, that is disposable or continuously replaced, is a potential path toward that goal.

8.6.8 Slurry-Free Pads

Slurry-free (or *fixed abrasive*, *FA*) CMP pads have also been developed. In these pads the traditional slurry-and-pad consumable-set is replaced with a pad in which the abrasive is attached to the pad. The main benefits of slurry-free pads are higher degrees of planarization and lower cost-of-ownership. The reduction in the cost-of-ownership is due to a number of factors. First, there is the reduced cost of consumables. Second, the cost of slurry handling is eliminated. Third, there is no longer a need for a slurry-distribution system. Finally, the problems of slurry-stability and solution-variation are eliminated. To date, most development work on slurry-free pads has centered on ILD and STI applications, but programs for slurry-free copper-CMP are also underway.

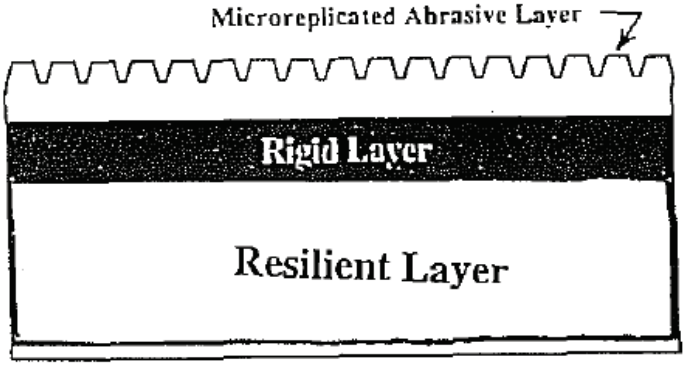


Fig. 8-63 Slurry-free (Fixed abrasive) pad structure.

A pad supplied by the 3M Co. embeds the abrasive material in the pad itself, in the form of a precise micro-replicated abrasive surface and a bi-component sub-pad system (Fig. 8-63). The abrasive surface can be constructed in one of several configurations chosen for specific performance criteria. Figure 8-64 shows two of the surface features currently in use.[49] For dielectric applications the slurry-free pad is used with a simple basic solution. For metal-CMP, a

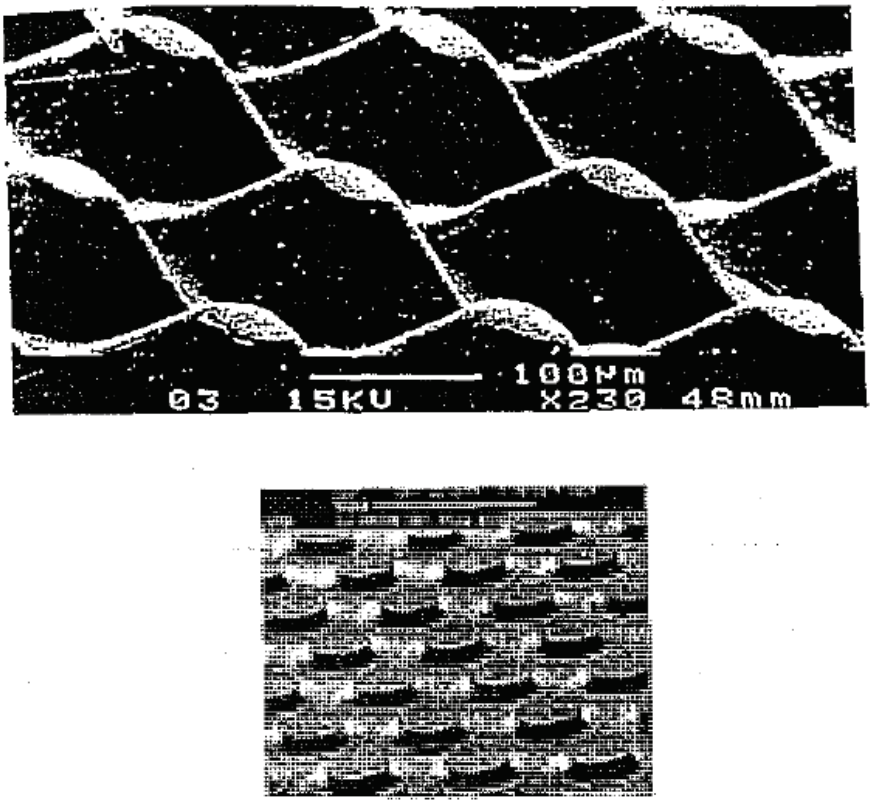


Fig. 8-64 Examples of slurry-free pad surface structures. Courtesy of 3M Co.

a specially formulated oxidative solution is needed, with a pH value of 2 to 7.

These pads are composed of four layers: a micro-replicated abrasive; a rigid layer; a resilient layer; and a self-adhesive backing. These slurry-free pads can be used either on a moving-platen (rotary or linear) or a fixed-platen. In early 2000 Applied Materials and 3M entered a strategic agreement to jointly develop a slurry-free pad for the Applied Mirra® multi-platen rotary CMP-tool. 3M also has a marketing alliance with pad supplier Rodel. SpeedFam-IPEC also announced an effort to develop a slurry-free pad for their rotary/orbital CMP tools. There is some concern about the consistency of slurry-free pads for rotary CMP tools throughout the pad life. As the pad wears, the surface area tends to change, which may cause a non-stable process.

These concerns are avoided in web-format pads for fixed table, roll-to-roll pads. The use of a slurry-free pad in a fixed-table (platen) CMP system was pioneered by Obsidian Corp. In Obsidian's web-based tool, a roll of abrasive pad material is spooled across a platen and incremented after each wafer. No slurry or pad-conditioning is required.[50] (Fig. 8-65). However, one concern about this slurry-free approach is the uniformity of polishing performance from roll-to-roll.

Slurry-free pads have demonstrated excellent planarization capability. This capability allowed a slurry-free STI CMP-process performed on an Obsidian 501 CMP tool to demonstrate exceptionally tight nitride ranges and minimal dishing (Fig. 8-66).[51] That is, the removal of nitride during the trench oxide CMP process is very uniform across the wafer, with only a 250-Å variation in the remaining nitride films after the polish step. The dishing in the field oxide regions was also very small, with the maximum dishing being less than 400-Å.

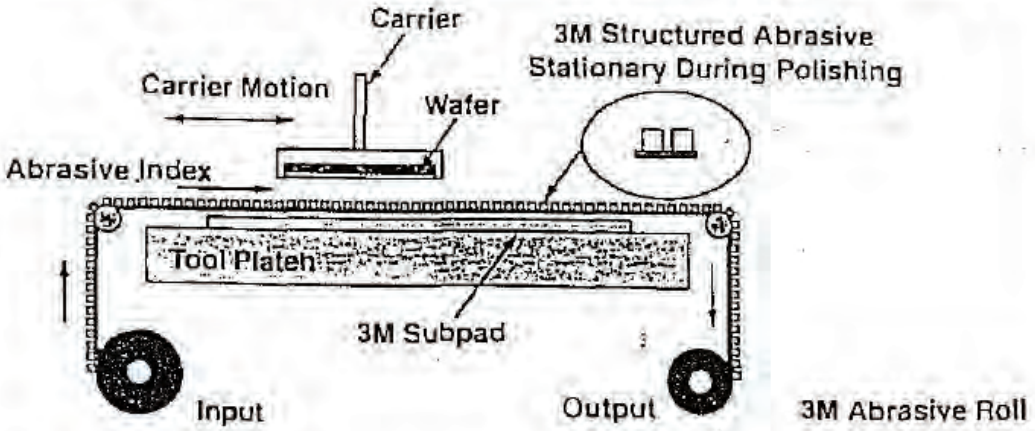


Fig. 8-65 Schematic of linear web format CMP tool, with indexing, slurry-free pad.

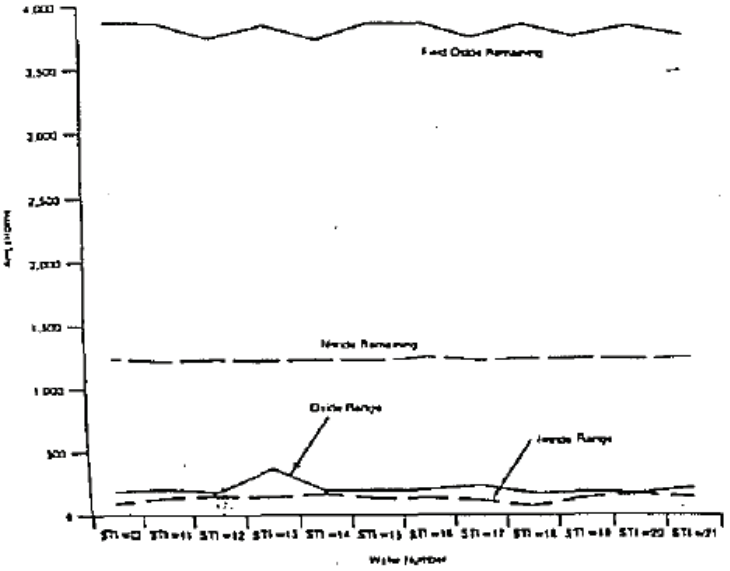
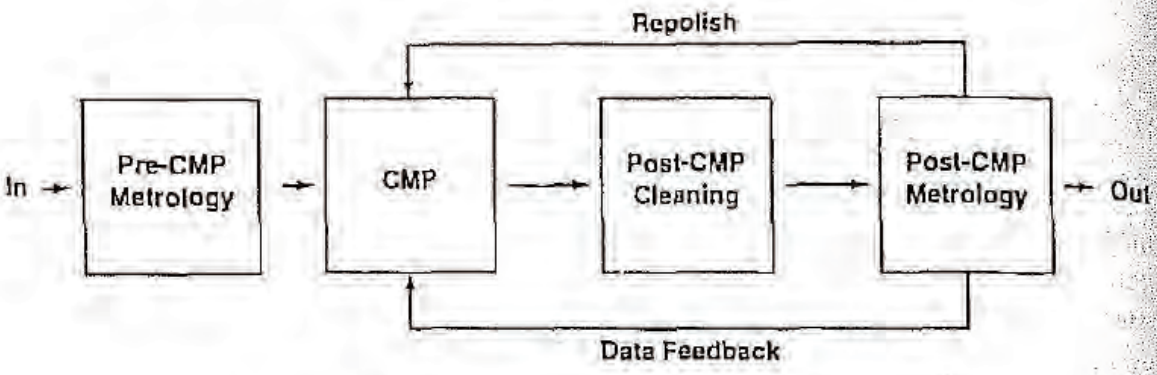


Fig. 8-66 Fixed abrasive polishing results on Linear CMP tool (roll-to-roll pad). Test performed on the Obsidian 501 CMP tool at the 3M Applications Lab. Courtesy 3M Co.

8.6.9 Endpoint-Detection in CMP

Since CMP is a process for reducing thickness at selected locations on a wafer, it is necessary to be able to determine when the process has reached its *endpoint*. There are two types of processes that need to be controlled: 1) polishing a dielectric to a known final film thickness; and 2) polishing an inlaid, usually multi-layered, metal film and stopping on the surrounding dielectric (after all of intra-inlaid metal is removed). The control of the first of these processes, that of polishing an oxide, can only be accomplished by making measurements on the film to be polished before the CMP process, and again after the wafer has been polished for some pre-determined time (based on the pre-polish film measurements). This is termed an *in-line endpoint-measuring* process. Another approach is to measure the film thickness as the polishing process is proceeding. This is called a *real-time, in-situ endpoint-detection* process. At the present time it is possible to control the second type of CMP process listed above (but not the first) by using *in situ endpoint-detection*. Nevertheless, both techniques will be described.

In the *in-line endpoint-measuring* method used for controlling dielectric-CMP processes, a pre-CMP measurement of the dielectric film to be polished is performed.[82] This measurement is made over an area containing dense device features to determine the thickness, T_O, shown in Fig. 8-17. Based on the



Polishing followed by measurement to confirm the thickness is the standard cycle in the CMP module.

Fig. 8-67 *In-line* endpoint detection method used in CMP (i.e., polishing, followed by measurement to confirm thickness - and re-polishing if necessary - is the cycle used with *in-line* endpoint detection).

measured value of T_o, the polishing time (t_{final}) needed to reduce the film to a final required thickness, T_{final} (see Fig. 8-17) is calculated. The value of t_{final} is chosen from CMP calibration data based on applied pressure, rotation speed, and average pad-life-degradation. This produces a value of T_{final} centered within the allowed process window. Various factors may alter the actual value of T_{final} realized at t_{final} (e.g., by as much as 20%, 1-σ, during process development). If the material removed by t_{final} exceeds the allowed limit, the wafer must either be scrapped or more dielectric material must be deposited. If the remaining thickness is excessive, the wafer can be returned to the polisher for more removal. In fact, a typical manufacturing approach to oxide-CMP in which in-line endpoint-detection is employed, is to deliberately *underpolish* the SiO_2 layer. The film is then measured and polished a second time, using an adjusted polishing time. The cycle is repeated until the measured thickness is within specification limits. This procedure is often carried out on a look-ahead, or "pilot" wafer for each lot (Fig. 8-67), and the remainder of the lot is polished according to the time calculated when polishing the pilot wafer.[53]

This loop is the basic approach used in dielectric CMP (Fig. 8-68), but it has some drawbacks. The most significant are that the technique is time-consuming, measurement-intensive, and reduces the overall time the polisher can be kept in use. In addition, there is a risk of wafer breakage when being removed for measurement. Some thickness metrology tools are designed to be placed inside the polishing tool housing as another subsystem (e.g., the ITM® tool of Nova

Measuring Instruments). A central handling mechanism transports the wafer to the measurement position and then to a scrubber, then back to polish or to the output cassette. Multiple points on the wafer can be measured with a small optical head. This head (rather than the wafer stage) is moved, providing about 5 times faster movement.

The use of a *real-time, in situ endpoint-detection* method allows the problems of in-line endpoint-detection to be circumvented. *In situ* endpoint-detection is a way of identifying the completion of a polish step without having to move the wafer outside of the tool platform. *In-situ* endpoint-detection would permit the user to stop at the desired remaining thickness without having to re-polish. This would also improve throughput and increase the process window. Several methods have been developed to perform *in-situ* endpoint detection in CMP, including: 1) front-side laser interferometry or spectrometry; 2) optical-sensing of multiple-points on the backside of the wafer; and 3) monitoring the current

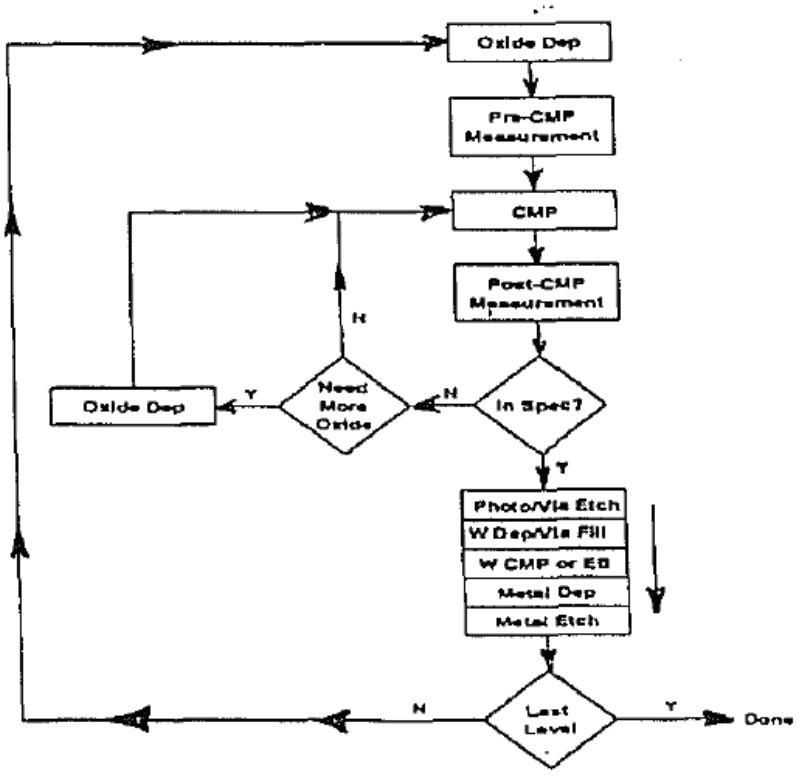


Fig. 8-68 Implementing the *in-line* endpoint-detection method used in CMP.

supplied to the motor of the wafer carrier. Unfortunately, as noted above, no satisfactory *in situ* methods yet exist for polishing homogeneous dielectric films to a known final thickness.

The *wafer-frontside laser-interferometry technique* is used by Applied Materials in their Mirra® tool, the *spectrometer reflectivity technique* on the Lam Research Teres® tool and in the Speedfam Auriga® tool. The name of the Applied Materials endpoint system is the *in-situ removal monitor* (ISRM).[54,55] Each polishing platen has a red laser and a detector mounted on the polishing table. A window in the polishing pad allows the laser to be aimed at the top surface of the wafer at 45° to the wafer surface. A detector is used to sense the reflected light (also collected at 45° from the surface). As the laser/detector passes under the wafer, the laser light passes through the clear window in the polishing pad, and reflections are sampled. The intensity of the reflected light is measured as a function of time (the light-intensity vs. time curve in Fig. 8-69).[56]

A typical reflectance curve for polishing Cu/Ta films on an oxide dielectric is shown in Fig. 8-70. After the maximum reflectance is reached, the process starts to break through the Cu layer to the semi-transparent Ta layer. At that point a sharp drop in reflectance occurs. Eventually, the copper filling the trenches continues to reflect a small constant amount of light.

The Applied Materials Mirra® CMP tool uses sophisticated signal processing/detection algorithms to reduce noise from the slurry and pad surface condition (Fig. 8-71). In development tests, endpoint accuracies to within 20-nm were reportedly achieved. In W-CMP polishing, the ISRM system measures the reflectivity of the metal film and signals the transition from metal film to oxide. An over-polish time is specified to ensure complete removal of the liner film.

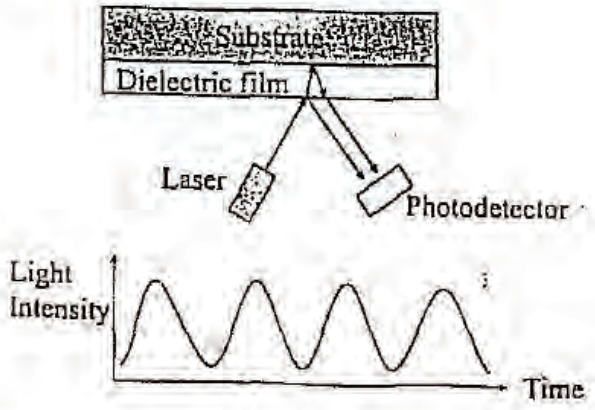


Fig. 8-69 *Wafer-frontside laser-interferometry* endpoint-detection for dielectric CMP.

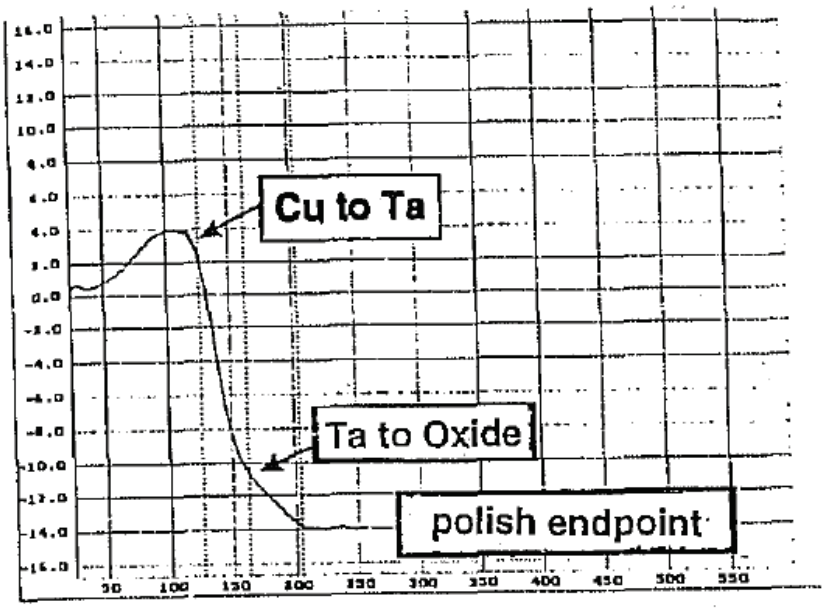


Fig. 8-70 Typical reflectance curve for polishing a Cu/Ta/oxide film.

In the Lam Research Teres® tool, a *multiple-wavelength, spectrometry-based endpoint-detection system* is employed. It can continuously monitor the thickness of transparent films as they are being polished. It can also measure the thickness of very-thin metal films. Thus, during a Cu CMP-process, the endpoint of the Cu polishing step can be sighted before all the Cu is removed. This allows a set of process conditions that rapidly removes Cu to be initially used. When the point of complete Cu removal is detected, the polish conditions can be switched to those that result in a slower Cu removal rate. This allows the Cu to be completely

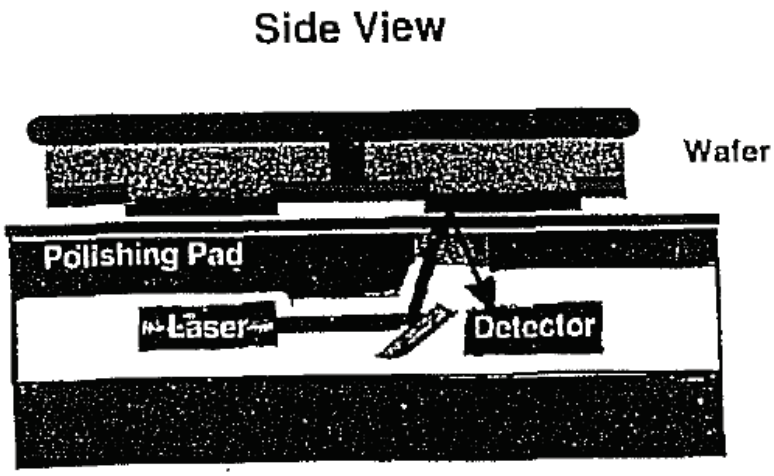


Fig. 8-71 Schematic of the Mirra® laser-reflectometer light path. Courtesy Applied Materials.

removed over the underlying Ta liner film without significantly eroding the exposed Cu in the trenches. In the Speedfam Auriga® multi-head polisher, a halogen lamp is used as the light source and diode arrays are employed to detect the light reflected from the surface of the wafer being polished. Endpoint in W-CMP is determined from an intensity change in the reflected-light as the exposed surface changes from tungsten to oxide.

The *optical monitoring of the wafer backside* as an endpoint-detection technique for monitoring the oxide-CMP process has been developed by Luxtron.[57] In this approach, the wafer's backside is illuminated with infrared radiation (to which silicon is substantially transparent). The radiation is therefore able to sense the film thickness on the wafer's front side as it is being polished.[58] By mounting several sensors onto the wafer carrier, multiple points on the wafer can be monitored. The film thickness is measured using the basic principles of optical interferometry as was described in earlier paragraphs. For wafers with no metal on the front surface, this technique is quite straightforward. However, if there is some metal on the front surface (e.g., in a multilevel-metal process), the situation becomes more complicated, because the metal will block some of the incident light. Fortunately, there is generally sufficient space between device features to allow light to pass. Figure 8-72 shows a schematic of a single sensor probe. Two optical fibers connected to an infrared emitter and detector are mounted in a modified wafer-carrier of a commercial CMP tool. Electrical communication from the sensors is accomplished via a wireless link. As many as four sensors can be mounted in a single carrier.

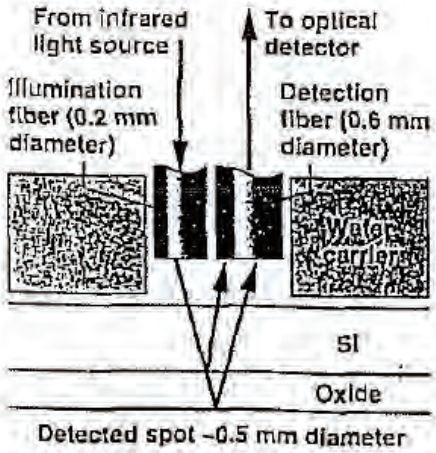


Fig. 8-72 Schematic of an optical sensor embedded in a wafer carrier.[57] Wafer is backlighted with IR light to measure films on the front surface. Reprinted with permission of Semiconductor International.

The *motor-current-monitoring technique*[57] is a production-proven endpoint-detection method that works well when polishing a material down to a stop layer (such as the polishing of a CVD-SiO_2 film on a silicon nitride stop layer in shallow-trench isolation,[58] and the polishing of W over oxide in W-plug CMP processes[57]). It is also relatively easy to implement. Circuitry such as a *current shunt* or a *Hall-effect probe* is used to monitor the current supplied to the motor that rotates the wafer-carrier (Fig. 8-73a). Since the wafer-carrier is driven at a constant rotational speed to maintain a constant polishing rate, the drive-current is varied to compensate for any load changes on the motor. This makes the motor-current sensitive to frictional changes at the wafer surface. The largest changes occur when one material has been polished away, exposing a layer which has different polishing characteristics (e.g., that could function as a CMP-stop layer, see also Fig. 8-73b). Since there is no layer breakthrough in oxide-CMP (as can be seen in Fig. 8-25a), the motor-current-monitoring technique is not appropriate for such applications. Luxtron and Aplex are vendors of such motor-current endpoint controllers (e.g., the Luxtron Xinix Model 2450®).

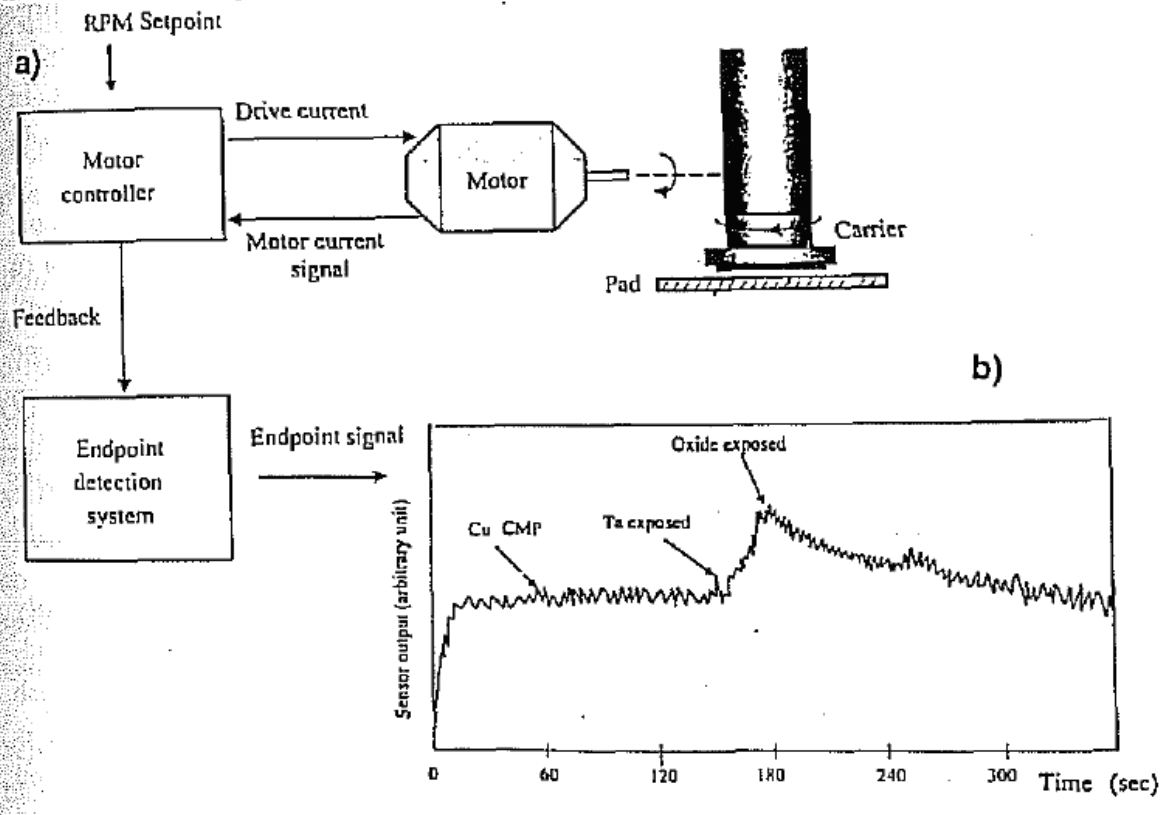


Fig. 8-73 a) Motor current endpoint detection. b) Motor current output during Cu/Ta/oxide CMP process. Courtesy of Aplex.

8.7 CLEANING ISSUES IN CMP

CMP is an inherently dirty process that can introduce yield-reducing defects to the wafer. That is, after a polishing step, the wafer-surface contains slurry particles, abraded pad-material, and debris that was removed from the wafer-surface. Residual slurry specifically introduces foreign particles, metal contaminants, and chemicals; all of which are undesirable for the remaining steps of chip fabrication. If the slurry particles alone were allowed to remain on the wafer, such contamination could obstruct electrical connections, create voids during subsequent oxide- or metal-deposition steps, or cause corrosion of exposed metal areas. Some early resistance to the adoption of CMP stemmed from such contamination concerns. As a result, significant effort has been expended to develop post-polish cleaning techniques so wafers returned to the cleanroom after CMP are not contaminated.[60]

To prevent CMP-based contamination, wafers are subjected to post-CMP cleaning steps that return the wafer-surface to an acceptable level of cleanliness. While the main objective is to provide a wafer-surface that is free of slurry particles, the cleaning process must also remove organic-residues and trace metal ions. The cleaning process must do this without introducing further defects. In the case where inlaid metal is exposed (e.g., in a W-plug CMP process or a in Cu-damascene CMP process), corrosion must also be controlled or avoided.

Since CMP is a wet process, CMP cleaning-technologies are mostly aqueous. The post-CMP wafers are kept wet as they are moved from the polishing tool to the cleaning station. The initial wafer-rinsing (after the polishing is completed), is done while the wafers are still on the polishing pad. Spray rinsing is also done as soon as the wafers are lifted from the pad. If wafers are exposed to air for even one or two minutes before such spraying (e.g., while waiting to be unloaded), subsequent cleaning becomes difficult (and sometimes impossible). Immediate removal of metal-polish slurries is also critical, because these can continue to react and remove material even after polishing has ceased.

Current models of CMP polishing tools either have an on-board (i.e., fully integrated) cleaning station, or they have the capability to interface with a stand-alone cleaner (in the form of a cassette-to-cassette track with stations for brush scrubbing [one or more], rinsing, and spin-drying). If a stand-alone cleaner is employed (as was common in early CMP processes), the wafers are transferred from the polisher to the cleaner by a human worker. If the cleaner is integrated with the polisher, (the most common method in the latest generation of CMP

tools), automated wafer transfer is used (i.e., the wafer is moved from the polisher to the cleaner by a robot).

To prevent additional wafer and slurry contamination, the number of airborne particles should be minimized by filtering the air fed to the CMP tool with the same HEPA filtering techniques used for the cleanroom environment. The cleanliness of the DI water used to rinse or submerse the wafers during cleaning is also an issue. Tubs or baths in which wafers are kept wet must be continuously refreshed with new DI water. The CMP tool should automatically rinse all surfaces that come in contact with slurry, including the polishing table, drain tubs, carriers, and pad conditioner. These areas should also be cleaned periodically to prevent dried slurry buildup. As mentioned previously, slurry is composed of very small abrasive particles, and filtering by conventional methods is not easy. Nevertheless, such filtering is needed because large agglomerates can develop within the slurry and have serious device-defect consequences.

8.7.1 Mechanisms of Particle Adhesion and Removal from Surfaces

Particles are deposited on wafers as aerosols from the air or as particles present in the liquids in which wafers are immersed. A description of the techniques used to remove them (e.g., after a CMP polishing step) rests on an understanding of how such deposited particles adhere to wafer surfaces. Thus, a short description of this phenomenon follows.

If a solid particle wanders close enough to the vicinity of a solid surface, the *Van-der-Waals force* (*VdW*) can attract and capture it. The particle is then *physisorbed* on the surface. The Van-der-Waals force is universal, and is thus effective in trapping all types of solid particles (and gas molecules, as well). Other forces that physisorb particles include *static charge on the particle*, and *capillary adhesion* (the force exerted by a liquid film between the particle and the surface), but the VdW force is usually the strongest of these. In fact, it can be so strong that the particle may elastically (and sometimes even plastically) deform so that the area of contact between the particle and the surface is increased. This makes such binding even stronger. Nevertheless, particles held by VdW forces can be removed by the techniques described here.

There are three methods used to remove physisorbed particles: 1) *electrical repulsion of the particle and the surface*; 2) *chemically assisted removal* (particle dissolution, particle oxidation [and then dissolution], and lift-off by slight etching of the surface); and 3) *mechanical dislodgement* (sliding and rotating the particle

by the application of a shear force). These mechanisms are described in more detail in the next sections.

Note that cleaning by immersion-only is not effective for post-CMP cleaning, due to the particle adhesion forces and boundary layer effects. That is, merely rinsing in DI-water or immersing the wafers in a wet-chemical solution has not been found to adequately remove slurry particles from the wafer-surface.

8.7.1.1 Electrical Repulsion of the Particle and the Wafer Surface: Besides the particle adhesion forces discussed above, other inter-particle (or particle-to-surface) forces exist which can be either attractive or repulsive. The most important of these forces are those involving the charge near the particle surface. That is, the outer-most charge *near* the surface of a particle most influences the way a particle will interact with other particles or surfaces. Particles and surfaces immersed in a liquid can be charged by the adsorption of ions from the solution. The charge thus adsorbed on the surface is balanced by an equal but oppositely charged layer in the adjacent liquid, resulting a so-called *electric double layer*. (Fig. 8-74) The potential associated with this adsorbed charge on a particle or surface is termed the *Zeta potential* (ZP) and is also depicted in Fig. 8-74. Since the particle or surface adsorbs ions from the liquid, the ZP of the particle is also

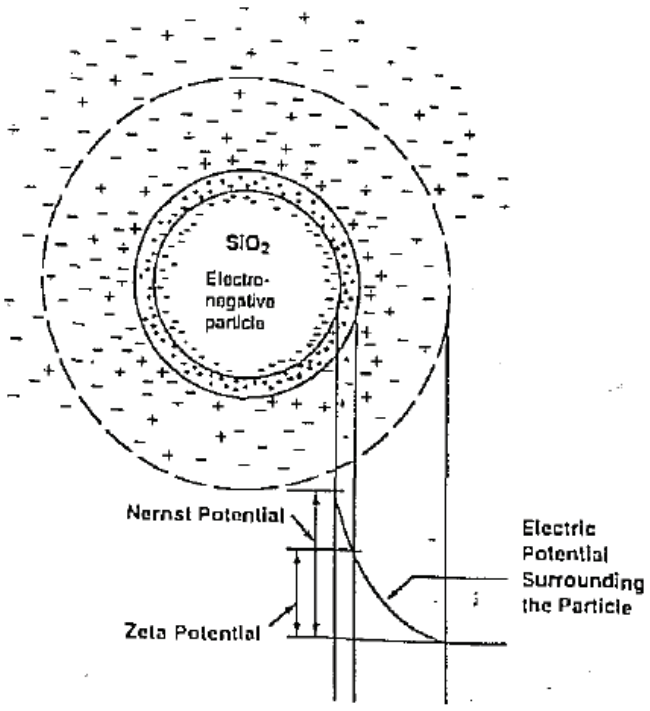


Fig. 8-74 Slurry chemistry electrostatic layers surrounding an SiO_2 particle (including a depiction of the electric dipole layer).

consequently dependent on the pH of this surrounding media. For example, Fig. 8-75 shows the ZP of a variety of particle types (SiO_2, Al_2O_3, W, poly-vinyl alcohol [PVA], and Si_3N_4), as a function of the pH of the liquid in which they are immersed. The silica particles have a negative potential of –30-mV if they are immersed in a liquid with neutral pH (i.e., pH = 7). Also note that their ZP (as well as those of the other particles shown in Fig. 8-75) becomes strongly negative as the pH is increased above pH = 7. (The ZP is also affected by the addition of a surfactant to a liquid, and thus the ZP of a solid-surface can likewise be manipulated by the addition of surfactant to the liquid.)

To prevent slurry particles from sticking together, it is best to immerse them in a solution that will stabilize their respective ZPs to either a positive or a negative value. This will ensure that the particles will repel each other (i.e., this repulsion - see Fig. 8-76 - will counteract such attractive forces as the Van-der-Waals forces). Consider the case of silica particles in a silica-based slurry. Since silica particles tend to have negative values of ZP, the net-negative electric charge near the particle-surface produces a force that tends to make the silica particles repel each other. However, to ensure that these repulsive forces are strong enough to overcome the opposing attractive forces (such as VdW), a stabilizing chemical that increases the pH of an aqueous solution (e.g., NH_4OH) is added to the slurry.

From the data shown in Fig. 8-75 it can be seen that to keep the silica particles in a slurry from agglomerating, they should be immersed in a liquid with pH values > 7 (basic). On the other hand, for slurries containing alumina particles, one might want to consider using a slurry solution with a low pH (i.e., < 7 [acidic]), because in such an acidic solution the alumina particles exhibit a positive ZP.*

To remove particles from a surface, consider the case of removing silica-slurry particles from a wafer-surface covered with oxide (i.e., a condition that exists at the completion of an oxide-CMP step). Since the ZP of the SiO_2 surface is also negative (at neutral pH), it is preferable to use a cleaning chemistry with a high pH (i.e., > 7). This causes the silica slurry particles adhering to the oxide

* We noted earlier that the CMP process also consumes large volumes of DI water. Work has thus been carried out to develop methods of treating the CMP wastewater so that it can be re-used. One phase of such reclaim treatments involves removing the slurry particles from the wastewater. For the case of W-CMP waste water, it was found that by a simple adjustment of the pH of this water, the slurry particles would agglomerate into large clumps, making their capture and removal quite easy.

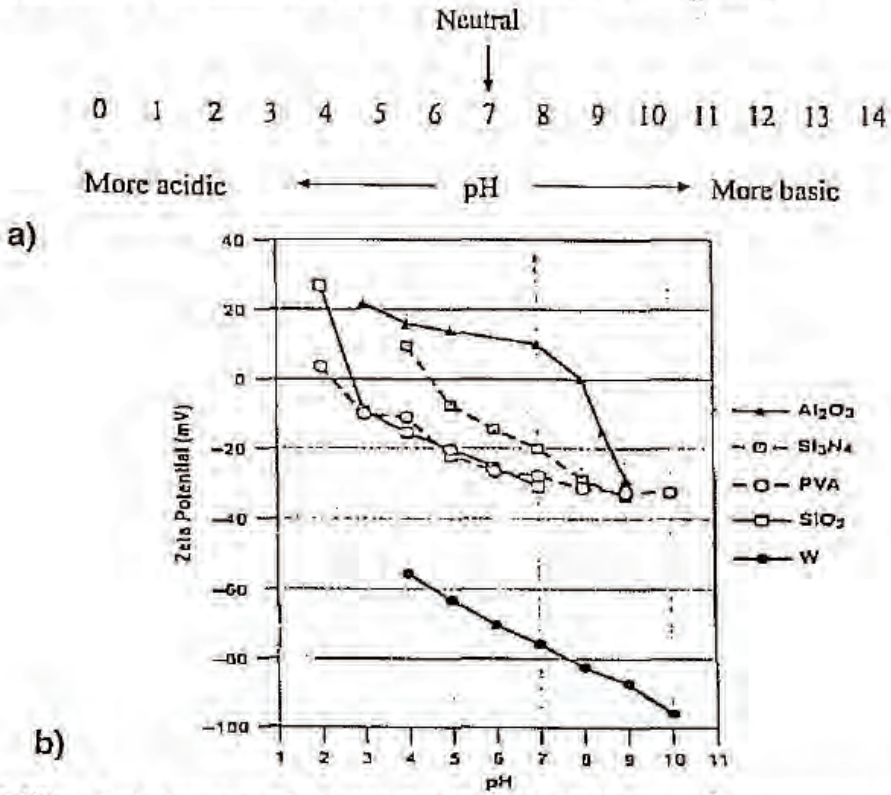


Fig. 8-75 a) Relationship of pH and acidity and alkalinity. b) Zeta potentials versus pH for particles of various types of slurry particles and materials in an aqueous solution.

surface to acquire additional negative charge, which increases the repulsive component of the force between particles and the surface, making any par-ticles adhering to the surface easier to remove. Note that in most cleaning solutions, chemistries are extremely dilute (i.e., 1:50, chemical:DI water).

It should also be mentioned that for many CMP processes, including W-CMP and Cu-CMP, the exposed wafer-surface at the conclusion of the polish-step contains more than one type of material (i.e., not just oxide). For example, when the CMP-process for polishing W-plugs (or Cu-interconnect-lines) is completed, more than 50% of the exposed wafer-surface area consists of oxide, and the remainder consists of metal. Nevertheless, since the majority of the exposed surface is still oxide, the most effective cleaning chemistries will still be those that remove slurry particles from oxide-surfaces. Hence, cleaning solutions used to remove silica-slurry- particles from these surfaces should have pH values > 8.

The same considerations apply to the cleaning of slurries that use alumina particles. That is, as shown in Fig. 8-75, alumina particles have a positive ZP if they are immersed in a solution with a low (acidic) pH (i.e., < 8). Thus, if an acidic cleaning solution were to be used, the alumina particles would be attracted to the electronegative oxide surface. To prevent this situation, the cleaning

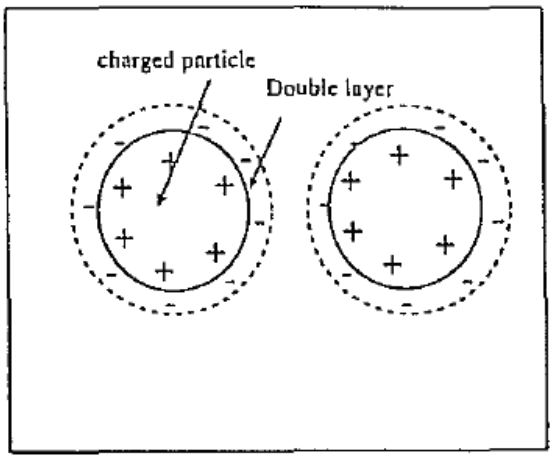


Fig. 8-76 Electrostatic repulsion keeps particles separated by repulsion of charged double layers.

solution must again have a high pH (> 8). This makes the ZP of the alumina particles negative, so they are repelled from the oxide surface on the wafer.

Figure 8-77a compares brush-cleaning results after oxide-CMP with and without NH_4OH in the cleaning solution. Figure 8-77b compares the results after W-CMP with and without NH_4OH. In both cases, the use of NH_4OH (which raises the pH of the solution) is seen to reduce defects by as much as 75%.

8.7.1.2 Chemically Assisted Particle Removal: The *chemically assisted* particle removal techniques are implemented in some wet-cleaning procedures (Fig. 8-78). For example, in the SC-1 step of the RCA clean (see Vol. 1, 2nd Ed. Chap. 5), several chemically-assisted particle removal mechanisms are believed to work in tandem. That is, the SC-1 solution removes some types of particles by dissolution (i.e., organic particles). For the case of cleaning bare silicon wafers, it removes other types of particles by slightly simultaneously oxidizing and etching the Si surface. That is, it undercuts the oxide under a particle, which reduces the contact area between the particle and the surface. This makes the particle more prone to being lifted off.

8.7.1.3 Mechanically Assisted Particle Removal: The *mechanical removal of particles* generally proceeds by momentum transfer. That is, shear forces impact a particle and dislodge it by either sliding it off the surface, or by rotating it until it detaches. As particle sizes decrease, they become more difficult to dislodge by mechanical force. Although the Van-der-Waals-adhesion force may get smaller, the electrostatic forces increase, and the efficiency of transferring translational momentum also rapidly declines. Mechanical particle removal techniques have traditionally involved: 1) mechanical brush scrubbing; and 2) vibrational

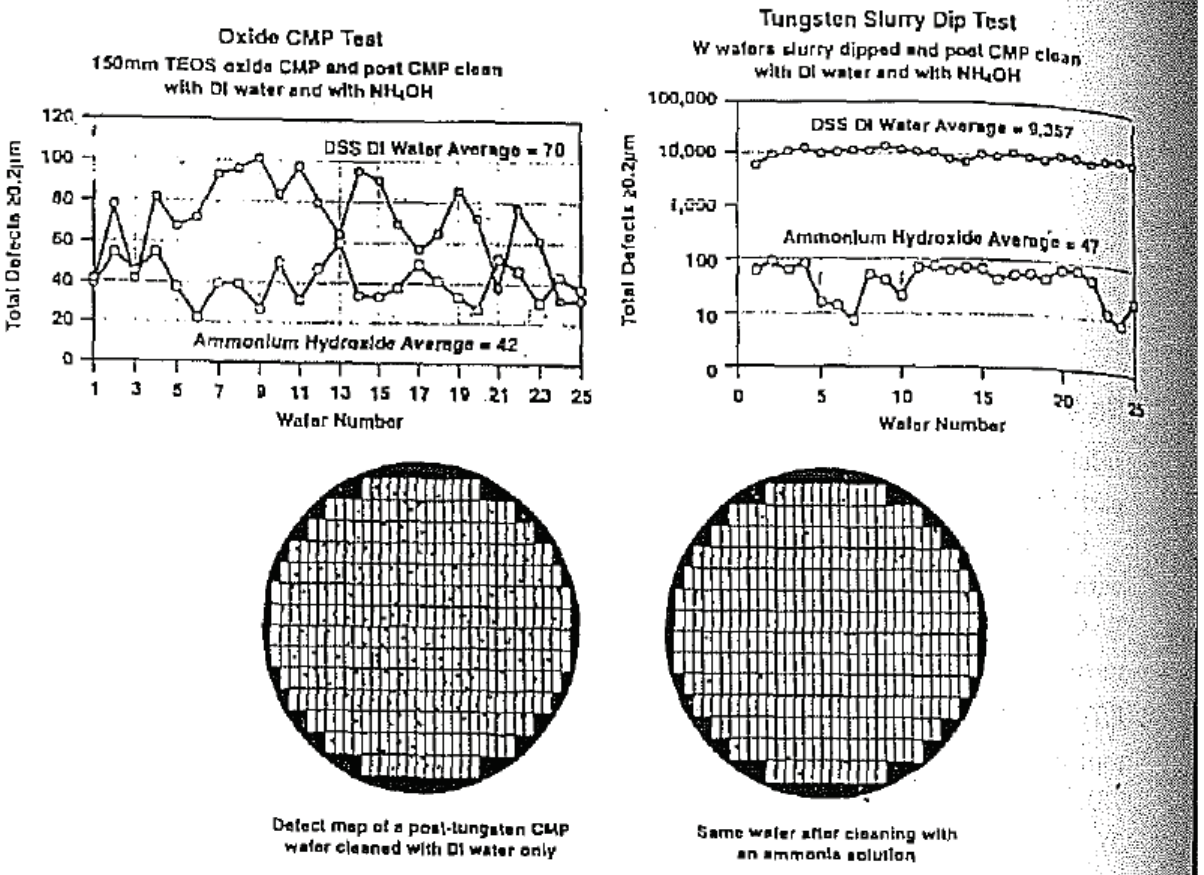


Fig. 8-77 Brush cleaning results comparing use of H_2O(DI) vs. NH_4OH: a) Oxide-CMP cleanup; b) W-CMP cleanup.

(*megasonic*) action. In fact, these particle removal mechanisms are employed in tandem in post-CMP cleaning processes. An explanation of how this is done will be provided in the upcoming sections.

8.7.1.4 Steric Stabilization for Suspending Particles in a Solution: Particles can be prevented from attaching to each other, or from settling on a surface, by the phenomenon of *steric stabilization*. That is, a molecular layer is attached to the particle surface (often a *dispersant* like a polymer or a surfactant is used for this role). This attached layer causes *entropic repulsion* as the particles approach each other (Fig. 8-79). This technique is a second method used to prevent settling and agglomeration in CMP slurries. If this mechanism is used to stabilize a slurry, care must be taken to ensure that the additive does not adversely affect any of the other slurry properties.[61]

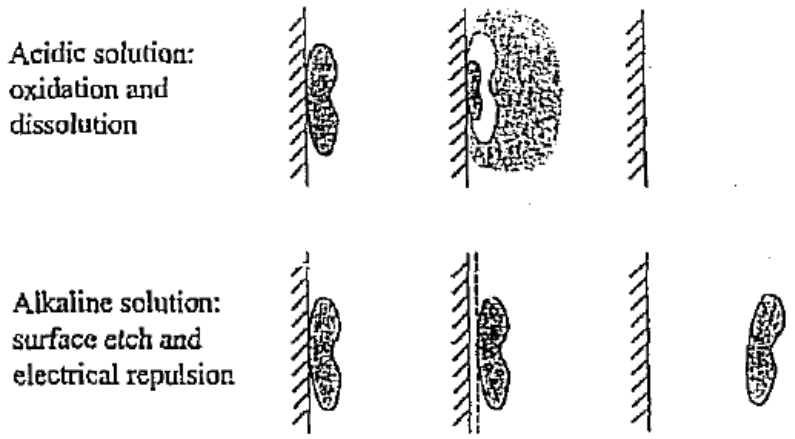


Fig. 8-78 Wet-chemical mechanisms that remove particles from surfaces: a) in a powerful oxidizing agent; b) in an alkaline solution. Courtesy of Semiconductor International.

8.7.2 Practical Particle Removal Processes: Brush Scrubbing

The most widely used method of post-CMP-cleaning involves brush scrubbing the wafers while they are sprayed with appropriate chemical solutions (or while they are immersed in these solutions). Brush cleaning systems have long been used in many wafer-cleaning applications. Despite its name, it actually uses a hydrodynamic drag to exert a removal force on the surface-particles (mechanical dislodgment). Brush bristles do not contact the particle or the surface, but rather act as oars or paddles that push liquid across the wafer surface.

While the brush scrubbing technology is not new, some modifications had to be made to allow a stand-alone brush scrubber to be integrated with a CMP

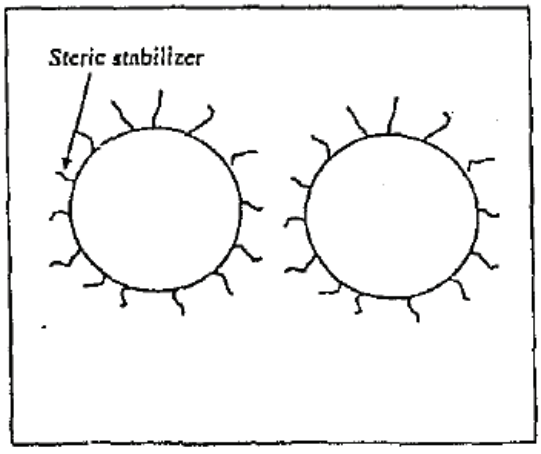


Fig. 8-79 Steric-stabilization of particles in aqueous solution. Particles are stabilized by attaching a molecule to the surface. The molecular layer causes an entropic repulsion as the particles approach each other.

polishing tool. For example, the current technology of choice for both oxide- and metal-CMP employs simultaneous double-sided brush scrubbing (see Fig. 8-80).[62] This simultaneously scrubs both the front and back of the wafer. Another change involves the fact that the scrubbing equipment must function reliably in the presence of both basic slurry chemistries (e.g., KOH and NH_4OH), and acidic cleaning chemistries (such as citric acid and HF), in order to optimally remove both metals and mobile ions. Alumina-bearing slurries used for tungsten-polishing more readily adhere to the wafer surface and are generally more difficult to remove. Particles as small as 0.12-μm have reportedly been effectively removed by brush scrubbing.

Particles are dislodged and they then stay suspended in the liquid until they are pushed over the edge of the wafer into the drain, as liquid is continuously delivered to the wafer surface. Dislodged particles, however, must be prevented from re-adhering to the wafer surface or the brush material. This is done by controlling the electrostatic forces between the particles and the wafer or brush surfaces. As described earlier, particle re-adhesion to both the wafer surface and the brush is suppressed by keeping the pH values of the cleaning liquid greater

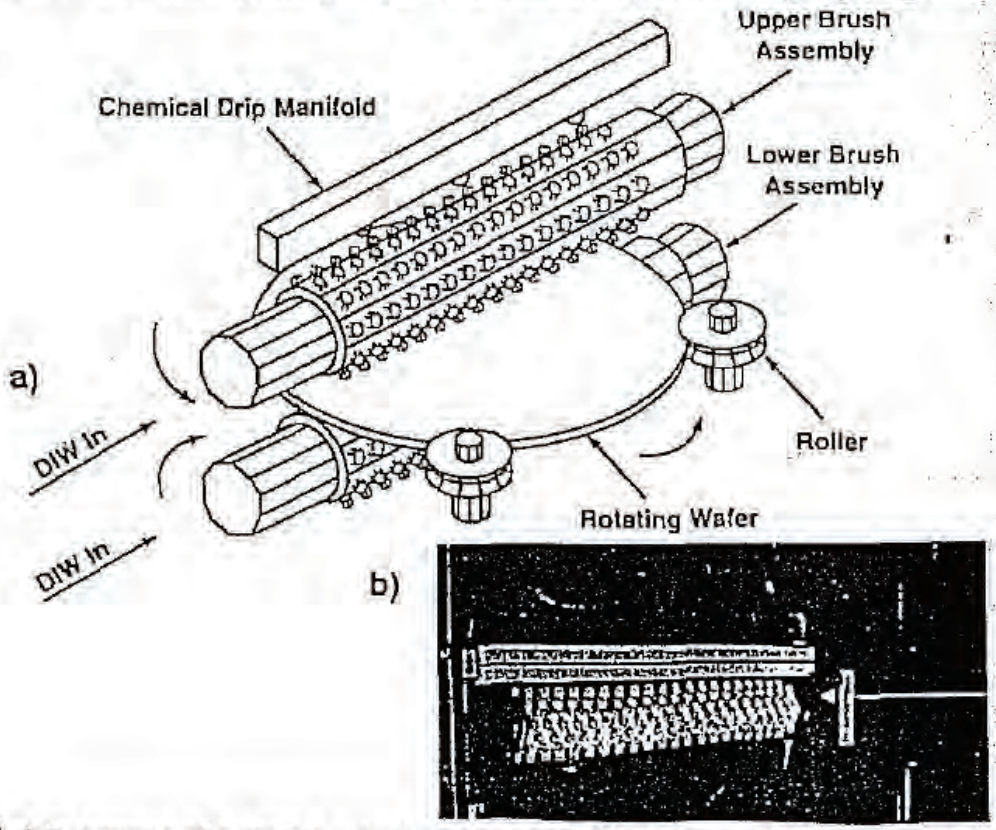


Fig. 8-80 a) Schematic of a double-sided brush scrubber.[65] Courtesy of Lam Research; b) Photograph of a PVA brush. Courtesy of Solid State Equipment.

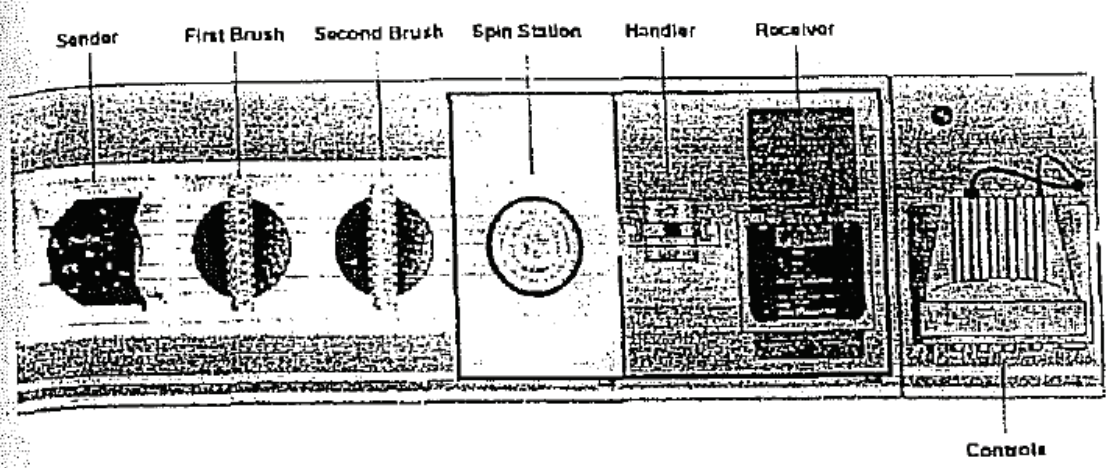


Fig. 8-81 Example of a post-CMP cleaning tool. Courtesy of Lam Research.

than 8. The post-CMP cleaning tool typically consists of a load station, a brush scrubbing module (or modules), a rinse-and-dry module, and an unload station (Fig. 8-81). Isopropyl alcohol and a heat lamp may be used to speed up drying.

8.7.2.1 Brush Scrubbing to Remove Particles in Post-Oxide-CMP Processes:

When brush scrubbing wafers after an oxide-CMP process, a cylindrical polyvinyl-alcohol (PVA) brush (with or without nodules that protrude from its surface) is rotated across the surface of the wafer (Fig. 8-80), as the wafer is continuously flushed with the cleaning solution. The PVA brush-material is a soft, sponge-like material that is chemically resistant and highly compressible. This material allows the mechanical component of cleaning to occur without damaging the oxide surface. The brush is thus brought into direct contact with the wafer surface. Unlike rigid brushes such as nylon (commonly used in the 1970s), PVA does not cause scratches when brought in contact with the wafer surface.[63]

The PVA material is a sponge-like material that has an open structure consisting of interconnecting cells that allow the brush to be constantly flushed with DI water during wafer cleaning. These softer brushes clean more efficiently when they are compressed against the wafer. PVA brushes eventually wear out and need replacement. The average use rate is 20,000 wafers for post-oxide CMP cleaning, and 10,000 wafers for post-metal CMP cleaning.

It has been reported that higher brush speed can improve the particle removal efficiency without introducing scratches. For post-oxide CMP-cleaning (in which the CMP process uses silica-based slurries), the cleaning solutions use mixtures of DI-water and NH_4OH (and often H_2O_2, as well) to decrease silica-particle adhesion to the exposed oxide surface of the wafer.

One problem has been identified with respect to post-CMP cleaning, namely the difficulty of removing particles from non- or partially-planarized recesses. That is, as shown in the AFM image of Fig. 8-82,[64] a slurry particle resides inside a 0.7-μm contact hole with a 3:1 aspect-ratio after brush cleaning with dilute NH_4OH, and similar results were obtained with megasonic cleaning. It was observed that the ratio of the step-height to the horizontal-dimension of the feature is crucial for the cleaning capability of a specific cleaning technique.

8.7.2.2 Brush-Scrubbing to Remove Particles in Post-Metal-CMP Processes: In post-metal-CMP cleaning,, brush-scrubbing is also effective, but the process is modified from the one used for post-oxide-CMP brush-scrubbing. This is necessary because the post-metal-CMP process must: 1) leave the surface of the metal without an oxide layer; 2) leave the metal smooth; and 3) not attack the metal. In addition, when cleaning wafers after a W-CMP-step, the wafers come from an acidic environment (2–4 pH) to a water-rinse and PVA-brush-clean. Some PVA materials cannot endure this chemistry well and must be frequently replaced. Other PVA materials (believed to be free of unreacted residues that hydrolyze in the acidic environment) appear to render them stable over an operating pH range of 2–12. These materials are reportedly compatible with post-W-CMP scrubbing.[66] It is reported that rinsing the wafers with dilute NH_4OH during post W-CMP scrubbing is also highly effective in removing alumina slurry-particles.[67] An additional benefit of using NH_4OH is that it prevents loading of the PVA brushes by alumina particles or tungsten oxides.

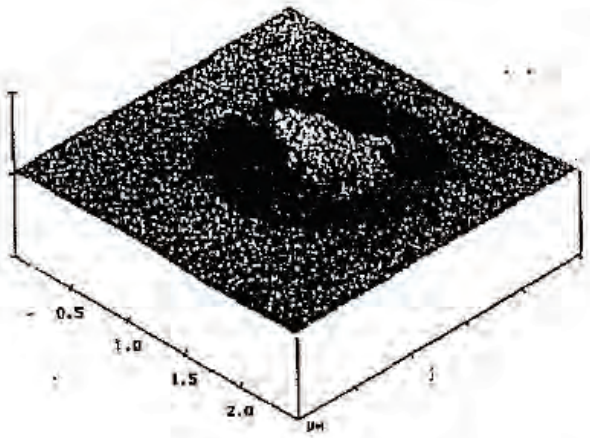


Fig. 8-82 AFM image of an agglomerated-silica slurry-particle in a contact hole.

Cleaning wafers after copper CMP presents serious challenges, including the following:

1. *The CMP-clean-step must not allow Cu-contamination of the fab cleanroom area*: That is, the Cu-CMP cleaning process must remove Cu from the front, back, and beveled edge of the wafer. However, the copper oxides formed and removed as part of the CMP process do not dissolve in DI water and may be easily transferred to the PVA brushes. If the brushes become "loaded" with these copper oxides, they may transfer them to subsequently processed wafers, thus producing severe Cu-contamination in the fab. Proprietary chemicals have been developed to dissolve the copper oxides and prevent such brush loading.[68] In the On-Trak systems brush-scrubbing tool, such a chemical is delivered to the wafer surface through the porous core of the brush itself (Fig. 8-83).
2. *The cleaning chemistry must not corrode the Cu*: If NH_4OH is used in the cleaning solution, great care must be exercised. This is because ammonia is known to complex with copper and cause surface roughness and corrosion products.[69]
3. *The Cu surface must be protected during the cleaning step*: Copper has a soft, porous oxide-layer through which chemicals can diffuse and cause further oxidation/corrosion. Protection of the copper-surface while cleaning the dielectric surface free of metals may be accomplished by using inhibitors.[70]

Note that more information on CMP of Cu is found in Chap. 16.

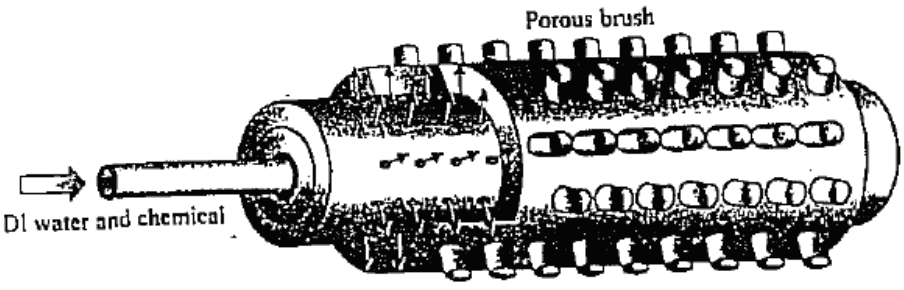


Fig. 8-83 *Through-the-brush* Chemical Delivery for post-CMP cleaning.

8.7.3 Vibrational Scrubbing to Remove Particles (Megasonic)

Brush scrubbing, while effective in removing particles, may cause scratching and brush shedding (which contributes additional particles). Vibrational scrubbing of wafers is an alternative non-contact and brushless technique for removing particles (Fig. 8-84a). Megasonic immersion cleaning systems for CMP using high-frequency sonic waves (0.7-1-MHz) are commercially available. The model of megasonic particle removal is shown in Fig. 8-84b. The wafers are immersed into an SC-1 solution contained in a megasonic tank. Sonic energy from acoustic transducers is directed parallel to the wafer surface, and this reduces the liquid boundary layer that surrounds a particle. The particle is thus subjected to larger fluid-acoustic-streaming velocities and energy from cavitation. These detach the particle from the surface.[71] The particles in the liquid must then be removed by passing the liquid through a continuous filtration system. This keeps them from remaining in the liquid where they could wander into the vicinity of the surface and be re-captured. Such filtration should also remove any particles generated by the megasonic system itself. (That is, particles shed by its own gaskets and transducer-bonding-material when the transducers are vibrating.)

The most common megasonic systems are batch-cleaning-systems. However, one recent report described a single-wafer megasonic-system that used a dilute basic-chemistry and cleaned each wager for 15-60 sec.[72] Cleaning with this system after an oxide-CMP-step yielded wafers with only 50 defects per wafer, well below the levels obtained when using brush scrubbing.

Megasonic cleaners may also offer cleaning improvement when topography such as alignment marks, scribe lines, or W-keyholes are encountered. That is, as discussed in the section dealing with *Brush Scrubbing for Oxide-CMP*, when particles are trapped within high-aspect-ratio recesses on the wafer surface, brush scrubbing is not effective in removing them. On the other hand, megasonic cleaning does a better job of dislodging particles in recessed features. Another advantage of megasonic cleaners involves cost. That is, cost is reduced since brush changes are not needed. The Applied Materials Mesa® combines brush scrubbing with a megasonic cleaner to remove particles in deep recesses. Other vendors, such as Verteq, also offer non-contact, non-scrubber CMP cleaners.[73]

8.7.4 Secondary (or Buff) Polish Procedure as a CMP Cleaning-Step

The polishing process itself can play a role in achieving low particulate levels at the conclusion of the CMP process. That is, by using a secondary buff-polish (i.e., polishing with a soft pad) and following this buff-polish with a high-pressure DI-water-rinse, fewer particles remain on the wafer after the subsequent

cleaning step. As a result, a second platen for buff-polish has become a standard feature for commercial polishers. It is also reported that if DI-water is used in the slurry (i.e., instead of a basic solution), particles counts are further reduced as the pressure during the buff-polish is increased.[74]

8.7.5 Post-CMP-Cleaning of Metallic Contaminants

Besides the removal of particles, post-CMP cleaning processes must also ensure that no metallic contaminants remain on the wafers when they are returned to the cleanroom area of the wafer fab. Since some commercial slurries are KOH-based, an effective clean must be employed to remove surface metals, especially alkaline metals. It has been reported that at least several-hundred angstroms of the top-oxide must also be removed to ensure that alkaline metal levels are adequately removed from the wafer-surface. Consequently, an "HF-deglaze step" is usually employed to reduce metallic contamination (and also any defects embedded within, or close to, the surface).

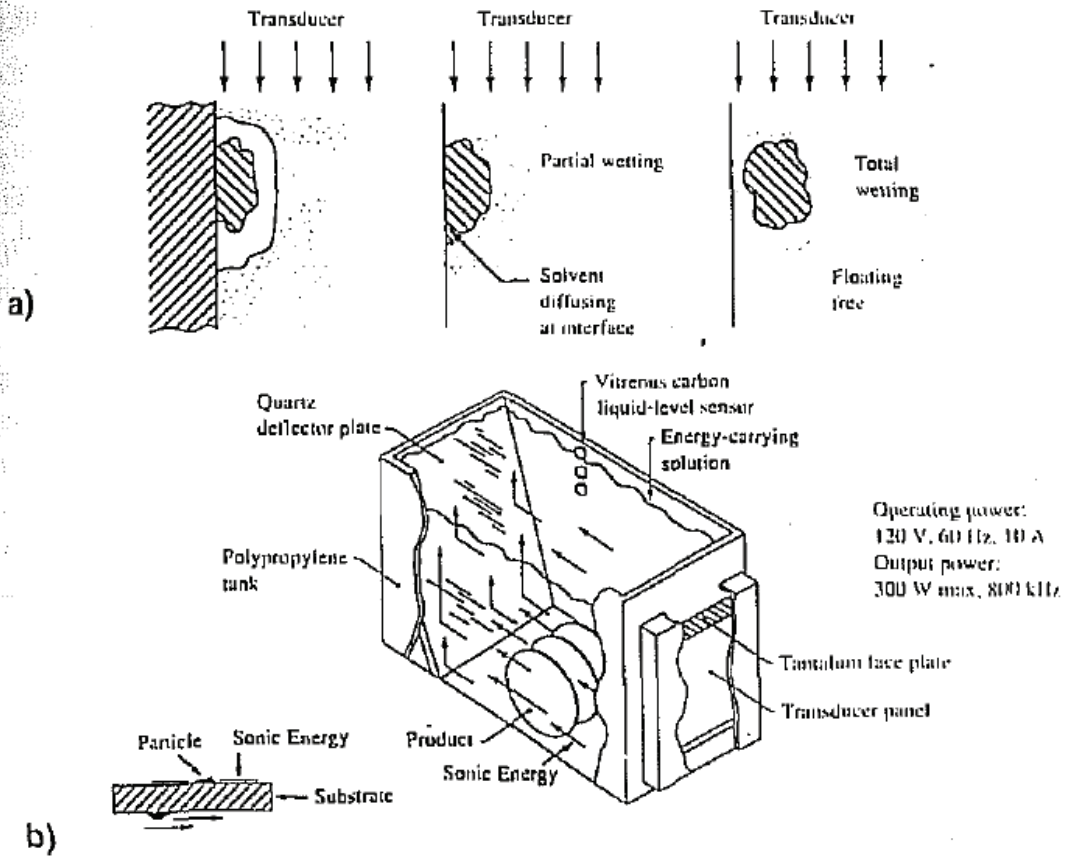


Fig. 8-84 a) Model for the mechanism of particle removal by megasonic scrubbing; b) Schematic of megasonic cleaning tank.

8.7.6 CMP Cleaning Equipment

On-Trak Systems (a division of Lam Research), supplies stand-alone double-sided scrubbing units for the post-CMP cleaning, as does Dainippon Screen. Other LAM post-CMP cleaning tools (e.g., the Synergy Integra® model, Fig. 8-85), are designed to be integrated with the polishing tools of other CMP vendors (including Applied Materials, SpeedFam, and Strasbaugh). Lam Research held about 70% of the post-CMP cleaning-tool market-share in 1999.

However, the Mirra® can also be integrated with an AMAT cleaning tool (the Mesa® Fig. 8-86).[75] It consists of four separate process modules: a megasonic bath; two-double-sided brush-scrub modules; and a spin-rinse-dry module. Each module cleans both sides of the wafer, and all handling is done by edge grip. Another fully integrated cleaner in a CMP tool is shown in Fig. 8-87.[76]

CMP processes are generally located adjacent to the cleanroom, connected through a bulkhead load, similar to etchers or CVD systems. In some cases, the CMP-tool-configuration consists of a polisher and stand-alone cleaning system with a robot that transfers wafers between them. A problem with this configuration being adjacent to the cleanroom is the potential for water and slurry spillage (and then leakage into the cleanroom) as transfers are made. A better solution is a *dry-in/dry-out CMP system*, in which the polisher and cleaning system are completely integrated. Wafers are planarized, cleaned, dried, measured, and returned back to the cassette located in the cleanroom.

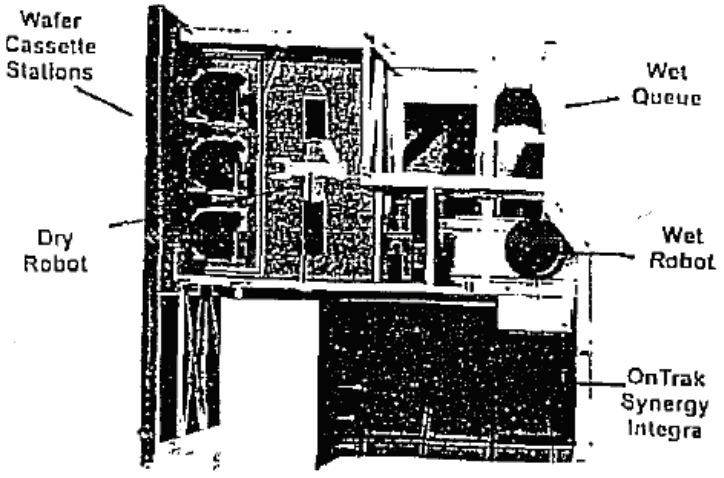


Fig. 8-85 Schematic of the Lam Research Synergy® post-CMP cleaning tool. Courtesy Lam Research.

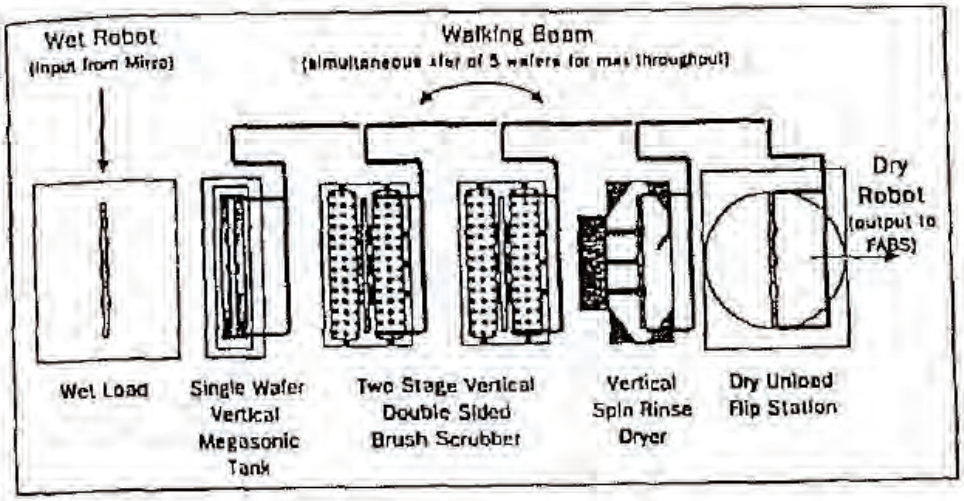


Fig. 8-86 Mirra Mesa® cleaning module layout. Courtesy of Applied Materials, Inc.

8.8 CMP METROLOGY

There are a number of metrology requirements in CMP. First, it must be determined when the CMP process is completed (*endpoint detection*, Sect. 8.5.3.4). Second, the film-thickness after CMP must be measured, including thickness-uniformity across a wafer. Third, the degree of film-planarity must be quantified after CMP. Finally, films subjected to CMP must be inspected for defects and damage. The latter two CMP metrology topics are discussed here.

Film thickness after CMP (and before CMP in in-line endpoint detection), is performed with a number of different techniques, including: spectrophotometry, ellipsometry, and beam-profile reflectometry. One problem associated with these techniques is that of *cycle-skipping* (or *order-skipping*). Since the signal obtained from the thickness-measurement is periodic, the same measured-data can fit more than one thickness-value. For oxide, the resulting errors are multiples of about 240-nm, enough to lead to incorrect conclusions about the post-CMP oxide thickness (which might cause properly processed wafers to be scrapped). Different strategies by each measurement technique are employed to avoid such problems. Ellipsometers use additional wavelengths, spectrophotometers rely on better signal-to-noise data-collection, and beam-profile-reflectometry is combined with spectrophotometry to avoid cycle-skipping errors.

One factor that impacts *thickness uniformity after CMP* is the uniformity of the wafer itself. It is observed that the *within-wafer-nonuniformity* (*WIWNU*) of the film removed by a CMP-process increases linearly with an increase in the

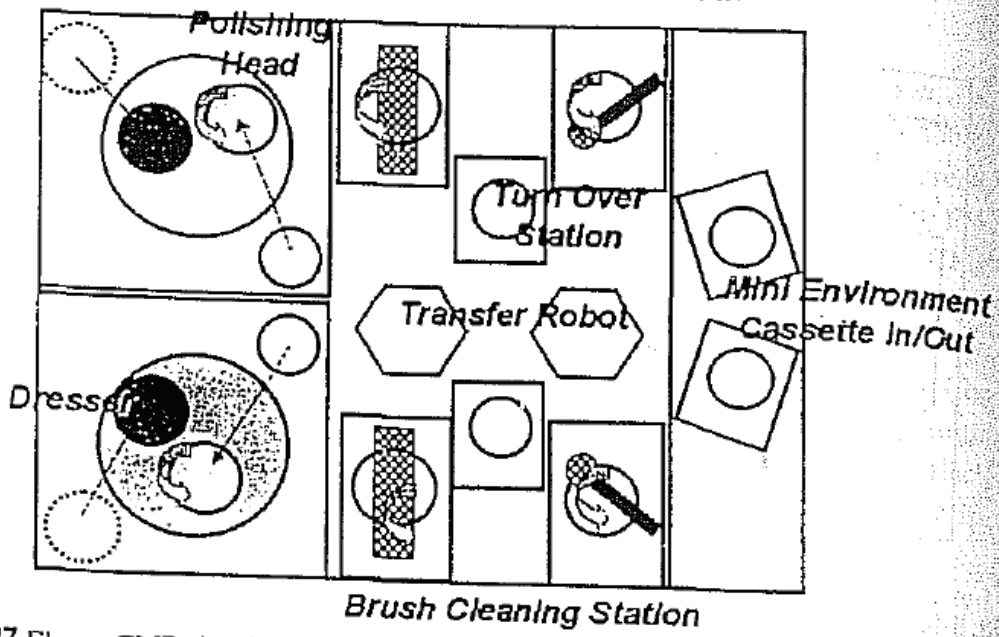


Fig. 8-87 Ebarra CMP cleaning module. Courtesy of Ebara.

total-thickness-variation, TTV, of the wafer.[77] In addition, wafers may not retain their initial flatness as they undergo the various steps of the process sequence. In particular, stress arising from thin-film-deposition processes can deform a wafer into a concave-shape (negative bow) or a convex-shape (positive bow). Although bow is reduced by the pressure applied to the wafer during CMP, some remains (as the degree of bow can vary from wafer-to-wafer). In addition some residual local-warpage persists after the wafer has been pressed against the wafer platen. Thus, the CMP-process may polish the edges faster than the center in a concave wafer. Wafers with lower pre-polish-bow reportedly had a smaller WIWNU than wafers with higher pre-polish-bow. As a result of the above issues, measurement of the wafer TTV as received from the wafer suppliers has been recommended.[77] Some CMP users have set an empirical TTV-cut-off-value of 5-μm for 200-mm-wafer pre-polish-sorting. Capacitive-gauging-tools are used to perform such TTV and bow measurements. More than 16,000 data points can be collected on a 200-mm-wafer in about one minute, and the results plotted to give a high-resolution map of the wafer geometry. Such gauging-tools are available from ADE Corp.

The *degree-of-planarity* achieved with a CMP-process is typically measured on a single-die or a few dice on a wafer with profilometry or atomic-force-microscopy. Such planarity can be location- and pattern-dependent. Within-die-planarity is influenced by the tendency of CMP to polish smaller, individual-features faster, and larger and densely-packed-features more slowly. Thus, CMP-thickness-measurements made with optical-thickness-tools commonly include

several features at each selected site (from the smallest that the tool can conveniently measure, to a large [field] area where the polishing-rate reaches its lowest limit).

8.8.1 Detection of Defects Associated with the CMP Process

Detection of defects on post-CMP wafers is another important aspect of CMP metrology.[78] This importance arises from the fact that CMP is probably the most yield-impacting of all IC process steps given its intrinsically "dirty" processing environment (which includes chemically corrosive slurries, nano-particulate abrasives, and polish by-products). The defects associated with the CMP process can take the form of: 1) *damage*; 2) *contamination*; and 3) *particles*.

The acceptable density of defects is becoming lower for each device generation. In 0.18-μm CMOS-technology the SIA Roadmap calls for a defect-density of 0.14-defects/cm^2, and the minimum-detectable-particle-size is decreased to 0.06-μm from 0.08-μm for the 0.25-μm generation. In oxide-CMP, defects can be separated into two categories: 1) *foreign material* (such as residual-slurry, surface-particles, embedded-particles, and residual-metal); and 2) *voids in the surface material* (such as microscratches and dishing). *Microscratches* can occur when a particle is dragged across the wafer surface, causing a shallow void in the oxide layer. Figure 8-88 shows the defects arising from oxide-CMP processes.[79]

In metal-CMP, defect types include: metal-loss, puddles (which can cause shorts if they remain between metal lines), surface-voids, residual-slurry, surface-particles, metal-filled-micro-scratches (which can cause shorts), and plug-holes from which metal has been removed. Figure 8-89 depicts a variety of defects associated with the W-CMP-process.[79]

The detection signals caused by such defects are obscured to some extent by noise sources associated with the CMP process. The three prevalent noise sources that impact defect inspection at post-CMP processing are: 1) layer-thickness-variation; 2) grain-noise; and 3) pattern-noise. *SiO_2-layer-thickness-variation* is caused by nonuniform-polishing across the wafer. *Grain-noise* arises from the surface-roughness caused by the polycrystalline structure of materials beneath the oxide being polished (e.g., tungsten, aluminum, and polycrystalline-Si). The noise occurs from randomly-distributed pyramidal-peaks at the surface of the poly-crystalline-films. These scatter the detection-signal in a non-consistent manner so it is not completely subtracted when two adjacent-die are compared. *Pattern-noise* arises from metal-features beneath the oxide (which distort the incoming signal by causing background-scattering). Defect-detection-tools

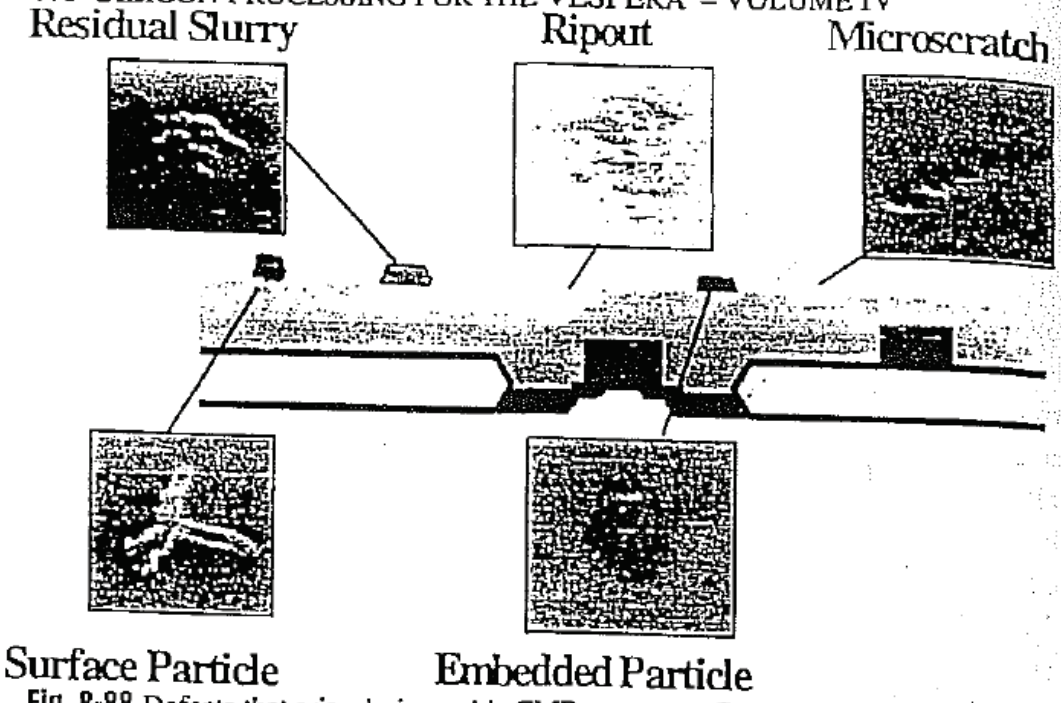


Fig. 8-88 Defects that arise during oxide CMP processes. Courtesy of KLA-Tencor.

employ image-processing algorithms to limit false alarms caused by such noise (i.e., that would indicate the presence of a defect where none actually exists).

Two illumination modes (*bright-field* and *dark-field*) are commonly used in the laser-scattering defect-counters employed for in-line automated-defect-inspection (such as the KLA-Tencor Surfscan AIT® model). Since reflected light is used in *bright-field inspection*, one of the main challenges for CMP-metrology is to reduce the interference between reflection at the oxide-surface and reflections that arise from underlying features. Some inspection tools use an ultra-broadband illumination-source (in place of a monochromatic-source) to suppress such interference effects. In *dark-field systems* only scattered light is collected, while the spectrally reflected light is rejected. Dark-field systems with a low-angle of illumination appear to work best for detecting slurry-particles on an oxide-layer over a grainy sub-layer (Fig. 8-90).[79] Dark-field systems with optical repetitive-pattern-filtering are reportedly more effective for detecting defects over dense repetitive arrays.[78]

Some 0.18-μm device-inspections for detecting CMP-defects may require use of SEM technology (such as the KLA-Tencor 2710). Because of their high resolution and large depth-of-focus, SEMs can locate ultra-small-defects hidden between dense geometries or in high-aspect-ratio areas. Small CuO particles ~0.1-μm from a dual-damascene process are easily detected using SEMs.

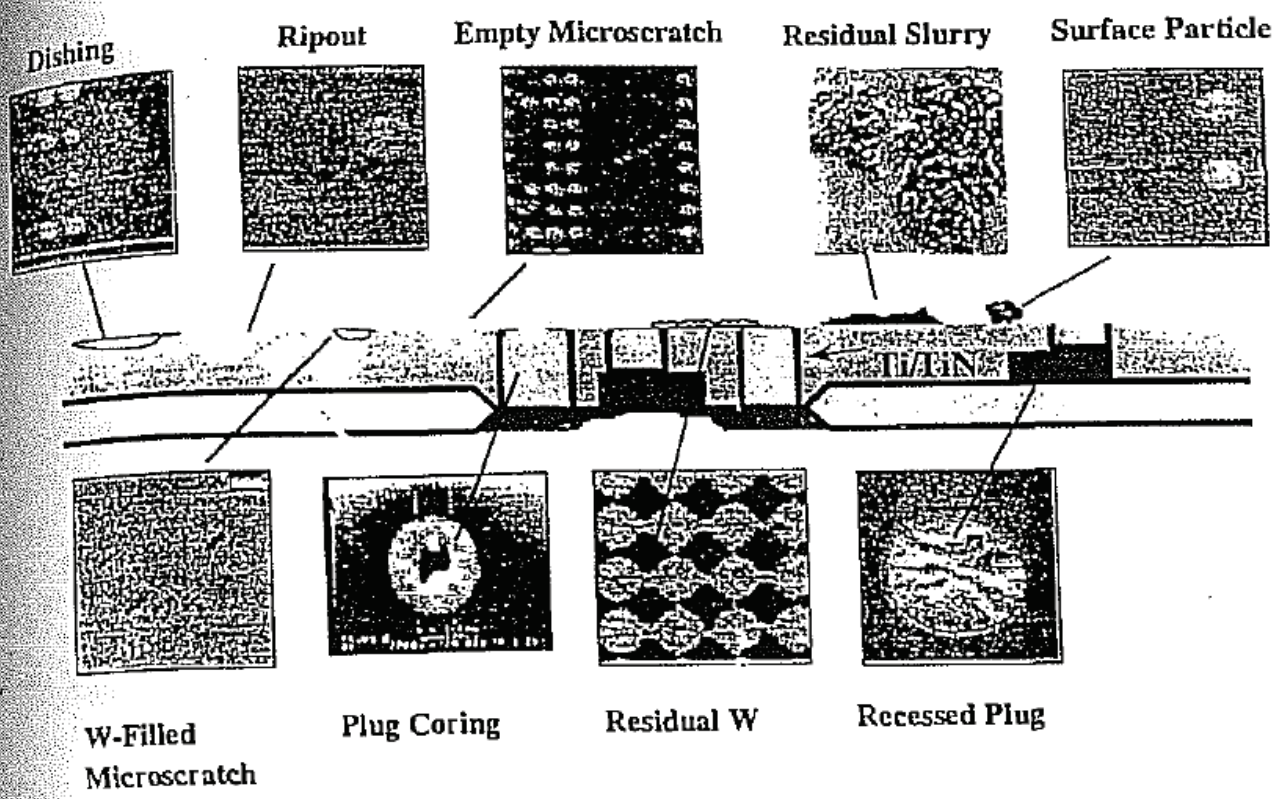


Fig. 8-89 Defects that arise during W-CMP processes. Courtesy of KLA-Tencor.

8.9 CMP POLISHER TOOL RELIABILITY

The CMP process creates a harsh set of working conditions for the polishers. Some have compared the operation of a CMP-polisher to "flying a helicopter in a sandstorm."[80] In the early days of CMP, production polishing-tools could have a mean-time-between-failure (MTBF) as small as a few days. Some of the problems of early CMP tools included the following:[39] insufficient vacuum for wafer handling; wafer sensor failure, misalignment of the load cup and sender and receiving cassettes, wafer release on polish pads (resulting in wafer breakage), equipment hangup (causing slurry to dry on the wafer-surface, and leakage of slurry and de-ionized water because of lack of quality control on valves and rotating unions. To make the systems more reliable, tools were designed to protect moving mechanisms from the abrasive and corrosive slurries (including sealing of parts and using air and water labyrinths).

Exposed materials were made with corrosion-resistant materials. Improved robotics for wafer loading, unloading, and handling were implemented. New system components were subjected to accelerated-testing to simulate one-year of

production in three months. As a result, MTBF rates in CMP tools have been continuously improving. The goal of CMP tool suppliers is to build polishers with a MTBF of at least 400 hours (about 16 days).

8.10 CMP SYSTEMS AND PROCESS INTEGRATION

The CMP process requires a host of tools and consumables (Fig. 8-91). When CMP first came to the IC industry, each component of the CMP process was supplied by a separate company. The burden of integrating the various components fell to the user. As the market for CMP tools and consumables grew, equipment and materials suppliers realized the need for a better CMP-infrastructure. They understood that to gain market share, CMP users should be offered an off-the-shelf, "one-stop-shopping" package, perhaps with guaranteed process results. One way to accomplish this was to form strategic partnerships with other (non-competitive) suppliers. Most polisher suppliers included slurry distribution systems, cleaning tools, pads, slurries, endpoint detection, and film thickness measurement tools from other vendors as available options compatible with their polishing tools. Some companies (including Lam Research, Applied

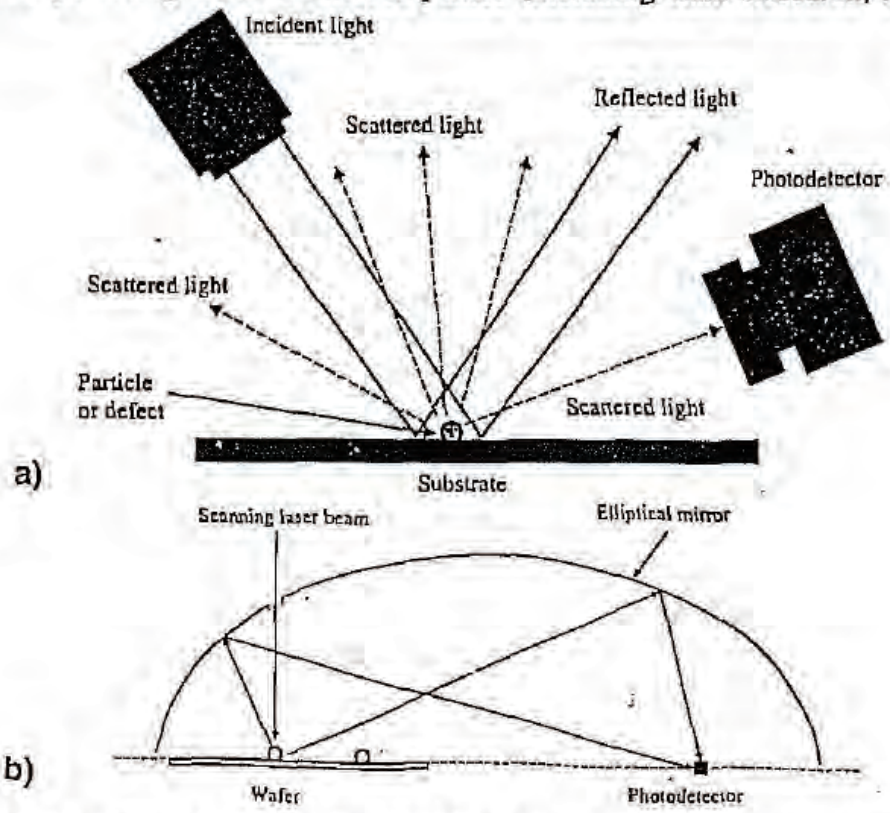


Fig. 8-90 a) Particle detection by Darkfield Light-Scattering; b) Light-scattering particle detector with an elliptical mirror.

Materials, SpeedFam-IPEC, Ebara, Strasbaugh, and Peter Wolters) designed CMP systems as cluster tools. Such systems integrate polishing, endpoint detection, cleaning (and perhaps inspection) into a single system. The Lam Research Teres® and the Applied Materials Mirra Mesa® are examples of such cluster CMP tools.

These integration trends have been accompanied by business consolidations as well. Speedfam and IPEC (two large polishing tool suppliers) merged in 1999. Lam Research acquired On-Track Systems (a maker of polishing tools and post-CMP-cleaning-systems). Applied Materials announced it would acquire Obsidian, Inc., a company that had developed a fixed-table CMP tool which uses a slurry-less pad technology. As CMP technology matures it is likely that such consolidations will continue.[82]

8.11 CMP OF VARIOUS MATERIALS

Here the issues of oxide-CMP and W-CMP are described in detail. The subject of Cu-CMP is briefly introduced in Sect. 8.11.3 here, but more details on Cu CMP are provided in Chap. 16 (that deals with the process technology of Cu in ULSI).

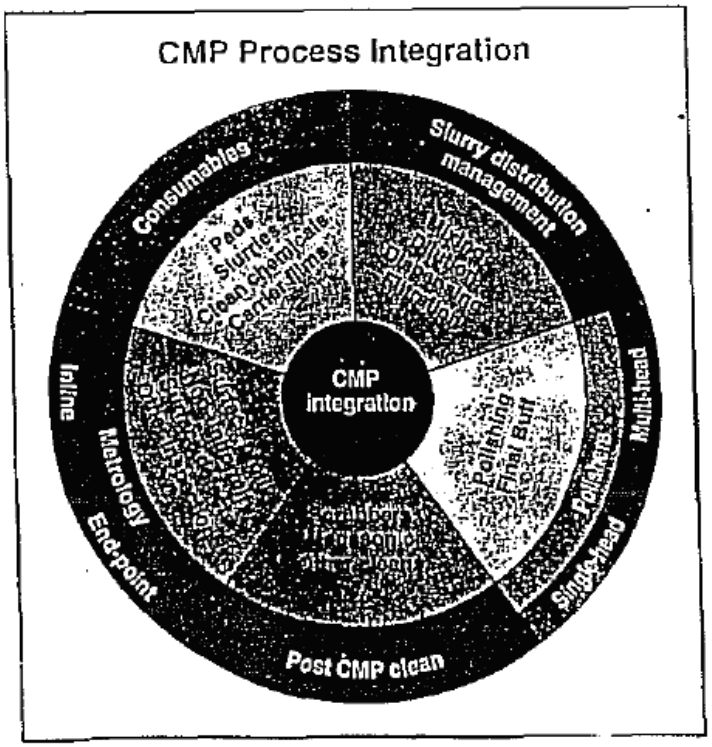


Fig. 8-91 Many components of CMP must be successfully integrated to develop a production-worthy CMP process. Reprinted with permission of Semiconductor International.

Note that the CMP issues associated with shallow-trench-isolation (STI) are mentioned briefly in this chapter in Sect. 8.12 (which deals with the WIDNU phenomenon of *dishing*). However, additional details concerning CMP-issues in STI are found in Chap. 9 (which is dedicated to the topic of STI).

8.11.1 CMP of Oxide Interlevel and Intermetal Dielectrics (ILDs and IMDs)

A blanket-dielectric-layer (typically an oxide-like film) is deposited (generally by CVD) atop polysilicon after gate-patterning. This layer is called an *interlevel dielectric* (ILD). Likewise, a blanket dielectric-layer (again, usually of silicon oxide) is CVD-deposited onto patterned metal-features, and such layers are called *intermetal dielectrics* (IMD). Topography is produced when oxide films are deposited by CVD onto etched polysilicon and metal patterns (see Fig. 8-1a). Many techniques have been attempted to planarize this topography (including: reflow for ILDs; and resist-etchback and spin-on-glass for IMDs). CMP has replaced these planarization techniques because they can only provide local-planarization (see Fig. 8-1d), whereas CMP can provide global-planarization (see Fig. 8-1e). However, unlike CMP for STI-applications, there is typically not a polish-stop-layer, and repeatable control of the polish-rate and uniformity is required. As noted previously, process variations (such as pad-age, slurry batch-changes, machine-calibration-drift, pad-conditioner-age, and chamber-to-chamber dielectric-film-property-variations) can all result in significant removal rate and process non-uniformity.

The industry standard pad for ILD and IMD CMP has been the IC-1000/Suba-IV stacked pad. This pad consists of a hard-cast polyurethane-pad (IC-1000) on top of a soft, compressive felt-pad (Suba-IV). The hard-top-pad is needed for planarization, and the soft-pad underneath improves the WIWNU (compared to polishing without the soft underpad). The drawback for including the soft underpad is a lower degree of planarization. Total stack compressibility can be modified by varying the thickness of the two pads.

Another issue of ILD-CMP is that doped silicon-oxides, such as PSG, and BPSG are softer glasses that undoped silicon-oxides (USG). As a result, they exhibit faster removal rates than USG. Furthermore, doped ILD-films exhibit higher potassium and calcium levels on the wafer-surface after scrubbing and cleaning than do USG films. This occurs because the phosphorus in the PSG and BPSG films acts as a gettering species for metal-impurities in the oxide-film. An HF-dip in the post-CMP-cleaning -equence can reduce the level of such metallic-contamination on the PSG surface (as well as slurry-residue). However, soft films, such as PSG tend to have high post-CMP defect-counts, and the HF etches

microscratches isotropically, and enlarges their size. Thus, the post-CMP-cleaning-processes of PSG and BPSG ILDs need to be optimized to simultaneously reduce metallic-contamination and defectivity (e.g., by performing an oxide buff, which is very effective in reducing microscratching).[82]

8.11.2 CMP of Tungsten

Tungsten (W) CMP has been developed to replace the W-etchback step of the W-plug formation sequence (see Chap. 6, Vol. 1, 2nd Ed.). The W-CMP process, however, requires a surface that is planar prior to via-formation and metal-deposition (a condition not required in processes that form W-plugs by W-etchback). W-CMP has been described as a balance between the electrochemical oxidation of W to form a blanket passivation layer WO_3 on the entire W surface. and the mechanical removal of that reaction-product by abrasive slurry particles. Tungsten, being one of the refractory metals, tends to form strong oxygen bonds under oxidizing conditions. This oxide is so dense that the metal is protected from further chemical attack and the metal is "passivated."

Polishing of tungsten films raises three key issues. The first involves selection

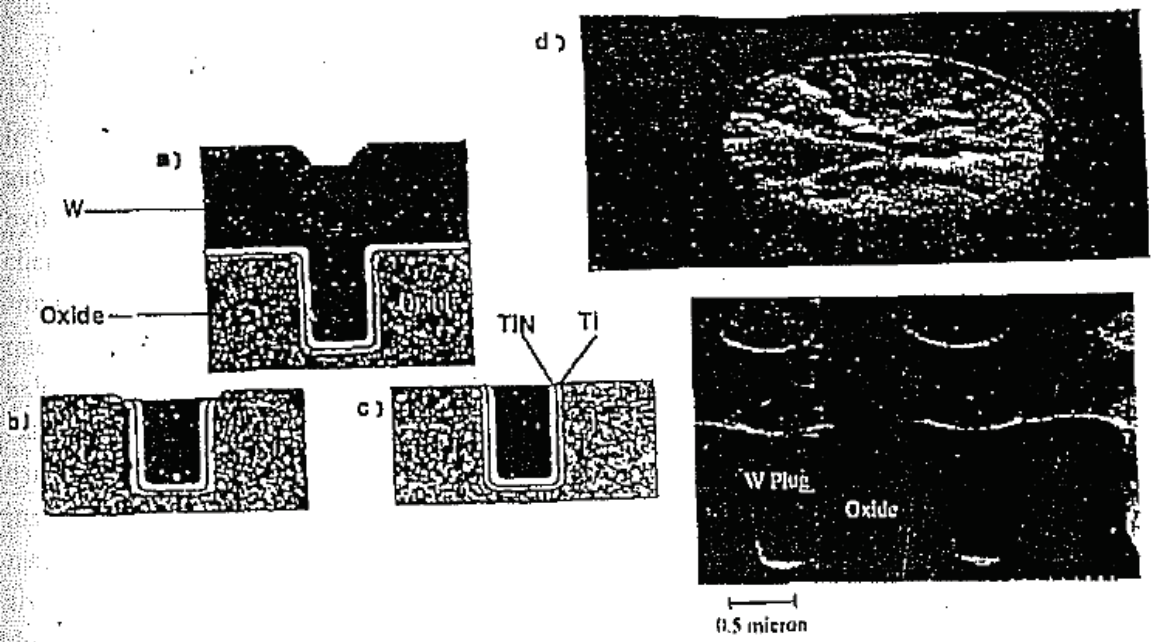


Fig. 8-92 W-plug formation using CMP: a) W deposition over TN/Ti barrier layers; b) CMP of W and TiN/Ti leaves recessed plugs; c) Oxide buff eliminates recess.[84] Reprinted with permission of Solid State Technology. Published by PennWell.; d) SEMs of W-plugs after W and TiN/Ti CMP steps and oxide buff step.[85] Reprinted with permission of Semiconductor International.

of an appropriate slurry. Tungsten polishing at IBM was originally done with a slurry-mixture of potassium ferricyanide and potassium phosphate (having a pH adjusted to about 5.0 to 6.5), together with colloidal-silica or dispersed-alumina, Al_2O_3.[83] (Alumina-based slurries have come to be the more commonly used.) However, potassium ferricyanide is highly toxic, and its disposal causes severe environmental impact. As a result, this chemical has been replaced, with the most widely used substitutes being ferric nitrate {$Fe(NO_3)_3$}, potassium iodate (KIO_3), and hydrogen peroxide (H_2O_2).[84] An example of commercially available slurries for W-CMP are Cabot's SEMI-SPERSE® WA-355 and SEMI-SPERSE® FE-10.

The second key issue of W-CMP arises because blanket-W-films are deposited over a liner-film of TiN/Ti. When removing the W-film, this liner-film must also be polished away. A high removal-rate of the TiN/Ti layer is essential to prevent erosion of the oxide underlying the metal-layers, and to also prevent plug-recessing (Fig. 8-92a). However, polish-rates of TiN/Ti are typically three to eight times slower than that of W (which is removed at rates exceeding 350-nm/min). It is also necessary to have a high selectivity to the underlying oxide to obtain local planarity of the plug-surface and to prevent excessive loss of the surrounding oxide by dishing. To remove the liner-film, *over-polishing* is typically used (i.e., so named because it causes some recessing of the W-plug as shown in Fig. 8-92b). During this step the field-oxide is also damaged to some degree.

To solve the problem of having to polish two-different metals, a multi-step polishing sequence is often used in W-CMP processes (typically a three-step CMP process is needed).[84,85] For example, in the first step, the W-layer is removed using one platen of a polishing tool. In the second step, the wafer is transported to a second platen where the liner-film (e.g., TiN/Ti) is polished away. Finally, a third oxide-buff-step (on yet a third platen) removes the oxide-damage by polishing away a small amount of the now-exposed underlying oxide (using a silica-based slurry that minimizes surface-roughness and facilitates the removal of alumina slurries). This also restores the level of the W-plug to that of the oxide-surface by eliminating the recess (Fig. 8-92c). Such CMP-W-plug-processes reportedly produce higher yields than processes which employ a W-etchback-process.[86,87] The higher yields with W-CMP are attributed to improved defectivity and superior final plug-morphology.

The third problem associated with CMP of tungsten involves the voids or seams that can occur when the blanket-W-film fills the contact or via holes. If such voids exist, they will be exposed during the CMP process. These voids become regions in which slurry can be trapped. Slurry chemistry can also attack a seam along a W-plug center and enlarge the seam to form a void or widen/expose

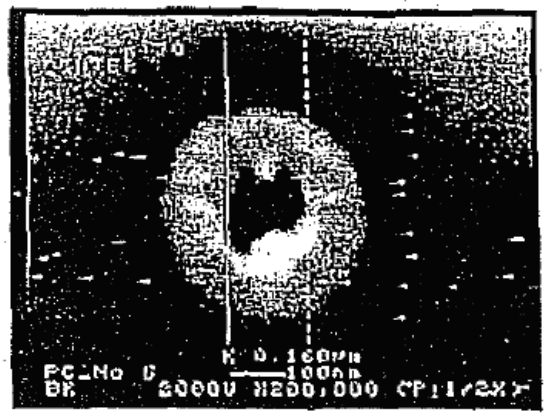


Fig. 8-93 Plug-coring.

an existing void, a phenomenon termed *plug-coring* (Fig. 8-93). Prevention measures include developing W-deposition-processes in which void-formation does not occur, and the use of less aggressive slurry chemistries (which will have less tendency to attack weak seams).

8.11.3 CMP of Copper

The CMP of copper (Cu) interconnect films involves a whole set of unique challenges,[88,89] and some of these issues are described here

The removal-rate of Cu by CMP is about 350-nm/min with a selectivity of 150:1 to the underlying oxide. However, Ta is commonly used as the liner-layer for Cu, and removing it by CMP represents some difficulties. Cu is relatively soft and oxidizes easily with standard chemistry (such as hydrogen peroxide or hydrogen persulfate). On the other hand, Ta is very hard and is also much more chemically inert. This typically leads to large differences in selectivity when simultaneously polishing Cu, Ta, and the underlying oxide (i.e., the Ta removal-rate is only about 30-nm/min). This leads to significant *dishing* of wide Cu and oxide features (the term *dishing* is defined in Sect. 8.12). Damascene CMP applications require low selectivity between the primary conductor-material and barrier-films (1:1) and a high selectivity between the primary conductor and insulator (> 100:1) to essentially stop the process. Since Ta is the harder and more inert metal, it is thought to be removed primarily by abrasion. An aggressive oxidizer could be used, but this would create the possibility of polishing through the Cu and removing device features. Work has been proceeding to find an oxidizer that selectively attacks the Ta without removing Cu.[90]

Another related problem in Cu/Ta-CMP is the electrochemical-interaction between Cu and barrier-materials during CMP.[91] Two dissimilar metals in

simultaneous electrical-contact and contact with an electrolyte-solution constitute a galvanic-cell. This chemistry frequently leads to anodic-corrosion. Such a situation arises near the end of a polishing cycle of Cu and Ta (when the barrier-metal becomes exposed in the field-regions, and Cu is exposed in the trenches). Cu in the trench can be removed (sometimes entirely), even if the trench is initially well-filled with Cu. Obviously, this effect must be suppressed to enable CMP of Cu/Ta films to be successfully performed.

Use of multi-step polishing-processes has also been pursued for CMP of Cu. As in W-plug-CMP, a three-step damascene-Cu-CMP process uses three platens, each with its own slurry mixture. Copper is removed with the first platen, Ta layers with the second, and an oxide-buffing step is carried out on the third. It must be ensured that all of the Cu in the field-regions is removed during the first step, since the removal-rate of Cu in the second is deliberately kept very small. In another approach used by Speedfam, multiple slurries are simultaneously delivered through-the-pad to each of four autonomous polishing modules. In this way, a multi-step Cu CMP process can be carried out without transferring wafers from one platen to another (which increases throughput and yield).

For multi-step Cu-CMP processes one of 3 or 4 types of slurry might be best for the second polishing step. The selection of a particular slurry can depend on the liner-layer being used and the type of ILD material beneath the metal film. A large number of companies are developing slurries for Cu-CMP, hoping to capture a significant market share of this important emerging process.

8.11.4 CMP of Aluminum

CMP of aluminum is being examined for use in damascene-Al-interconnects. One of the challenges facing Al-CMP is that the metal is soft and susceptible to scratching, but its oxide Al_2O_3 is hard and adheres tightly to the metal. This makes Al_2O_3 difficult to remove. If a strong oxidizer is used, the entire Al-film (not only its elevated regions) tends to disappear. It has been shown that if a correct set of process conditions is established (e.g., an appropriate combination of pad, slurry type, and applied pressure), Al films can be planarized with CMP. While removal-rates of about 200-nm/min have been demonstrated, resolution of the issues concerning the Al-CMP process had not yet occurred by 1997.[92]

8.11.5 CMP of Polysilicon

Planarization of polysilicon layers by CMP is carried out in some IC fabrication applications. These include the following:

1. Removal of the polysilicon from the wafer surface when a blanket-polysilicon-film is deposited to fill the trenches of trench-capacitor based

DRAMs. (The issues of poly center-seam attack observed in some RIE etchback-processes used to remove the poly do not occur if a poly-CMP process is used. In addition, the polysilicon slurries are highly selective to oxide or nitride hard-mask materials and thus, are self-stopping);

2. Topography that exists when a polysilicon-film is deposited over an STI structure may need to be planarized. As shown in Fig. 8-94, the topography near the edge of an STI structure is often not perfectly planar. Since the polysilicon deposition process is inherently conformal, the non-planar surface of the STI-structure is replicated in the overlying poly film. In some applications it is advantageous to remove such undulations.

Polishing of polysilicon-films is done with a CMP-process similar to that used to planarize oxide.

8.11.6 CMP of Photoresist

There are some applications where the ability to CMP photoresist (PR) is beneficial. One example involves the fabrication of trench-based DRAM capacitors in embedded-DRAM technology. In the fabrication of the deep trenches used in such DRAM-capacitors, a sacrificial PR-layer is spun onto the wafer to control the height of the surface on the wafer and to facilitate alignment when building other device structures in the vicinity of the trench array. Both within-wafer and within-die non-uniformities of such sacrificial-PR-layers are important, as schematically shown in Fig. 8-95. These PR layers were found to

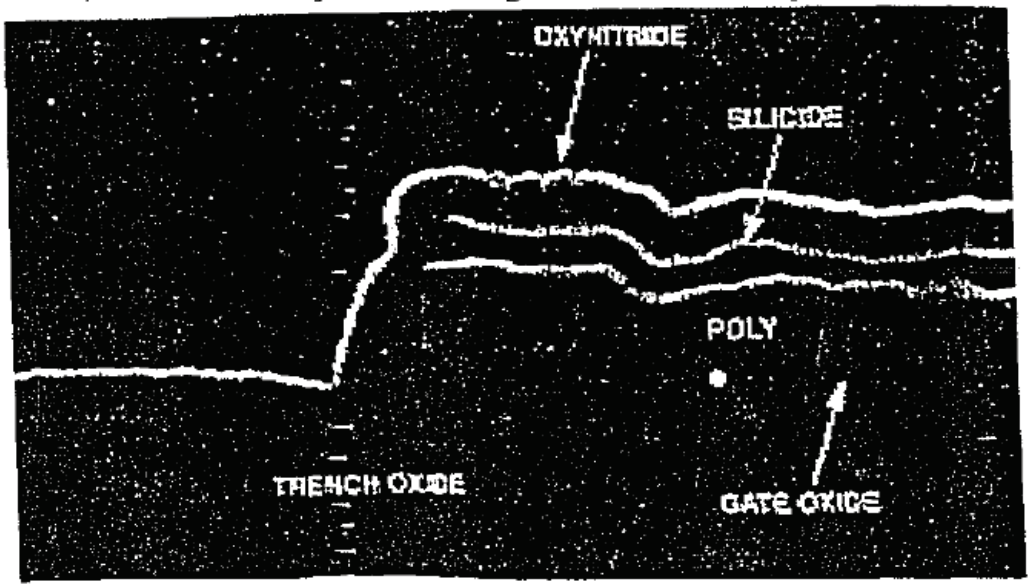


Fig. 8-94 CMP of polysilicon.

be successfully polished with a low-cost CMP process that utilized a slurry-free pH compensated solution of DI water (i.e., no slurry particles).[93]

8.11.7 CMP of Low-*k* Dielectrics

Other materials besides CVD-SiO_2 are being investigated for use as intermetal dielectric films (especially those with a lower dielectric constant - low-*k* dielectrics). These materials are discussed in detail in Chap. 14, but here the issues of low-*k* dielectric CMP are considered. That is, to be used in damascene interconnect structures, such low-*k* materials must be compatible with CMP.

Many of the CMP issues involving low-*k* dielectrics are common to both the subtractive-Al and damascene interconnect schemes. Most of the low-*k* stack options intend on keeping the low-*k* material encapsulated, never exposing it to the CMP process and post-CMP clean. In these cases, the investments in processes, materials, and equipment for silicon-oxide-based CMP systems will carry over and maintain their value. Film-adhesion is fundamental to all on-chip materials, particularly during CMP, with high friction and shear forces at work. Most proposals include capping the low-*k* dielectric with a traditional layer of silicon-oxide (or in some cases, with silicon nitride). The low-*k* materials must adhere well to both the cap-layer on top and to the underlying layer below (which is also likely to be silicon-oxide or silicon-nitride). In addition to adhesion, the compressibility of low-*k* films will make it possible for CMP processes to cause some local flexing within the interconnect stack. Such flexing could lead to cracking of the more rigid cap layers and produce localized delamination. In addition to unplanned exposure of the low-*k* film to slurry and post-CMP cleaning chemicals, the cap fragments could lead to scratching. Consequently,

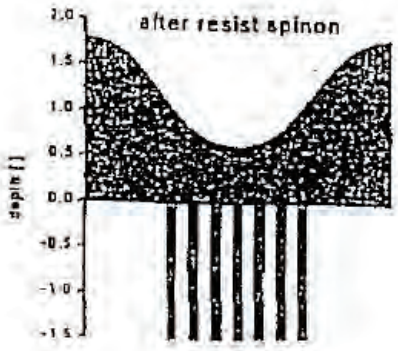


Fig. 8-95 CMP of photoresist.

polishing of cap layers on low-k is one of the earliest tests performed before a new low-k material can be considered integration-worthy. Yet another low-k CMP problem involves the low-thermal-conductivity of such films compared to silicon-oxide. Frictional heating in the CMP process may give rise to hot spots, possibly resulting in localized-stress and cap-delamination. High-linear-speed equipment being developed for greater planarization efficiencies would exacerbate this problem. In the extreme, local temperature excursions may further cure the low-k film, resulting in non-uniform dielectric properties.

Up to the end of 2000, there was little published information dealing with the issues of carrying out CMP of low-k materials in a volume-manufacturing environment. Part of the difficulty lies in the fact that process and consumables requirements may vary widely, depending on the specific low-k material chosen. One published report on low-k dielectric CMP involves the CMP of fluorinated oxide (F-SiO_2).[94] It was found that both mechanical-abrasion and chemical-erosion are intensified by fluorine-doping (relative to undoped TEOS oxides). This happens because of the reduced hardness and chemical reactivity of fluorinated oxides. One way to avoid these problems is to cap the F-SiO_2-layers with a CVD-SiO_2-layer. This cap-layer must be thick enough so that the dielectric-stack is planarized by CMP. In such cases the F-SiO_2-layer is not reached. Dielectric-erosion due to CMP-dishing must also be less than the thickness of the cap-layer. Furthermore, the cap-layer increases the overall dielectric-constant of the IMD-stack. In addition, depositing the stacked-film is a multi-step-deposition process. Another approach is to pursue the development of more chemically-stable low-k-materials, and such efforts are continuing.

A second study of direct polishing of low-k-dielectrics using a MnO_2-slurry reported that the deep scratches formed with CMP using a conventional silica-based slurry were reduced from a 20-nm depth to a 2-nm depth. Residual MnO_2 slurry particles can also be removed using H_2O_2 with a pH of 6, which does not attack the fluoro-carbon low-k material being polished (a poly-AE layer, FLARE®).[95]

8.12 MISCELLANEOUS ISSUES CMP

8.12.1 Dishing and Erosion

Dishing, another CMP issue, can represent a significant problem when a film of one type of material is being polished down until it remains only in recessed (or trench regions - Fig. 8-96). In such cases the layer under the film being polished

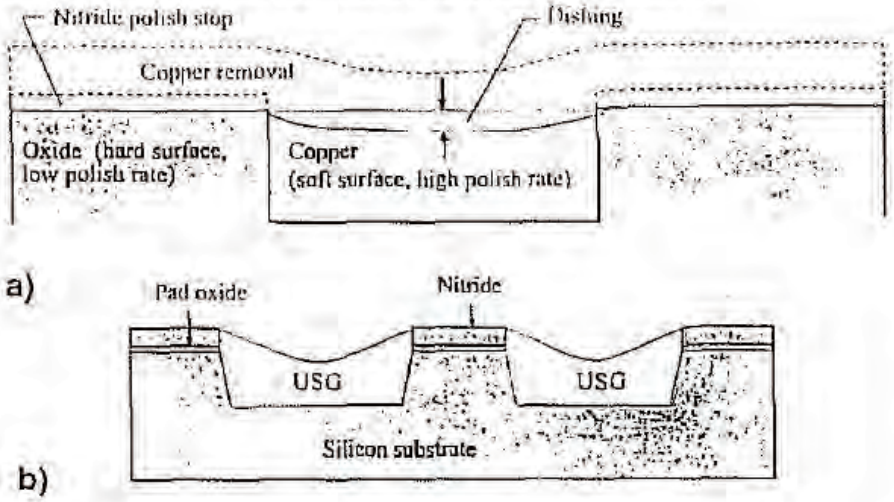


Fig. 8-96 Dishing-effect, shown here for: a) CMP of C in oxide-trenches; b) CMP of oxide in Si-trenches during formation of STI.

becomes exposed to the CMP process during the over-polish time (used to ensure that the film is completely removed over all regions of the wafer by the CMP process). Ideally, such underlayers should act as perfect polish-stop-layers. However, during the over-polish time, the film being planarized can continue to be polished away (especially in wide areas), causing it to be significantly thinned in such areas. One case where this occurs is in the CMP of a CVD-oxide-layer, deposited to fill trenches in the course of fabricating shallow-trench-isolation (STI) structures. A silicon-nitride-layer is typically selected to act as the polish-stop-layer. (See Chap. 9 for a discussion of STI technology.) During the over-polish time, the oxide in the trenches continues to be removed. The degree of thinning of the oxide in such areas (*dishing*) is known to be highly dependent on feature size, with the dishing-effect becoming larger as the width of the field-oxide is increased (Fig. 8-97). For example, in some processes dishing can cause the field-oxide thickness to decrease by more than 50-nm when the field-oxide width exceeds 100-μm.[96]

Dishing occurs because the polishing-pad is not perfectly rigid. If the pad were to exhibit perfect rigidity, CMP of the exposed oxide-regions would produce a perfectly planar surface (because the pressure on the exposed recessed regions would be zero). Since actual pads exhibit some flexibility, above-zero polishing-pressures are produced in recessed areas, especially in the middle of wide exposed-regions. Thus, oxide is removed from these regions, and a "dished" region exists after planarization. Dishing must be minimized to maintain planarity for subsequent wafer processing. It has remained a serious impediment to developing damascene processes (as is *erosion*, which is covered later).

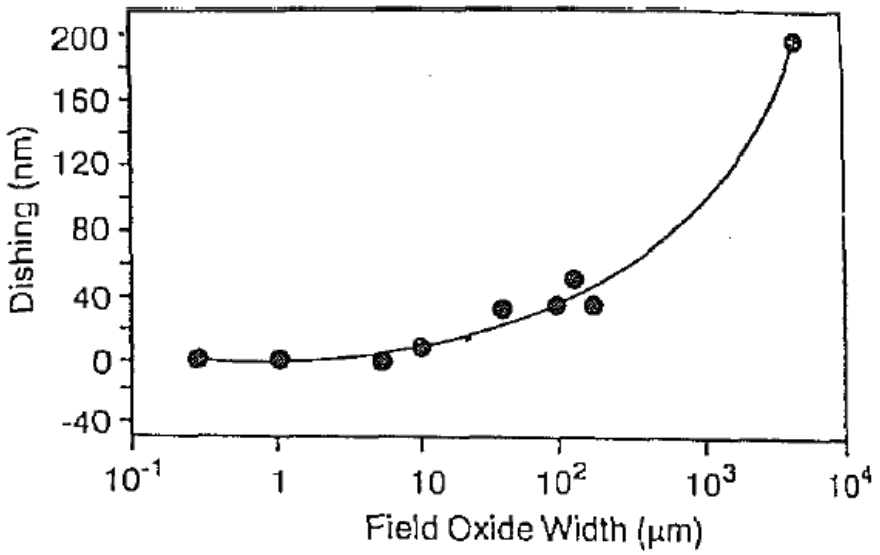


Fig. 8-97 Dishing versus field-oxide-region width for a specific oxide CMP process.

A number of approaches have been proposed to decrease dishing arising from the CMP-process used to fabricate STI, including: 1) a combination CMP/RIE process; 2) dummy structures added in the field-oxide regions of STI structures; 3) adding an extra nitride-layer on the trench-dielectric after it is deposited;[103,104] 4) adding an extra nitride-layer onto the dielectric after it has been deposited *and* using a reverse block-mask;[105] and 5) implementing a direct-polish CMP-process using either a linear-polishing-tool, or a rotating-polishing-tool with a harder polishing-pad and an efficient endpoint-detection-system. Techniques #3 and #4 are described in Chap. 9 (STI), and the others here.

In the *CMP/RIE-approach*, the oxide is polished by CMP to a point where the nitride is not yet exposed and the surface is thus still completely planar (no dishing has yet occurred). Then RIE is used to etch the oxide-layer, with an endpoint-algorithm used to stop the etching just as the nitride is exposed to plasma.[97] The drawback of this method is that it increases the complexity of the process used to planarize the trench-oxide.

In the *dummy-active-region-approach* some areas of the trench are not etched. Instead, as shown in Fig. 8-98, regions of Si remain in the center of the trench. This essentially reduces the trench-width so that dishing effects are decreased to an acceptable degree. The CMP-removal-rate over regions that are densely packed with active-areas is also slower than over regions containing small isolated areas. Thus, to expose the nitride CMP-stop-layer over such densely packed regions, the nitride over small isolated active areas gets stripped, causing damage to Si in areas beneath the stripped nitride. Thus, dummy-active-regions

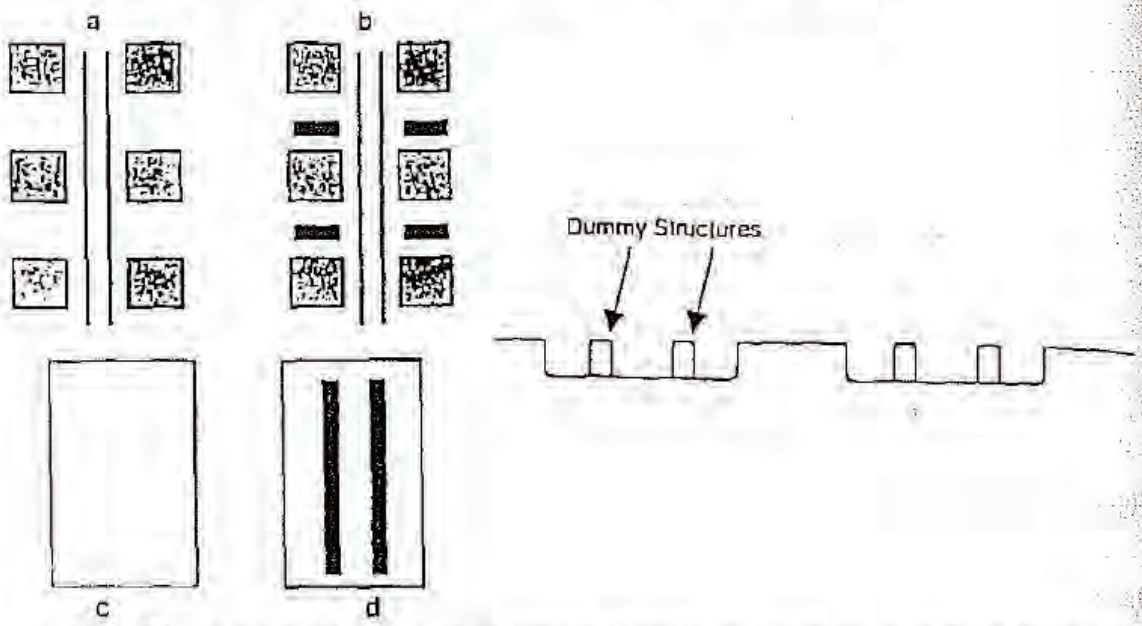


Fig. 8-98 Dummy (b) and dishing support features of small (a) and large wafer (c) features can prevent dishing.[89] c) Cross-section of dummy features in a trench.

must also be added to areas around isolated active regions. Finally, since dishing also occurs when wide areas of metal are polished (e.g., bond-pads, power-busses, and wide interconnect-lines), such dummy structures can also be added to these features. The technique, however, poses electrical, design, and space constraints, with the main issue being that several times as many structures must be added within a die over previous designs. This complicates the design and mask-making process, and significantly raises defect levels. Thus, some see this approach as merely a stop-gap measure.

The implementation of a *direct-polish CMP-technique* has also been pursued to avoid the drawbacks of the above methods. When a direct-polishing-process involves the use of a *linear-polishing-tool*, higher pad velocities (300–500-ft/min versus 34-ft/min for rotary-pad speeds) and lower polishing-pressures (3–5-psi versus 7-psi for rotary polishers) are used to suppress dishing. A highly-selective slurry (with cerium-oxide particles dispersed in an aqueous suspension of material that inhibits silicon-nitride removal) is also employed. This set of process conditions has been reported to significantly reduce dishing-effects (maximum oxide-removal across a die was 20-nm) and decrease the nitride-removal over isolated active-regions.[98] Decreased dishing using a direct-polish approach has also been described using *rotary-polishing-tools*. To achieve decreased dishing the correct set of process conditions and CMP consumables must be selected. That is, dishing effects are reduced by using a harder polishing

pad, a relatively non-selective slurry, an efficient endpoint-detection method, and an optimized CMP process.[99] As also discussed in Sect. 8.6.8, the use of slurry-free pads holds the promise of also reducing dishing.

The dishing problem also occurs during other CMP processes as well. As noted in earlier, dishing can occur when polishing ILD and IMD layers. Furthermore, as will be discussed in Chap. 16 (which deals with Cu interconnect processing issues), dishing and erosion of the ILD layers can occur when the Cu (and Cu barrier layer) are being polished back. Refer to Chaps. 15 and 16 for more information on these topics.

Erosion is an effect that occurs during CMP of Cu or W which results in the thinning of the ILD layer. That is, erosion is defined as the difference in the ILD thickness before and after the metal-CMP-step. The Cu (or W) film is also thinned, but its final thickness is reduced by the sum of the dishing plus erosion that has occurred. Erosion occurs because there is also a non-zero removal rate during the over-polish of the metal.

The extent of erosion depends heavily on the pattern-density (but not on metal linewidth, as is the case with dishing). Erosion is worst in regions of dense metal pattern (Fig. 8-99). This is so because when the polish begins the metal (e.g., Cu) covers the entire wafer. In regions of high Cu-pattern-density there is a lower

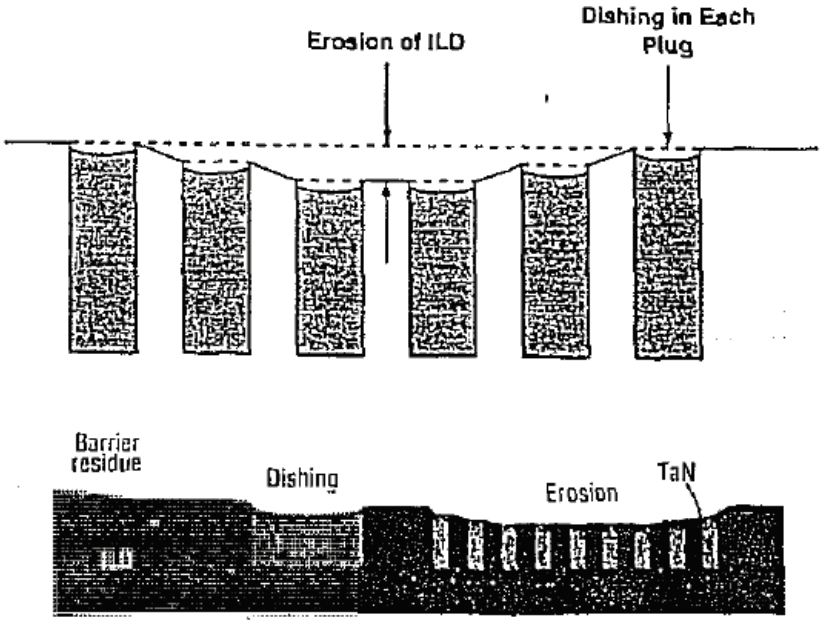


Fig. 8-99 Erosion of the ILD that occurs during over-polish of Cu during CMP.

aerial-density of Cu contacting the pad ("up-areas" of Cu). This is because the Cu in low areas (the "down-areas" of Cu) does not contact the pad. Thus, there is a higher pressure on the "up-areas" of Cu (and consequently higher polish-rate) than in regions where the Cu pattern-density is lower. Because of this difference in polish-rates, Cu is cleared in regions of high pattern-density sooner than in regions of lower pattern-density. Thus, the ILD in regions of high pattern-density receives a longer over-polish, adding to increase in erosion with pattern density.

8.12.2 Thickness Non-Uniformity Within a Wafer After CMP

Rigid polishing-pads provide good planarity within a die (chip), but poor thickness-uniformity across a wafer, for reasons explained here. If wafers were perfectly flat, this problem would not occur. A film on the surface of a flat wafer would be polished to a uniform thickness across the wafer (Fig. 8-100a). But not all wafers are uniformly flat. For example, if a wafer is tapered and the pad is rigid, the film on the thicker end of the wafer will be polished more than on the thinner end (Fig. 8-100b). If the wafer thickness undulates (with some regions thicker than others), a film covering regions where the wafer is thicker will be polished more (and will end up thinner) than a film on the thin parts (Fig. 8-100c). Softer (i.e., more compliant) pads produce improved uniformity across-the-wafer thickness, but at the price of degraded planarity. As discussed earlier,

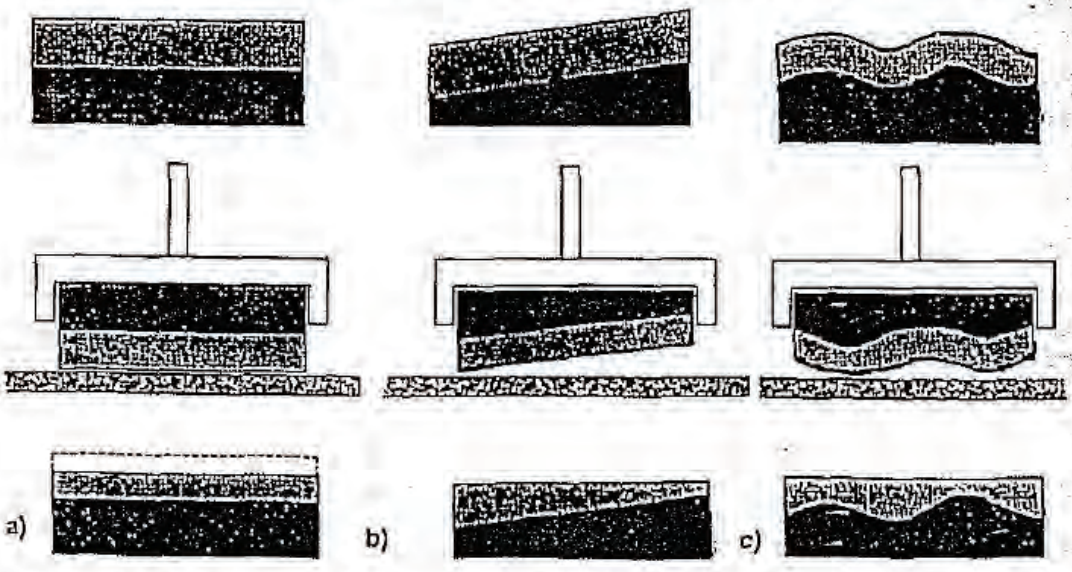


Fig. 8-100 If a rigid pad is used for CMP: a) A film on a flat-wafer is uniformly polished; b) A film on a tapered-wafer is polished thinner over the thicker end of the wafer; (c) For wafers with variations in wafer-thickness across the wafer, the film is polished thinner over the thicker wafer regions.

one solution is to use a composite pad (a rigid-pad backed by a softer one). Another approach is to use a linear polishing tool with a flexible (e.g., air) bearing beneath the polishing belt. This allows the belt to conform better to the wafer-surface, and reportedly produces improved across-the-wafer thickness-uniformity (and significantly less dishing).

8.12.3 Economic Considerations: Throughput and Cost of Ownership

While CMP is an enabling process for deep-submicron IC-fabrication, it is also a relatively expensive one. The cost of a single CMP-pass in 1999 (without a cleaning process) ranged from \$5 to \$8, depending on the material being polished. The post-CMP cleaning-step adds about another \$1 to \$2 per wafer. The main contributors to the cost of CMP include the polishing (and cleaning) tools, the slurry, the test- and monitor-wafers, and the need to do off-line measurements. The CMP-facilities, labor-costs, utilities, DI-water, pads, and waste-disposal are additional expenses. The 2002-target for the cost of a CMP process on 300-mm wafers is \$5.56/wafer/pass.[101]

The price of a fully-integrated polishing-tool in 1999 ranged from \$2–2.5 million, with the cleaning-module costing about \$500,000. Such tools are capable of polishing 50-60 200-mm wafers-per-hour. Slurries are also expensive. Oxide-slurries cost about \$10/gallon, tungsten-slurries about \$30/gallon, and Cu-slurries from \$40–50/gallon. One gallon is consumed for about every ten wafers.

Test- and monitor-wafers are needed not only during the set-up of the process, but also to check tool-performance for every lot. Typically, the removal-rate and the within-wafer-non-uniformity are monitored for each lot to ensure the tool is operating within spec. Some IC manufacturers have reported that up to two hours per shift are needed to process and measure the monitor-wafers and product test-wafers off-line, and that this reduces the CMP-equipment-capacity by nearly 50%. This is the reason that *in situ* measuring techniques are so attractive.

When the *cost-of-ownership* (CoO) of various CMP-tools are being compared, the three parameters that have the most significant impact on this cost are:

1. Throughput
2. Mean-Time-Between-Failure (MTBF)
3. Availability

The *throughput* of a CMP-tool is perhaps the most important parameter that contributes to the cost-of-ownership. It is defined as the quantity of wafers that can be completely polished within a given period of time (e.g., wafers-per-hour). A throughput of 50-60 wafers-per-hour is considered to be a good throughput

range. This parameter is highly dependent on the number of polishing platens and carrier heads. The 2002 performance-target for CMP-tool-throughput is 75 wafers-per-hour-per-tool.[102]

The *mean-time-between-failure* (*MTBF*) is a measure of a tool's reliability. It is defined as the mean-value of the number of hours that a specific tool can run continuously, non-stop, without breaking down. Most CMP tool-manufacturers seek to design their tools with an MTBF of at least 400 hours.

The *availability* of a CMP tool is the actual time the tool can run product wafers (i.e., the time when a tool is being qualified or tested, or is shut down for preventative-maintenance is subtracted from the total-time a tool is in service). This parameter reflects the ease with which a tool can be serviced or maintained

It should be added as a final note that the test-time is not always taken into account when calculating CoO. However, the amount of time required to qualify a CMP-tool for processing and the *mean-time-to-test* (*MTTT*), can also have a noticeable impact on CoO. If one assumes that a CMP-tool needs to be qualified once during each shift (i.e., three times a day), and that each qualification test (i.e., off-line thickness-measurement) takes about 30 minutes, this means that the MTTT is 1.5 hours per day, or 10.5 hours per week. This 10.5 hours could be considered lost CMP-processing-time. The shorter is this time, the higher the CMP throughput-number, which would lower the CoO.

REFERENCES

1. L. Rothman, *J. Electrochem. Soc.*, **127**, 2216 (1980).
2. C. Jang *et al.*, *Proc. 4th Intl. IEEE VMIC Conf.*, Santa Clara CA, p. 357, 1987.
3. S. Morimoto *et al.*, "Characterization of the Chemical-Mechanical Polishing of Intermetal Dielectric Films," *J. Electrochem. Soc.* 1993, p. 449.
4. J.M. Steigerwald, S.P. Murarka, and R.J. Gutmann, *Chemical Mechanical Planarization of Microelectronic Materials*, Wiley & Sons, New York, 1996.
5. M.A. Fury, "The Early Days of CMP," *Solid State Technol.*, May 1997, p. 81.
6 B. Davari *et al*, "A New Planarization Technique, Using a Combination of RIE and Chemical Mechanical Polishing (CMP)," *Tech. Dig. IEDM*, 1989, p. 61.
7. R. DeJule, "CMP Grows in Sophistication" *Semicond. Internatl*, Nov. 1998, p. 56.
8. M. Tomozawa, "Oxide CMP Mechanisms," *Solid State Tech.*, July 1997, p. 169.
9. F. Preston, "Theory & Design of Plate Glass Polishing Machines," *J. Soc. Glass Tech.*, **11**, 214 (1927).
10. Paper TuB-A4, *Tech. Dig., Topical Meeting on the Science of Polishing*, Optical Soc. of America, 17 April 1984.
11. C.W. Liu *et al.* "Characterization of CMP Process Based on Nanoindentation Measurement of Dielectric Films," *J. Electrachem. Soc.*, **142**, p. 3098 (1995).

12. S.R. Runnels and L.M. Eyman, "Tribology Analysis of CMP," *J. Electrochem. Soc.*, 144, p. 1698 (1994).
13. W.T. Tseng, "Re-examination of Pressure & Speed Dependence of Removal rates during CMP processes," *J. Electrochem.. Soc.*, 144, p. L15 (1997).
14. B. Zhao and F.G. Shi (1999). "CMP: Threshold Pressure and Mechanism," *Electrochemical Solid State Letters* 2, p. 145 (1999).
15. G.B. Shinn *et al.*, "Chemical-Mechanical Polish," in *Handbook of Semiconductor Manufacturing Technology*, Y. Nishi and R. Doering, Eds., Marcel Dekker, New York, 2000, p. 415.
16. "New Model Could Vastly Improve CMP Processes," *Semiconductor International*, May 1999, p. 21.
17. L. Cook, "Chemical Processes in Glass Polishing," *J. Non-Crystalline Solids*, 126, 160 (1990).
18. W.-T. Tseng *et al.*, "CMP of Fluorinated Silicon Dioxide: Is it Necessary and Feasible?" *Solid State Technology*, February 1997, p. 64.
19. W.J. Schaffer and H.W. Fry, "BPSG Improves CMP Performance for Deep Submicron ICs," *Semiconductor International*, July 1996, p. 205.
20. T. Izumitani and S. Harada, "Polishing Mechanism of Optical Glasses," *Glass Technology*, 12, 131 (1971).
21. *ibid. Ref. 17.*
22. J. Olsen and F. Moghadam, "Planarization Techniques," in *Handbook of Multilevel Metallization for Integrated Circuits*, Eds. S.R. Wilson, C.J. Tracy, and J.L. Freeman, Noyes Pubs., Park Ridge, NJ, 1993, Chap. 6, pp. 426-438.
23. W.J. Patrick *et al.*, "Application of CMP to the Fabrication of VLSI Circuit Interconnections," *J. Electrochem. Soc.*, 138, June 1991, p. 1778.
24. P.A. Burke, "Semi-Empirical Modeling of SiO_2 CMP Planarization," *Proc. VMIC* 1991, p. 379.
25. J. Schlueter, "Controlling Contamination During CMP Processes," *Solid State Technology*, May 1996, p. S8.
26. F. Kaufman., "Chemical Mechanical Polishing for Fabricating W Metal Features as Chip Interconnects," *J. Electrochemical Society*, 138, Nov.1993, p. 3460.
27. M. Pourbaix, *Atlas of Electrochemical Equilibria in Aqueous Solutions,"* Nat. Assoc. of Corrosion Engr., Houston TX (1974).
28. J. Gisler, "Enhancing CMP Performance, *Semicond. Internatl*, July 2000,p. 150.
29. C.C. Yu., *Advanced Metallization for ULSI Applications Proceedings*, Oct., 1993.
30. A.E. Braun, "CMP Battles Low-k Fundamentals," *Semiconductor International*, October 2000, p. 72.
31. K. Witt and L. Cook, "Getting an 'Edge' with CMP," *Semiconductor International*, October 2000, p. 75.
32. Courtesy of Rodel Inc.

33. Courtesy of Rodel Inc.
34. H. Nojo, *et al.*, "Slurry Engineering for Self-Stopping & Dishing-Free SiO_2 CMP," *Tech. Dig. IEDM*, 1996, p. 349.
35. A.E. Braun, "Slurries and Pads Face 2001 Challenges," *Semiconductor International*, November 1998, p. 65.
36. T. Hara *et al.*, "Direct CMP of Low-*k* Organic by MnO_2 Slurry," *Mtg. Abs. Electrochem. Soc.*, Fall 1999, Abs. 1269.
37. J. McCann and C. Smith, "Bulk Slurry Distribution for Integrated Manufacture," *Semiconductor International*, November 1996, p. 94.
38. G. Basim *et al.*, "Effect of Particle Size of Chemical Mechanical Polishing Slurries for Enhanced Polishing with Minimal Defects," *J. Electrochem. Soc.*, September 2000, p. 3523.
39. I. Ali *et al.*, "Chemical-Mechanical Polishing of Interlevel Dielectric: A Review," *Solid State Technology*, October 1994, p. 63.
40. J.H. Golden *et al.*, "Evaluating and Treating CMP Wastewater," *Semiconductor International*, October 2000, p. 85.
41. H.-J. Kim *et al.*, "Physical and Chemical Characterization of Re-Used Oxide CMP Slurry," *Mtg. Abs. Electrochem.. Soc.*, Fall 2000, Abs. 477.
42. *ibid. Ref. 4.*
43. Courtesy of Rodel Inc.
44. Courtesy of Rodel Inc.
45. A.E. Braun, "Slurries and Pads Face 2001 Challenges," *Semiconductor International*, Nov. 1998, p. 65.
46. M.R. Oliver *et al.*, "CMP Pad Surface Roughness and CMP Removal Rate," *Mtg. Abs. Electrochem. Soc.*, Fall 2000, Abs. 487.
47. I. Ali and S.R. Roy, "Pad Conditioning in Interlayer Dielectric CMP," *Solid State Technology*, June 1997, p. 185
48. C.-Y. Chen *et al.*, "Operational Aspects of Chemical Mechanical Polishing: Polishing Pad Profile Optimization," *J. Electrochem. Soc.* October 2000, p. 3922.
49. Courtesy of 3M Company.
50. Courtesy of 3M Company.
51. Courtesy of 3M Company.
52. W.L. Smith *et al.*, "Film Thickness Measurements for Chemical Mechanical Planarization," *Solid State Technology*, January 1996, p. 77.
53. A.S. Lawing, "Pattern Dependent Polish Rate Behavior in Oxide Chemical Mechanical Polishing," *Mtg. Abs. Electrochem. Soc.*, Fall 2000, Abs. 493.
54. E. Worthington, "New CMP Architecture Addresses Key Process Issues," *Solid State Technology*, January 1996, p. 61.
55. R. DeJule, "Advances in CMP," *Semiconductor International*, November 1996, p. 88. 56. Courtesy of Applied Materials

57. H.E. Litvak and H.-M. Tzeng, "Implementing Real-Time Endpoint Control in CMP," *Semiconductor International*, July 1996, p. 259.
58. H.E. Litvak, *U.S. Patent No. 5,499,733*. Issued March 19, 1996.
59. H. Litvak, *Proc of the SEMI Taiwan CMP Seminar* Sept. 1995, Taipei, p. 123.
60. *ibid.* ref. 25.
61. J. Skrovan, K. Robinson. "Planarization slurry including a dispersant and method of using same." *U.S. Patent 5,827, 781*. Oct. 27, 1998.
62. F. Preston, "Theory & Design of Plate Glass Polishing Machines" *J. Soc. Glass Tech.*, **11**, 214 (1927).
63. V.B. Menon, R.P. Donovan, in W. Kern Ed., *Handbook of Wafer Cleaning Technology*, Noyes Publications, Park Ridge, N.J., 1993.
64. K. Mikhaylich *et al.*, "Evaluation of Contact and Non-Contact Cleaning for CMP Applications," *Mtg. Abs. Electrochem. Soc.*, Fall 1999, Abs. 1268.
65 D. Hymes *et al.*, "Brush Scrubbing Emerges as a Future Wafer-Cleaning Technology," *Solid State Technology*, July 1997, p. 209.
66. T.L. Myers, "Post-Tungsten CMP Cleaning," *Solid State Tech.* Oct. 1995, p. 59.
67. J.M. de Larios *et al.*, "Post-CMP Cleaning of Oxide and Tungsten Applications," *Semiconductor International*, May 1996, p. 121.
68. D. Hymes *et al.*, "The Challenges of the Copper CMP Clean," *Semiconductor International*, June 1998, p. 117.
69. K. Osseo-Asare, K.K. Mishra, "Solution Chemical Constraints in the CMP of Copper: Aqueous Stability Diagrams for the Cu-H_2O and Cu-NH_3-H_2O Systems," *J. Electron. Materials*, **25**, p. 1599 (1997).
70. Y. Feng, *et al.*, "Corrosion Protection of Copper by a Self-Assembled Monolayer of Alkenathiol," *J. Electrochem. Soc.*, **144**, p. 55 (1997).
71. A.A. Busnaina and F. Dai, "Megasonic Cleaning," *Semiconductor International*. August 1997, p. 85.
72. A. Busnaina *et al.*, "Single Wafer Non-Contact Post-CMP cleaning Using DI Water and Dilute Chemistry," *Mtg. Abs. Electrochem. Soc.*, Fall 1999, Abs. 1267.
73. M. Elissa *et al.*, "Post Copper CMP Cleaning: A Non-Contact Method," *Mtg. Abs. Electrochem. Soc.*, Fall 1999, Abs. 1270.
74. I. Ali *et al.*, "Investigating the Effect of Secondary Platen Pressure on Post-CMP Cleaning," *Microcontamination* **12**(10), p. 45, 1994.
75. R. Davenport and B. Brown, "Integrated CMP Clean Methods," *European Semiconductor*, July 1999, p. 36.
76. Courtesy of Ebara.
77. Y. Zhang *et al.*, "Wafer Dimensional Analysis for Chemical Mechanical Planarization," *Solid State Technology*, July 1997, p. 179.
78. J. Reynolds *et al.*, "Defect Inspection Technologies to Meet the Challenges of Advanced CMP Processes," *Solid State Technology*, March 1998, p. 39.

79. Courtesy of KLA-Tencor Inc.
80. *ibid., Ref. 55.*
81. *ibid., Ref. 39.*
82. W. O'Mara, "The Inevitable Shakeout in the CMP Equipment Market," *Solid State Technology*, May 1998, p. 84.
83. F. Kaufman *et al.*, "Chemical Mechanical Polishing for Fabricating Patterned W Metal Features as Chip Interconnects," *J. Electrochemical Society*, Vol. 138, November 1993, p. 3460.
84. K. Wijekoon *et al.*, "Tungsten CMP Process Developed," *Solid State Technology*, April 1998, p. 53.
85. I. Kim *et al.*, "Optimized Process Developed for Tungsten CMP," *Semiconductor International*, November 1996, p. 119.
86. S. Roy, "Advanced Metal & Intercon. Syst. for ULSI Applications," 1995, p. 733.
87. *ibid. Ref. 29.*
88. J.-F. Wang "CMP of Cu Multilevel Interconnects," *Procs. VMIC* 1995, p. 505.
89. R. DeJule, "CMP Challenges Below Quarter Micron," *Semiconductor International*, Nov. 1997, p. 54.
90. D.R. Evans, "Electrochemical Interaction Between Copper and Barrier Materials During CMP," *Mtg. Abs. Electrochem. Soc. Mtg.*, Fall 1996, Abs. No. 504,
91. A. Beverina *et al.*, "Copper Photocrrosion Phenomenon during Post-CMP Cleaning," *Electrochem. Soc. Letts.*, March 2000, p. 156.
92. A.R. Seturaman *et al.*, "Copper vs Aluminum: A Planarization Perspective," *Semiconductor International*, June 1996, p. 177.
93. J. Nuetzel *et al.*, "Conventional and Slurry Free CMP of Polymer Layers for Gbit DRAM," *Mtg. Abs. Electrochem. Soc.*, Fall 1999, Abs. 1232.
94. W.-T. Tseng, "CMP of Fluorinated Silicon Dioxide: Is it Necessary and Feasible?" *Solid State Technology*, February, 1997, p. 61.
95. *ibid. Ref. 36.*
100. S.J. Fang *et al.*, "Optimization of the CMP Process for Pre-Metal Dielectrics," *J. Electrochem. Soc.*, February 2000, p. 682.
101. T. Marbeiter, *et al.*, "An Update: Transition to 300-mm CMP," *Semiconductor International*, November 1998, p. 76.
102. *ibid. Ref. 101.*
103. M. Nandakumar *et al.*, "Shallow Trench Isolation for Advanced ULSI CMOS Technologies," *Tech. Dig. IEDM*, 1998, p. 134.
104. L. Peters, "Choices and Challenges for Shallow Trench Isolation." *Semiconductor International*, April 1999, p. 71.
105. G. Badenes *et al.*, "A New Dummy-Free Shallow Trench Isolation Concept for Mixed-Signal Applications," *J. Electrochem. Soc.*, 147, Oct. 2000, p. 3827.

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,

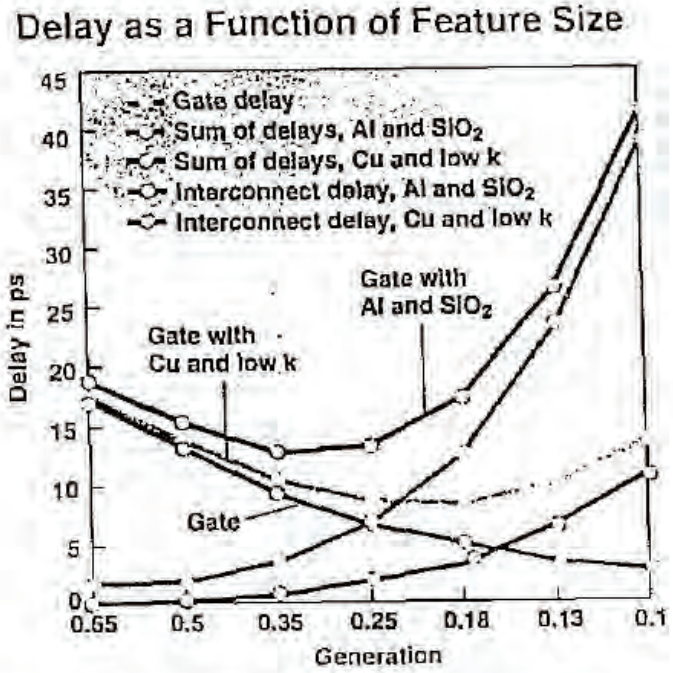


Fig. 12-1 RC-delay as a function of feature size.[1] (© IEEE 1997)

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]

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

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

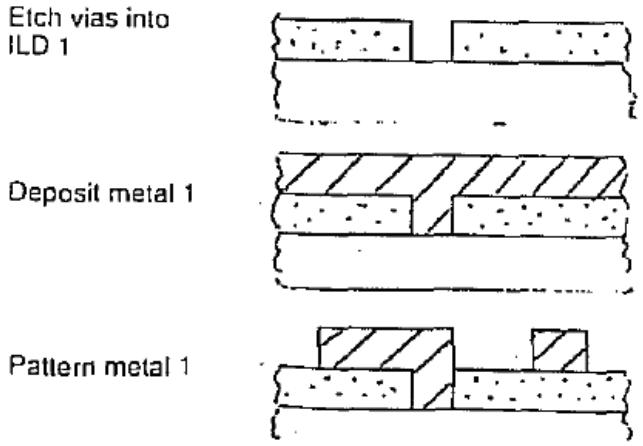


Fig. 12-2 Schematic of process steps used to produce a single-level-metal interconnect.

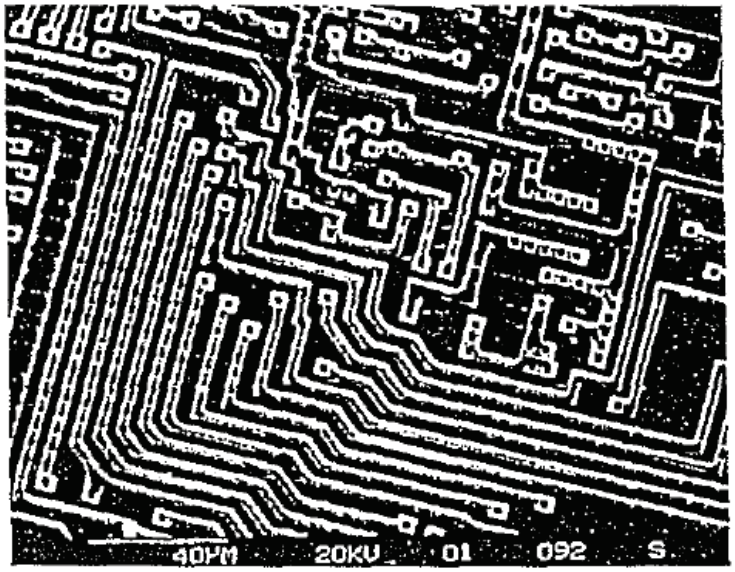


Fig. 12-3 SEM showing the patterned metal layer on a circuit with only one layer of metal.

technology. As a point of reference for materials parameters, the resistivity of pure Al is $\rho_{Al} = 2.7$-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-interconneet 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

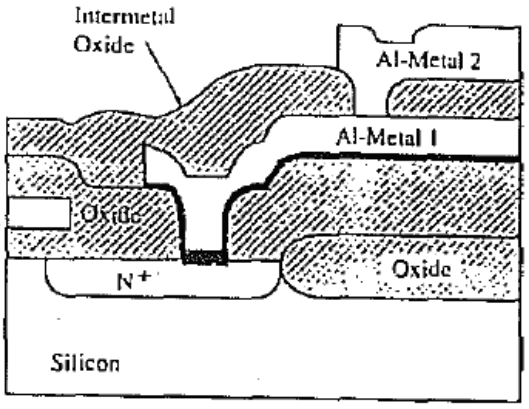


Fig. 12-4 Schematic of a double-level-metal structure used in early bipolar-ICs.

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.

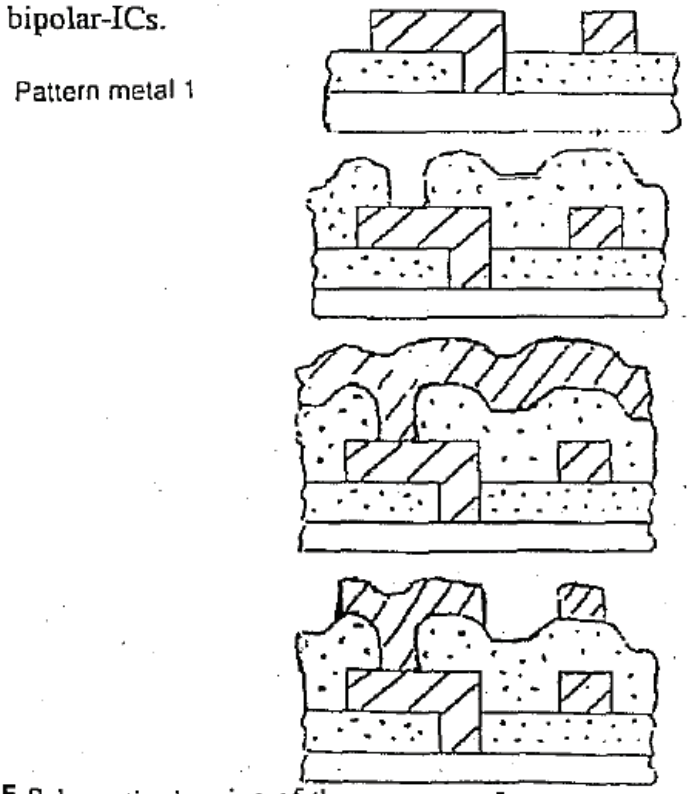


Fig. 12-5 Schematic drawing of the sequence of process steps used to produce a double-

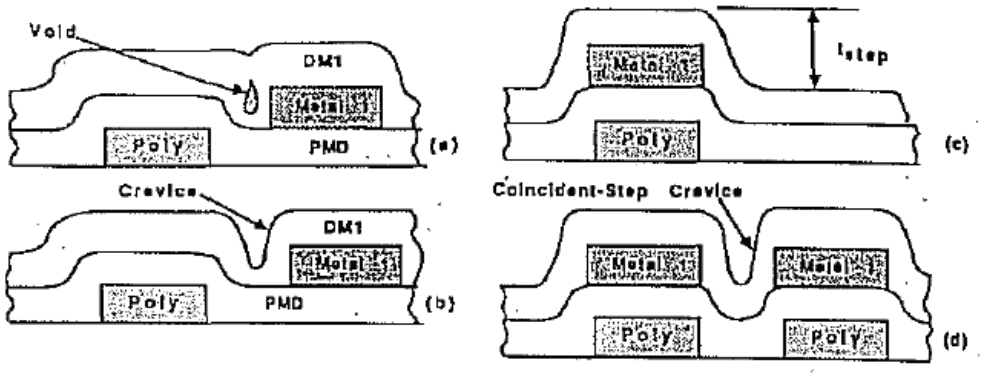


Fig. 12-6 Topography conditions that arise in MOSFETs that make it difficult to fabricate a double-level-metal structure without some kind of planarization method.

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

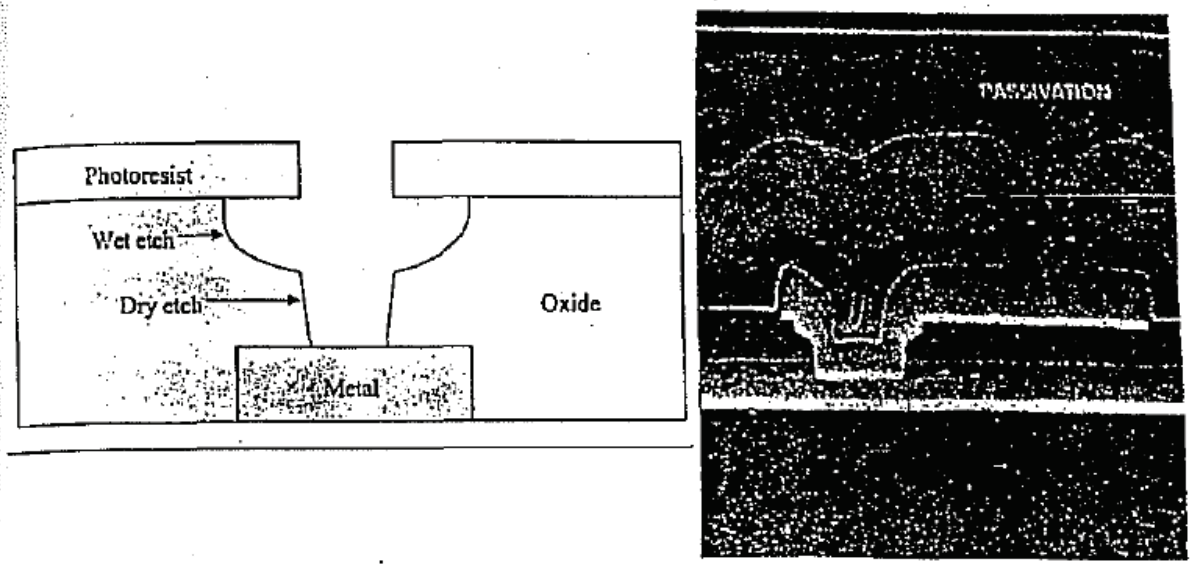


Fig. 12-7 a) Schematic of a via with a "champagne-glass" shape; b) SEM of actual via.

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.

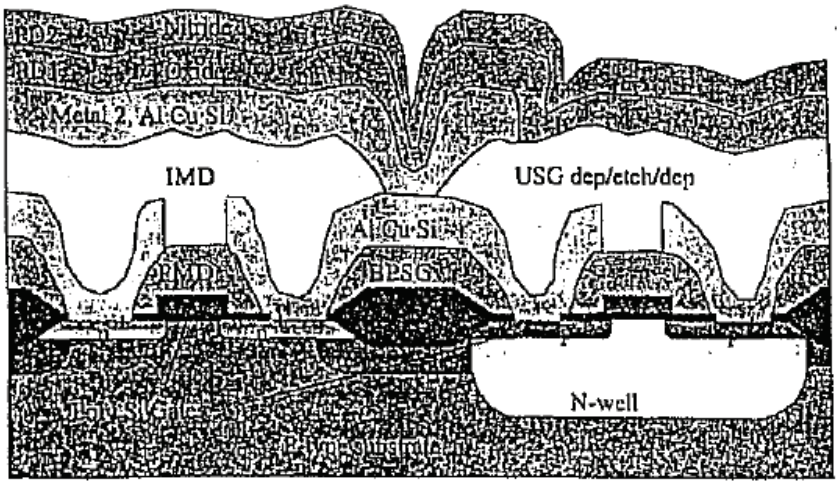


Fig. 12-8 Schematic of a double-level-metal structure in CMOS with partial-planarization.

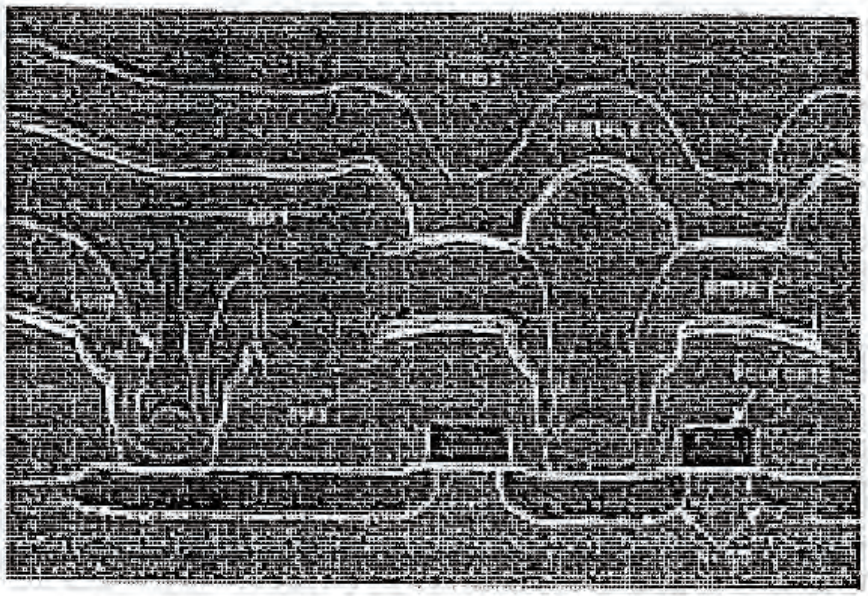


Fig. 12-9 SEM of an actual double-level-metal structure in CMOS with partial planarization.

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

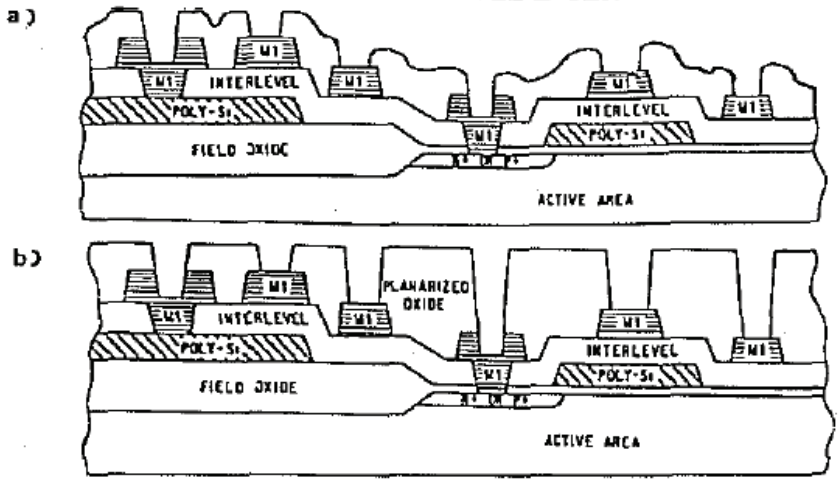


Fig. 12-10 a) If no planarization is used, the via and contact holes between adjacent conductor levels all have approximately the same depth; b) If planarization of the intermetal dielectric layer is achieved, the vias will vary widely in depth.

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

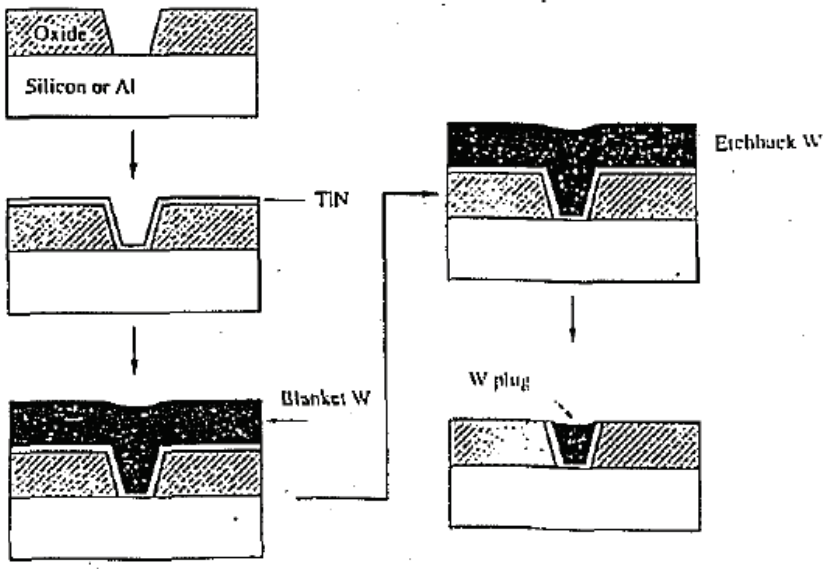


Fig. 12-11 Schematic of the process used to form W-plugs with CVD-W and etchback.

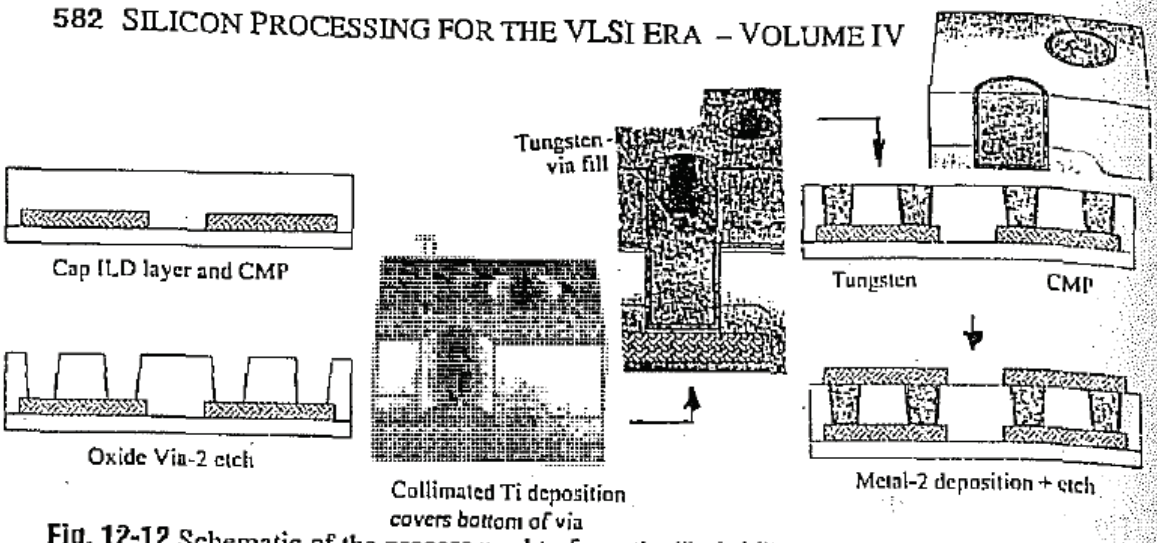


Fig. 12-12 Schematic of the process used to form the "hybrid" W-plug/Al-alloy structure.

"*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.

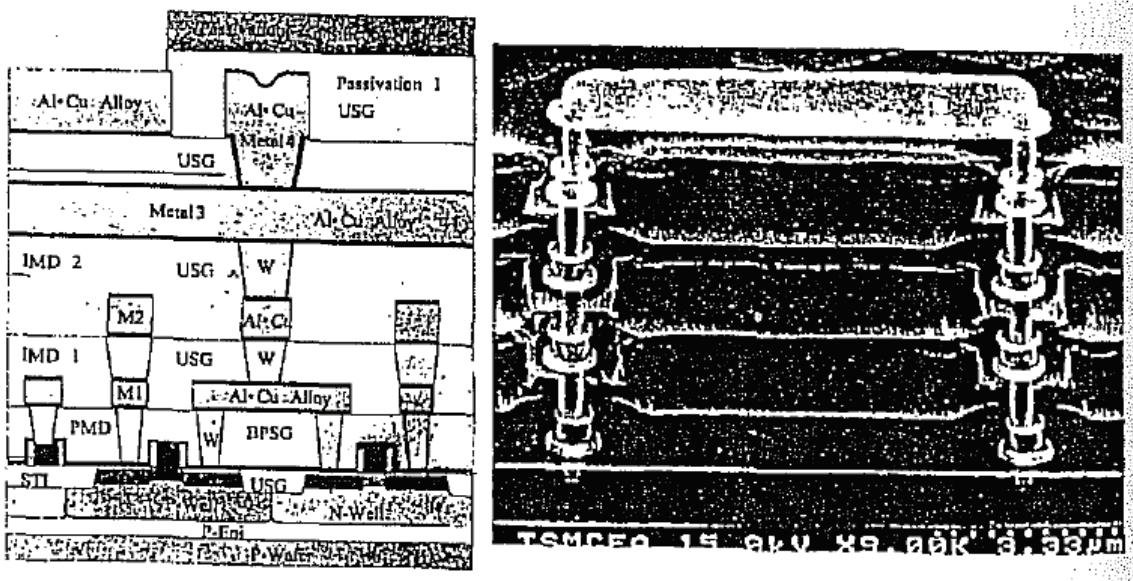


Fig. 12-13 a) Schematic of a W-plug/Al-alloy interconnect structure with stacked vias; b) SEM of an interconnect structure with stacked vias and five-levels of metal.

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.

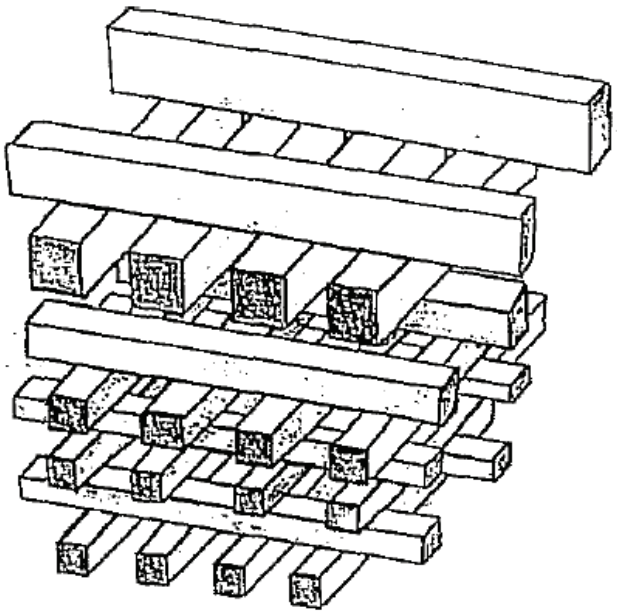


Fig. 12-14 Schematic showing the orthogonal architecture of multilevel interconnect lines.

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

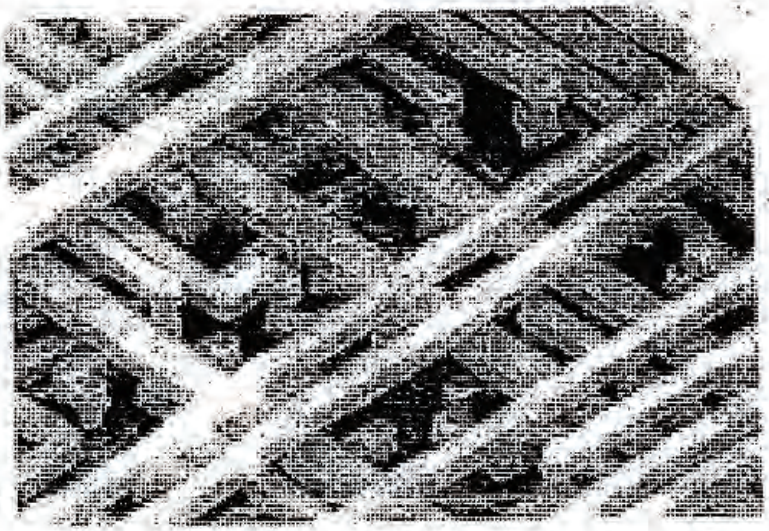


Fig. 12-15 SEM of orthogonal architecture of metal lines in a W-plug/Al interconnect.

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

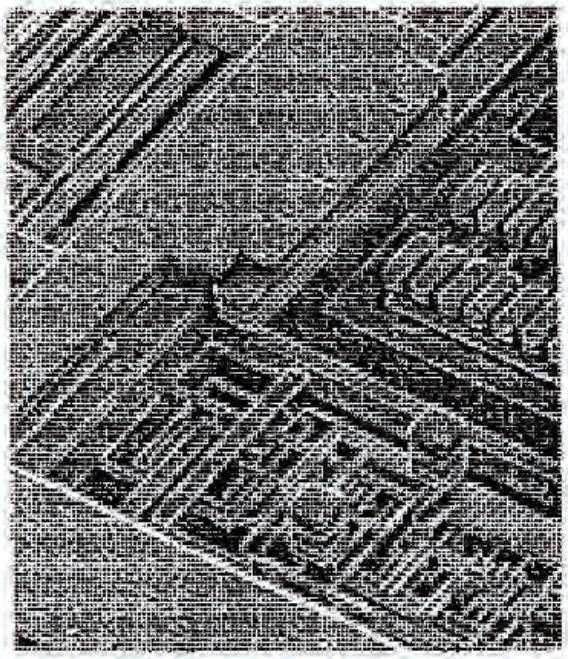


Fig. 12-16 SEM of orthogonal architecture of metal lines in a Cu interconnect.

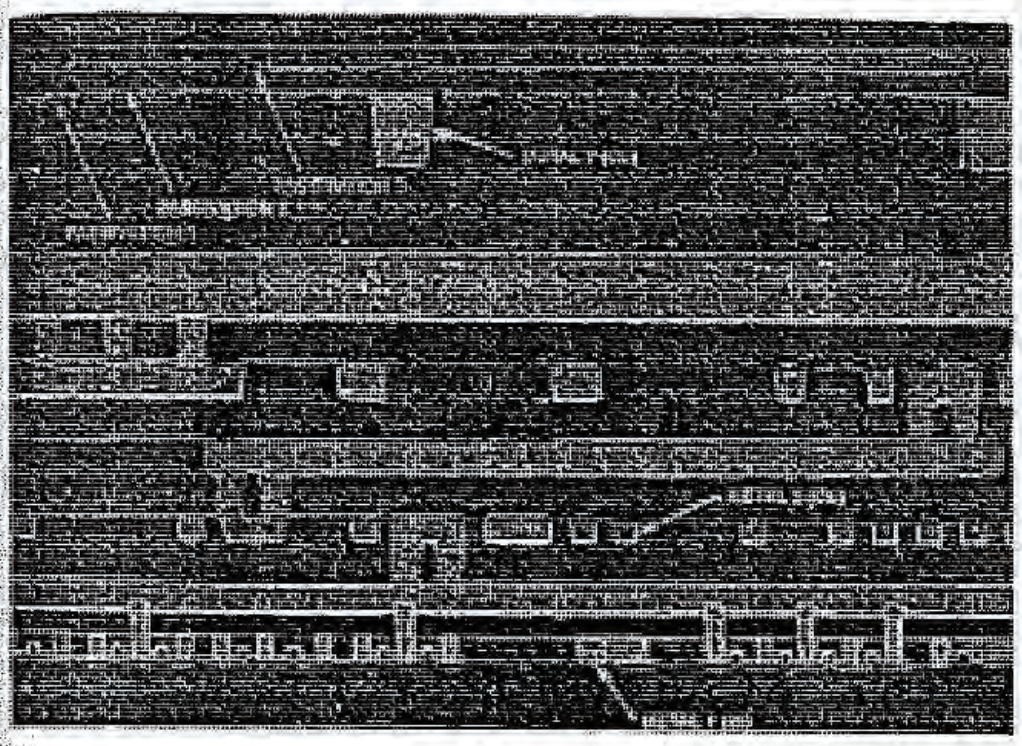


Fig. 12-17 SEM of a cross-section of a 6-metal-level interconnect.

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

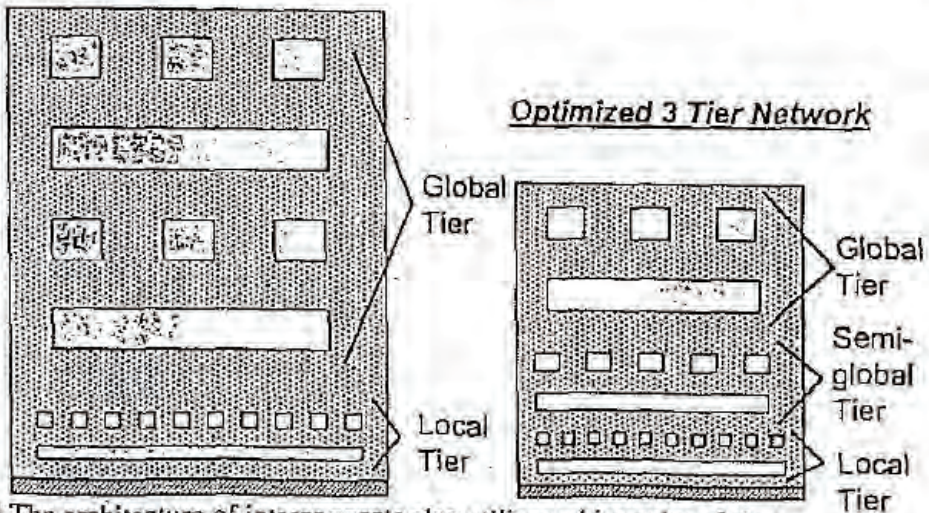


Fig. 12-18 The architecture of interconnects also utilizes a hierarchy of *tiers*.

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 ½ pitch	230	160	115	80	55	40
MPU gate length (nm)	140	85–90	65	45	30–32	20–22
DRAM ½ 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.

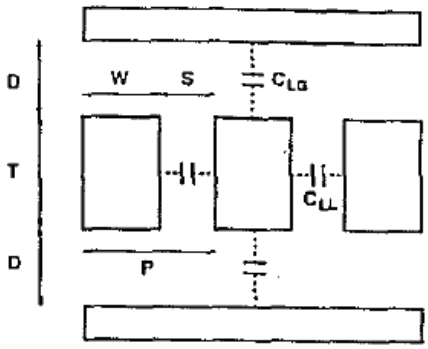


Fig. 12-19 Simplified *RC*-delay test-structure.[11] (© IEEE 1995)

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:

$$R_{int} = \rho l/wt = \rho l/apt \tag{12.2}$$

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:

$$C_{LG} = k\varepsilon_o wl/d = k\varepsilon_o apl/bt \tag{12.3}$$

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

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

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:

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

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

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

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

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

This equation can also be rewritten as:

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

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:

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

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:

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

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:

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

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

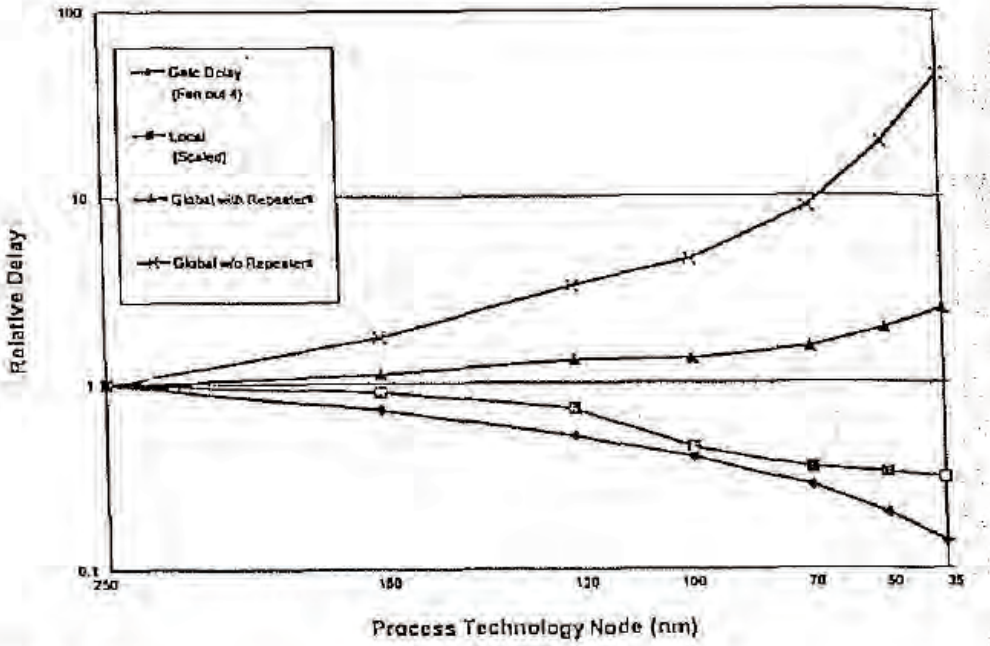


Fig. 12-20 Interconnect versus delay, from ITRS99.[3]

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:

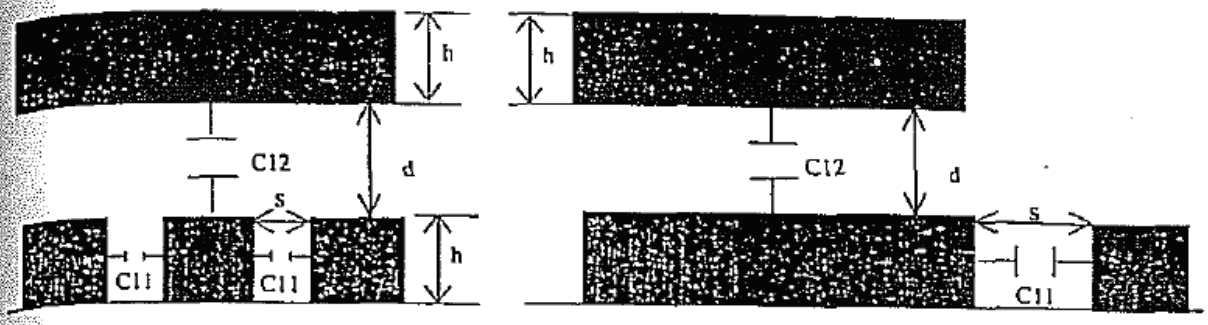


Fig. 12-21 Capacitance coupling contributions for interconnect systems: a) Dense interconnect case where major coupling is between minimum-spaced lines; and b) Wide interconnect lines where the major coupling is between metal-levels.

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

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.

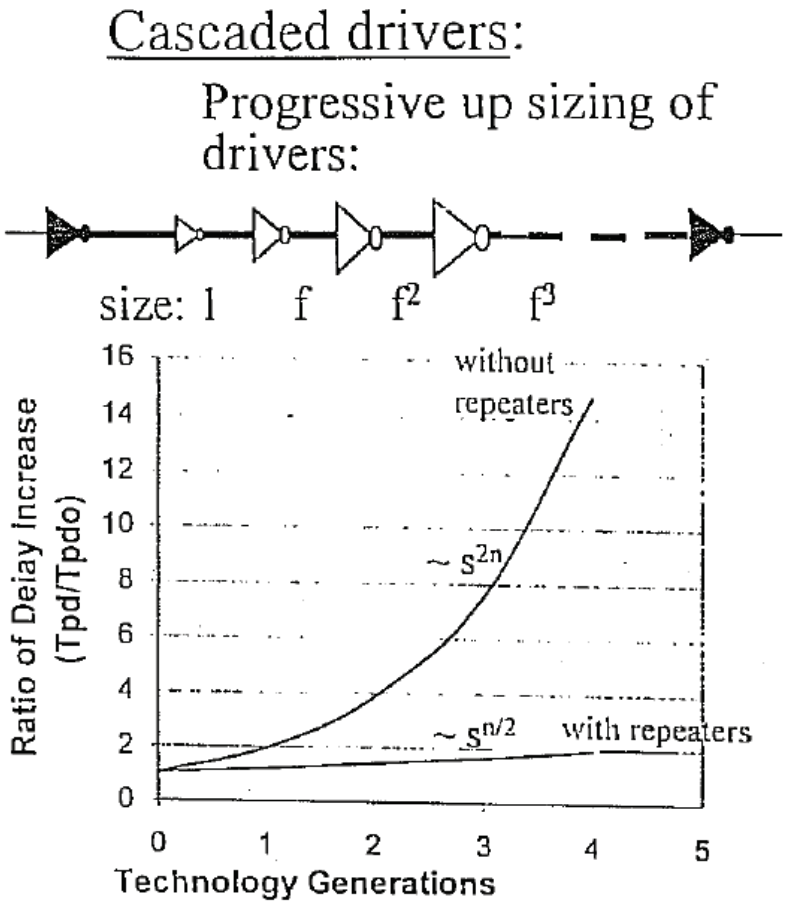


Fig. 12-22 Impact of repeaters on global propagation delay.[5]

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.

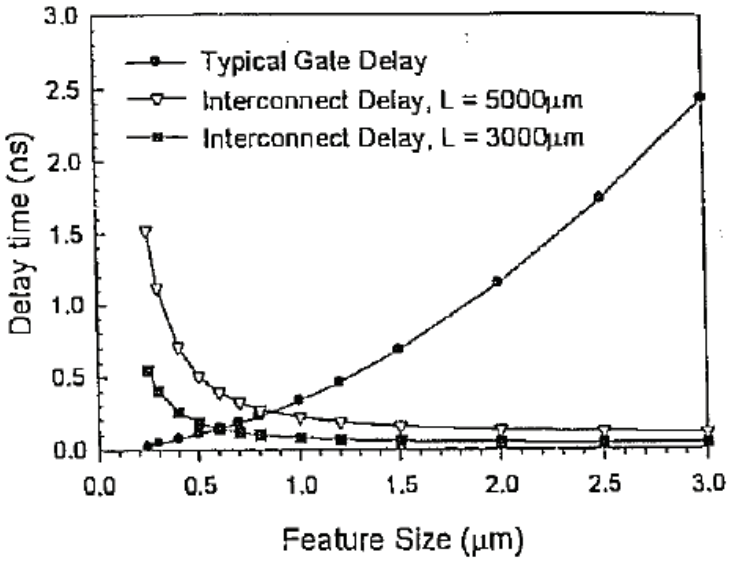


Fig. 12-23 Gate and interconnect delay versus feature size. Interconnect delay is shown for repeater spacings (L) of 3000-μm and 5000-μm.[14]

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:

$$P \sim f\, C_{total} V^2 \tag{12.16}$$

or

$$P \sim I_{total}^2\, R_{total} \tag{12.17}$$

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

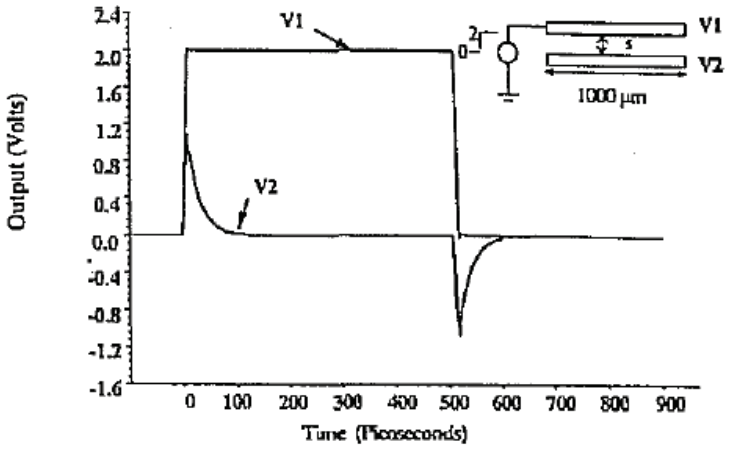


Fig. 12-24 Electrical modeling result showing the crosstalk voltage induced from a driven interconnect-line (V1) into an undriven line (V2) seeing a 20-psec waveform-edge. The lines were spaced at s = 1000-nm with a common length of 1-mm.[10]

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

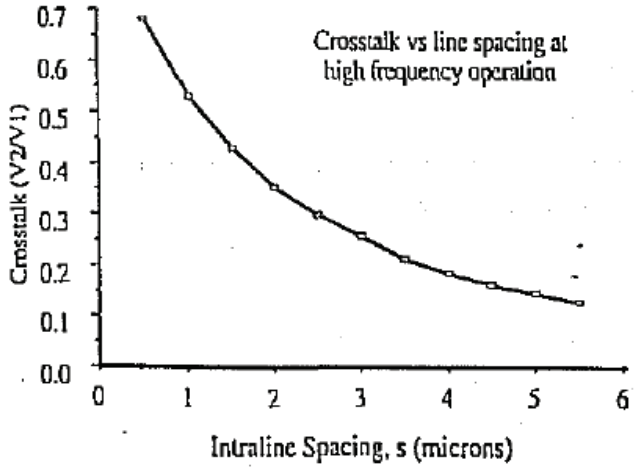


Fig. 12-25 Simulation results showing that the magnitude of the crosstalk between 2 interconnects separated by different line spacings, s.[10] Conditions described in Fig. 12-24.

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

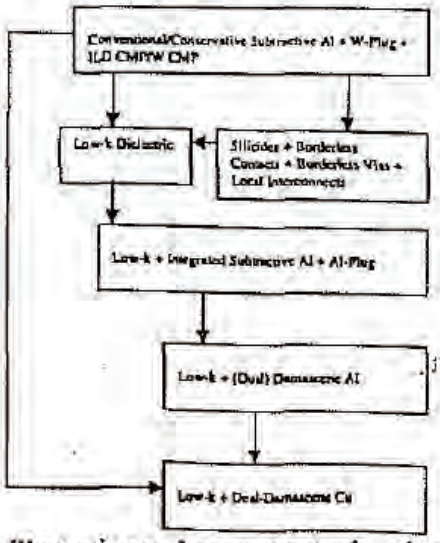


Fig. 12-26 Schematic illustrating an interconnect migration.[15]

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.

Chapter 15

DUAL-DAMASCENE INTERCONNECTS

As the submicron era evolved in the 1990's, several novel manufacturing methods were adopted to keep pace with Moore's law. One such approach, namely the formation of interconnect structures with *damascene processes* was chosen to fabricate IC interconnects. In this chapter we describe this technique.

15.1 DAMASCENE VERSUS SUBTRACTIVE INTERCONNECT STRUCTURES

The conventional way of fabricating each conducting layer of an interconnect structure of an IC is a so-called *subtractive process*. In such a subtractive process, a blanket metal film is deposited onto an insulator-covered wafer surface (with contacts - or vias - having been previously etched in the insulator layer - Fig. 15-1). Next this metal film (e.g., Metal-1) is patterned by etching - to create metal interconnect lines. Finally a dielectric film is deposited over these etched metal lines. Vias are then etched in the overlying dielectric film to allow connection passages for the next layer of metal that is to be deposited onto the dielectric film covering the Metal-1 lines. This subtractive approach requires several process capabilities: 1) It must be possible to effectively etch the metal and; 2) It must be possible to deposit both the dielectric and metal layers with a deposition process that has good gap-filling capability. In some interconnect schemes (especially as the ICs are scaled below 0.5-μm), it becomes difficult if not practically impossible, to meet these goals. For example, for technology nodes smaller than 0.5-μm it is desirable to replace Al with Cu as the intercon-

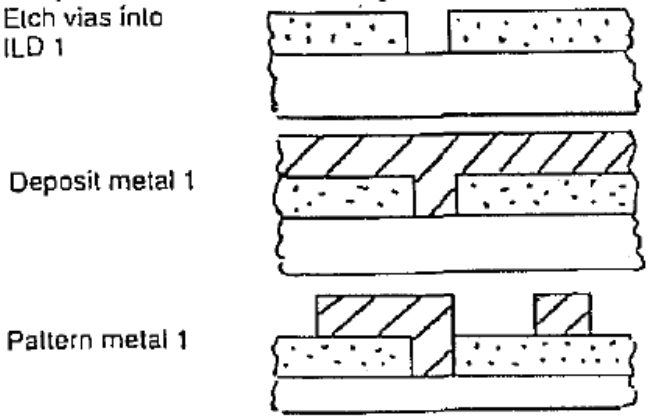


Fig. 15-1 Subtractive approach to forming interconnect structures for ICs.

nect conductor (i.e., because Cu has a higher conductivity than Al, and also exhibits superior electromigration characteristics). However, since it is difficult to etch Cu, an interconnect scheme that does not require Cu etching would be preferable. In addition, as the spaces between metal lines become smaller, it becomes progressively harder to fill such narrow gaps with an insulator film. Again, a significant advantage would be gained if it were possible to use a process that does not require gap-filling when insulator films are deposited. To achieve the above goals, an alternative interconnect fabrication technology to the subtractive method has thus been developed, called the *damascene method*.

15.2 DAMASCENE STRUCTURES AS ALTERNATIVE DEEP-SUBMICON INTERCONNECTS

In the damascene approach to forming interconnects on ICs, a trench is cut into the dielectric layer and then filled with metal. (The term *damascene* is derived from the decorative practice of creating metal inlays, as developed by the artisans of Damascus in the Middle Ages.) This method thus provides two major advantages over the subtractive method of forming interconnect structures:

1. In the damascene approach, no etching of the metal layers is required. The metal-etch process of the subtractive approach is replaced with a metal chemical-mechanical-polishing process (CMP). The metal-CMP step removes any metal deposited above the top surface of the trench (including the regions where the metal is deposited on the dielectric surface). This leaves a planarized dielectric surface with metal inlaid in trenches of the dielectric. Since the CMP process leaves a globally planarized surface, *the need for filling narrow gaps during the dielectric deposition process is eliminated.* The only gap-filling capability required by the damascene approach is during the deposition of metal into the recesses of the dielectric. Since metal etch (especially the etching of Cu) and dielectric gap-fill are viewed as two of the greatest challenges in the drive to smaller dimensions, elimination of these steps represents a major benefit. On the other hand, filling of even narrow, high-aspect-ratio trenches with metal (i.e., with W by CVD, or Cu by electroplating) is not as difficult.

2. There is only a bottom interface for Cu (versus top and bottom interfaces for Al-alloys with W plugs). This helps reduce overall via resistance.

We will also see that by using a *dual-damascene* process, the number of processing steps is smaller than for the subtractive approach. Hence, use of dual-damascene interconnects may also result in lower manufacturing costs. The

damascene approach was pioneered by IBM, and began to appear in commercial IC products at about the 0.25-μm node.[1]

15.3 SINGLE-DAMASCENE INTERCONNECT STRUCTURES

Damascene interconnects can be formed with either a *single-damascene* or *dual-damascene* process. The dual-damascene process offers the advantage of reducing the number of process steps, and thus most of our discussion focuses on it. However, the single-damascene process is always carried out on at least the first level of a multi-level interconnect structure of an IC. For this reason, it is useful to begin with a description of the single-damascene approach.

In the *single-damascene process* (see Fig. 15-2 [and Fig. 12-10 in Sect. 12.1.4]), the interlevel dielectric layer (ILD) is planarized after it is deposited. Next, a contact (or via) is patterned with a lithography step, and then the contact opening is etched into the ILD layer. After completion of the contact-hole etch step, a metal plug is formed in this opening - typically by blanket deposition of the metal, followed by an etchback or CMP step to form the plug (e.g., a W-plug). Thus, the surface of the dielectric containing the metal plug after such an etchback or polish step is globally planar (see Fig. 8-92). The metal lines of the interconnect level are formed next. This can be done using a subtractive process, as described in Sect. 15.1, or with another single-damascene process.

Assuming a single damascene process is applied twice (first to create plugs, and second to create metal lines), the second time it is employed involves the following process steps (Fig. 15-2). Another dielectric layer is deposited over the

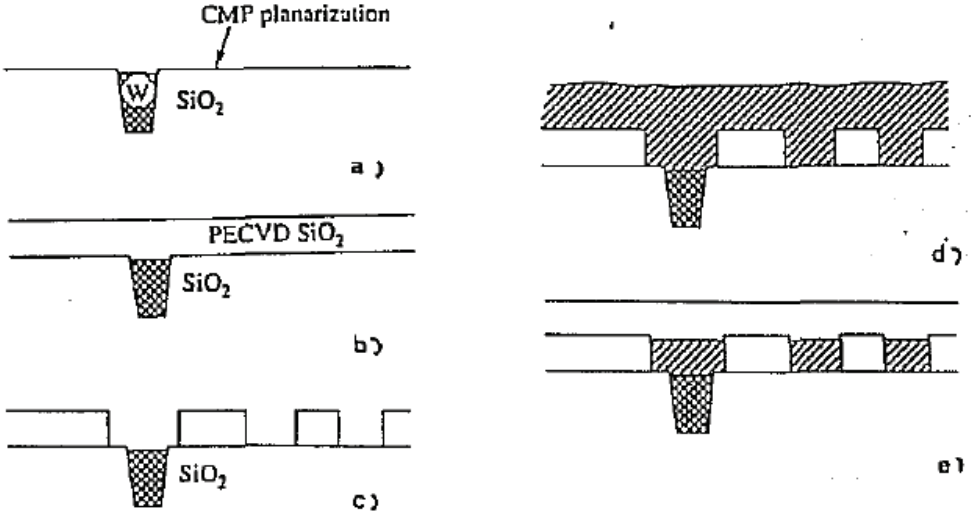


Fig. 15-2 A *single-damascene* process (applied twice) to form one level of an interconnect structure: (a) W-plugs formed after the SiO_2 is planarized; (b) IMD is deposited by CVD; (c) A trench is etched in the IMD; (d) Metal is deposited to fill trenches; (e) CMP metal.

underlying, plug-containing layer. This overlying dielectric film therefore also remains planarized. A second lithography step is then used to form the pattern of a trench, which is aligned to the previously fabricated via plug. With this aligned trench pattern acting as a template, a trench or groove is etched into the planarized overlying ILD. After the resist pattern is stripped, the trench is filled with a metal (such as aluminum or copper).[2] This metal is polished back by CMP to create a metal line embedded in a dielectric trench, and in contact with the underlying plugs. Finally, another dielectric layer is deposited onto the dielectric layer containing the metal-filled trenches. This dielectric layer will be the one through which the next level of vias (or holes to the contact pads) will be created.

It is worth emphasizing several important aspects of the single-damascene process: 1) If CMP is used to form both the plug and the metal in the trench, no metal etch process is needed; 2) No gaps have to be filled by the dielectric-film-deposition processes; 3) The process still contains about the same number of steps as the subtractive interconnect process (i.e., the two polish steps replace the via and metal etch steps of the subtractive process, and the two metal-deposition steps replace the metal-plug-deposition step and main-metal-line deposition step of the plug/metal-line subtractive interconnect process); and 4) Since the first-level interlevel dielectric layer is planarized before the contact holes are etched, *contact holes* will generally have different depths (e.g., those going to the substrate will be deeper than those going to the gates). Thus, to allow all contact holes to be opened, the shallower ones will expose the underlying material to the plasma etching reactants during the time the deeper contacts are being opened. To prevent the underlying material in the shallow contact holes from being attacked during this time, an etch-stop layer is often deposited after the salicide process is completed (typically using Si_3N_4 as such an etch-stop film). After the contact etch is completed, the nitride etch-stop layer is removed, Since the thickness of this layer is uniform on all regions, it can be etched away with a much shorter overetch time. This method allows all of the contact holes in a single-damascene process to be opened without damaging the underlying material on any part of the die.

15.4 INTRODUCTION TO DUAL-DAMASCENE INTERCONNECT PROCESSES

In a *dual-damascene (DD) structure*, only a single metal deposition step is used to simultaneously form the main metal lines and the metal in the vias. That is, both trenches and vias are formed in a single dielectric layer. The vias and trenches are defined by using two lithography steps (Fig. 15-3). Trenches are typically etched to a depth of 4000-5000-Å, and the vias are typically 5000-

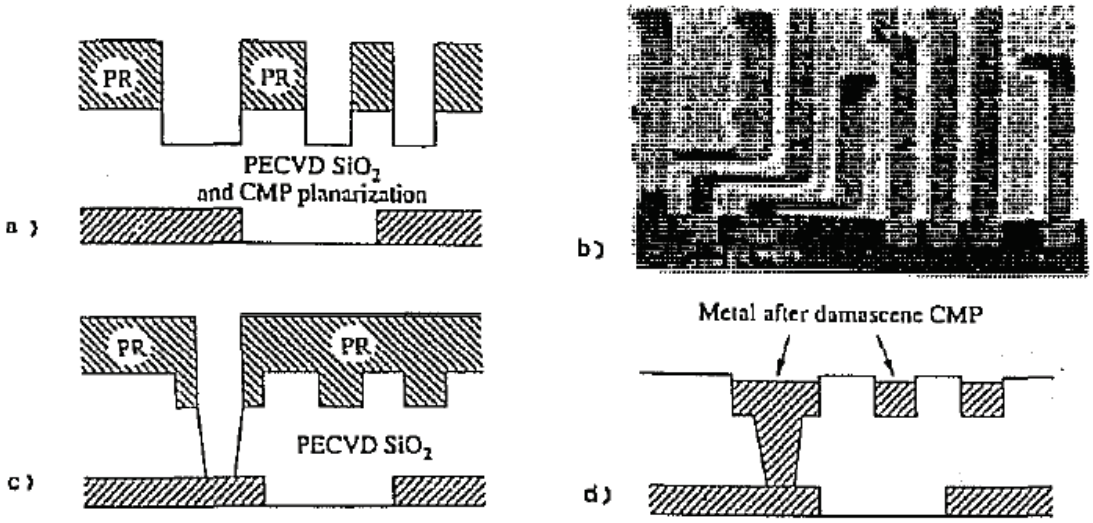


Fig. 15-3 *Dual-Damascene* process: (a) IMD is deposited by CVD and planarized by CMP. Trench is defined by PR #1 and then etched; (b) SEM of the etched trench after PR #1 has been stripped; (c) Vias are defined by PR #2 and then etched, using PR #2 to protect other regions of the IMD from etching; (d) Metal is deposited to simultaneously fill trenches and vias. CMP is used to remove excess metal. The sequence is repeated for the next level of metal.

7000Å-deep. After the via and trench recesses are etched, the via is filled in the same metal-deposition step that fills the trench. After filling, the excess metal that is deposited outside the trench is removed by a CMP process, and a planar structure with metal inlays is achieved. As in the single-damascene process, once a planarized surface is achieved, it is no longer necessary to perform CMP on the dielectric layers. Thus, one of the CMP steps needed for each metal level in the subtractive interconnect process is also eliminated.

It should also be mentioned that in early dual-damascene processes, the trench dimension was somewhat wider than that of the vias. This made the alignment of the via to the trench somewhat easier. However, with time, the via and trench widths have been made equal. (At the 0.18-μm node, the via and trench widths are 0.22-0.24-μm wide, while at the 0.13-μm node they will probably be ~0.18-μm wide.) Thus, modern dual-damascene processes are more sensitive to misalignment between the trench and via. However, this overlay sensitivity problem is mitigated to some degree by the nature of the damascene process, making it possible to maintain high interconnect packing density. That is, the alignment of the trenches to the vias in the dual-damascene sequences is an easier task than in the subtractive interconnect process. In the latter, aligning metal lines

to the underlying contact holes (or vias) must be done after the opaque, shiny, blanket-metal film has been deposited. In damascene approaches, aligning the trenches to the vias is done through the transparent dielectric film, which allows tighter design-rule tolerances to be used when performing this alignment.

When comparing the single-damascene process to the simple dual-damascene process described above, one observes one metal-deposition step and one CMP step are eliminated in the latter (as well as one dielectric-deposition step). The reduction in the number of processing steps is another of the benefits that have driven the development of dual-damascene processes. Note that in single-damascene structures, plugs are typically filled with W, but in the dual-damascene-interconnects the metal that fills the holes is Al or Cu. The technology of filling of vias and trenches with Cu is described in Chap. 16. A dual-damascene process using Al was described by workers at IBM in 1998.[3]

15.5 THE THREE DUAL-DAMASCENE PROCESS SEQUENCES

Three different fabrication sequences have been developed to produce dual-damascene structures:

1. Trench-First Dual-Damascene
2. Via-First Dual-Damascene
3. Self-Aligned (or Buried-Via) Dual-Damascene

It should be noted that the *via-first* and the *trench-first* sequences are the processes currently being used in mainstream production. The trench-first sequence was probably the first one to be adopted, but the via-first sequence is currently the approach that has gained the greatest acceptance (and will likely remain so going forward). As a result, we will focus on these two sequences, and will point out the advantages and drawbacks of each one. While the third approach (the self-aligned process), has not been implemented in production (for reasons we will also explain), we will provide a brief description of it as well. However, whichever sequence is used, the final damascene structure is the same (see Fig. 15-4).

Before outlining the three dual-damascene processes, it should be mentioned that the dielectric layer in which the trenches and vias are etched actually consists of at least three layers of dielectric (see Fig. 15-8b). That is, SiO_2 (or a low-k dielectric is deposited to a depth of 5000-7000Å (and this will be the layer in which the *vias* will be etched). Next, a thin silicon nitride layer - or other material (which will serve as an *embedded etch-stop layer*), is deposited (about 30-nm thick). Finally, another SiO_2 layer (or a low-k dielectric layer) is deposited

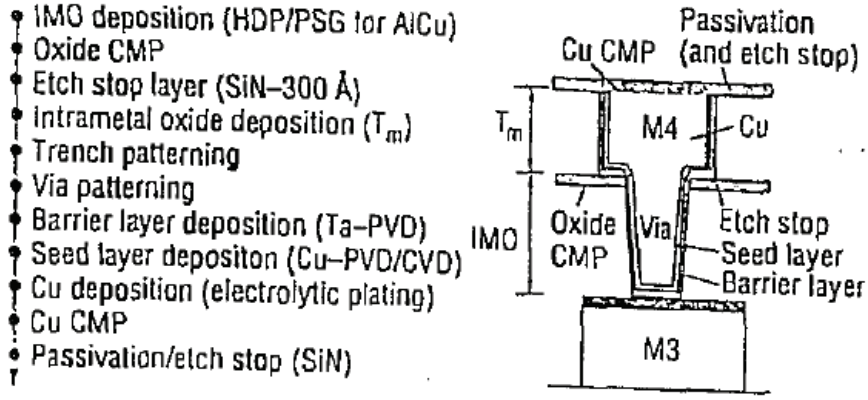


Fig. 15-4 A suggested dual-damascene-based Cu metallization flow and resulting cross-sectional structure.[8] Reprinted with permission of Solid State Technology. Published by PennWell.

(~4000-5000-Å thick). This will serve as the layer in which the *trenches* will be etched.

In the *trench-first* sequence (Fig. 15-5), the trench patterns are defined in the ILD first. That is, after spinning on the resist, the trench-pattern-mask is used to expose the resist. The trench is then produced by etching the dielectric down to the embedded etch-stop layer. After the trench-etching process, the first resist layer is stripped. A second resist layer is then spun on, and the via-pattern-mask is used to create openings in this resist layer - aligned to the trench that was etched previously. The resist protects the other parts of the wafer surface (including the etched regions of the trenches) so that the vias can be etched without further etching the dielectric in the trenches.

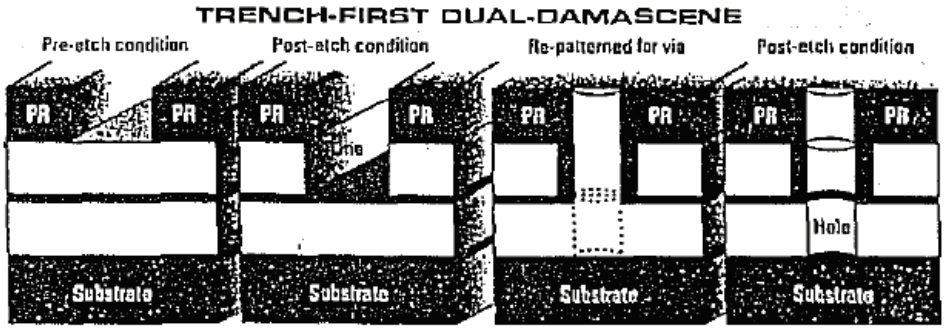


Fig. 15-5 In the trench-first approach, the vias are patterned & etched after trenches are etched.

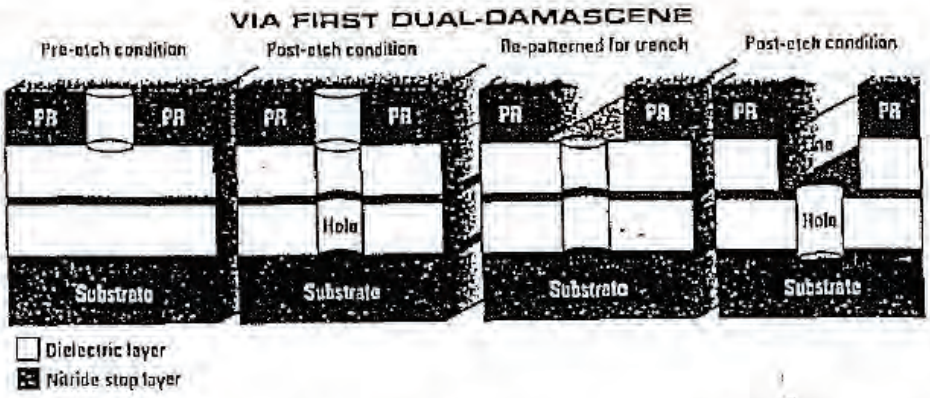


Fig. 15-6 In the via-first approach, the trench is patterned and etched after the via is formed.

The chief disadvantage of the trench-first sequence involves the fact that the vias must be patterned after the trench etch. That is, since resist is applied as a liquid onto the wafer surface, it fills recessed regions as water fills a lake. Hence, the top surface of the resist is planar. This means that the regions of resist over the damascene trenches is quite thick. Resolving fine features in thick resist is harder than in thinner resists. For features smaller than 0.25-μm the process latitude becomes too small for this to be a practical manufacturing process. Thus, while the trench-first sequence could be used for early dual-damascene processes (0.35-μm and some 0.25-μm processes), it is difficult to extend it to smaller technology nodes (i.e., in order to able to use it at such small dimensions, it may be necessary to implement a thin imaging-resist-layer process, as discussed later).

In the *via-first approach* the vias are defined first in the ILD, followed by patterning the trenches. As noted above, this is currently the most common method of fabricating dual-damascene interconnects (and is schematically shown in Fig. 15-6). The sequence of forming the damascene recesses in this approach begins by exposing the *via patterns* with the first mask. After etching the vias completely through the entire dielectric stack (except not through the barrier layer at the bottom of the dielectric stack) and stripping the resist, a second mask is used to pattern the trenches. The trenches are then created by etching the dielectric down to the embedded etch-stop layer. The barrier layer at the bottom of the vias is protected from further etching during the trench-etch either by resist or a BARC layer (applied for the trench-etch step) that floods the vias, or by using an etch process that is highly selective to the bottom etch-stop material (i.e., so that the etch-stop layer at the bottom of the via is not removed during the trench-etch step). After the resist is stripped and the etch-stop layer at the bottom of the via is removed by dry-etching, the metal that fills both the vias and the

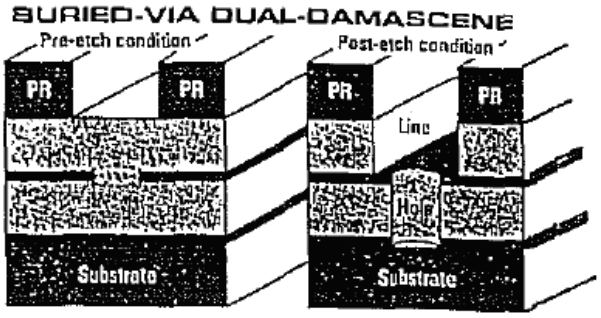


Fig. 15-7 In the self-aligned, or buried-via approach, a hard mask is patterned and etched. The trench and via can be etched in one step.

trenches can be deposited. After deposition, it is polished back to create the dual-damascene structure.

In the *self-aligned* dual-damascene process (Fig. 15-7), the via pattern is created in the embedded etch-stop layer with a lithography-and-etch sequence that is performed before the top dielectric layer of the stack is deposited. A special etch process is required for this step. After the via pattern is thus etched, the top dielectric layer is deposited. Finally, the trench mask is aligned to the via openings in the embedded nitride layer, and both the trench and via are opened with a single etch step. This sequence requires nearly perfect trench-to-via alignment, otherwise the via may no longer be round (but instead only half-moon shaped), resulting in high via resistance. In addition, an etch-process with high nitride-to-oxide selectivity is required. For these reasons this scheme has not found use in production.

15.6 THE EVOLUTION OF DUAL-DAMASCENE FROM SiO_2 TO LOW-k DIELECTRIC STACKS

In the first generation of dual-damascene processes (i.e., at the 0.25-μm node), Cu was used together with a conventional undoped PECVD-SiO_2 dielectric (with a PECVD-SiN layer beneath the SiO_2, which acts as both a Cu diffusion barrier and an etch-stop when the vias are etched, Fig. 15-8a).* The nitride and oxide layers may be deposited sequentially in the same tool (e.g., a PECVD reactor).

* This SiN layer is deposited first, and as noted, serves as a barrier against the diffusion of Cu between metal levels and also as an etch-stop in the dielectric etch that defines the vias. Because SiN has a high value of k, this layer should also be as thin as possible to avoid adding significantly to the interlevel capacitance. At the same time, the SiN layer must be low-stress, dense, and pinhole–free, so that its diffusion-barrier properties are not compromised. High-density is also critical to the nitride layer's role as an etch-stop.

This Cu/SiO_2 structure gave improved interconnect performance over the Al/W-plug subtractive process because Cu interconnects (i.e., with Cu in both the lines and vias) have a lower resistance than Al lines combined with W-plugs. Thus, the RC delay term was reduced through a decrease in R. A significant problem that occurs with using this dielectric structure is that the trench formation has no etch stop. Since the trenches are etched to an intermediate point part way through the dielectric layer, within-wafer etch-non-uniformity translates directly into within-wafer non-uniformity of the trench depths.

In second generation dual-damascene processes (down to about the 0.18-μm node), the CVD-SiO_2 material was replaced by a first generation low-k material, namely a PECVD-SiO_2 layer doped with fluorine (*fluorinated silicate glass, FSG*) whose $k = 3.5$. (FSG has many properties similar to SiO_2 and thus requires the smallest number of process integration changes.) A SiN layer, embedded halfway through the thick dielectric layer, was also added to the dielectric stack to serve as an etch-stop layer when the trenches are etched. However, since the k value of SiN is $k = 7.0$, the embedded-etch-stop SiN layer increases the net-k value of the resulting dielectric stack. In fact, the overall capacitance of the FSG dielectric stack with an embedded-SiN etch-stop layer was about the same as that of first generation SiO_2 layers - but it had the benefit of being able to ensure the trench is only etched halfway through the dielectric layer. The corresponding drawback is that fabricating a multi-layer dielectric stack adds cost and complexity to the manufacturing sequence.

The next step in the evolution of the dual-damascene dielectric stack was to incorporate a second generation low-k material ($k = 2.7$) into both the upper and lower segments of the dielectric stack, as this now allows the wiring capacitance to be reduced. (Note that the wiring capacitance is reduced by about 11% when a SiLK/SiN dielectric stack is implemented, compared to the wiring capacitance obtained when using a FSG/SiN stack.*) The first dual-damascene structure that incorporated a 2nd generation low-k dielectric in the stack (together with Cu), was announced by IBM in April 2000. It was intended for use in the IBM's Cu-15

* In a paper published in late 2001, an alternative dielectric stack to the SiLK/SiN stack was proposed. That is, it used FSG and a new embedded layer material (SiCN), whose k value of 5 is lower than that of SiN. The capacitance of this stack was about 9% lower than that of the FSG/SiN stack, making its wiring capacitance comparable to the SiLK/SiN stack. However, the authors claim that the SiLK/SiN stack is mechanically weak, and tends to fail during bonding and packaging processes, whereas the FSG/SiCN stack is more robust.[5]

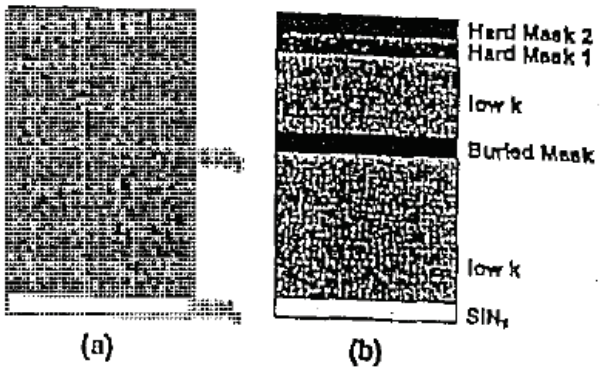


Fig. 15-8 a) First generation dual-damascene ILD stack, consisting of SiO_2 and a bottom SiN layer. b) Later generation dual-damascene ILD stack with low-k dielectric, embedded etch-stop layer, and hard-mask/capping layer (SiO_2).

0.13-μm CMOS technology.[4] The low-k dielectric material selected for this process was SiLK™ from Dow Chemical (see Chap. 14). The dual-damascene process is used to form the #2 through #5 interconnect levels of the 5-level-metal interconnect structure used for this technology node.

It should also be noted that when a 2nd generation low-k dielectric replaces the FSG (or SiO_2) material in the stack, a hard mask must also be added to the top of the stack. The material chosen by IBM for this hard mask was SiO_2, giving the dielectric stack a structure as shown in Fig. 15-8b. SiO_2 was selected for this role because it has a significantly lower k-value than SiN. In addition, this SiO_2 hard mask remains in place during metal deposition, so it can later serve as a polish-stop during metal CMP. Since use of an SiO_2 hard mask maintains the same materials system encountered during the CMP of metal and a PECVD-SiO_2 dielectric stack (Cu/TaN/SiO_2), the developed consumables, equipment, and processes for Cu-CMP on plain SiO_2 should be extendable to Cu-on-SiO_2-capped low-k. However, since this oxide cap becomes a primary supporting structure against material loss by dielectric erosion during CMP, the allowable dielectric erosion (see Sect. 8.11.1) in a capped dielectric stack is dramatically decreased.

It should also be noted that in some cases different materials may be used for the dielectric at the via level than for the trench level. Since most of the capacitive coupling issues that a low-k material helps solve are encountered between adjacent metal lines, the *trench dielectric* will be where the lowest-k

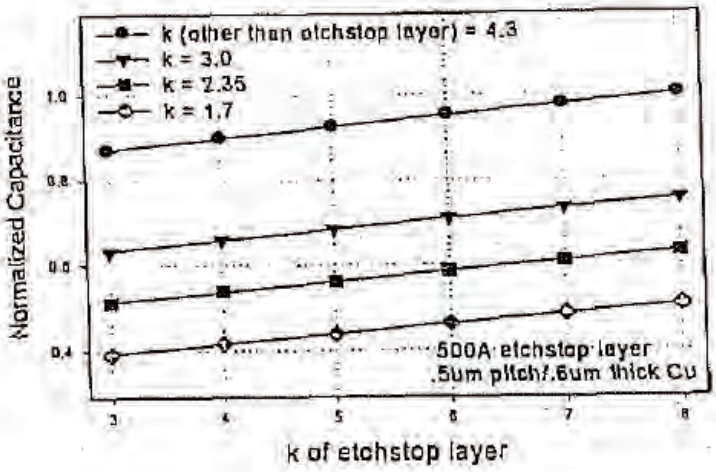


Fig. 15-9 Capacitance impact of the *k*-value of the etch-stop value in a dual-damascene structure for various dielectric constants.[32]

materials are first used. Since the via dielectric is less critical from a capacitance perspective, implementing a low-*k* material there may be delayed.

The use of SiN as an embedded etch-stop layer increases the capacitance. Figure 15-9 shows the effect of a 500-Å SiN on dielectric stacks with different low-*k* dielectrics.[32] To allow the capacitance of the dielectric stack to be reduced even further, a new material has been introduced to replace SiN for the embedded-etch-stop and hard-mask roles. This material is SiC (with $k = 4.5$). Figure 15-10 shows how much the use of Si-C instead of SiN reduces the effective *k*-value of the damascene dielectric stack.[33] Applied Materials introduced this material in 1999 under the name BLOk (barrier low-*k*).[6,7] In

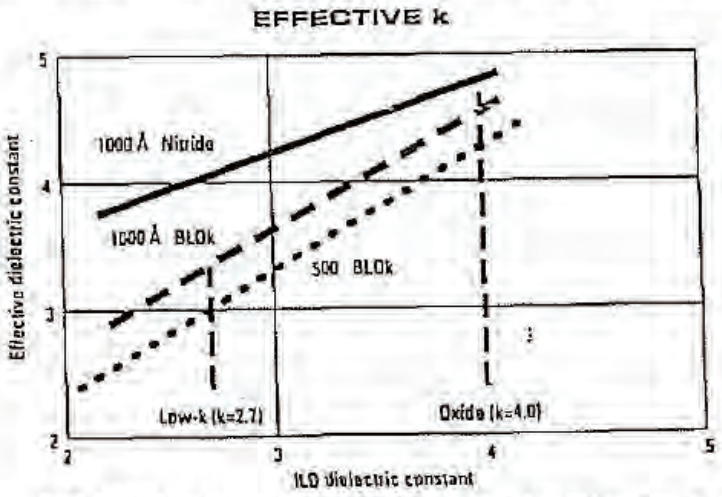


Fig. 15-10 The SiC-based dielectric barrier impacts the effective *k* value much less dramatically than the traditional nitride film. Courtesy Applied Materials.

addition, this material offers twice the etch selectivity of nitride to oxide-like films, has better Cu diffusion barriers than nitride, and shows good adhesion ot SiO_2, FSG, and organic-based low-k dielectrics. Furthermore, it exhibits the optical properties needed to make it behave like an anti-reflective coating in DUV imaging. The film is deposited in an HDP-CVD chamber using trimethylsilane at 350-400°C. The resulting Si-C film can also be etched using a "soft etch," which means it can also be used as the capping layer for Cu (i.e., at the bottom of the via of the via-first dual-damascene structure). Some companies have reportedly incorporated the Si-C films into damascene dielectric stacks which employ undoped SiO_2 or FSG, while others are using it with stacks containing lower-k films. Another vendor that supplies a tool and a process to deposit such an Si-C film is Novellus.

It is thought that the Si-C material will replace SiN as the etch-stop layer in low-k dielectric stacks for the 0.13-μm node. However, as devices are scaled even further, it is predicted that even the k value of 4.5 of Si-C will prove to be too large. Thus, it has been projected that at the 0.1-μm node, the embedded-etch-stop layer will be removed entirely. The ramifications of removing the embedded etch-stop layer will be discussed in a later section.

15.7 TRENCH-FIRST DUAL-DAMASCENE PROCESS FLOW

Here we present an example of a trench-first dual-damascene process flow, using undoped SiO_2 as the main dielectric material and Cu as the main interconnect metal. We assume that an STI isolation structure is used, so that the substrate is flat, up to the time of gate-stack formation. We will also see that both single-damascene and dual-damascene processing are used. Many variations of this example damascene-interconnect structure have been published.[8,9,10,11,12] Some advanced interconnect structures may elect to use both conventional W-plugs and etched Al-wire interconnects for some metal levels, and Cu-damascene for others. One such example is given in Ref. 13.

The fabrication sequence of this interconnect structure is begun by carrying out two single-damascene process sequences to form both the Metal-1 lines (Cu) and CVD-W plugs in the contact holes. The first process sequence is started by depositing a dielectric layer over the gate stack. This film serves as the *pre-metal dielectric* (PMD) layer in this structure, and it is polished flat with a CMP step. Vertical contact holes are etched into the PMD layer. A CVD-W process is then used to fill the contact holes. The W is then planarized, so that W-plugs are formed such that their top surface is flush with the top surface of the PMD.

The second single-damascene process is then performed. That is, an additional

dielectric film is deposited onto the PMD layer, and trenches (aligned to the W-plugs) are etched into it, deep enough so that the tops of the W-plugs are cleared of dielectric. A liner film (e.g., TaN) and a Cu seed-layer are sequentially blanket-deposited deposited by PVD. The main Cu film is then blanket deposited by electroplating. The Cu film (and TaN liner layer) are polished off of the top surface of the dielectric film with CMP, leaving the Metal-1 Cu lines inlaid in the grooves. The main challenge here is to obtain the same polishing rate on the tantalum nitride (which is a hard refractory-metal-containing nitride), and the copper (which is comparatively soft). This completes the formation of the Metal-1 and contact plugs by two single-damascene processes. The surface of the wafer is flat for the next dielectric film deposition.

The subsequent layers of metal are formed by a dual-damascene process sequence. The sequence begins by depositing a thin, low-stress silicon nitride layer. This layer forms a top encapsulation layer on the level-1 inlaid Cu lines, and also acts as an etch-stop layer when the vias are etched in the dual-damascene sequence. Next, a dielectric layer (here, undoped SiO_2) is deposited by a CVD process to a thickness required for the via-hole depth. Since the surface is flat, gap-filling by this layer is not required. Then, the embedded etch-stop layer is deposited by CVD (typically PECVD-silicon-nitride with a thickness of 30 nm). Finally, the inter-metal-dielectric stack (IMD) is completed by depositing a second dielectric film, whose thickness is determined by the next-metal-level thickness. (Note that the multi-layer dielectric film, containing two nitride etch-stop layers adds cost and complexity, as well as increasing the interlevel capacitance of the interconnect stack.) Photoresist is applied to the wafer surface. The trench pattern in the resist is formed by exposing the resist through a mask containing this pattern (Fig. 15-5). An etch step is performed next, with the etch creating the trenches in the upper part of the IMD stack. The trench etch is stopped by the silicon-nitride etch-stop layer embedded in the dielectric stack. The resist is then stripped, and a second layer of resist is applied. Another mask (this one containing the via pattern), is used to expose the second resist layer, with the via aligned to the previously etched trench-pattern. The via pattern is then etched in the lower part of the IMD stack. This may be a two-step etch process, as etching the nitride etch-stop layer at the bottom of the via to expose the Cu of Metal-1 may require a separate plasma chemistry than the one used to etch the underlying oxide material. Note that in practice, the trench width may be only slightly larger than the via width.

After the second resist layer is stripped, the formation of the metal line is begun. Again, a TaN (or Ta or TiN) liner and a Cu seed layer are sequentially

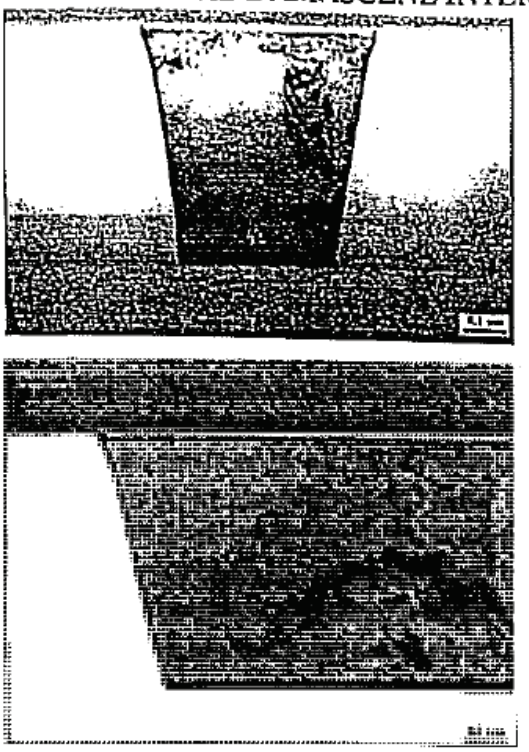


Fig. 15-11 SEM of an in-laid copper line in a dual-damascene interconnect-structure.[34] Reprinted with permission of Semiconductor International.

deposited. This may require an ionized-metal-plasma sputtering process, as the via hole is quite deep and has a high aspect ratio. The main part of the Cu interconnect is then deposited by electroplating. Note that both the via and the trench are filled simultaneously. A metal-CMP process is then used to remove the excess metal. A passivation/etch stop layer (silicon nitride) is deposited onto the top of the layer containing the grooves (trenches + vias) in the dielectric. This completely encapsulates the Cu with diffusion barriers. The entire interconnect structure is finally covered with an SiO_2/SiN passivation layer. Figure 15-11 shows an in-laid Cu-line in a dual-damascene interconnect structure.[34]

15.8 VIA-FIRST DUAL-DAMASCENE PROCESS WITH SiO_2 DIELECTRIC STACK

Many of the issues of forming the vias and trenches in the via-first process are common with the trench-first process and were described above. These include use of a single-damascene process to form the first layer of metal, and technologies for filling the damascene recesses with metal and polishing it back with CMP. We also assume that the via-first dual-damascene structure is being formed on the first level of metal, which has been covered with a SiN layer (that acts both as an etch-stop to the via etch and a barrier layer covering the top surface of the Metal-1 copper layer, to prevent its diffusion into the Si substrate).

However, there are also issues particular to the via-first sequence, and we will discuss them here.

For the purposes of opening this discussion, we assume that the dielectric stack consists of CVD-SiO_2/SiN/CVD-SiO_2. (The implications of using low-k material instead of SiO_2 in the stack were described in Sect 15-6.) The vias in this sequence are, of course, patterned first, using the via mask to align the vias to the Metal-1 lines produced previously (Fig. 15-6). The etch process used to open the via needs to be a multiple-step process because the first etch mechanisms must be *non-selective* (so that they can etch both SiO_2 and the embedded SiN etch-stop layer at about the same rates). However, the nitride capping layer at the bottom of the via should not be removed, and thus the etch mechanisms operating when the vias are close to being opened must be *selective* (to nitride).

After the via s are opened (but with the barrier-SiN layer at the bottom of the vias remains intact), the first layer of resist is stripped. Note that since SiO_2 is used in the dielectric stack, it can withstand the resist-stripping process (i.e., no extra hard mask is needed, as will be the case when most low-k materials are used in the ILD stack). The second layer of resist is then applied, and the trench mask is aligned to the opened vias. Note that the nitride at the bottom of the via must also remain intact throughout second (i.e., trench-etch) process.

Three methods are used to prevent removal of the nitride during trench etch:

1. The resist applied for the purpose of patterning the trenches will flood the open vias. If this resist is not exposed during the trench-patterning exposure step (i.e., because it is so deep beneath the top surface of the resist layer, that some may not develop after the exposure step. Thus, it remains to protect the via bottoms). However, due to pattern-density issues discussed later, such resist remaining in the vias may vary in thickness across the circuit. Thus, this approach will not always work.

2. A trench-etch process is used that is highly selective to the SiN that remains at the bottom of the open vias.

3. A bottom-anti-reflective coating (or BARC) is spun-on to partially (or completely) fill the open vias. This BARC remains after the resist is developed to protect the nitride at the via bottoms against the trench etch.[15] While this approach adds process control difficulties and process complexity (as the BARC must be applied, and separately removed later), this has been the method most widely adopted for preventing nitride-barrier attack during trench etching. We will continue the discussion of this approach in Sect 15.9.

In any case, the trench etch is carried out to create the trenches, stopping on the embedded SiN etch-stop layer. At that point the second resist layer is stripped.

After the second resist layer is stripped, the nitride at the bottom of the via must finally be opened. However, this procedure itself can cause some difficulties. First, the SiN-removal-step should be a "soft" etch. That is, the ions that bombard the surface being etched must not have enough energy to sputter Cu during the inevitable overetch time. If such sputtered Cu ends up on the via sidewalls it could be trapped on the wrong side of the subsequently deposited barrier-layer (exactly where you don't want it!). Another potential problem involves the possibility of slightly misaligning the via and the edge of the Cu film below. In this case a deep slit can be etched in the dielectric under the nitride. This slit would be difficult to fill with the barrier material, again leading to copper-diffusion problems. The way to avoid such slit-formation in the oxide, is to use a nitride-etch-process that is highly selective to oxide.

Another desired goal of the trench etching process is to reduce the handling of wafers between the various steps of the trench etching process (and the variability in waiting time between these steps). That is, the process actually contains four etch steps: 1) organic ARC open (typically required for DUV resists); 2) trench etch; 3) resist strip; and 4) nitride removal from the bottom of the via. Etching tools are now being made available that allow all four of these steps to be carried out in a single chamber, and use of such tools may be critical to achieving consistent device yield and lower overall cost-of-ownership.

15.9 PROCESS INTEGRATION ISSUES OF THE VIA-FIRST DUAL-DAMASCENE PROCESS WITH LOW-*k*-DIELECTRIC STACK

The process flow of forming each level of metal in a via-first dual-damascene interconnect system consists of the following steps:

- DEPOSIT Dielectric Stack (including the SiN barrier-layer over inlaid Cu);
- LITHO Step I (Via Patterning);
- Via ETCH;
- RESIST-STRIP I;
- Post-Via -Etch-Residue CLEAN;
- LITHO Step II (Trench Patterning);
- Trench ETCH;
- RESIST STRIP II;
- Post-Trench-Etch-Residue CLEAN;
- ETCH Barrier Film at Bottom of Via;

- Pre-Metal-Deposition CLEAN;
- Deposit Barrier Metal and "Seed" Cu Layer by PVD;
- Deposit Cu film by ELECTROPLATING;
- Post-Cu-Deposition ANNEAL;
- Remove Excess Cu/Barrier Metal (and do an SiO_2 Buff) by CMP;
- Post-CMP Residue CLEAN;
- CLEAN Top Surface of Cu Before Deposition of SiN Barrier/Etch-Stop Layer.

It can be seen that this process sequence contains: 1) One Dielectric-Film-Stack Deposition sequence; 2) Two Lithography processes; 3) Two Resist-Strip processes; 4) Three Etch processes; 5) Five Cleaning processes; 6) One CMP process; 7) One PVD Deposition sequence; and 8) One Electroplating (and Anneal) process.

Although many of these processes were mentioned earlier, there are some additional important process integration issues that should be discussed when considering the overall "via-first Cu/low-k dual-damascene" process flow. Thus, we will discuss these relevant process integration details here (Fig. 15-12). However, we will cover only those involved with Processes 2 through 5 above (since Process 1 is discussed in both Chap. 14 and in Sect. 15.6, and Process 6 is covered in Chaps. 8 and 16, and Processes 7 and 8 are described in Chap. 16).

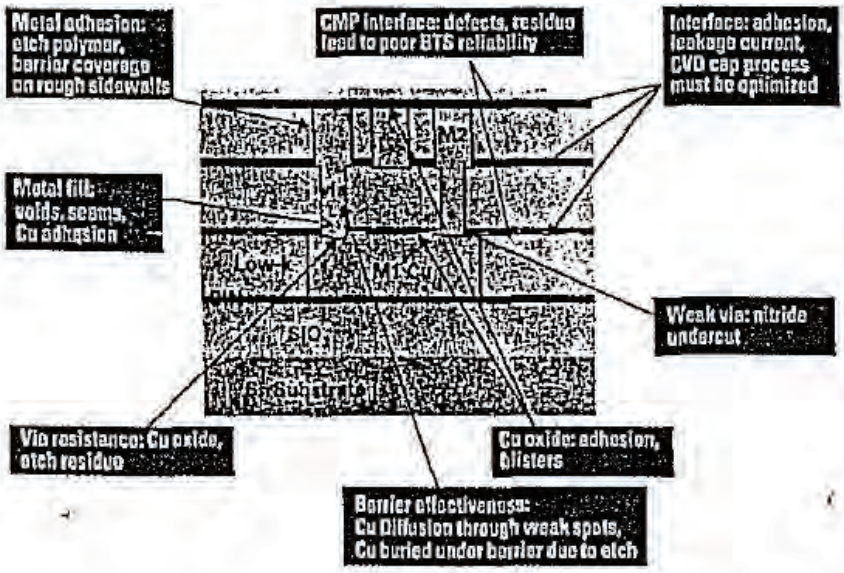


Fig. 15-12 Possible failure types associated with damascene processing pertain to film adhesion, barrier and metal fill capability, copper oxidation and CMP defects.

Besides the technical issues, there are those involving economics. In the face of increasingly difficult integration, the rise in the number of process steps per layer, and the 300-mm conversion, there will be the daunting economic challenges of containing interconnect processing cost, which in 2001 accounts for about 50% of the total wafer processing cost.[16]

15.9.1 Via-First-Damascene Integration Issues: Lithography

With the application of Cu/low-k dual-damascene integration, lithography and etch become more difficult because of the complex dielectric stack and intermediate topography. Here we consider the issues involved in carrying out lithography steps for the leading dual-damascene scheme, namely the *via-first approach*. There are two lithography processes needed in the via first approach, and each as its own challenges. They are: 1) the *via-patterning lithography step*; and 2) the *trench-patterning lithography step*.

Note that the issue of misalignment of the trench pattern to the via pattern is less severe for the via-first approach than for the trench-first approach. That is, as shown in Fig. 15-13b, even if the trench pattern and the via pattern are somewhat misaligned, since the via is already completely formed during the via etch process, it makes complete contact to the metal layer beneath it (provided it has been properly aligned to the underlying metal line). In the trench-first process, however, if the via is misaligned to the trench pattern, during the second etching step (via-etch) the entire via pattern will not be transferred to the bottom of the dielectric stack. This will cause the via resistance between the 2nd and first levels of metal to increase. Nevertheless, alignment of the via pattern to the underlying metal level below remain critical for both the via-first and trench-first processes, for reasons described in Sect. 15.9.2.3 (see also Fig. 15-13a).

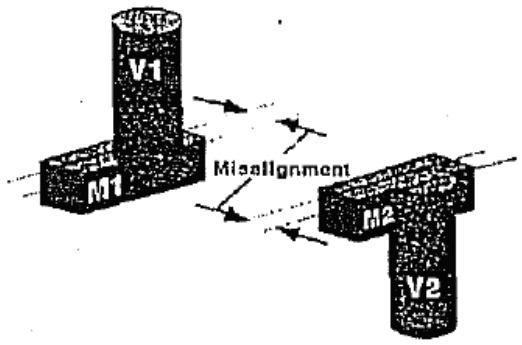


Fig. 15-13 Via-first damascene: a) Misalignment of the via to the underlying metal line is a problem; b) Partial misalignment of the trench to the previously formed via does not cause a problem because the via is completely etched through the ILD stack film.

15.9.1.1 Via-Patterning Lithography Issues: The major concern for the via-patterning lithography step is the ability of the resist to withstand the via etch process. This problem becomes worse as the vias are scaled because thinner resists have to be used to image smaller features, and the etch resistance of 193-nm resists is known to be inferior to that of 248-nm resists. The use of a bilayer resist process (as discussed in Sect. 15.9.1.3 below) may be required, since the imaging resist-layer contains silicon, making it more resistant to dry-etch attack than conventional DUV resists.

15.9.1.2 Trench-Patterning Lithography Issues - DUV Resist Poisoning: As is described in Chap. 6, DUV resists are sensitive to contamination by trace quantities of amines, a problem referred to as *DUV-resist poisoning*. While the most common source of such contamination is amines in the processing ambient, another source is amines in materials that are in direct contact with the resist. Some low-k dielectrics contain such amines (nitrogen compounds). For example, the SiOCH films (formed by PECVD using trimethylsilance and N_2O) entrap NH_3 products in the film as they are deposited. Such amines, if near the surface of the SiCOH films, can diffuse into and poison the DUV resist which is in direct contact with the dielectric surface. If an oxide hard mask is present on the top of the dielectric stack, the resist poisoning problem will not occur during the via-first etch. However, during the trench etch, the resist will come into contact with the exposed sidewalls of the SiOCH film under the hard mask, and resist poisoning can then occur during the trench-etch process. Performing an oxygen-plasma pre-treatment of the via sidewalls (perhaps as a part of the via-first resist stripping process) has been proposed as a way around this problem. Such a pre-treatment can leach out the amine compounds from the SiOCH surface on the sidewalls of the via. However, there is some concern that such a plasma pre-treatment may also take carbon out of SiOCH films and reduce their k-value.

As noted in Sect. 15.9.3, a plasma-treatment of the sidewalls of nanoporous low-k films with an NH_3 plasma has been proposed as a way to strengthen their surfaces against subsequent resist–stripping process of the via-patterning step. However, such a plasma-step can also trap amines in the film. These amines can later diffuse out and poison the resist during the trench-patterning step. (Figure 15-14 shows such a poisoned resist in the via after trench patterning.) Amines can readily diffuse in most low-density and nanoporous films, with diffusion lengths of hundreds of microns. The deposition processes, as well as etch and strip, must therefore be designed to avoid exposing the low-k film to any source of nitrogen. As mentioned above, if an SiO_2-capped low-k film is used, resist

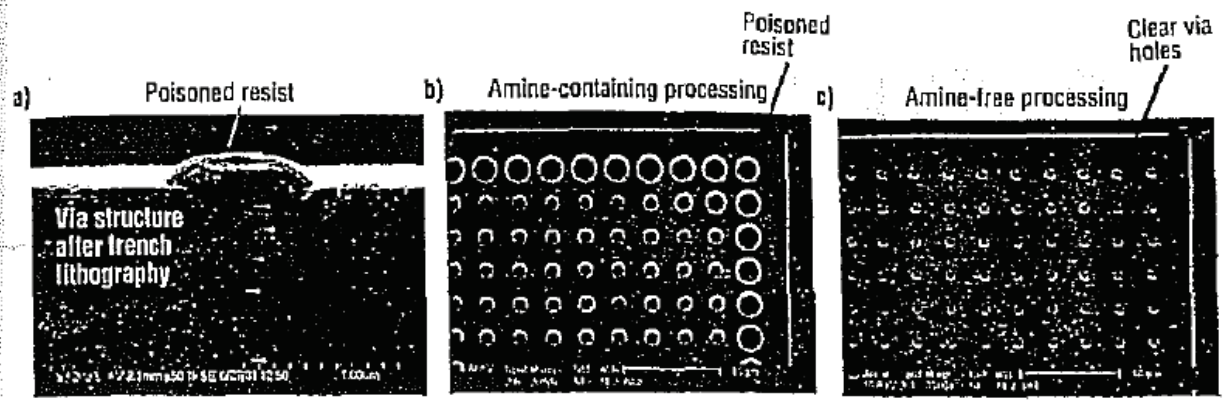


Fig. 15-14 DUV photoresist poisoning is indicated after trench patterning in a) cross-section, and in b) plan view. Poisoning was prevented in c) when the dielectric film was not exposed to nitrogen-containing plasma.[29]

poisoning may not be observed at the via lithography step, but poisoning can occur at the trench lithography step as a result of the trench-patterning resist coming in contact with exposed via sidewalls.[17]

15.9.1.3 Trench-Patterning Lithography Issues - Trench CD Control Issues: To be able to meet the stringent CD-control requirements (i.e., for trench width) for 0.18-μm technologies (and smaller), the resist film spun on a wafer should have the same thickness everywhere (to within ±100Å – see Vol. 1, 2nd Ed. Chap. 12). However, in the via-first damascene approach, since the resist floods the vias that were previously etched, the local thickness of the resist over the top surface of the oxide can be impacted by the presence of these vias. Namely, as shown in Figs. 15-15, and 15-16, the resist is considerably thinner in regions where a dense array of vias exists, compared to regions where only isolated vias exist. In addition, near the edge of such an array of vias, the resist thickness varies significantly, as also shown in these figures. Since higher resolution is generally

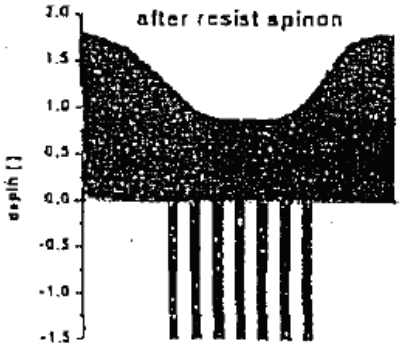


Fig. 15-15 Non-uniformity of resist layer within a chip due to pattern density variations.

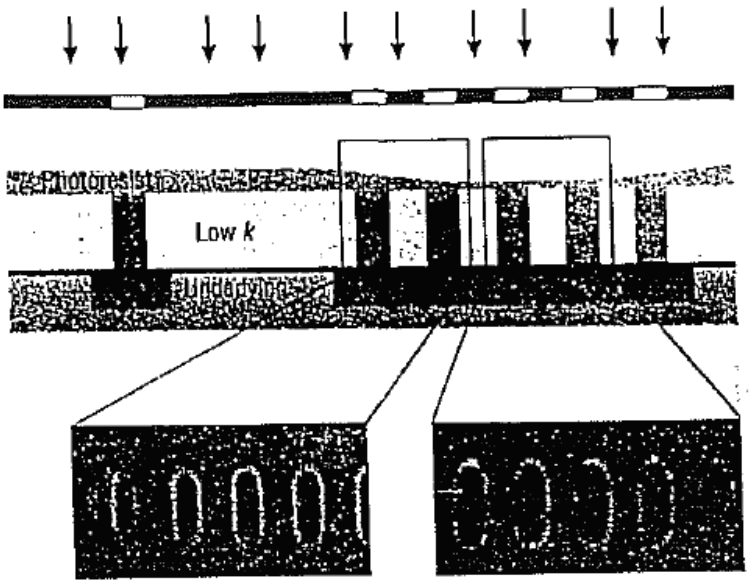


Fig. 15-16 Via-pattern density variations can lead to significant resist thickness non-uniformity during trench patterning in the via-first approach, and subsequent differences in trench widths. a) Trench widths vary at the edge of the via array. b) Trench widths do not vary at the center of the array of visas, where the local resist thickness is uniform.[20]

possible for thinner resist thicknesses, as the resist gets thicker the resolution degrades (e.g., the trench width in this case gets narrower). Figure 15-16 shows that the trench width in trenches patterned at the edge of an array of vias varies, compared to the trench widths patterned over vias at the center of the array. We will see in Sect. 15.9.1.5 that using bi-layer resist process (in which a planarizing underlayer produces more uniform resist thickness over the entire wafer) is a technique being explored to overcome this problem.

15.9.1.4 Trench-Patterning Lithography Issues - BARC Veils: Using a standard DUV lithography process for the trench patterning lithography step, a bottom antireflection coating (BARC) is applied. As noted earlier, the BARC serves two purposes. First, it improves pattern resolution by reducing substrate reflections. Second, the resulting partial fill of the vias with a BARC (Fig. 15-17b)[18] helps protect the SiN barrier layer from being attacked during the trench-etch step. This allows minimization of the barrier layer thickness. The BARC that partially fills the vias remains in place during trench etching and is removed later.

However, it is important to be able to control the depth to which the BARC fills the vias. (The BARC material that fills the via in this procedure is referred to a *via plug*.) That is if the via plug is not recessed below the buried SiN layer in

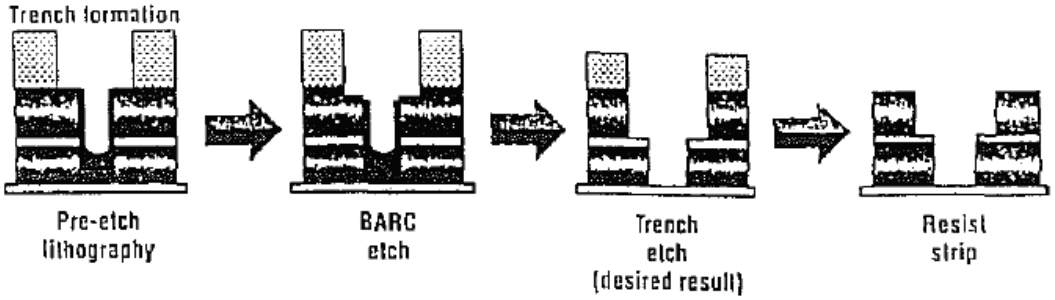


Fig. 15-17 Using a BARC to partially fill the vias prior to trench-PR deposition.[18]

the dielectric stack, "veils" (or "fences') may be created around the via opening at the bottom of the trench.

How such veils may arise is shown in Fig. 15-18. Here the BARC material (via plug) fills the via to a height that is above the top of the buried embedded-etch-stop SiN layer. As the trench is etched, the sidewalls of the via plug above the SiN layer are exposed to the etch ambient that removes the dielectric material to form the trench (Fig. 15-18a). The upwardly-protruding sidewalls of BARC can be chemically converted by an exposure to the plasma etching processes. This altered BARC-sidewall material may not be removable by the same resist stripping step that also removes unexposed BARC material. Thus, it may remain behind in the form of a "veil" surrounding the via opening. Figure 15-18b shows

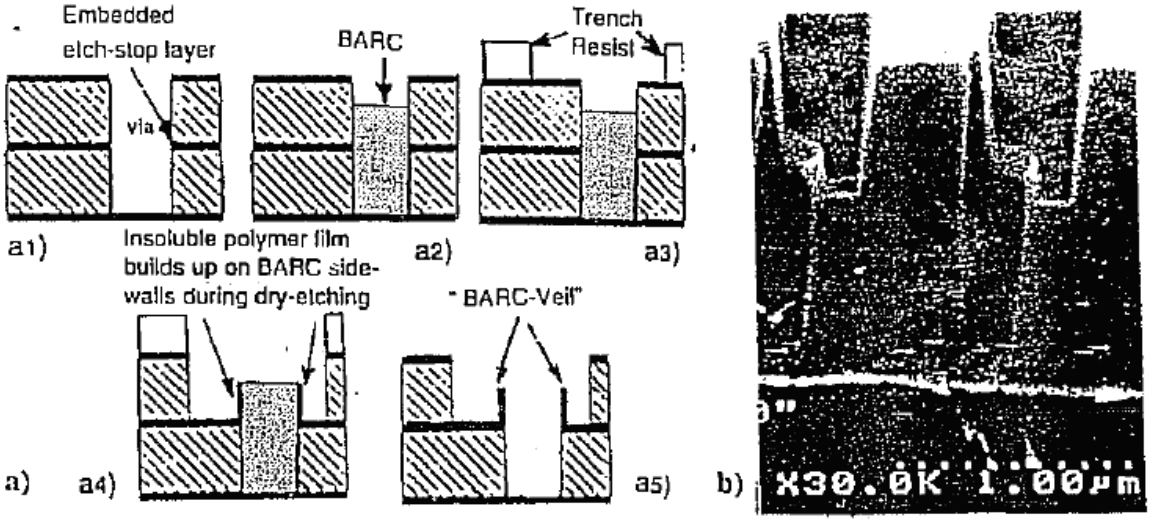


Fig. 15-18 a) BARC-veil formation mechanism in the Via-First Dual-Damascene process: a1) Etch Via, strip resist; a2) Via is filled with BARC to a level greater than that of the embedded etch-stop layer; a3) Deposit and pattern Trench photoresist; a4) Etch Trench. (*Note that during the dielectric etch step, neither the resist nor the BARC is attacked, BUT, an insoluble polymer builds up on the sidewalls of the BARC pillar that sticks up above the etch-stop layer.*); a5) After Trench resist and BARC are stripped, the insoluble film that was created during the dry-etch on the BARC-pillar sidewalls, remains behind as "BARC-Veils." b) SEM of BARC via-veils after trench-patterning and resist-strip.[18]

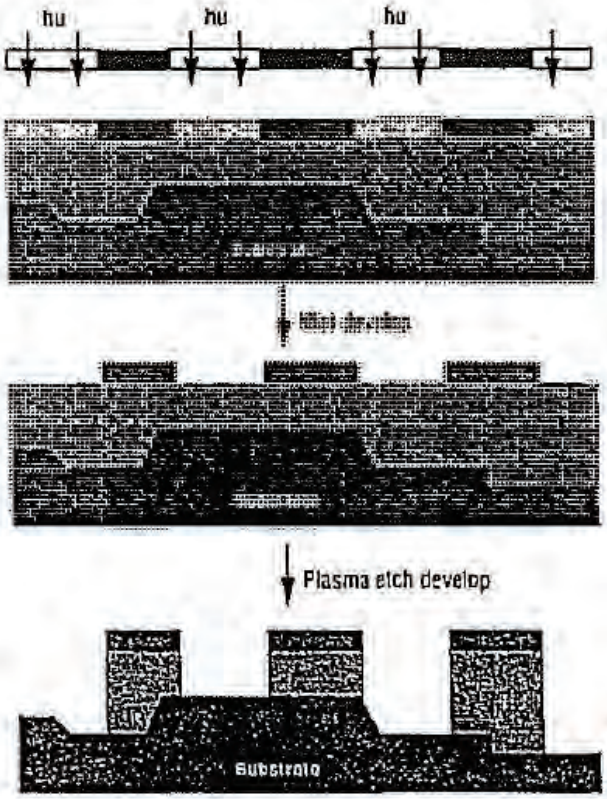


Fig. 15-19 TIS-2000 process flow with a thick planarizing underlayer and a silicon-containing imaging layer.

a SEM of such veils that remain in the trench around the via opening. Filling the trench/via openings with metal in the presence of such veils is problematic, and thus it is important that they be prevented from being allowed to form. To ensure that veils are not created by the trench etch step, it is necessary to control the degree to which the BARC fills the vias.

However, optimizing the BARC-fill of damascene vias is difficult, and often requires considerable process development and materials testing. It is possible to fill vias of different pitch equally by careful selection of the right BARC and coating conditions, but this only works if all the vias have the same size. A new BARC optimization also has to be done for each via diameter.

15.9.1.5 Bi-Layer Resist Process: To overcome the issues of resolution, critical-dimension control over arrays of vias, and the problem of via-plug depth-control that occur when the trenches are pattered in the via-first dual-damascene method, an alternative lithography approach has been proposed, namely a *bilayer, thin surface-imaging scheme* (Fig. 15-19) This method is based on a bilayer resist system developed by Arch Chemicals, called TIS-2000. The technique uses a

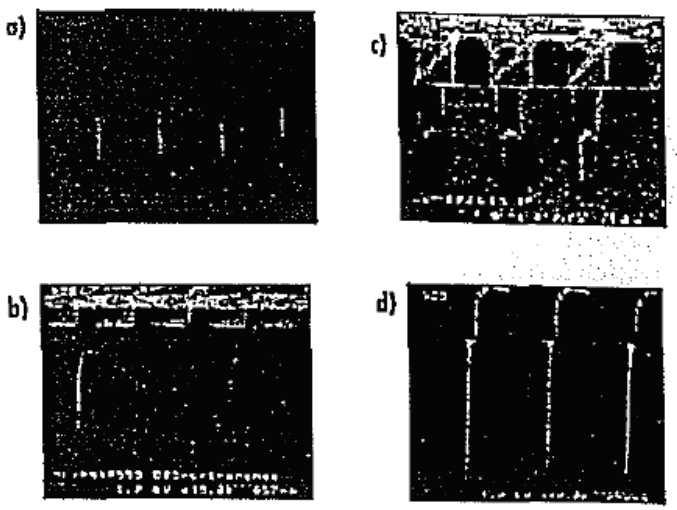


Fig. 15-20 SEMs of cleaved dual-damascene structures patterned with the TIS-2000 process: a) before exposure with vias filled; b) after wet development; c) after partial etch of underlayer and trench; d) after full underlayer etch (top resist still in place).[19]

thin silicon-containing top-imaging layer on top of a highly planarizing fully absorbing underlayer (chemically similar to a novolac resist). The processing sequence is similar to a process with an organic BARC. While this technique was developed for 248-nm lithography, it appears that it can be extended to 193-nm lithography (by using an alternate imaging layer and underlayer).

The thin imaging-layer on a planarized layer allows for very high resolution. The silicon present in the imaging-resist-layer also gives it a very-high etch resistance when the underlayer is dry-etched (developed). The underlayer effectively planarizes the topography of the via-containing dielectric stack (Fig. 15-20a).[19] In this figure (in which the underlayer fills 0.25-μm vias), the vias are completely filled and the top of the underlayer is flat without any topography caused by the vias underneath. The thickness of the underlayer over the substrate in the via-array region is about 350-nm, which is enough to insulate the imaging layer from any optical effects by substrate reflections (i.e., the underlayer also acts as an effective BARC layer). Figure 15-20b shows an array of vias after the *imaging layer* has been exposed and wet developed.

The next step is a dry-develop process done with an SO_2/O_2 plasma. This transfers the pattern created in the imaging-resist-layer to the underlayer. Dry-developing the underlayer to the dielectric surface below defines the trench pattern. At this point the vias are still completely filled with the underlayer material. If we stopped the develop step at this point, and proceeded with the etch step to remove the dielectric material which creates the trench, veils would again arise (in the same way they occur if the via plugs made from BARC material

filled the vias too full). Thus, the dry-develop step is extended so via-plugs (now consisting of underlayer material instead of BARC) are recessed. (Since trench open-areas are large compared to via areas, an optical endpoint signal can be used to determine when the recess etching begins, providing good control.)

Figure 15-21 shows the patterned trench-resist and the recessed via-plug (whose level is below the top of the buried-nitride etch-stop layer). When the trench dielectric is now etched, no veils remain as residue within the trenches. Figure 15-22 shows the trench and via after trench-etch and resist-stripping. The trench and via profiles are vertical and veils are not present. Since the via-plug covers the bottom SiN barrier-layer during the trench etch, the trench-etch process does not have to be made selective to the nitride. This also allows a trench etch process with higher etch rates to be used.

Comparison of trench lithography at the edge of large via arrays shows another advantage of the top-surface bilayer-imaging technique. As is shown in Fig. 15-23, while trenches over vias at the edge of dense array vary considerably in width for standard DUV lithography (Fig. 15-23a), those patterned with TIS-2000 show no variation in width.[20] This is because the underlayer of the TIS completely fills the vias. Thus, the resist thickness is constant and there is little (or no) trench-width variation.

15.9.2 Via-First-Damascene Integration Issues: Dielectric Etching

The plasma etching and resist/polymer strip steps are among the most critical for the successful integration of a *low*-k dielectric. The etch process must be able to

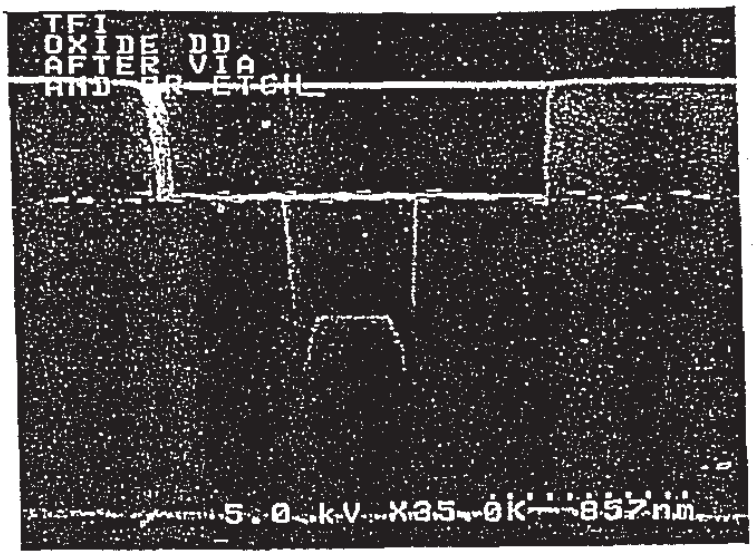


Fig. 15-21 Trench patterning using TIS-2000 bilayer resist. The dry-developed underlayer with recess etch brings the level of the via plug along the sidewalls below the nitride etch stop, which is important for the subsequent trench etch.[18]

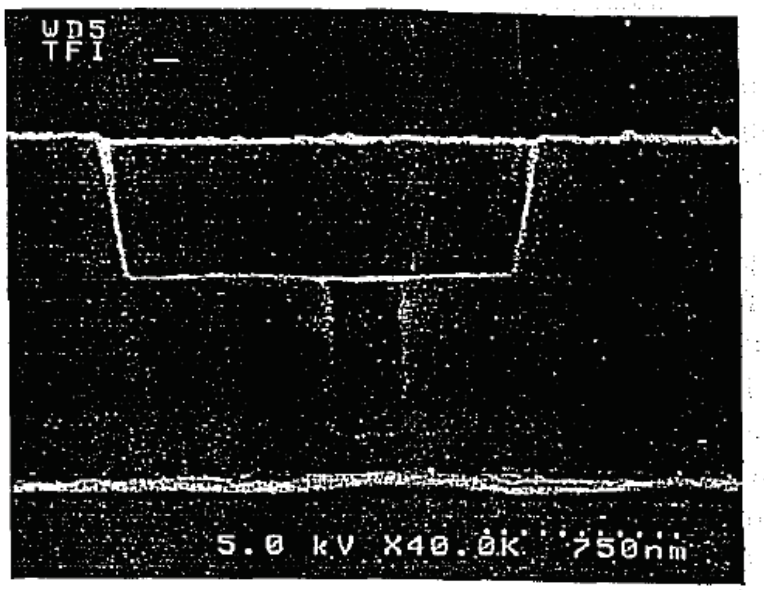


Fig. 15-22 Fully etched and stripped trenches over a via using the TIS-2000 bilayer resist process. The dual-damascene structure is nearly ideal in that there are no veils or faceting of the via due to trench etch.[18]

effectively perform the following tasks: 1) Remove multi-layers of dielectric; 2) Maintain a notch-free profile for optimum subsequent metal deposition; 3) Achieve a high selectivity to resists (which is especially important as resists are getting thinner and 193-nm resists are less resistant to plasma etching than those used for 248-nm lithography); 4) Retain the low-k value during the etch and strip processes so that ultimate device speed is not compromised; 5) Achieve high selectivity to etch-stop layers, (especially since we want to make them as thin as possible to minimize the stack's effective k-value); 6) Minimize microtrenching at the etch front; and 7) Minimize aspect-ratio-dependent etch (ARDE) effects.

15.9.2.1 Via Etching for Low-k Processes: In the "via-first" dual-damascene process sequence, the vias are patterned and etched before the trenches. The via-

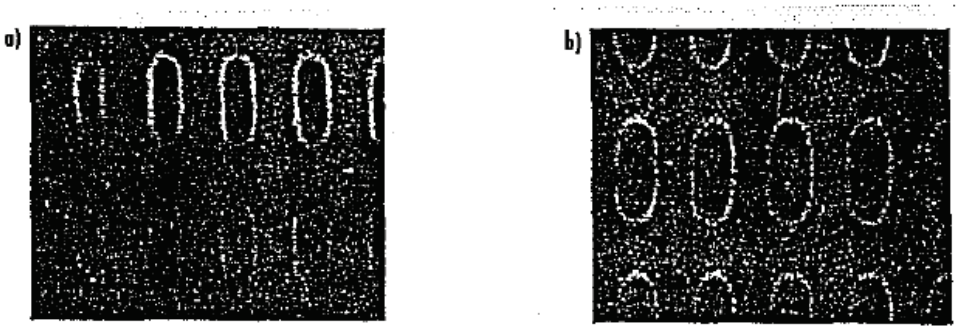


Fig. 15-23 Trenches over vias near the edge of a dense array: a) trench widths vary considerably for standard DUV lithography; b) the same area with TIS-2000 resist.[18]

etch process has some challenges, one of which was described in Sect. 15.9.1 (the resist must be able to withstand the deep via-etch process). To open the via, the following materials must be etched by the via-etch process in a low-k dielectric stack: 1) BARC; 2) SiO_2-hard-mask; 3) upper low-k dielectric layer; 4) thin SiN (or SiC) embedded etch-stop layer; and 5) lower low-k dielectric layer. The etch step itself will therefore have to employ a multiple sub-steps (Fig. 15-17a).[18] The first such sub-step should be a non-selective etch, so that the BARC, SiO_2 hard-mask layer, low-k dielectric layers, and the embedded-etch-stop layers are all etched at about the same rate. This would ideally produce a via profile that is free of notching and bowing. This first sub-step would be terminated prior to removing all of the second low-k dielectric layer. The etching sequence would then switch to the second sub-step. The etch characteristic of this second-sub-step would be that it is highly selective to etching oxide and not nitride. This would allow it to remove the remaining lower low-k dielectric material from all of the vias without attacking the SiN etch-barrier layer at the bottom of the vias (i.e., a sufficiently long overetch time would be permitted to ensure that the lower low-k material was entirely removed from all of the vias).

While the above process is the desired goal, an even more complex etch process might be needed to produce a via-profile free of notching and bowing. That is, the etching characteristics of various low-k dielectrics are different from SiO_2, regardless of whether a spin-on organic-based material (such as SiLK® or FLARE®), or an SiCOH-CVD dielectric (such as Back Diamond,® Coral,® or Flowfill®) is used.

For example, a three-sub-step etch process might be needed for a SiO_2-hard-mask/SiLK stack. That is, fluorocarbon chemistries with high bias power might be used for the first-sub-step to etch the top SiO_2-hard-mask material and the embedded SiN layer (because this process can disrupt the strong Si-O and Si-N bonds), but when the SiLK layers are etched, a second-sub-step using oxygen as the active etchant is used. However, in an HDP tool, the high degree of gas-phase dissociation leads to high concentrations of atomic oxygen, which may produce some bowing of the via profile. Also, hard-mask undercut after via (and trench) etching can occur (Fig. 15-24), which would make it more difficult to achieve void-free metallization.[21]

When the low-k dielectric is a SiCOH-CVD film (having a higher porosity than normal CVD-SiO_2 films), the gas mixture for etching these low-k materials might be mixtures of CF_4 (the principal source of F atoms) and CH_2F_2 (a polymerizing gas for improving selectivity and providing sidewall passivation). Using such an etch procedure has reportedly allowed 0.3-μm vias to be etched in

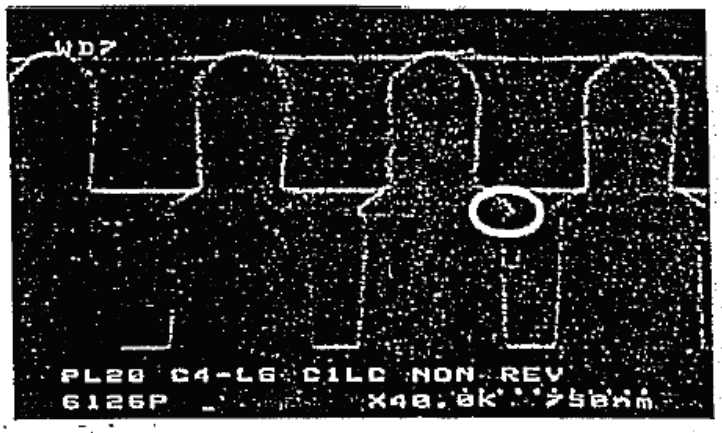


Fig. 15-24 Etched via profiles in SiLK® low-k dielectric showing hard-mask undercut.[21]

such stacks with the desired via profiles (Fig. 15-25).[22] The choice of gas mixture depends on the precise application being considered. That is, since these films contain some C and H, if their composition is changed (i.e., the more C, the lower the k-value), it will be necessary to re-tune the etch process in order to meet the selectivity and etch requirements.

15.9.2.2 Trench Etching: Following the via-etch-step in via-first, dual-damascene

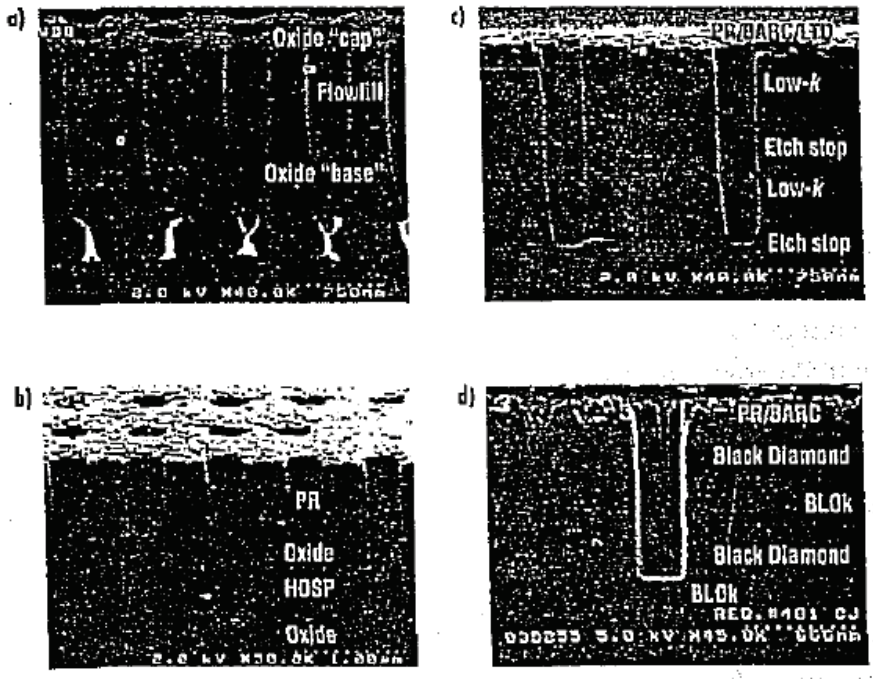


Fig. 15-25 Etching of: a) 0.3-μm vias through Trikon's low-k Flowfill (k = 2.9); b) 0.4-μm vias through Honeywells HOSP (k ~2.5); c) 0.28-μm vias through a Trikon dual-damascene stack (k ~2.8); and d) 0.29-μm vias through a dual-damascene stack containing Applied Materials Black Diamond film (k ~2.6) and BLOk (k ~4.5) etch stop.[22]

applications, trenches are etched down to the buried etch-stop layer. The issues involving protecting the barrier nitride at the bottom of the via (exposed during the previous via-etch step), from being etched during the trench-etch-step were previously considered in Sect. 15.9.1.

When a 2nd generation low-*k* dielectric is used, the top surface of the stack generally employs an SiO_2 hard-mask and a BARC to improve lithography resolution. Thus, the trench-etch process is also a multi-step procedure, involving: BARC-open; SiO_2-hard-mask etch; and low-*k* etch (down to the embedded etch-stop layer - Fig. 15-17b). Following the trench etch, the resist must be stripped, the BARC removed from the vias, and the barrier layer at the bottom of the vias etched away.

In early via-first, dual-damascene process flows, these steps were performed in several discrete tools. However, in the interests of improving cycle time (less handling and waiting), tool vendors are seeking to provide equipment that will allow an integrated-trench-etch process. That is, BARC opening, SiO_2-hard-mask etch, low-*k* dielectric etch, resist/BARC strip, and SiN barrier-layer open would all be carried out *in situ* in a single tool and a single process run.

15.9.2.3 Etching the SiN Barrier Layer at the Bottom of Vias: After the trench is completed and the second resist layer stripped, the nitride at the bottom of the via must be opened. As noted earlier, this etch process can cause some problems. First, the SiN etch step should be a "soft" etch. That is, the ions that bombard the SiN being etched should not have enough energy to sputter Cu during the inevitable overetch time. If such sputtering causes Cu to end up on the via sidewalls, it could be trapped on the wrong side of the subsequently deposited Cu-barrier-layer. Thus, any such sputtered Cu would have to be removed before the barrier metal was deposited. While a conventional dip in dilute HF is efficient for sidewall Cu contamination cleaning, this inevitably also cause some CD loss.

Another problem arises if the via is misaligned with the edge of the Cu line beneath it. In that case, during any overetch time of the nitride, a narrow slit would be etched in the dielectric alongside the Cu line (see Fig. 15-26).[31] This slit is hard to fill with barrier material, again leading to unwanted Cu diffusion. This issue can be addressed by employing a nitride etch process that is highly selective between nitride and oxide, and between nitride and Cu.

15.9.3 Via-First-Damascene Integration Issues: Resist Stripping

With the introduction of low-*k* and ultra-low-*k* porous dielectrics, existing resist-strip technology cannot be used without at least some modifications. Dry-etching in an oxygen plasma has become the most widely used stripping technology

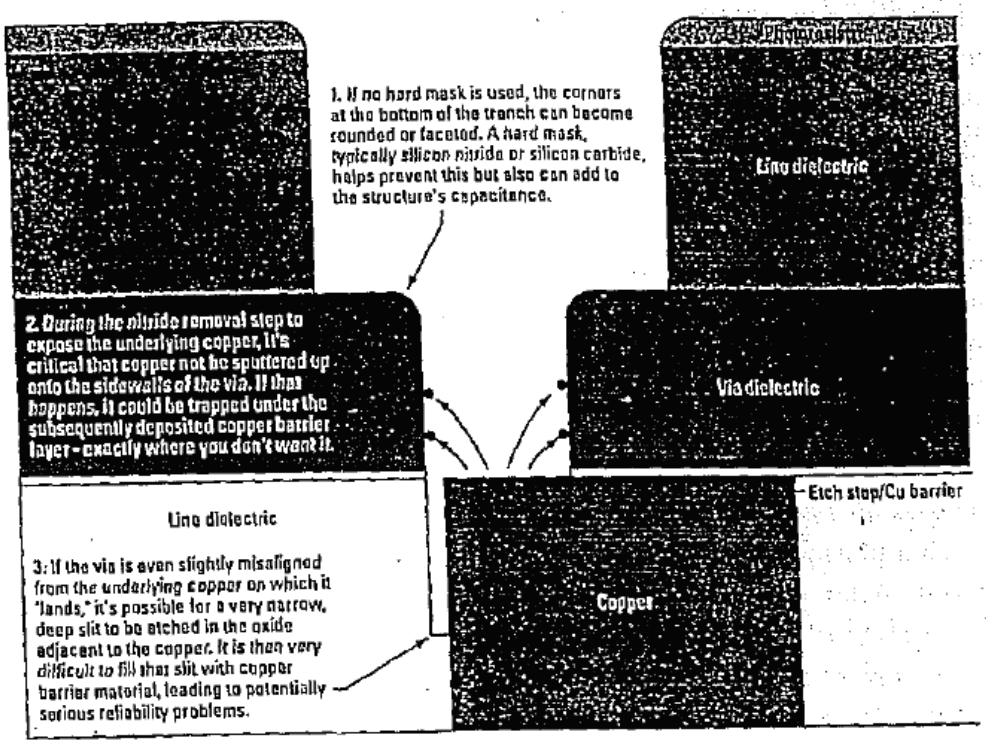


Fig. 15-26 Common problems in the via-first approach include rounding of the corners at the bottom of the trench; sputtering of the Cu onto the via sidewalls during nitride removal; and hard-to-fill slits etched next to the underlying Cu in the case of misalignment.[31]

(especially for back-end-of-line [BEOL] processes). However, many of the low-k dielectric materials are also etched by oxygen plasmas, and in some cases their k value may be degraded by reactive oxygen atoms as well. Here we discuss the issues facing resist stripping in such applications.

To begin, as noted earlier, SiO_2-hard-masks will typically be used on the top of the low-k dielectric stacks. These layers will actually serve multiple functions, including: a) as a hard-mask during via- and trench-etching; b) as a polish-stop layer for the Cu-CMP process; and c) as a dielectric anti-reflective coating (DARC). The use of a hard mask is mandatory if organic low-k films are used (since they are attacked by the same chemistries used when stripping the resists after an etch step). Even with a hard mask, however, difficulties arise when etching organic low-k materials. That is, if an etch-process is used that is based on an O_2 plasma, very reactive oxygen radicals are present in the plasma. These isotropically attack organic low-k dielectrics, causing bowing of the via and/or trench profiles (as shown in Fig. 15-27).[23] To stop such attack of the dielectric layer, sidewall-passivating gases must added to the O_2 etch gas. In one paper,

ethylene was added to the O_2 etch gas when an organic low-k material was being etched. Undercutting of the hardmask was prevented, and good CD control was maintained, as the hardmask itself was not eroded.[24]

Even if an HSQ or MSQ-based low-k material is used instead of an organic film (see Chap. 14), the resist stripping process is still a challenge. That is, while the resist is being stripped, the oxygen radicals will react with the exposed low-k material of the trench sidewalls. These reactions convert the Si-CH_3 bonds in the MSQ, and the Si-H bonds in HSQ films, back to Si-O bonds. This causes the k-value of the dielectric to rise back toward that of SiO_2 (~4.0). As a result, an alternative resist stripping process is needed. That is, an alternative anisotropic ashing process must be available.

One method of making low-k materials more robust against attack by oxygen radicals is *plasma treatment of the exposed low-k material surfaces.* An example of such a plasma treatment was reported to have carried out on an organic low-k material. After exposing this dielectric to an NH_3 plasma for 10 minutes, a nitrided layer is formed on its surface. The nitrided low-k layer exhibited lower leakage current after an ashing plasma exposure than did the untreated material. The film also acted as a better barrier against Cu diffusion.[25] In another paper, three such NH_3-plasma treatments in were performed on a dual-damascene via-first etched dielectric stack consisting of HSQ/PECVD-SiO_2/HSQ. The first treatment was carried out prior to depositing the embedded etch-stop PECVD-SiO_2 layer on the first HSQ segment of the stack. This improved adhesion of the embedded etch-stop layer to the HSQ. The second treatment was carried out after the second HSQ layer was deposited and the trench etch was completed. This purpose of this treatment was to strengthen the exposed HSQ sidewall surfaces against erosion during the resist stripping step.[26]

Carbon-doped low-k materials such as Coral® and Black Diamond® are ozone

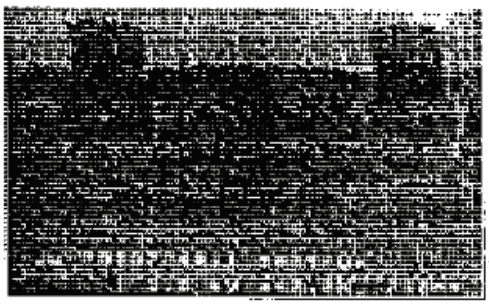


Fig. 15-27 Cross-section SEM picture of single damascene structure made in 3M's low-k dielectric after strip and etch in a low-pressure, low-temperature strip.[23]

resistant. Resist stripping in processes that use these dielectric materials might potentially be carried out with a DI ozone strip process.

15.9.4 Via-First-Damascene Integration Issues: Cleaning

Cleaning issues are critical to producing dual-damascene interconnects with high yield and reliability. The cleaning challenges are especially severe for dual-damascene structures compared to front-end-of-line (FEOL) processes, because more aggressive chemicals can be used on silicon and thermal oxides. That is, the strong acids and bases in the FEOL flow effectively remove contamination and resist residues after critical FEOL steps. However, metals in back-end-of-line (BEOL) cannot be subjected to such chemicals since they are susceptible to chemical corrosion. Fortunately, the CMP process used in dual-damascene interconnects tends to remove particles on the dielectric surface during polish, another inherent advantage over the subtractive interconnect approach. However, other residual films must still be removed, as will be discussed here.

15.9.4.1 Post-Dielectric-Etch-and-Resist-Strip Residue Removal: After photoresist is stripped with a dry-stripping step (ashing), residue from the etch and the ashing step often remains on the via and trench sidewalls, and must be removed. Typically this is done with wet-chemical residue cleaners, in a separate cleaning step after the dry-strip step. One goal (as a way to improve cycle time), is to integrate these two steps into a single tool.

As long as SiO_2 was the material used in the dielectric stack, the post-etch residue removal step was easier, and proprietary, organic solvents sold by such vendors as EKC Technologies, ASCI, J.T. Baker, and Olin could effectively remove the residues (Fig 15-28). However, with the introduction of low-k dielec-

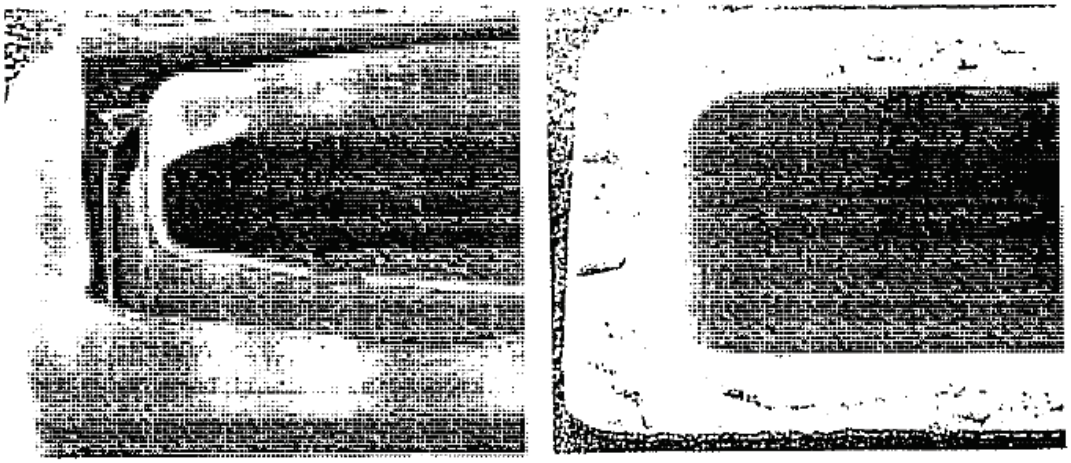


Fig. 15-28 Aqueous residue removers specifically target organometallic and inorganic residues left after ashing. Courtesy of J.T. Baker.

trics, the situation has changed. New cleaning chemicals must be used, and suppliers are working on chemistries for residue cleaning in Cu/low-k applications. However, a major issue in pursuing this goal is that it has not yet been decided exactly which low-k materials will be used for this application, and what are all of the material properties of the candidate materials. In addition, it appears that there will be a continuous transition from one material to another over the next few years, as ever lower-k materials are integrated into damascene structures. However, some general trends can be foretold.[27]

First, it must be ensured that residue-removal chemistries are compatible with specific the low-k dielectric and the metallurgy on the wafer. This is complicated by the lowering of the porosity of the new low-k materials and the thinning of the dielectric layers. Furthermore, since each low-k choice requires different plasma etching processes and chemistries to etch vias and trenches into the low-k materials, the nature of the residues can be different. Thus, no one product will likely to be able to meet all Cu/low-k applications. Nevertheless, it appears that some form of wet cleaning after etch and ashing will still be needed until below 0.13 μm. Fundamental barriers to wet cleaning approaches do not appear to be reached until well below 0.1 μm. At some point, however, there are concerns about liquids being able to go into and out of small, high-aspect-ratio features. At that time a dry-approach (or the use of supercritical fluids – see Sect. 15.9.4.5) may need to be used.

15.9.4.2 Pre-Metal-Deposition Clean: The pre-metal-deposition clean is carried out just before the barrier layer metal and Cu seed layers are deposited. The Cu film at the bottom of the vias must be free of organic residues and any native Cu oxide in order to ensure that low-resistance contact between the deposited barrier metal and the underlying Cu is fabricated with high yield. This clean step normally involves sputter etching of the wafer in the same tool in which the barrier layer metal and the seed Cu layer are sputter-deposited. Argon ion gas is used, and a dual-frequency rf sputtering module with a high-density plasma (HDP) source is required (so that a high degree of directionality can be obtained with the sputter-clean process).

During this pre-metal-deposition cleaning process, copper ejected from the lower layer can be re-deposited onto the via sidewalls, increasing the risk of line-to-line leakage. By 1999, however, the magnitude of this problem had not yet been quantified. It is hoped that the quantity of copper re-deposited will not be sufficient to degrade device performance. Nevertheless, if the problem becomes worse than expected, alternate via-clean methods will need to be developed that

selectively remove the etch residue and copper oxides without removing any copper.

Another special problem occurs when trying to integrate Cu with a fluorine-containing dielectric material, such as FSG. That is, the F bonds in many FSG films are reactive (particularly with water), and they can be broken during other process steps to produce free fluorine. At the interface between the FSG film and the barrier metal, Ta can react with such free fluorine to form volatile by-products at elevated temperatures. The resulting weakened interface can fail when subjected to the stress of an overlayer (such as that produced by the SiN diffusion barrier layer). A number of solutions to this problem have been proposed, including: use of a TaN barrier material (which is less reactive than Ta); using a more stable FSG film; or using a hydrogen-plasma treatment before depositing the Ta to remove mobile fluorine from the surface.

15.9.4.3 Post-CMP Clean: The post-CMP clean step is a critical one for Cu CMP. The bevel and backside of the wafer will most likely be exposed copper-containing solutions during the CMP step. Any copper remaining on the bevel or backside of the wafer risks contaminating other tools within the fab. As in the electroplating process, simple cleaning methods are available to remove residual copper, but the risks associated with copper cross-contamination demand careful monitoring of the post-CMP clean (see also Chap. 8).[28]

15.9.4.4 Clean of Top-Surface of Cu Prior to Diffusion-Barrier-SiN-Deposition Process: After the CMP step of the metal, and post-CMP clean, a SiN layer is deposited on top of the damascene structure. As noted earlier, this SiN film serves as barrier to keep the Cu from diffusing into the dielectric layer above it, and also as an etch-stop layer for the via etching step. However, the SiN may fail to adhere to the Cu surface if that surface is not properly cleaned. That is, the SiN may not adhere to the Cu interface if even a small amount of residue is left on the Cu surface following CMP. It turns out that there are two classes of such contaminants that must be removed from the Cu-surface: 1) residual organics that are present in the Cu-CMP slurry; and 2) native oxide that has formed on the Cu surface.

The residual organics can be removed by a pre-treatment with an *in situ* exposure to an aggressive hydrogen plasma just before depositing the SiN layer. Another technique is to remove such organics (corrosion inhibitors, usually BTA) from the CMP process. In addition, Cu forms a native oxide that can interfere with SiN adhesion. That is, if this oxide is not completely removed from the Cu surface, the Cu/SiN interface will be weak. Thus, a cleaning step

(consisting of a copper oxide reduction step using the same hydrogen plasma used to remove the residual organics) is recommended.[29]

15.9.4.5 Supercritical Fluids as Wafer Cleaners:* The temperature at which the vapor pressure above a pure liquid reaches one atmosphere is known as the normal boiling point – for water it is 100°C. In an open container, liquid water's temperature cannot be raised above 100°C because it would cause the vapor pressure to rise above one atmosphere, exceeding ambient pressure conditions. However, in a sealed container, liquid water can be heated to higher temperatures. As the water is heated, its density decreases (through thermal expansion), and simultaneously, the vapor density in the container increases.

This heating can be continued until the liquid's density has been so reduced – and that of the vapor so increased – that both densities are equal. This point is known as the *critical temperature*, and the pressure of the vapor at that point is called its *critical pressure*. Since the temperature and density of the water inside the sealed container are equal everywhere, thermodynamics dictates that the critical pressure inside the container also be equal everywhere. A fluid brought to conditions above its critical temperature and pressure (the *critical point*) is called a *super-critical fluid.*

Carbon dioxide (CO_2) is a material that is particularly attractive as a practically useful super-critical fluid. This is because it is non-toxic, non-flammable, and inexpensive. In addition, its critical conditions are easily achievable with existing process equipment. This unique set of characteristics has led workers to examine the possibility of using supercritical CO_2 instead of other solvents in such semiconductor processes as cleaning and photoresist stripping.

The attractiveness of using supercritical CO_2 as a replacement for traditional solvents in cleaning and stripping applications stems from the combination of gas- and liquid-like properties that it possesses. Table 15-1 compares the physico-chemical properties of a typical organic fluid in it's liquid, supercritical, and gaseous states. It can be seen that the supercritical fluid state possesses *gas-*

Table 15-1 PHYSICO-CHEMICAL PROPERTIES of TYPICAL ORGANIC FLUIDS

	Diffusivity (cm^2/sec)	Viscosity (mPa-sec)	Density (kg/m^3)
LIQUID	10^{-5}	1	1000
SUPERCRITICAL FLUID	10^{-3}	10^{-2}	700
GAS	10^{-1}	10^{-2}	1

* Written by James Rubin, Los Alamos National Laboratories. The material originally appeared in Semiconductor International. Reprinted with permission

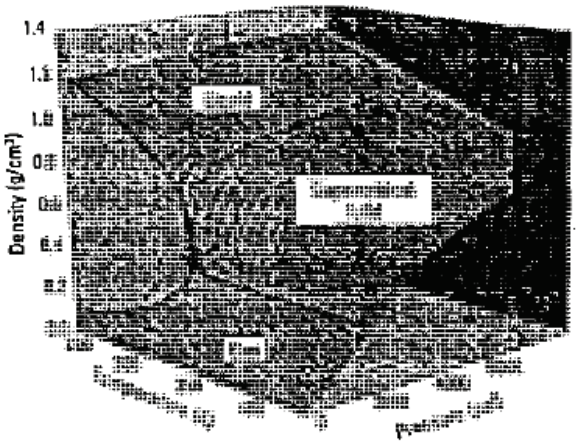


Fig. 15-29 Pressure-temperature-density surface for pure CO_2. The critical point is the small circle at the intersection of the contour lines.

like values of viscosity and diffusivity, and *liquid-like* density.

Figure 15-29 shows the density-pressure-temperature surface for pure CO_2. The critical point for pure CO_2 is 31°C (304K) and 1072 psi. To a first approximation, a fluid's solvent power is proportional to its density. Thus, high liquid-like densities achievable in supercritical fluids allow for liquid-equivalent solubilities. Figure 15-29 also indicates that small changes in pressure or temperature near the critical point result in large density changes. This implies that the density – and hence the solvent power – of a supercritical fluid is "tunable."

Figure 15-30 is a schematic pressure-temperature phase diagram for CO_2, incorporating the process flow for a closed-loop supercritical fluid treatment system. No uncontrolled waste streams can exit such a closed-loop system. All materials solubilized in the supercritical fluid are retained in the separation vessel

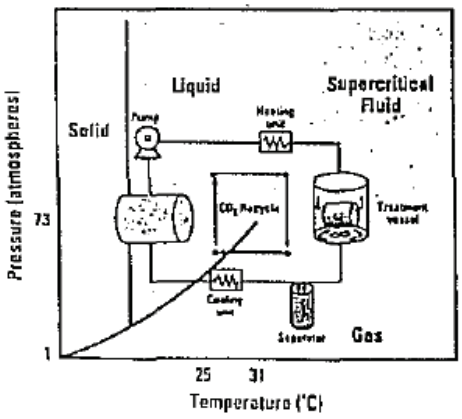


Fig. 15-30 Pressure-temperature phase diagram of pure CO_2 with a superimposed flow diagram for a closed-loop supercritical fluid treatment process.

for analysis, recycle, treatment, and/or disposal. In addition, these materials are concentrated in the separator, greatly reducing waste volume.

Several companies are working on commercializing tools that employ supercritical fluids in cleaning processes, including Supercritical Systems (a TEL-America company), and GT Equipment Technologies.

15.10 VIA-FIRST DUAL-DAMASCENE DIELECTRIC STACK WITH NO EMBEDDED ETCH-STOP LAYER

Eventually, the increase in the k-value of the damascene dielectric stack due to the embedded etch-stop layer (whether it is composed of SiN or SiC material) becomes too great. At that point, it may be necessary to use a damascene dielectric stack without an embedded etch-stop layer. Some predict that this approach will be needed for the 0.1-μm technology node.

The elimination of the etch-stop layer, however, produces some significant process integration issues. First, since there is no etch-stop layer, etch systems will have to have extremely good control of uniformity to ensure all trenches are of the same depth across the wafer, and from wafer-to-wafer. This could be difficult as there is no endpoint indicator, and the depth of the trench can only determined by the time of the etch step. Making the problem even more difficult is that low-k materials may not lend themselves to that kind of control. Second, if a hard etch-stop layer is not present, the top corner of the via can become rounded during the trench etch process (due to faceting, see Vol. 1, Chap. 11 - Physical Vapor Deposition, and also Fig. 15-26). A second issue involves the creation of microtrenches at the corners of the trenches themselves. This phenomenon is likely to be caused by enhanced ion-bombardment-flux at the corners of the trench (due to ions reflecting at shallow angles off of the sidewalls). Such microtrenches would provide locations that are poorly covered by the Cu diffusion barrier films. This could give rise to Cu diffusion through the barrier layer at such points where the barrier layer is thinned (or is even discontinuous).

15.11 PROCESS INTEGRATION ISSUES OF DUAL-DAMASCENE DIELECTRIC STACK WITH ULTRALOW-k DIELECTRICS

If damascene dielectric stacks are fabricated with ultra-low-k materials, additional process integration problems are encountered. That is, special techniques must be developed to protect these porous, less-robust materials during ashing and CMP. Two novel damascene processing steps have been proposed to handle such fragile films. The first is termed *half-via etch*. In this approach, the via is again etched first, but only half of the via is opened with this

process, down to the depth of the trench. Then the rest of the via is opened during the trench-etch step. This provides a simpler etch and alignment process and makes the lithography problem of resolving small feature sizes easier.

Another change may involve returning to the trench-first etching sequence. With a porous material, the via cannot be etched first because, after the via is etched and you begin etching the trench, problems will arise because the resist that flooded the vias will begin to be absorbed by the porous low-*k* dielectric.

REFERENCES

1. P. Singer, "Making the Move to Dual Damascene Processing," *Semiconductor International*, August 1997, p. 79.
2. B. Luther *et al.*, "Planar Copper-Polyimide Back End of the Line Interconnections for ULSI Devices," *Proceedings of the 1993 VMIC Conf.*, p. 15.
3. R. Iggulden *et al.*, "Dual-Damascene Aluminum for 1-Gbit DRAMs," *Solid State Tech.*, November 1998, p. 37.
4. P. Singer, "IBM to Use Dow's SiLK in Low-*k* Process," *Semiconductor International*, May 2000, p. 44.
5. H. Aoki *et al.*, "Robust 130-nm-Node Cu Dual Damascene Technology with Low-*k* Barrier SiCN," *Tech. Dig. IEDM* 2001, p. 76.
6. "New Barrier/Etch-Stop for Copper Damascene," *Semiconductor International*, July 1999, p. 28.
7. S.-M. Cho *et al.*, "Low Dielectric Constant Material Deposited by HDPCVD for Barrier and Etch Stop Applications in Cu Damascene Structure," *Abs. Mtg. Electrochem. Soc.*, Spring 2000, Abs. 231.
8. X.W. Lin and D. Pramanik, "Future Interconnect Technologies and Copper Metallization," *Solid State Technology*, October 1998, p. 63.
9. B. Zhao *et al.*, "A Cu/Low-*k* Dual Damascene Interconnect for High Performance and Low Cost Integrated Circuits," *Dig. of Tech. Papers 1998 Symp. on VLSI Technology*, p. 23.
10. Y. Morand *et al.*, "Copper Integration in Self Aligned Dual Damascene Architecture," *Dig. of Tech. Papers 1997 Symp. on VLSI Technology*, p. 31.
11. S. Venkatesan *et al.*, "A High Performance 1.8V, 0.2 μm CMOS Technology with Copper Metallization," *Tech. Dig. IEDM*, 1998, p. 769.
12. D. Edelstein *et al.*, "Full Copper Wiring in a Sub-0.25 μm CMOS ULSI Technology," *Tech. Dig. IEDM*, 1998, p. 773.
13. M. Igarashi *et al.*, "The Best Combination of Aluminum and Copper Interconnects for a High-Performance 0.18-μm CMOS Logic Device," *Tech. Dig. IEDM*, 1998, p. 829.
14. P. Singer, "Tantalum, Copper, and Damascene: The Future of Interconnects," *Semiconductor International*, June 1998, p. 90.

Ramon CA, p. 183, 1999.
16. L. Peters, "Leaders Describe "Surmountable" Copper and Low-*k* Hurdles," *Semiconductor International,* September 2000, p. 25.
17. T. Mountsier *et al.*, to be published.
18. B. Gadsen *et al.*, "Top Surface Imaging Improves Copper Process Resolution," *Solid-State Technology*, March 2001, p. 77.
19. M. Neisser *et al.*, "Applying a Thin Imaging Resist System to Substrates with Topography," *Solid State Technology,* August 2000, p. 127.
20. M. Maenhoudt *et al.*, "248 nm and 193 nm Lithography for Damascene Patterning," *Solid State Technology,* April 2001, p. S15.
21. M. Moussavi "Advanced Interconnect Schemes Toward 0.1 μm," *Tech. Dig. IEDM* 1999, p. 611.
22. D.J. Thomas *et al.*, "Meeting the ITRS Roadmap Challenges with Low-*k* Dielectric Etching," *Solid State Technology,* March 2001 p. 108.
23. S. Vanhaelemeersch *et al.*, "Dry Etch & Clean Aspects for Advanced Integration of Low-*k* Dielectrics," *Mtg. Electrochem. Soc.*, Spring 2000, Abs. 224.
24. G. Rajagopalan *et al.*, "Development of Damascene Etch Processes for Organic Dielectric Materials," *Abs. Mtg. Electrochem. Soc.*, Fall 1999, Abs. 702.
25. I.-C. Deng *et al.*, "Using NH_3 Plasma Treated Spin-On Low-*k* Film as Barrier Metal-Free Dielectric to Suppress Copper Diffusion and Improve Its Ashing Resistance," *Abs. Mtg. Electrochem. Soc.*, Spring, 2000, Abs. 225.
26. T. Fukuda, *et al.*, "0.5-μm Pitch Copper Dual-Damascene Metallization Using Organic SOG (k = 2.9) for 0.18 μm CMOS Applications," *Tech. Dig. IEDM* 1999, p. 619.
27. A. Braun, "Photostrip Faces 300 mm, Copper and Low-*k* Convergence," *Semiconductor International,* September 2000, p. 88.
28. D. Hymes *et al.*, "The Challenges of the Copper CMP Clean," *Semiconductor International,* June 1998, p. 117.
29. V. Shannon, "Techniques to Improve Cu/Low-*k* Integration," *Solid State Technology,* September 2001, p. S22.
30. J. Rubin, "Supercritical Fluids as Wafer Cleaners," *Semiconductor International,* September 2000, p. 84.
31. P. Singer, "Dual-Damascene Challenges Dielectric Etch," *Semiconductor International,* August 1999, p. 69.
32. S. Russell *et al.*, "Integration Challenges of Ultra-Low-*k* Dielectrics," *Adv. Metallization and Integrated Systems for ULSI Applications,* 1997, p. 289.
33. L. Peters, "Low-*k* dielectrics: Will Spin-On or CVD Prevail?" *Semiconductor International,* June 2000, p. 108.
34. A. Braun, "CMP Battles Low-*k* Fundamental Barriers," *Semiconductor International,* October 2000, p. 66.

Chapter 16

COPPER INTERCONNECT PROCESS TECHNOLOGY

As described previously in Chap. 12, interconnects in deep-submicron ICs comprise a multilevel structure of fine wiring located on the top of the transistor circuitry of logic and memory chips. To avoid significant degradation of circuit speed, on-chip interconnections should permit rapid signal transmission among the various parts of the circuitry. Ever since the development of the integrated-circuit in the early 1960's, the materials most widely used for the wiring structures have been aluminum as the conductor (or eventually an aluminum-copper alloy for better reliability), and silicon dioxide as the insulator. In 1997 the transition to copper as the conductor was initiated with the announcement by IBM that they had developed a technology for forming copper interconnect lines and vias for IC applications. This was one of the most important changes in materials that the semiconductor industry has experienced since its inception. Since that breakthrough, AMD reportedly applied Cu to its microprocessor circuits at the 0.18-μm technology node, enabling it to introduce the first GHz-rated microprocessor in March 2000, ahead of competitors. In addition, circuit designers have also started to explore innovative layouts to take full advantage of the multiple benefits of Cu (as opposed to simply replacing Al with Cu in an existing design). Thus, it appears that the Age of Copper is dawning!

16.1 WHY COPPER FOR DEEP-SUBMICRON IC INTERCONNECTS?

As noted above, copper is now poised to take over as the main on-chip conductor for all types of ICs.[1,2] Early adopters of Cu technology put copper in production at the 0.25-μm node. Recent market-research studies indicate that for the 0.18-μm generation about 5% of the interconnect structures will be implemented with Cu. It also appears that most Cu in production (or pilot-line production) in 2001 was at the 0.18-μm and 0.13-μm technology nodes. For some manufacturers, 0.13-μm is a second-generation of Cu, for others 0.13-μm is the entry point for placing Cu into production (Fig. 16-1). It is predicted that Cu will represent about 60% of the interconnects for the 0.1-μm generation. The reasons why copper is displacing Al as the dominant conductive material for deep-submicron IC interconnects are explained next.

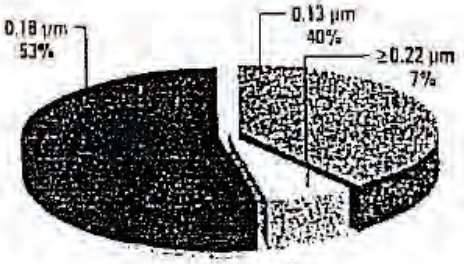


Fig. 16-1 Market analysis of device manufacturers using pilot production in 2001.[39]

16.1.1 The Advantages of Copper Interconnects

There are several critical advantages that copper interconnects provide over the traditional aluminum interconnects that have been used ever since ICs were invented in 1961:

1. Copper has a resistivity that is approximately 40% lower than Al (ρ_{Cu} = 1.7-$\mu\Omega$-cm versus ρ_{Al} = 2.7-$\mu\Omega$-cm), and 50% lower than the Al:Cu-alloys used in IC interconnects ($\rho_{Al:Cu(0.5\text{-at}\%)}$ = 2.9-3.3-$\mu\Omega$-cm). By referring to Eq. 12-7 for the interconnect *RC*-time-delay, it can be seen that *RC* scales directly with ρ. Specifically, if copper can be used instead of aluminum-alloys, the resistance of the wiring structures can also be proportionately reduced. That is, a reduction in the wiring resistance (and thus also the *RC*-delay) of as much as 45% can be obtained when copper replaces Al as the core-interconnect-metallurgy material.

It is important to note that a more complete analysis of the wire resistance must take into account not only the core-material of the interconnect (Al:Cu or Cu), but also any other (and usually higher-resistivity) materials used to surround the core-conductor. A comparison between the effective resistivity ρ_{eff} of Al:Cu wires and Cu wires is illustrated in Fig. 16-2. That is, the Al:Cu lines in Fig. 16-2 are fabricated by etching an Al:Cu film that has both bottom and top Ti layers. During sintering the Ti-layers react with the Al to form $TiAl_3$ (whose resistivity is ~35-$\mu\Omega$-cm, and the thickness of the resulting $TiAl_3$-layers is about 3.5 times the original Ti-layer thickness). The Cu wires are formed by a damascene process (see Chap. 15). This requires that the bottom and sides of the Cu wires be clad with a thin layer of a diffusion-barrier material. While this barrier-layer film also has high resistivity, it does not react with the Cu (unlike Ti, which reacts with Al), and this material also serves as an adhesion-layer. In any-case, it is clear from Fig. 16-2 that copper offers a very significant reduction in the effective resistivity for both present and future technologies compared to aluminum.

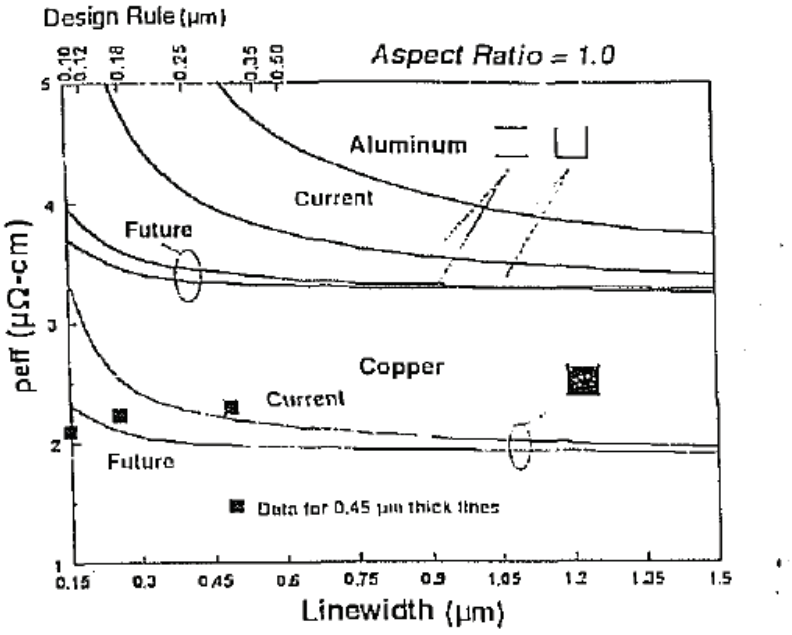


Fig. 16-2 Effective resistivity for aluminum and copper interconnects as functions of linewidth and design-rule generation. Computational results (*solid lines*) have been obtained for various combinations of conductor material and barrier location shown in the right side of the figure. Experimental points correspond to published results from real structures.[46]

This lower effective-resistivity of Cu-wiring can be exploited in one of two ways to reduce the *RC*-time-constant of Cu interconnect-systems:

i) The total resistance R_{int} of an interconnect line can be reduced (by retaining the same metal thickness [and cross-sectional area]), or;

ii) The side-by-side capacitance C_{LL} of interconnect lines can be reduced by decreasing the metal-film thickness. This allows R_{int} to remain the same as it was for a thicker Al line, but it reduces C_{LL}. Meanwhile, the same current is carried in a smaller-cross-sectional feature, without fearing EM-failure (see following paragraphs).

The primary clock-speed benefit derived from Cu is gained by using it in the upper-interconnect-levels of a multilevel-metal structure (where the conductor lengths can be of the same order as the chip size). As was shown in Chap. 12, Fig. 12-1, by utilizing Cu together with a low-k dielectric to replace Al and SiO_2, the overall delay (and power consumption) in technologies below 0.25-μm is dramatically reduced.[3]

As also discussed in Chap. 12, the interconnect RC-delay is predicted to become larger than the *RC*-delay of MOSFET transistors for technology nodes below about 0.5-μm. As an example, Fig. 16-3 shows that the *RC*-

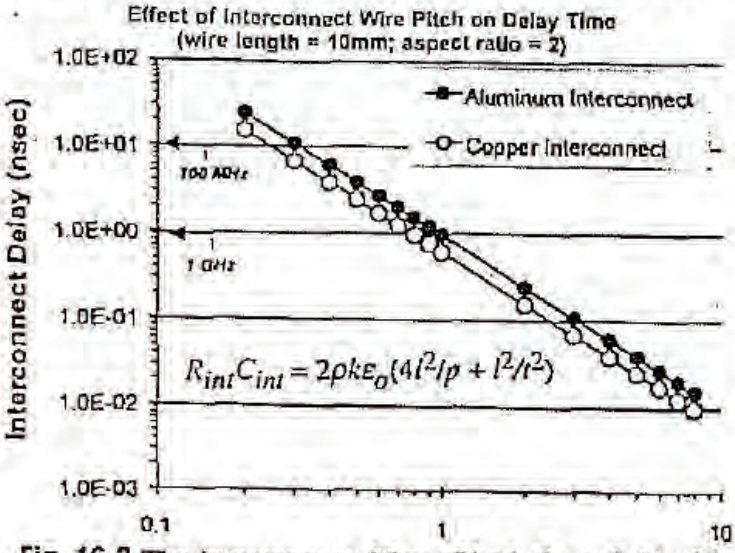


Fig. 16-3 The interconnect delay of both Cu and Al wiring has been calculated using Eq. 16-4 for *RC*. The delay-time in nanoseconds (1 nsec = 10^{-9} sec) is plotted versus the pitch of the wiring in microns (1 μm = 10^{-4} cm) for a line with an aspect-ratio = 2 and length l = 10-mm. The dielectric constant of the ILD has been set to k = 4.0, which is approximately equal to that of CVD deposited silicon-oxide. The graph shows that as the wire-pitch is decreased below 1-μm, the delay-time can become greater than the intrinsic switching-time of the transistors. The speed limit is increased by using Cu over Al.

delay of a dense array of 10-mm-long Al-alloy interconnect-lines with an aspect-ratio of 2 is longer than 1-nsec for pitches smaller than 0.8-μm. Since 1-nsec is equivalent to a frequency of 1-GHz, this implies that signal delays in IC wiring will limit circuit speed to about 1-GHz, (even though individual transistors may have faster intrinsic clock-rates). The interconnect thus represents a real limiter to high-performance ULSI chips – analogous to owning a sports car but only being allowed to drive it on a highway at rush hour. Switching to Cu is one way of improving this situation.

As was done in Chap. 12, Sect. 12-6, equations can be derived to calculate the interconnect delay-time $R_{int}C_{int}$. That derivation is repeated here, but for a simplified special case that exemplifies the impact of switching to copper. The following symbols apply to those aspects of the interconnect system depicted by the test-structure shown in Fig. 12-19.

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 the *length* of the metal lines is l, and the *resistivity* of the metal is 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. Using this approach, the resistance of a metal-line is calculated from:

$$R_{int} = \rho l/wt \qquad (16.1)$$

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

$$C_{LG} = k\varepsilon_o wl/d \qquad (16.2)$$

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

$$C_{LL} = k\varepsilon_o tl/s \qquad (16.3)$$

where ε_o is the permittivity of free space. For the $R_{int}C_{int}$ time-delay-constant, it follows that:

$$R_{int}C_{int} = 2\,R_{int}(C_{LG} + C_{LL}) = 2\rho(l/wt)k\varepsilon_o(wl/d + tl/s) \qquad (16.4)$$

To simplify the discussion, we assume that the pitch is equal to twice the line width ($p = 2w$), which also implies that $s = w$, and the metal-thickness is equal to the insulator thickness ($t = d$). In this case, Eq. 16.4 becomes:

$$R_{int}C_{int} = 2\rho k\varepsilon_o(4l^2/p^2 + l^2/t^2) \qquad (16.5)$$

Equation 16-5 is the relationship that is used here to calculate the value of $R_{int}C_{int}$. As also pointed out in Chap. 12, the dependence on ρ, k and l in Eq. 16-5 shows that $R_{int}C_{int}$ can be reduced by choosing metals with low-resistivity, insulators with low-k-values, and/or making line-lengths as short as possible.

Using Eq. 16-5, it is also possible to calculate the dependence of $R_{int}C_{int}$ on metal-pitch and metal-thickness. The results of such a calculation is illustrated in Fig. 16-4, which shows that for a given metal-

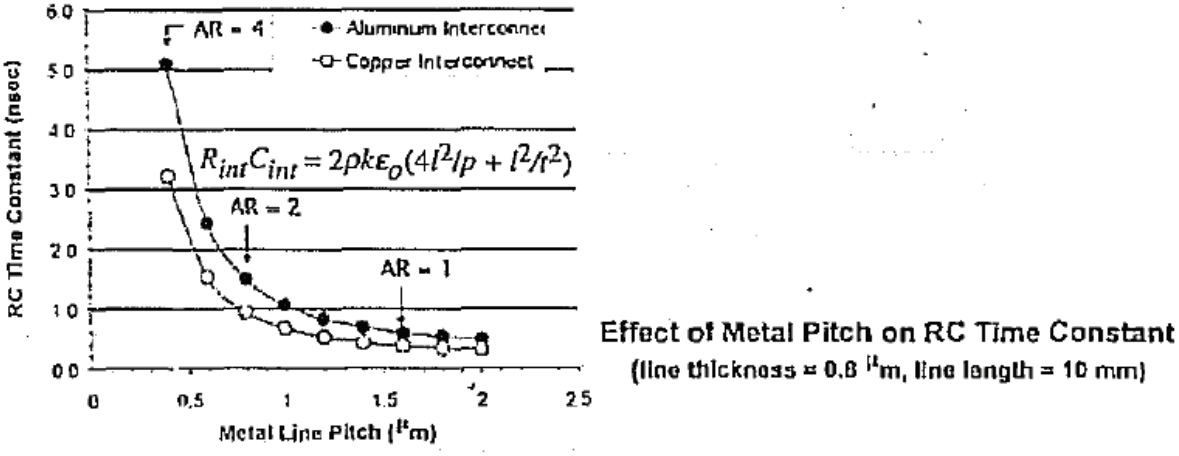


Fig. 16-4 The effect of Cu and Al metal-line-pitch on RC time-constant for a line thickness of $t = 0.8$-mm, a line-length of $l = 10$ mm for the case of both Cu and Al. Eq. 16-5 is shown on the graph and was used to calculate $R_{int}C_{int}$

pitch and line-width that $R_{int}C_{int}$ can be reduced by increasing the aspect–ratio of the line (which of course reduces R_{int} due to the cross-sectional-area increase). However, for aspect-ratios greater than 2, this reduction in the $R_{int}C_{int}$-product is offset by a corresponding increase in the intra-line capacitance C_{LL}. This fact (and the practical difficulty of fabricating metal-lines with very-high-aspect-ratios), is consistent with industry roadmaps predicting metal-line aspect-ratios of less than about 2.7 through the 70-nm node.

After IBM debuted Cu interconnects in 1997, several other companies in 1998 announced the development of Cu-multilevel interconnect systems that they would implement in commercially-available ICs.[4,5] The resistivity of composite copper interconnects in these reports (with the diffusion-barrier in place) did not exceed 2.2-$\mu\Omega$-cm, a 40% improvement over that of comparable Ti/Al:Cu lines.

2. Copper films exhibit much better electromigration (EM) characteristics and stress-void resistance than do Al films. Electromigration is an atomic-scale phenomenon in which the electrons that constitute electric-current collide with the metal-atoms of an interconnect line and push them in the direction of the current-flow. If this "*electron wind*" is strong enough, a significant migration of the metal atoms will occur. Over time this transport of material can lead to the electrical failure of the interconnect in one of two ways: 1) by a break in the line (causing an open-circuit) due to local depletion of material; or 2) by the formation of a short-circuit to an adjacent line due to local accumulation of material. The strength of the electron wind increases as the current-density is increased, and the upper-limit of current-density to prevent electromigration in Al:Cu interconnects is $5x10^5$ A/cm^2 (for a 20-year lifetime). This limit puts restrictions on circuit-design and chip-performance. (For example, a current of 5-mA in an interconnect-line of cross-section 0.5-μm x 1-μm (i.e., 0.5-μm^2 = $5x10^{-9}$ cm^2), represents a current-density $1x10^6$ A/cm^2, which is twice as great as the current-density-limit for Al:Cu-lines. Scaling of interconnect wires to increase functional-density and chip-performance inevitably leads to increased current-densities (see also Chap. 12), making Al:Cu interconnects less attractive for future technology structures.

Since Cu-atoms are more massive and more tightly bound together than Al atoms, interconnects made of Cu are generally more resistant to electromigration than those made of Al or Al:Cu (i.e., they exhibit 10-100x

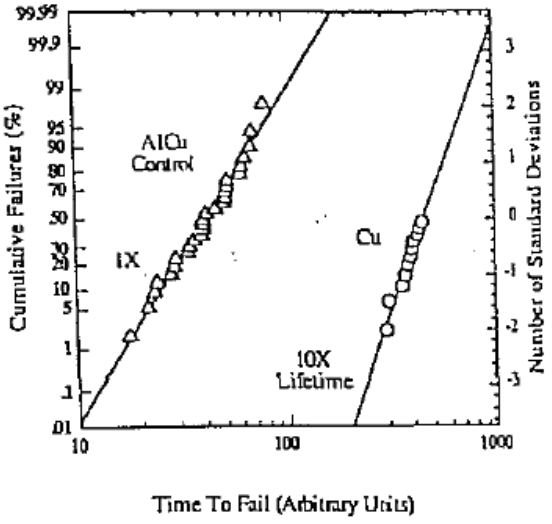


Fig. 16-5 Electromigration performance improvement using copper metallization.[5]

more EM-resistance). About an order of magnitude longer EM lifetimes are obtained for Cu lines of the same cross-section as Al lines, allowing Cu to deliver more reliability for a given interconnect design.[4] In addition to electromigration, stress-voiding (i.e., due to mass transport resulting from thermally-induced stress gradients), is another reliability concern.

Electromigration tests of 0.3-μm lines conducted at an elevated temperature (295°C) and a stress-current of 2.5×10^6 A/cm^2 showed a mean-time-to-failure (MTF) that was more than 10^3 times longer for Cu than for Ti/Al(Cu)/Ti lines (see Fig. 16-5). Excellent results were also obtained for stress migration testing.[6] Both sets of results argue strongly in favor of using copper interconnections for higher usable current densities and better overall performance.

Another benefit of Cu interconnects with respect to electromigration failures involves the current-flow in vias. As shown in Fig. 16-6, a small diameter via serves as the electrical bridge for current flow from metal-level-n to metal-level-n +1. A problem known as "current-crowding" arises because the current from the rectangular interconnect is compelled to flow through the much smaller circular-cross-section of the via. Due to the change of geometry and reduction in cross-sectional area, the current-flow in the via can be highly non-uniform, with local current-density much higher than in the interconnect line. If the via current-density exceeds the design specifications, excessive heating and/or electromigration can result. In turn, this can lead to circuit-reliability problems. Replacing W-plugs in

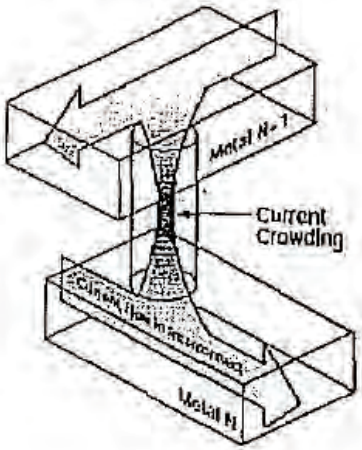


Fig. 16-6 Copper-filled vias can provide additional protection against structural failures associated with current crowding, thereby maintaining via reliability within designed specifications. In particular, the uncertainty in calculating the exact 3-D current-flow in the via is mitigated by the increased tolerance of Cu to electromigration at high current density.

the vias with Cu-plugs significantly reduces the electromigration problems observed in interconnects consisting of Al:Cu wires and W-plugs.

3. The third major benefit of employing copper (and some say it's the most important one) is that lower manufacturing costs can be obtained compared to using aluminum. The cost is lowered for two reasons.

First, because copper is difficult to etch, a new strategy called "damascene" or "in-laid" patterning was developed to form the Cu interconnect lines (see Chap. 15). IBM (which pioneered the process), claimed that the damascene approach requires 20-30% fewer steps than traditional subtractive patterning. Not only are there fewer steps, but dual-damascene eliminates or reduces the need for some of the most costly ones - including aluminum etch, and many of the tungsten-CMP and dielectric-CMP steps.

Second, using copper reduces costs because smaller lines can be employed to carry the same amount of current. Thus tighter packing-density can be achieved per level. This also implies that fewer levels of metal may be needed (see Fig. 16-7),[7] which significantly reduces manufacturing costs. This data indicates that for certain circuit applications only about half the number of levels are required if Cu is used instead of Al-lines and W-plugs.

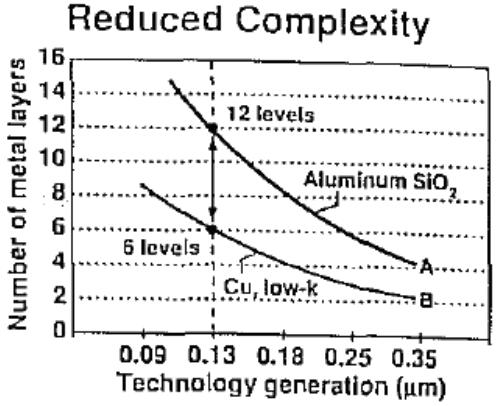


Fig. 16-7 One of the main benefits of copper, beyond the ability to increase chip speed and power consumption, is that the number of metal levels can potentially be reduced by as much as half.[1] Reprinted with permission of Semiconductor International.

4. The currently favored method for Cu deposition is electrolytic-plating, and this allows for much lower deposition temperatures that does Al sputtering or W-CVD. This enables the use of low-*k* dielectric materials whose thermal stability is lower than that of SiO_2.

16.1.2 Technological Challenges of Using Copper Interconnects

Although many powerful benefits can be gained if Cu is substituted for Al in the IC interconnect structure, a new set of technical challenges arise, including the following:

1. Cu does not form volatile copper fluorides and chlorides, making plasma etching of Cu impractical at temperatures below 200°C. Thus, another technique for forming Cu wires has been developed. Instead of patterning and etching Cu films, the Cu-films are first blanket-deposited over the entire wafer surface and into recesses in the dielectric layers. Then the Cu film is polished away by CMP in the flat areas, but is allowed to remain behind, in-laid in the recesses. This is the basis of *damascene* interconnect structure formation (as discussed in Chap. 15).

2. Unlike Al, Cu does not form a dense, self-passivating oxide on its surface. Thus, chemical reactions can continue to attack exposed Cu lines, making them more vulnerable to corrosion than Al lines.

3. Copper readily diffuses through SiO_2. If it is present in sufficient concentrations in a dielectric it can lead to shorts or leaky paths between conductors.

4. If the Cu manages to diffuse through the SiO_2 and then enters the silicon lattice, it will create deep impurity levels in the silicon bandgap. These levels act as *recombination/generation (R/G)* centers for free charge-

carriers. The presence of such R/G centers degrades transistor performance and causes *p-n* junction leakage. Copper can also precipitate at (or near) the silicon wafer-surface, and such metallic precipitates are known to cause defects in the gate-oxide as it is being grown. Cu is also a fast-diffuser in single-crystal silicon (i.e., it exhibits the highest diffusivity in silicon of all metal elements), and thus will quickly distribute itself everywhere within the bulk of the wafer (including in the device-regions near the wafer surface, where it can do the most harm). For these reasons, Cu is viewed as "poison" if it is present in the silicon lattice. As a result, special Cu processing areas are also set up in wafer fabs to keep trace amounts of Cu from contaminating any other areas of the cleanroom.*

5. Unlike Al (which reacts with SiO_2 to form a strong interface with excellent adhesion), Cu adheres very poorly to SiO_2.

6. The deposited copper films are polycrystalline in nature, and the preferred orientation of the crystal structure is a high (111)-orientation. Control of this orientation is important because resistivity, stress, adhesion strength, and electromigration resistance of copper layers are largely affected by the films crystal orientation.[17,53]

To prevent Cu from diffusing to the device areas of an IC, the interconnect structure must have the built-in reliability of several protective measures against this possibility. The primary line of defense involves totally encapsulating the Cu wires and plugs at all locations and at all levels to prevent copper from escaping. In dual-damascene interconnect processes, this is done in two-steps. In the first one, a *conductive* diffusion barrier-layer (e.g., Ta, TaN, or TiN) is deposited over the wafer after the trenches and vias of each level of metal have been etched in the dielectric stack, (but before the copper layer is put down). Thus, the insides of the trenches and vias of the damascene dielectric-stack are lined with a diffusion-barrier layer prior to deposition of the Cu into them. Such conductive barrier-layers must be effective against Cu out-diffusion and must simultaneously be thin enough to maximize the cross-section of the Cu-wires and plugs in the recesses. The second diffusion-barrier-layer is formed after the Cu-layer is deposited and the Cu structures have been polished with CMP. That is, a *non-conductive* diffusion-barrier-layer is deposited onto the wafer to seal the top-side of in-laid Cu structures (typically a CVD-silicon-nitride layer is selected to serve in this

* The use of Cu as an alloy in Al:Cu interconnects generates a completely different set of requirements for containing Cu atoms. That is, Cu-atoms alloyed with Al are immobilized by the aluminum matrix, which acts as a very efficient getter.

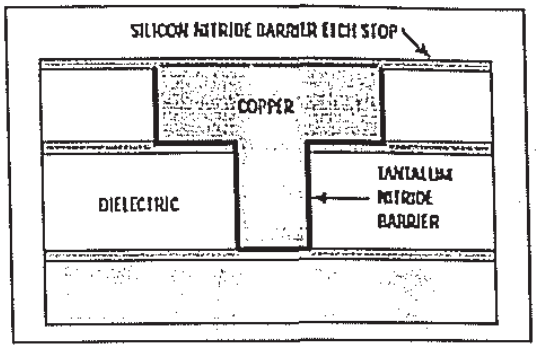


Fig. 16-8 Dielectric (silicon nitride) and conducting (tantalum nitride) barriers encapsulate copper in a damascene structure and protect against copper migration from interconnect structures to neighboring dielectrics and silicon.

role). Thus, Cu structures are ultimately enclosed on all sides by either a Ta (or TaN) or a silicon-nitride diffusion-barrier (Fig. 16-8).

The secondary on-chip line of defense is provided by W-plugs and phosphosilicate glass (PSG). The tungsten-plugs are used in the contact holes as a secondary barrier that keeps the first level of Cu-lines physically separated from the Si substrate (and this also means that a *single-damascene* process is used to form these W-plugs as well as the first level of Cu wires). PSG is known to getter metal atoms. The last level of protection on the chip itself is provided by a bottom silicon-nitride-layer, which further blocks the underlying silicon substrate from migrating Cu atoms.

While manufacturing protocols must also be modified to tightly control cross-contamination, the Cu-contamination problem does not seem to be as difficult as it first appeared to be. For example, fabrication tools cannot be shared between front-end and back-end processes, but copper manufacturing lines can be built in shared cleanroom facilities with aluminum interconnect manufacturing lines. Cleaning procedures must also be implemented to ensure that no detectable traces of copper are present on the bevels and backsides of wafers after processing.

Additional procedures have also been developed to heighten awareness of production workers and to contain copper contamination. For example, copper personnel in some fabs (especially maintenance technicians) wear cleanroom gowns that are used only in the copper bays, and they may not enter non-copper stations when wearing them. Specially colored wafer-carriers and conspicuous labels (indicating the presence of copper-products) are used to clearly identify wafers that have received copper processing and to segregate them from wafers that have not been subjected to copper processes.

Protocols also require that once a wafer has entered a copper bay it cannot return to a noncopper bay. In practice, however, wafers that leave the *front-end-of-line* (FEOL) after the contact holes to the silicon substrate have been opened, move to the *back-end-of-line* (BEOL), where the copper fabrication occurs. There is really no reason why the wafers should have to re-enter the FEOL part of the fab after that part of the process sequence is completed. Additional information on preventing cross-contamination from Cu is given in Sect. 16.7.

To counteract the problem of poor adhesion of Cu to oxide, the conductive diffusion-diffusion films under the Cu layers must also serve as adhesion promoters. The adhesion/diffusion layers used under the inlaid Cu features in damascene processes must therefore meet several criteria:

1. They must act as good adhesion-promoters for Cu onto SiO_2. That is, the adhesion/diffusion-layer materials should adhere well to oxide, and then the Cu films should adhere well to them.
2. They must be effective barriers against Cu diffusing through them under all post-copper-deposition temperatures.
3. They should exhibit low resistivity. That is, diffusion-barrier materials will be expected to exhibit resistivities of less than 1000-$\mu\Omega$-cm to meet the needs of advanced interconnect processes. This will keep the total resistivity of the composite Cu/barrier interconnect-structures below the resistivity of comparably-thick Al-alloy/barrier interconnect lines.
4. They should retain their diffusion-barrier properties when they are very thin, because they do not have the low-resistivity of Cu. That is, thinner diffusion barriers for Cu will be needed than for Al because in-laid-Cu interconnects require the barrier-layer to be formed on at least three sides of the Cu conductor (Fig. 16-9). To keep the lower-conductivity barrier-layers from impacting the resistance of the interconnect line, they must thus be relatively thinner for Cu lines.
5. They should protect the Cu from oxidation and corrosion.
6. They should be able to make a low-resistance contact to the copper layer of the metal-level under it, and this barrier-layer/Cu interface should exhibit good electromigration reliability characteristics.
7. They should provide a surface which will facilitate the nucleation of the Cu-seed layer deposited on it. Furthermore, the surface-condition and micro-

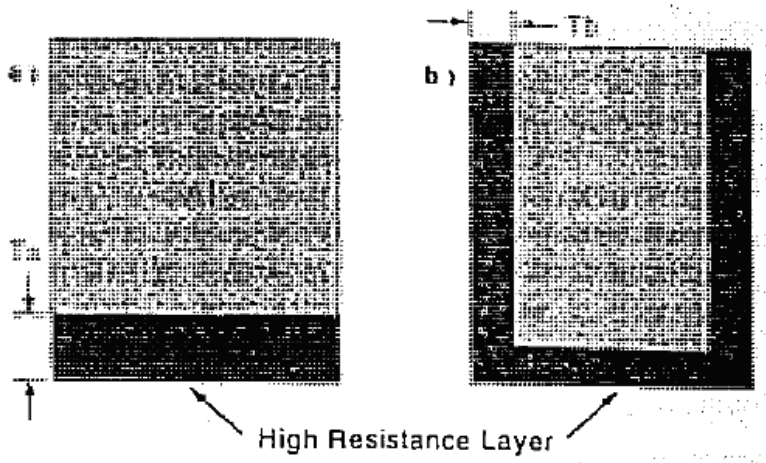


Fig. 16-9 Interconnect cross-sections showing the impact of higher resistance shunt or barrier layers: (a) Al interconnect; (b) Cu interconnect.[7] (© 1995 IEEE)

structure of the barrier-film can strongly affect the texture and grain-size of the deposited Cu (and these in turn, are critical in determining electromigration reliability).[50] Thus, the barrier-layer is also vital to seeding the desired Cu microstructure. Since the growth of Cu layers with a (111) preferred orientation provides Cu films with the best overall characteristics, the barrier-layer should enhance such Cu-structure.

In addition to the above materials-science issues associated with the use of copper, there are a number of difficult manufacturing challenges that will also have to be met when integrating Cu into an IC fabrication sequence. That is, getting a copper process (or any process) online and productive quickly is the key to profitability. The integration issues involved in bringing a Cu process to full-production are summarized in Fig. 16-10,[15] and they will be discussed in more detail in later sections of this chapter.

16.2 OVERVIEW OF COPPER PROCESS TECHNOLOGY

Before the details of depositing, polishing and cleaning copper IC interconnect structures are described, it is useful to give brief overview of the entire sequence of processes that are required to form a single-level of a copper dual-damascene structure. The sequence of process steps is as follows (and is also depicted in Fig. 16-11):

1. Non-Conductive Diffusion-Barrier-Layer Deposition
2. Dielectric Stack Deposition

3. Dual-Damascene Dielectric-Stack-Etch (Via, Trench, and Diffusion-Barrier-Layer Etch)
4. Conductive Barrier-Layer Deposition
5. Copper Seed-Layer Deposition
6. Copper Electrofill (and Annealing after Deposition))
7. Copper and Barrier-Layer CMP
8. Copper Clean
9. Non-Conductive Diffusion-Barrier-Layer Deposition.

The sequence starts by depositing a non-conductive diffusion-barrier-layer (e.g., a thin film of silicon nitride by PECVD) over the entire wafer (which has been planarized by a previous CMP step and may already contain one or more levels of metal). Next the remainder of the ILD stack is deposited (as described in detail in Chap. 15, Sect. 15.6). Using a lithography process, the vias and trenches of the dual-damascene structures are aligned to the previous layers of the circuit and are then etched with a dry-etch process. (This requires a sequence of two lithography and etch procedures, including stripping the resist after each etch step.) The SiN diffusion-barrier-layer at the bottom of the vias is also opened during the second of these etch steps, exposing the metal-layer of the previously formed level of the interconnect structure. Next, a conductive diffu-

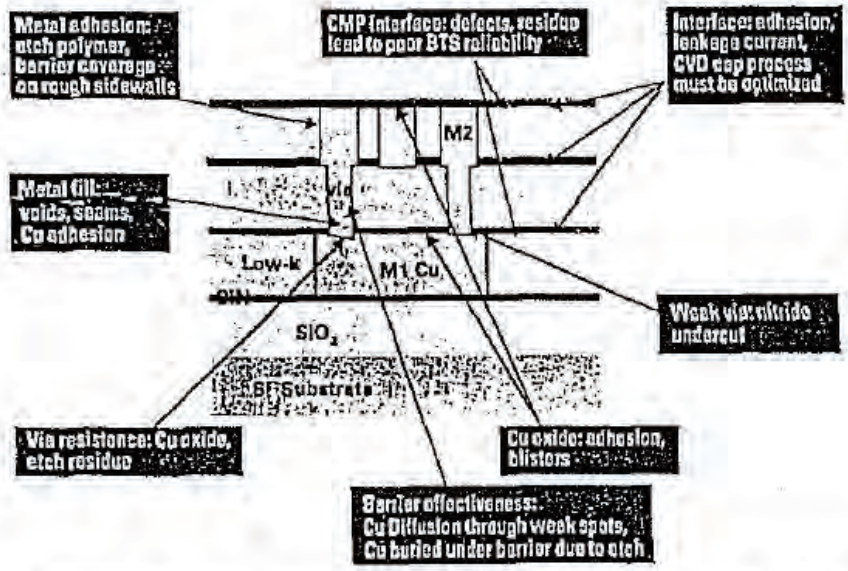


Fig. 16-10 Possible failure types associated with damascene processing pertain to film adhesion, barrier and metal fill capability, copper oxidation, and CMP defects.[9]

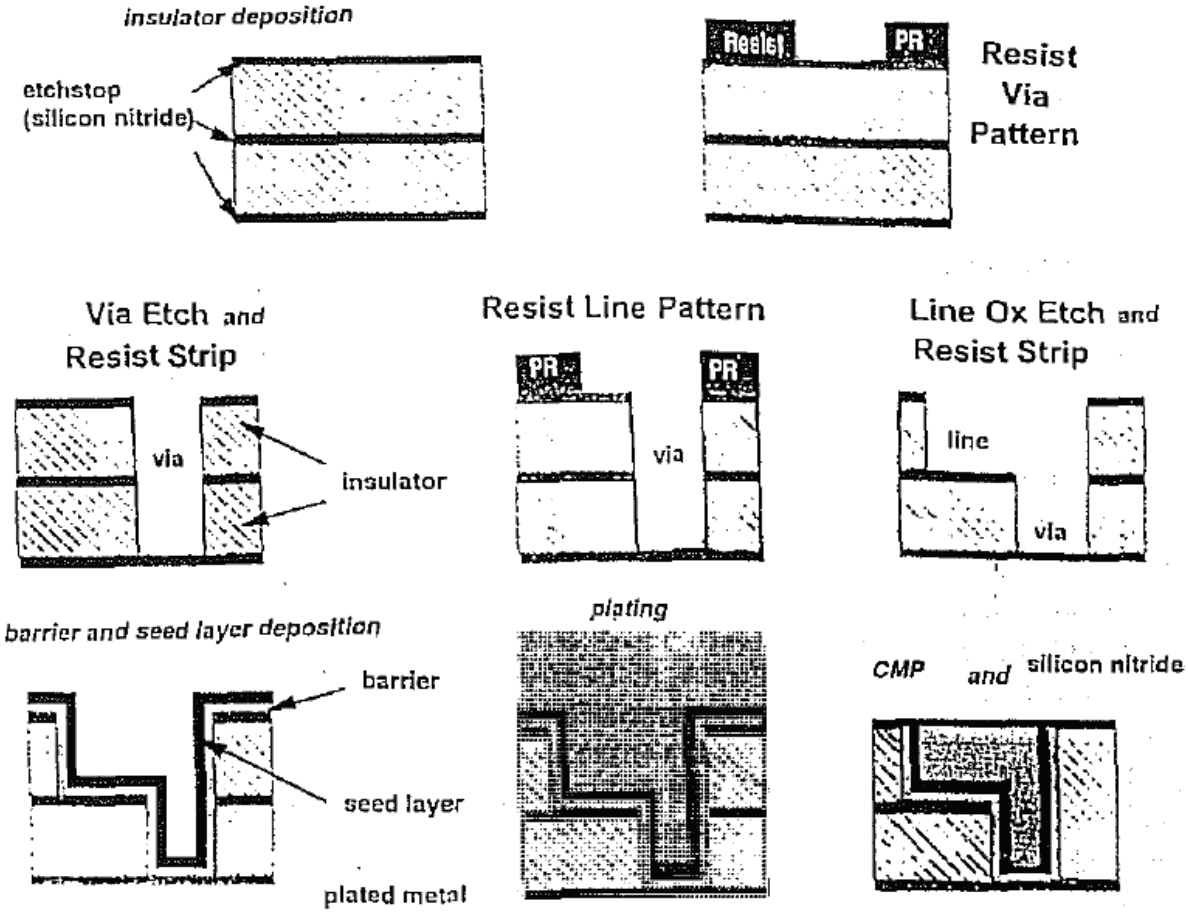


Fig. 16-11 Process sequence for forming a single-level of Cu metal in a dual-damascene interconnect structure (using a *via-first* approach).

sion-barrier/adhesion-layer is blanket-deposited onto the wafer, typically by PVD (with either a collimated-, a long-throw-, or [most typically] an ionized-metal-PVD process). Various materials have been selected to perform the role of this diffusion-barrier/adhesion-layer including Ta, TaN, TaSiN, TiN, TiSiN, WN, and WSiN (see Sect 16.3).

A copper seed-layer is deposited onto the diffusion-barrier/adhesion-layer to provide a low-resistance path for the copper electroplating process that comes next. The Cu-seed layer is also deposited *in situ* (i.e., without an air break) by one of same type of PVD-processes used to form the diffusion-barrier/adhesion layer film. Completing the process of filling the trenches and vias with Cu material is done with an electrolytic-plating technique. The key attributes of such Cu electrolytic-plating-processes include: a) the ability to fill the recessed structures without defects (i.e., no voids or seams should exist in the plated films after filling); b) the ability to achieve high-deposition rates; and c) the ability to

obtain uniform deposition-thickness across-the-wafer, and from wafer-to-wafer (e.g., ±2% across a 200-mm wafer with a 5-mm edge exclusion).

A CMP process is carried out next to remove any excess copper above the vias and trenches (called the *overburden*), but allowing metal to remain in the damascene recesses. During the CMP process, the diffusion-barrier-layer must also be removed. A key technological challenge involves achieving the same polishing rate on the diffusion-barrier layer, because such materials as Ta or TaN are hard, refractory metals (or metallic compounds), while copper is a comparatively much softer metal. After the CMP process is finished, it must be ensured that the wafer is cleaned, such that no copper residue remains on the oxide on the frontside of the wafer (nor on the wafer backside or beveled edges). Finally, another non-conductive diffusion-barrier film (e.g., CVD silicon-nitride) is deposited onto the wafer surface, which serves to cap-off the copper (i.e., this completes the encapsulation of the recessed-copper structures, ensuring they are surrounded by a diffusion-barrier-material on all sides).

The details of the first three steps of this sequence are described in Chap. 15, Sects. 15.6 and 15.9 (and thus will not be repeated here). The discussion that follows therefore begins with the Step 4 (Diffusion-Barrier-Layer Deposition, and the properties of such materials) and then proceeds to cover the details of the remaining steps of the process sequence.

16.2.1 The Transition from Aluminum to Copper Interconnects

The process of incorporating copper into the interconnect process of an IC technology may be done in one of two ways. Some manufacturers have decided to use Al:Cu (and W-plugs) for several of the lower levels of a multi-level metal interconnect structure, and limit copper to the upper levels. In one such example, Al:Cu was used in the four lower levels of the 6-level metal interconnect structure and Cu for the upper two levels (Fig. 16-12).[10] Another process described by TSMC also uses four aluminum levels of metal and two of copper.[11] The interface between the top aluminum interconnect level and the lower Cu interconnect level can be implemented either with a W-plug or with a Cu-damascene via structure.[12]

On the other hand, other manufacturers decided to use copper for the entire interconnect structure (with the exception of W-plugs being used in the contact holes). An example of such a 6-level Cu-interconnect-structure (with W-plugs used to make contact to the Si-substrate) was described by IBM (see Fig. 16-13).[13] Figure 16-14 shows a SEM of a perspective view of a six-level metal Cu interconnect structure (see also Fig. 12-16 in Chap. 12).

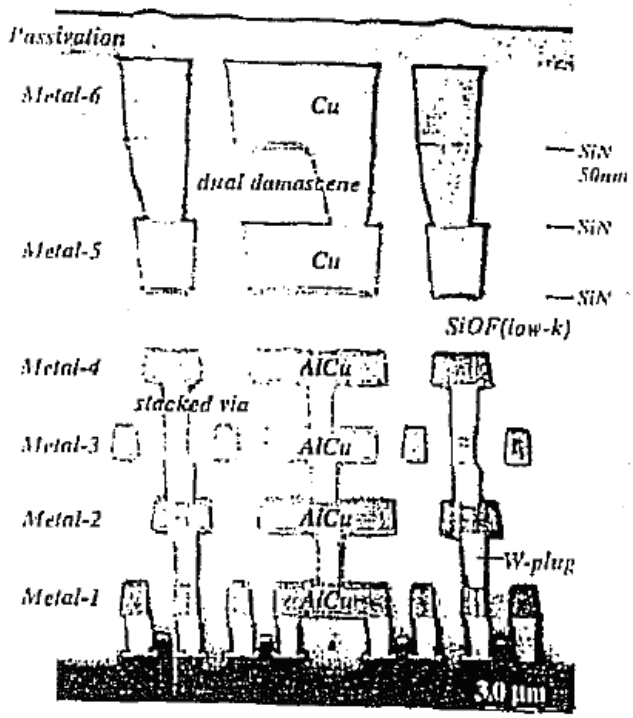


Fig. 16-12 Cross-section of six-level metal structure with the first four-levels being Al:Cu/W-plug interconnects and the top two levels of Cu metal.[10]

16.2.2 The Key Goal of Depositing Copper into Damascene Recesses

One of the most important aspects of depositing copper successfully into the recesses of a dual-damascene ILD stack is ensure it done without any voids or defects. The key to being able to do this is to fill the recesses with copper from

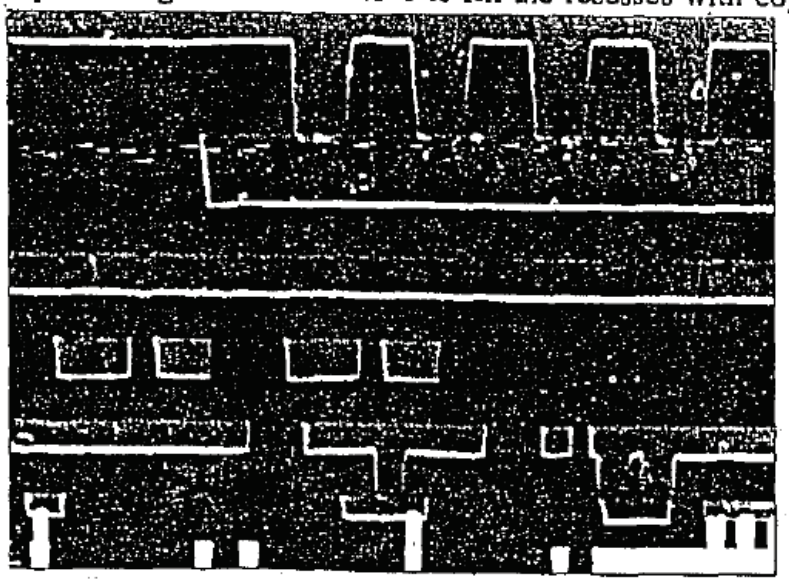


Fig. 16-13 Cross-section of six-level metal structure with all levels of Cu metal.[13]

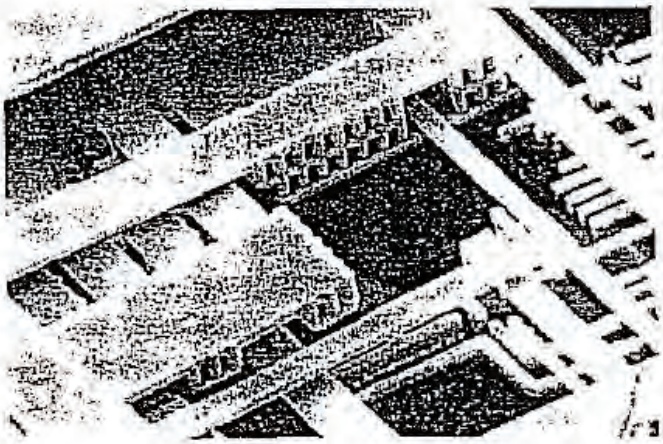


Fig. 16-14 An SEM of a perspective view of a six-level metal Cu interconnect structure.[32]

the bottom up, a technique called "bottom-up" filling, or "superfilling." However, successful superfilling is affected not only by the electrofill process but also by the dielectric etch process, the barrier-layer deposition-process and the seed-layer deposition-process. Thus, this topic has to be addressed in a piecemeal fashion, as will be done in the three upcoming sections.

16.3 BARRIER LAYERS FOR COPPER INTERCONNECTS

As discussed previously, to stop Cu from diffusing into the dielectric materials in which it is embedded (or into the Si substrate), it must be surrounded by a barrier-layer on all sides. The underlying layer is deposited prior to depositing the copper (as noted in Sect 16.2). In this section, the technology for forming these conductive barrier-layers is described. The subjects that will be covered are:

1. The impact of the barrier-layer on the Cu-electroplating process;
2. The technology of depositing barrier-layers into damascene recesses by PVD;
3. The material properties of tantalum (and TaN & TaSiN) barrier-layers;
4. The other candidate materials for Cu-technology barrier-layers (including CVD processes for depositing such materials).

16.3.1 The Impact of the Barrier-Layer Deposition Process on the Filling of Damascene Recesses with Copper

To recap, a barrier-layer for Cu interconnects must prevent Cu diffusion into the ILD and must assure good adhesion between the Cu and the ILD. Furthermore, the barrier must be thin, conformal, and have a low electrical resistivity. Deposition should be at a low temperature with few particles being produced.

The barrier-layer material should also be compatible with CMP of Cu and with the Cu electroplating process. Finally, there a number of other aspects involved with the deposition of the barrier-layer that impact the ability to completely fill damascene recesses with copper without voids. These include depositing a barrier-layer that is continuous and adequately thick everywhere along the recess surfaces (typically 25-30-nm-thick), and with little (or better yet, no) overhang. The PVD-process used to deposit the barrier-layer impacts these last two factors, and the PVD techniques for depositing seed and barrier layer is discussed next.

16.3.2 Introduction to the Technology of Depositing Barrier-Layers for Cu Interconnects

The conductive barrier-layer deposited into recessed damascene features before the Cu is deposited can be done by PVD or CVD. While PVD has been the most widely used method down to at least the 0.13-μm technology node, there may be a change to CVD at some point. One reason for such a technology replacement involves the formation of a film within a trench or contact hole by a *physical vapor deposition process.* Assume a contact-hole exists with the top face being open with the dimensions of a cube of side a. This represents a contact hole with an aspect ratio of 1:1. In such a case there are five inner surfaces of the cube that must be covered by material which would otherwise deposit on area a^2 on the top surface of the cube (i.e., if the top of the cube was not open, see Fig. 16-15). However, if the top of the contact hole is open, the total area inside the hole that must be covered now is $5a^2$. Thus, if the material that would have covered a^2

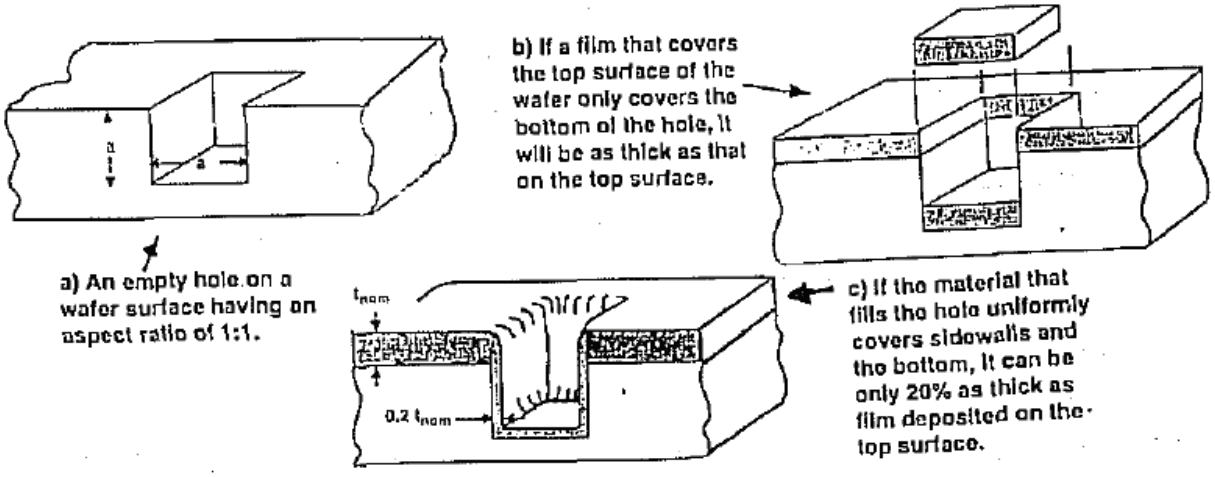


Fig. 16-15 One dilemma faced when forming films in high-aspect-ratio holes with PVD.

is evenly distributed onto all five insides of the contact hole, the film thickness will only be 20% as thick as that on the flat areas outside of the contact hole. Obviously, if the area of the hole-opening remains constant while its depth is increased, the aspect-ratio will also increase (e.g., if the square contact-hole opening-size remained at a^2, but the depth became 2a, the aspect ratio will increase to 2:1, and the area that will have be covered by the material of a^2 will now be $9a^2$). Thus, for the case of a physical-vapor deposition process, as the aspect-ratio of a recessed feature is increased, the material entering the feature must be spread over a larger area, making the thickness per unit area decrease. Furthermore, if the film on the upper flat regions must be etched or polished away after the hole has been filled, a much thicker film now has to be removed (which makes the etch or CMP process task significantly more time consuming and difficult). Deposition by CVD may be able to avoid this problem because CVD processes can be reaction-rate limited (allowing deposition to be almost 100% conformal at a very high-aspect-ratio, see Vol.1, Chap. 6). Thus, CVD may be employed for depositing barrier-layer films for Cu in the lower levels of deep-submicron interconnects (where the aspect-ratios continue to increase).

In the previous discussion it was noted that it becomes progressively harder to fill contact holes by PVD as the aspect-ratio increases. A related problem further increases the difficulty of filling high aspect-ratio-recesses. Recall that the sputtered-atoms undergo collisions as they are transported in the gas-phase in the space between the target and the substrate. In addition, sputtered atoms leave a planar-target surface from each point at which sputtering occurs in directions that roughly exhibit a cosine distribution. As a result of these two mechanisms, in a conventional sputtering apparatus the sputtered-atoms will arrive at the upper corners of the trench or contact hole on the wafer from a larger arrival-angle than they will at surfaces inside the recess. Thus, the film deposition-rate at the top corners of the recess will be higher than along its sidewalls or bottom. As a result, the top opening will begin to close off (i.e., an *overhang* will start to form), as shown in Fig. 16-16a. Thus, for this case, as time goes on (Fig. 16-16b) the deposition-rate onto the sidewalls and bottom will continually decrease. If the overhang gets thick enough, the hole will close entirely, leaving a void or keyhole.

Even if deposition does not continue long enough to completely close-off the opening, the overhang will shadow the bottom corners of the contact hole, where the arrival angle for sputtered atoms is already the smallest (see Fig. 16-16b). Thus, the deposited film will be thinnest at the bottom corners of the hole. Since a barrier-film must have a minimum adequate thickness along the entire bottom

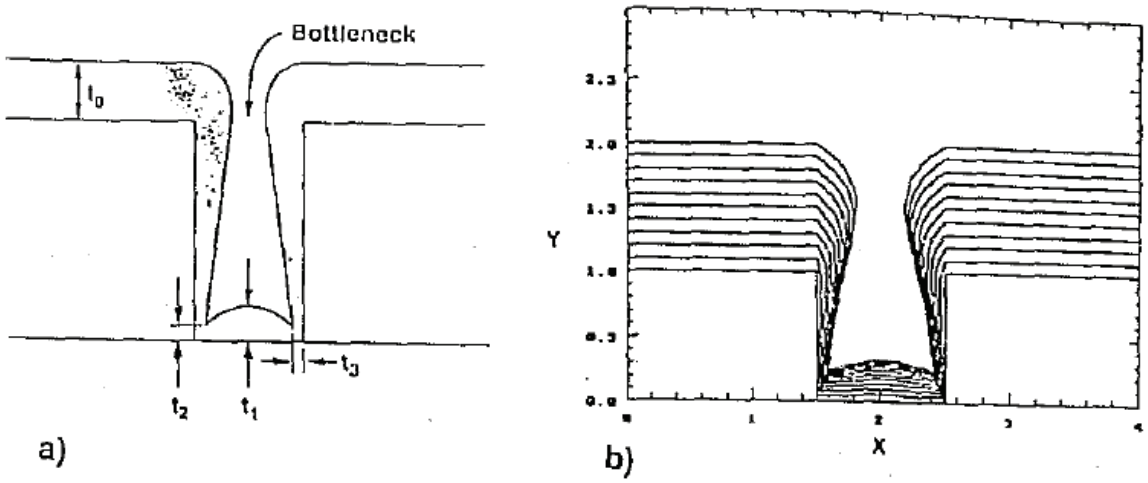


Fig. 16-16 a) Deposition of a sputtered film into a contact hole or via having a high aspect ratio. b) Development of a bottleneck due to the buildup of an overhang on the top corner.

of the hole in order to function effectively, such thin-spots at the corners can jeopardize the effectiveness of the barrier-layer.

Various methods have been explored to overcome this thinning of the film at the bottom corners of contact holes (including the use of CVD). However, modifications to the sputtering process have also been pursued. Two such early methods were *collimated sputtering* (in which a honeycomb baffle - i.e., a *collimator* - is inserted into the space between the target and the wafer to trap obliquely-directed vapor) and *long-throw-sputtering* (in which the target and wafer are separated by a distance 2–4 times longer than in a conventional magnetron source, and the plasma is also operated at a lower pressure). However, these two methods have proven to be inadequate for deep-submicron, dual-damascene applications. It is unlikely that they will even be commercially developed for 300-mm wafer processing. Thus, they will not be discussed further here (but more information on them is available in Vol.1, 2nd Ed, Chap. 11).

16.3.3 Ionized Magnetron Sputter Deposition

While the two sputtering techniques described above produce adequate barrier-layer films for contact-holes and vias for 0.35-μm (and some 0.25-μm) CMOS technologies, they appear to fall short for other 0.25-μm (and smaller) contact holes and vias when the aspect-ratio exceeds about 3:1. Thus, an alternative to collimated (or long-throw) sputtering, namely *ionized magnetron sputtering*, or *I-PVD* (also termed *ion metal plasma* [*IMP*] by Applied Materials), has been developed to deposit the next generation of ULSI barrier-layers with PVD. In this method, a large fraction of the neutral sputtered-metal-atoms are ionized while in flight between the target and the wafer. A bias on the wafer then causes an

electrical acceleration of these ionic species to the wafer-surface with a controlled energy and direction (i.e., perpendicular to the wafer surface).[14,15,16] There have been several sputtering-source configurations developed to implement the ionized magnetron sputtering technique.

Figure 16-17a shows a schematic drawing of an ionized magnetron sputtering chamber used in the first commercially available I-PVD tools. A conventional magnetron source (powered with a dc power-supply) remains in place to sputter neutral metal-atoms from the target. (The magnetron produces a plasma and dark space that is confined near the target surface.) However, an rf-powered (and water-cooled) coil is now also located within the chamber (Fig. 16-17b). This coil inductively couples energy from a second 13.56 MHz rf-power supply into a secondary plasma downstream of the magnetron plasma. By increasing the pressure in the chamber to 20-40-mtorr (up from the 5-mtorr of a conventional magnetron sputtering process), and by increasing the depth of the secondary plasma to 8–10 cm (by increasing the distance between the cathode and the wafer chuck), the number of collisions experienced by the sputtered atoms before they

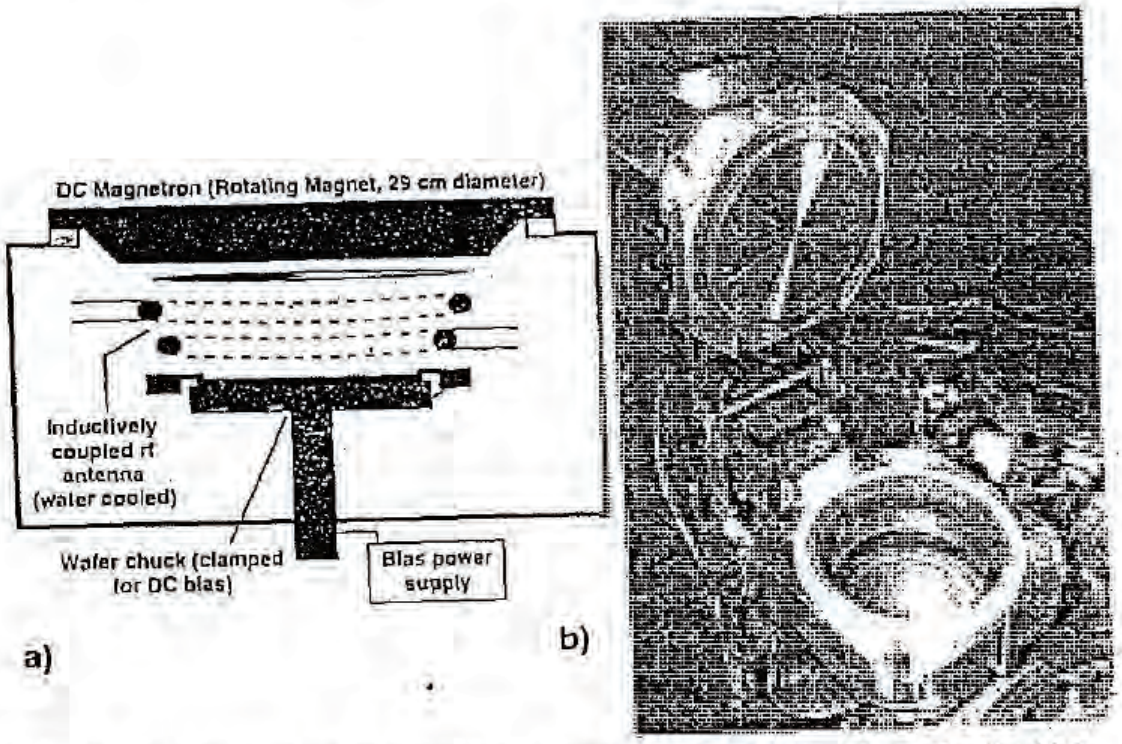


Fig. 16-17 a) Schematic of a magnetron sputter chamber fitted with an rf-coil which allows an ion magnetron sputtering process to be implemented. b) Photograph of an Applied Materials IMP system showing the internal coil. (Courtesy of Applied Materials)

arrive at the wafer (as well as their transport time through the gas phase), is increased. This longer transport time, and the fact that the ionization potential of metal atoms is much lower than that of Ar atoms, allows a much higher fraction of the metal atoms to be ionized by this secondary plasma (e.g., for a Ti-sputtering process, typically 50-70% of the sputtered Ti-atoms are ionized). A medium-density secondary-plasma is thus created in the region between the cathode and the wafer ($5x10^{11}$-10^{12} ions/cm^3).

The chief advantage of the I-PVD approach can now be exploited. That is, by applying a negative-bias to the wafers, the ionized metal-atoms can be accelerated toward the wafer surface (at voltages ranging from a few to several-hundred volts, depending on the specific application). Thus, the metal ions will strike the surface exactly at normal incidence to the wafer surface, regardless of their original trajectory. In addition, the third (substrate-bias) power-supply allows the kinetic energy of the metal ions striking the wafer to be controlled. Using an rf-power-supply to produce this bias overcomes the problems with insulating wafer surfaces (or backsides). Note that the ion-flux onto the wafer results in power dissipations ranging from about 25 watts to several hundred watts, which calls for some degree of wafer cooling.[16]

The bottom-coverage of high-aspect-ratio holes with I-PVD is better than that observed with collimated sputtering. That is, 40–80% bottom coverage has been achieved on 0.18-μm holes with aspect-ratios of up to 8:1.[17] Note that while the bottom-coverage is high, the sidewall-coverage is much thinner (although the film on the sidewall is continuous, see Fig. 16-18).

Several other advantages of I-PVD processes should be noted. First, much better target-utilization is obtained than with collimated sputtering (as almost all of the sputtered atoms arrive at the wafer surface, instead of some being intercepted by the collimator). Sputtering tool throughputs are therefore also increased. Second, the high flux of metal ions (as well as some Ar ions) directed perpendicularly onto the wafer surface appears to re-sputter the atoms that would otherwise form an overhang at the top-corners of a recess. This keeps the top opening of contact holes and vias from closing up. Figure 16-18 shows the coverage of the sidewall and bottom of a high-aspect ratio hole with I-PVD. Third, the negative-bias applied to the wafer not only causes the ionized metal atoms to preferentially strike the bottom of the via, but it allows the kinetic energy of these metal (and Ar) ions bombarding the wafer to be adjusted. For appropriate bias levels, re-sputtering from the feature base can be produced, which redistributes sputter-deposited material onto the lower sidewalls (Fig. 16-19). This is exactly where the lowest PVD coverage occurs in collimated and

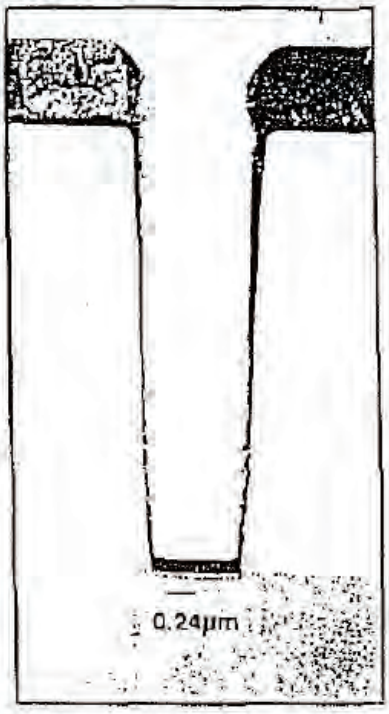


Fig. 16-18 SEM showing the coverage of the sidewall and bottom of a high-aspect ratio hole deposited by I-PVD. (Courtesy of Applied Materials)

long-throw PVD-processes. The process of local re-sputtering can also be used to tailor the conformality of the deposit to a greater degree, and this has extended the use of PVD for depositing barrier-layer and seed-layer films.

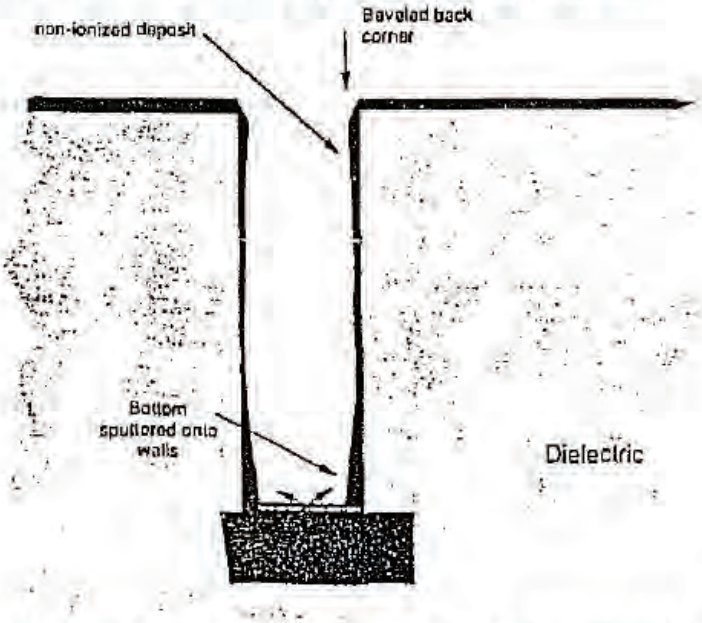


Fig. 16-19 Schematic drawing of I-PVD deposition of a high-aspect-ratio recess showing the effects of re-sputtering of the deposited film.

While the concept of I-PVD has the many advantages described above, the internal coil-configuration has some practical difficulties. First, the coil can get coated with the sputtered metal. If this gets too thick, the film may flake off and produce particles. Second, material from the coil itself can be sputtered, with some material ending up being re-deposited on the wafer. While a copper coil can be used to circumvent this problem when depositing Cu-seed-layers, this does not provide a fix when depositing Ta or Ti barrier-layers. Instead, a coil made of Ta (or Ti) would be needed, but forming such coils (which must also accommodate the flow of cooling water) is not a trivial task. As a result, two other I-PVD source configurations have recently been commercially introduced.

16.3.4 Hollow-Cathode Magnetron Sputtering Source

The *hollow-cathode magnetron* (HCM™) used in the Novellus INOVA™ sputtering system (Fig. 16-20) employs a similar technique to ionize sputtered atoms as they are transported to the substrate.[42,55] A uniform plasma throughout the volume of the cup-shaped hollow cathode target provides sputtering from a large fraction of the area of the target surface. Due to the geometry of the cup-shaped target a significant fraction of the sputtered neutrals are re-deposited on the target surface, where they can be sputtered again (instead of depositing on the chamber walls). This provides the benefit of high target utilization. In addition, no internal rf-coil is needed to generate the high-density plasma that ionizes the sputtered metal atoms. Instead, the orientation of the permanent magnets allows

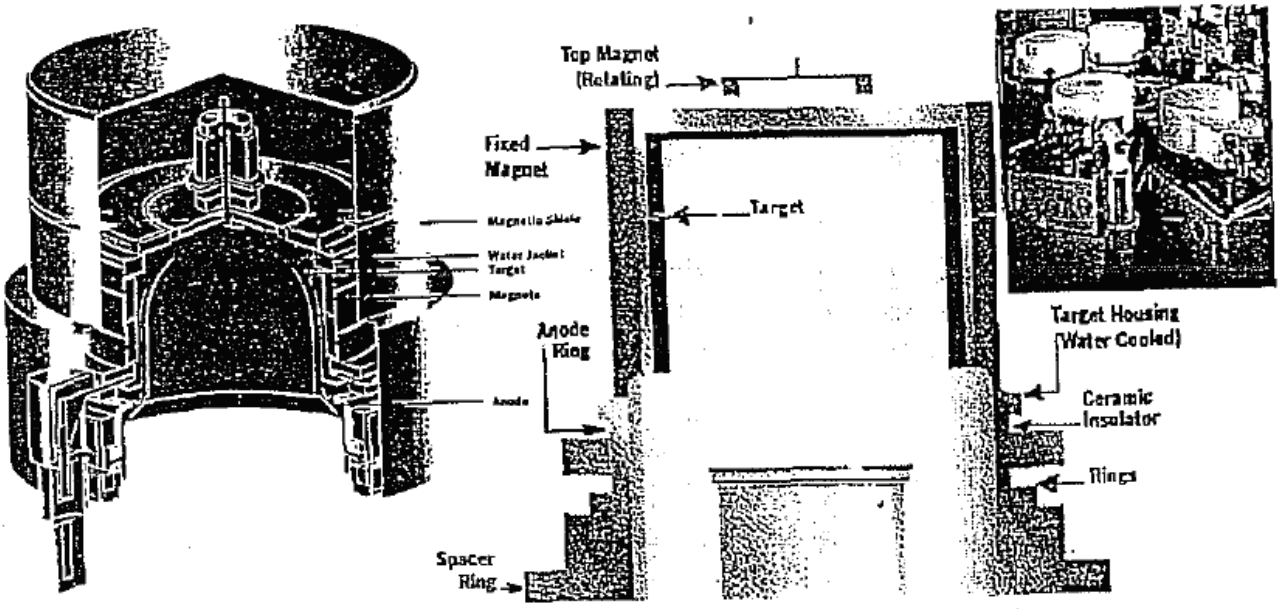


Fig. 16-20 Cutaway drawings of the hollow-cathode magnetron source used in the Novellus Systems's INOVA® sputtering system. Courtesy of Novellus Systems, Inc.

the electrons emitted by the target to efficiently ionize the sputtered neutrals (with the degree of ionization exceeding 90%). A bias on the wafer-chuck accelerates the ionized sputtered atoms toward the wafer in a direction perpendicular to its surface, allowing good coverage of high-aspect-ratio holes. The HCM can be used to sputter Ta, TaN, TiN, and Cu.

16.3.5 Self-Ionized-Plasma Sputtering Source

Another new sputtering source for implementing the I-PVD technique was recently introduced by Applied Materials. They refer to this source as *self-ionized-plasma* (SIP), and the internal rf-coil used in their IMP™ source to ionize the sputtered metal atoms is not needed. While few details about this source have been published, some of its aspects have been described.[18,19] The source operates at relatively low pressures (< 2 mTorr) with a target-to-substrate distance that is increased above the distance used in conventional magnetron sputtering sources. These measures increase the mean free path of the metal-ions and provide additional self-collimation to the metal-atom trajectories. The term *self-ionized-plasma* probably refers to the phenomenon in which some of the sputtered metal atoms that are ionized by the source, return to the target with enough energy to sputter other metal-atoms (and also produce additional electrons). These additional electrons, in turn, ionize more of the sputtered metal atoms, which keeps the plasma self-sustaining and produces more metal-ions. The phenomenon is apparently an efficient way of ionizing metal-atoms that have a small ionization energy (such as Cu). However, questions have been raised about whether it is effective for metals that have higher ionization energies (such as Ta).

16.3.6 Material Properties of Tantalum and Refining Tantalum Metal For Semiconductor Applications

The metal *tantalum* (or such tantalum-based compounds as TaN or TaSiN), have been the most commonly used barrier-layer materials for Cu-damascene interconnects. As such, it is useful to give some background about this element.

16.3.6.1 Material Properties of Tantalum: Tantalum is an uncommon (in nature) refractory metal, with an atomic-number of 73, an atomic-mass of ~181, a body-centered-cubic (BCC) crystal structure, and a steel gray color. It is an attractive Cu-barrier-material because of its high melting point (2996°C), and because it is immiscible with Cu. It is also a highly reactive metal that forms strong metal-metal and metal-oxygen bonds (much like titanium). Thus, it provides a low-resistance ohmic contact and good adhesion to the copper above (and to the

dielectric material below) it in the Cu-damascene metal scheme. Doping the Ta with a few percent nitrogen blocks grain-boundary diffusion-paths. More heavily nitrided Ta (produced by reactive sputtering of Ta in the presence of nitrogen) can also be used as a barrier-layer. Such TaN barrier-layers will be discussed further later.

16.3.6.2 Refining and Forming Tantalum Metal For Sputtering Targets: Tantalum exists in nature as an oxide Ta_2O_5. The principle ore of tantalum is *tantalite* (Fe, Mn)Ta_2O_5. Mines in western Australia are the largest producers of high-grade tantalite, and such concentrates are shipped to the tantalum metal manufacturers for refining.[20] The main market segments for refined tantalum are capacitor-grade powder (about 50%), mill-products and wires (16%), chemical compounds (such as tantalum carbides, 14%), and use as a metal-alloy additive (6%). High-performance tantalum products such as *sputtering targets*, represent less than 10% of the tantalum mill-product and wire business.

Tantalite ore is not the only source of tantalum. Nearly 25% of the total Ta production is recycled. In fact, if the market for high-quality sputtering targets outpaces the rate of increase in tantalite production, it might become necessary for wafer manufacturers to return spent targets to the Ta manufacturers in order to be assured of future availability.

The extraction and refining processes for tantalum conclude with a chemical precipitation reaction that yields Ta metal in powder form. The high-density, metallurgical-grade powders (and recycled Ta scrap) are used as feed for *vacuum melting*. (Because tantalum exhibits a great affinity for oxygen, nitrogen, and carbon, thermal treatments must be conducted in a vacuum or inert gas atmosphere.) For demanding applications (such as sputter deposition), the refined tantalum must possess a number of material properties that can be obtained by subjecting the tantalum powders to such vacuum melting processes.

Typically, tantalum ingots are melted several times to purge most of the impurities from the metal. Because the melting point of Ta is much higher than most materials, this is effective in volatilizing the interstitial gases (such as O, N, C) and many metal impurities. However, the other refractory metal elements (i.e., niobium, molybdenum, and tungsten) are not volatilized during vacuum melting, but remain as soluble impurities in tantalum. However, these refractory metal contaminants can be minimized through specialized front-end chemical extraction processes, resulting in extremely pure-Ta ingots. For example, the Performance Materials Division of Cabot Corp. (Boyartown, PA) offers Ta-ingots that are 99.999% pure.

The pure ingots consist of poly-crystalline Ta with small-sized grains. These ingots can be cold-worked (i.e., at ambient temperature) to cause a re-crystallization of the grain structure in the finished annealed product. The desired microstructure for Ta sputtering targets is a fine, homogeneous polycrystalline structure. Tantalum crystal grains predominantly have a (111) or (001) type structure. It is preferable to have a Ta ingot with a preponderance of (111) grains. This is because Ta sputtering-targets which contain large (001) clusters exhibit lower sputtering yields. The reason is that in such body-centered cubic (BCC) metals as Ta, the erosion-rate of (111) incident planes is predicted to be far greater than for (001) crystals.[21] The proportion of (111) to (001) grains is governed primarily by deformation history.

In summary, attaining microstructural and textural homogeneity in tantalum sputtering targets through the use of proper thermochemical processing is critical to assuring the performance of such targets. It should also be noted, that all of the refining and thermochemical processes are expensive. In turn, this also makes Ta sputtering targets quite costly (reaching $40,000 or more for some target configurations).[22]

16.3.6.3 The Properties of Tantalum (and TaN & TaSiN) as Cu-Barrier Layers: To prevent the drift or diffusion of Cu atoms, a barrier-layer should not react with Cu. Such a barrier-layer should remain thermally stable up to the maximum temperature that the wafer will be subjected to once the Cu has been deposited. Since this temperature is not likely to exceed 500°C, the barrier-layers should be stable to this temperature. Ta and TaN films are observed to remain thermally stable up to about 550°C, and thus they are capable of meeting this requirement for back-end processes.[23]

However, as noted in Sect. 16.1, the barrier-layer materials must also serve as adhesion layers for Cu. There is typically a tradeoff between barrier integrity and adhesion performance with Cu. If a thin-film of the barrier-layer material does not react with Cu at all, it may exhibit poor adhesion. If the material reacts too easily, it may not function as a good diffusion-barrier. For good barrier properties and good adhesion with Cu, an ideal barrier would react with Cu to some self-limiting extent.

Tantalum films, sputter-deposited using a Ta target in an Ar ambient, have been investigated for such Cu barrier-layer applications. For the case of Ta, an as-deposited Cu-layer on a Ta film displays an abrupt interface with Ta. However, a 3-nm thick amorphous layer is observed at the Ta/Cu interface after annealing for 1 hour at 400°C.[24] The mechanism for the amorphous layer

formation is not clearly understood because Ta has no equilibrium solubility with Cu at 400°C, and does not form compounds with Cu. Nevertheless, it is believed that the thin amorphous-layer improves the adhesion between Ta and Cu. In addition, the Ta film has the advantage of serving as a nucleation layer for Cu, enabling Cu films with (111)-orientation to be deposited.[25]

On the other hand, Ta barrier-layer films deposited by PVD are polycrystalline films. Thus, even though the reaction between Ta and Cu at 400°C occurs over only a short distance, the Cu may diffuse along the grain boundaries of the Ta film and escape into the neighboring dielectric material. Thus, by itself, Ta may not prove to be an adequate barrier-layer.

TaN is a related material to Ta, having a high melting point of 3087°C and a more dense microstructure. It is also typically deposited by PVD using a Ta target, but with nitrogen added to the argon gas in the sputtering gas ambient. Thus, as the reactive Ta is sputtered onto the wafer, it reacts with the nitrogen and forms a deposited film of TaN. The TaN material is more inert than Ta and does not react with Cu. In addition, TaN can be deposited as a polycrystalline film or an amorphous film by adjusting the nitrogen flow rate.

TaN films have been found to be better barrier diffusion-barrier films than Ta films. However, since TaN does not react with Cu, it exhibits worse adhesion behavior than Cu. For some Cu-CMP processes its adhesion properties are inadequate. To overcome the limitations that each film (i.e., Ta and TaN) exhibits when used alone, the implementation of bilayer films (TaN/Ta)[17] or trilayer films (Ta/TaN/Ta) have been investigated. These multilayer barrier-films reportedly offer better overall characteristics for Cu damascene interconnects.

TaSiN is another candidate for a Cu-barrier-layer material. It is formed by PVD of a TaSi target (with a low Si concentration) in an Ar/N_2 ambient. TaSiN is an amorphous refractory ternary material, and as such does not have any grain boundaries (which can serve as fast diffusion paths in polycrystalline films). It has been reported by Hara *et al.* that TaSiN exhibits better adhesion and barrier-layer characteristics than even TaN/Ta bilayers.[26,85] Furthermore, the same group reported that Cu-films deposited on TaSiN barrier-layers have a stronger (111) preferred orientation than do Cu films deposited on Ta or TaN layers.[27] A survey of over 75 papers from 1999-2000, indicated that TaSiN can maintain its diffusion-barrier properties for Cu to higher temperatures than Ta or TaN. However, since the Si containing ternary compounds (TaSiN and TiSiN) have higher resistivity, the Si industry tends to use TaN or Ta for Cu barrier-layers.[28]

16.3.7 Other Candidate Materials for Cu-Barrier Layers

Various other materials have been studied as Cu barrier-layers, and the most important of these (i.e., TiN, TiSiN, and WN) are discussed in this section. TiN is a good barrier-layer material for Al interconnects, and has been widely used for this application. While it also performs well as a barrier-layer for Cu, its behavior as an adhesion-layer and seed-layer for Cu is not as good as the Ta-based materials. CVD-TiN films have also been considered as Cu barrier-layer materials but they are reportedly less effective as diffusion barriers than PVD-deposited Ta and TaN films.[18,41] In addition, Cu films deposited on TiN have a weaker (111) structure than those deposited on Ta.[23] Hence, TiN deposited by either PVD or CVD has not been widely adopted for use in Cu interconnect metallurgy.[24] However, TiN deposited by ALD has been mentioned as a possible barrier-layer material for Cu. Work to deposit TiN by ALD has recently been reported.[29]

On the other hand, TiSiN-films deposited by CVD have been reported to exhibit significantly improved diffusion-barrier properties for Cu compared to TiN. The Si introduced into the TiN matrix bonds with the nitrogen (and this plays a role in producing improved barrier-layer behavior).[18] The TiSiN also has the additional benefit of good wetting to Cu, which has been correlated to improved Cu electromigration performance.[30] The CVD-TiSiN is deposited using TDMAT (tetrakis [dimethyl-amido] titanium) and silane as the precursors.[31,56] (Additional information CVD-TiN with TDMAT can also be found in Vol. 1, Chap. 6, Sect. 6.7.3.) The deposition of barrier-layers by CVD may become necessary for sub-0.1-μm technologies.

Tungsten nitride (WN) films deposited by CVD have also been considered as materials for Cu barrier-layers. Limited development of CVD-WN was described by Genus in 1999,[32] but they apparently decided not to commercialize this approach. Their decision may have been based on results indicating that such films are less effective than PVD-films of Ta or TaN as Cu diffusion-barriers.[18]

16.4 SEED-LAYER TECHNOLOGY FOR COPPER INTERCONNECTS

A *thin Cu seed-layer*, deposited by a method other than electroplating (e.g., PVD) must be deposited on top of the diffusion-barrier-layer.[9] This layer needs to be present to ensure that the subsequent Cu-electroplating operation can be performed successfully. Specifically, there are a number of requirements that the seed-layer must satisfy in order to enable the electroplating fill-process to be effectively carried out.[33] That is, if the seed-layer exhibits any one of a number of flaws, including poor morphology, inadequate sidewall- and bottom-coverage,

excessive overhang, or surface contamination, the electroplating results will be unsatisfactory (even if the best electroplating process is available). The details of the characteristics that must be possessed by the Cu-seed-layer are described as follows:

1. *Low-Resistivity* - At the start of the electroplating process, the seed-layer is the primary-conducting layer that carries current from the edge of the wafer to the center (allowing the plating current-source to contact the wafer only near the edge). Thus, the seed-layer should provide a low-resistance path from the edge to the center of the wafer, to make the current-density during the Cu-electroplating process uniform across the entire wafer (especially when the plating process is being initiated). Uniform-current density across the wafer-surface helps maintain a uniform plating-rate across the wafer.

2. *Film Continuity Along the Sidewalls and Bottom of the Trenches and Vias* - The electroplating process must also occur within the recessed features on the wafer surface. Since, the seed layer carries current from the top surface into the bottom of vias and trenches, the electrical path to the bottom of such features must be continuous and adequately thick. If there is insufficient seed-layer thickness at the bottom or sidewalls of the vias, this will lead to a slowdown in the plating-rate, which, in turn can cause a void during the Cu-fill process (Fig. 16-21a). To get adequate Cu-seed-

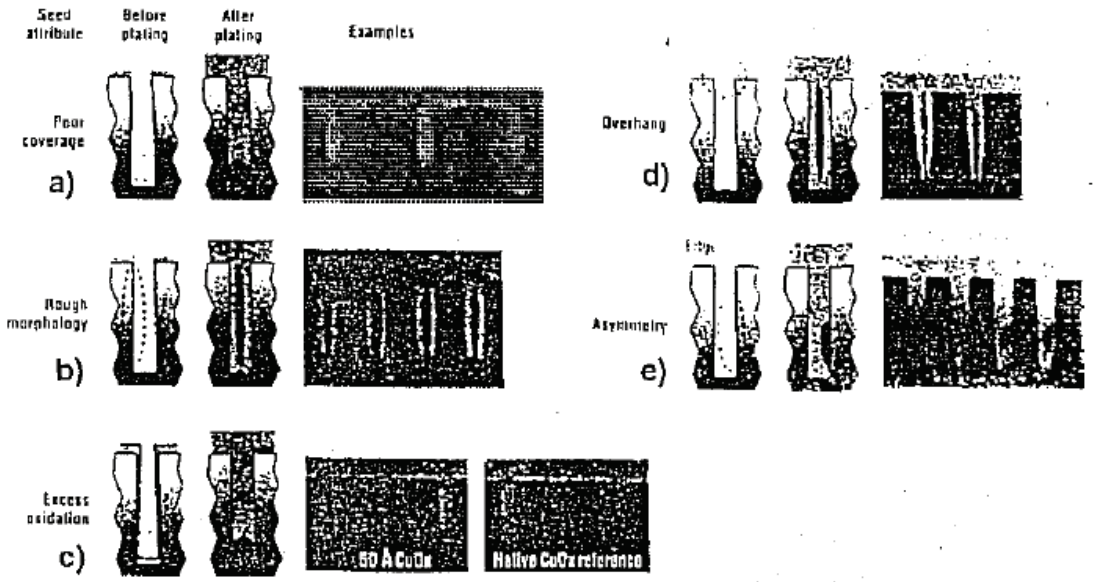


Fig. 16-21 Example set of illustrations and cross-sectional images showing key seed attributes and their impact on electroplating fill results.[33] Courtesy of Novellus Systems.

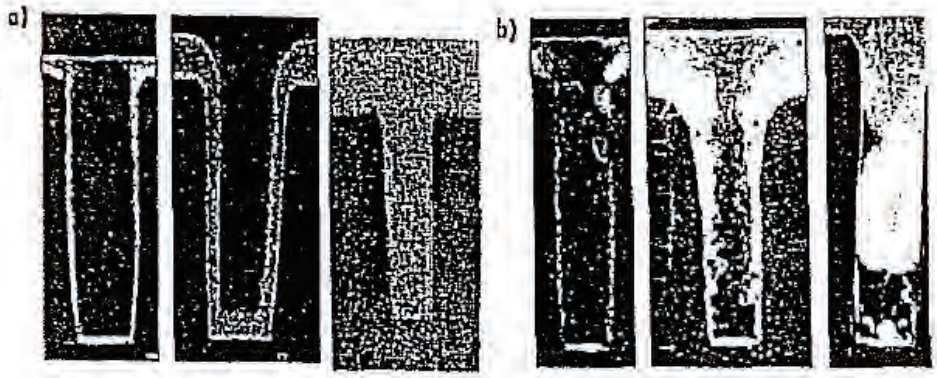


Fig. 16-22 Cross-sectional via images showing (left-to-right: a) Smooth seed coverage only, metal profile following partial fill, and complete fill by electroplating; b) Agglomerated seed coverage, metal profile following some electroplating, and resulting void at completion of electroplating.[33] Courtesy of Novellus Systems, Inc.

thickness along the sidewalls and bottom of deep-submicron, high aspect-ratio vias, either an ionized-PVD process or a CVD process is needed.

3. *Smooth Surface* - The seed layer must possess a smooth and continuous morphology. If the seed layer on the recess sidewalls is rough (or agglomerated) voids are also likely to form during plating. Rough sidewalls can lead to premature feature closure, and the discontinuity of an agglomerated film causes nucleation to fail at some locations. Either effect can give rise to creating a void (Fig. 16-21b). However, with a continuous, smooth seed-layer, plating readily occurs on all surfaces, thereby facilitating the subsequent as-desired, bottom-up filling (Fig. 16-22a).

It should be noted that the temperature of the wafer during the PVD process must be kept low to ensure that a smooth seed-layer surface is obtained (i.e., at room-temperature, or lower). Elevated wafer temperature during the seed-layer sputter-deposition process leads to a rough Cu surface (or even agglomeration, Fig. 16-23).

4. *Minimal Oxidation and Corrosion* – Oxidation and corrosion of the Cu seed-layer must be minimized both prior to plating and when the wafer is first immersed in the plating bath. Either effect can lead to seed-discontinuity and fill voiding. That is, the seed-surface must be largely free of oxide so that efficient charge-transfer can start to take place when the plating process is initiated. If the oxide on the PVD-Cu-surface is acid-soluble and sufficiently thin (i.e., in the thickness range of 10-20-Å), such initiation of the plating process is routine, and filling is complete.

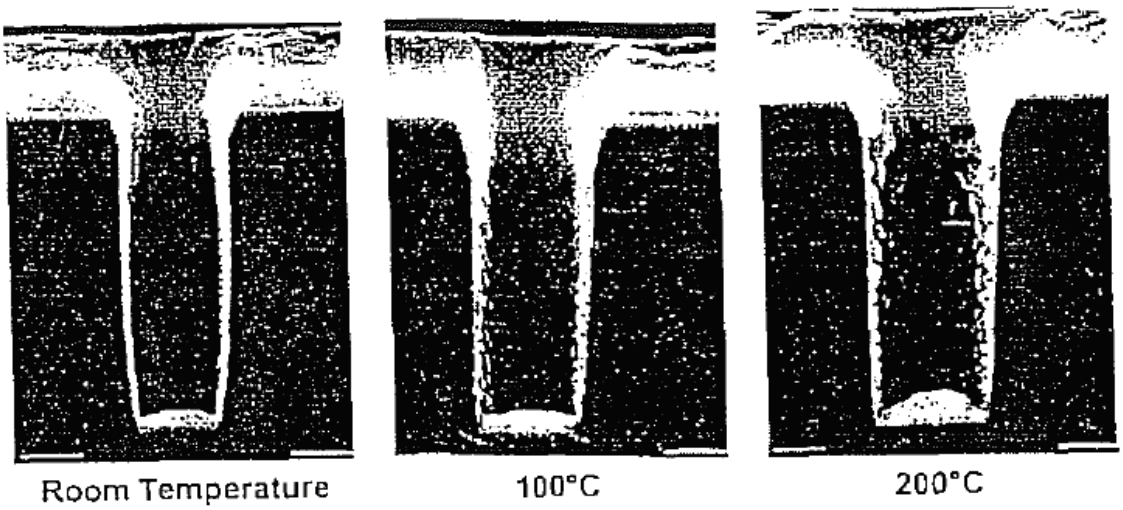


Fig. 16-23 Effect of wafer temperature on surface morphology of Cu deposited into vias by I-PVD. Courtesy of Applied Materials

However, when the cuprous oxides are thicker (i.e., 50-100-Å-thick), a greater degree of bottom voiding is observed (Fig. 16-21c).

5. *No Overhang at the Top Corners of Trenches and Vias* - Excessive overhang can lead to premature closure of the top of the recess during the plating process (Fig. 16-21d).

6. *No Asymmetry of Coverage Between Radially-Inward- and Radially-Outward-Facing Sidewalls* - Too thin a seed-layer on one side can impair nucleation and result in void-formation (Fig. 16-21e). That is, a much thinner seed-layer will form on that side than would be expected, and a void can thus result. This means that control of directional deposition during the PVD process must be maintained across the wafer plane from center-to-edge.

7. *Preferred Crystallographic Structure* – The crystallographic texture of the Cu seed-layer is influenced strongly by that of the underlying barrier-layer. This texture, in turn, determines the microstructure of the electroplated Cu deposited onto the seed layer. As noted previously, a (111)-orientation of the Cu is generally preferred because such films have the best overall film characteristics. A smooth Cu seed-layer-surface is also required to obtain this crystal orientation in the Cu layer that is subsequently deposited by electroplating.

In order to obtain a strong (111) Cu texture, it is also important that an *in situ* Cu-seed-layer deposition be performed. That is, the seed-layer

deposition process should be carried-out in the same cluster PVD-tool in which the barrier-layer PVD process is performed. This ensures that the surface of the barrier-layer (typically consisting of a reactive metal, such as Ta), does not form an oxide. If the barrier layer is exposed to air prior to Cu deposition, the resulting (111) Cu texture is much weaker than without an air break. This is probably attributable to oxidation of the barrier surface, and the resulting elimination of the pseudohexagonal surface that would otherwise allow for heteroepitaxial deposition.

Controlling the via shape itself is also important to ensure repeatable feature-filling. That is, the via-shape influences the filling by impacting the continuity of the seed-layer. Problems with the shapes of the vias often occur when trying to integrate an organic low-*k* ILD and Cu. That is, the via profiles of such materials frequently exhibit undercutting and notching as a result of the via-etch process (see Chap. 14). Such notches in the via sidewall profile can cause a discontinuity in the seed layer, and adversely impact the ability to completely fill the feature.

To get adequate seed-layer thickness into along the sidewalls and bottoms of high aspect-ratio recesses, either an ionized-PVD-Cu or a CVD-Cu process may be needed. (Ionized-PVD technology for such applications is discussed in Sect. 16.3, and additional information about secd-layer deposition with CVD is given in later paragraphs of this section.) As also noted in Sect. 16.3.2, the ionized-PVD process offers an additional advantage provided by the fact that a negative-bias can be applied to the wafer. This not only causes the ionized metal-atoms to preferentially strike the bottom of the via, but it allows the kinetic energy of these ions bombarding the wafer to be adjusted. For appropriate bias levels, this can cause re-sputtering from the feature base, which redistributes sputter-deposited material onto the lower sidewalls. The improvement in the conformality of the seed-layer thickness helps enhance filling of the recess without voids during plating. The impingement of the energetic ions on the top corners also helps keep the top opening from forming an overhang. Since the aspect-ratios of mid- and upper-level lines and vias is not likely to exceed 5:1, it is probable that ionized PVD techniques will continue to find use for these applications in future technology generations.

Unfortunately, as the feature sizes get smaller without a corresponding reduction in the ILD thickness, the aspect-ratio of the recesses gets larger. This is a situation that is likely to be encountered for the lower metal-levels of sub-0.1-μm interconnect structures. As noted in Sect. 16.3.1, for a PVD process the amount of material entering the recess must shared among all of the recess

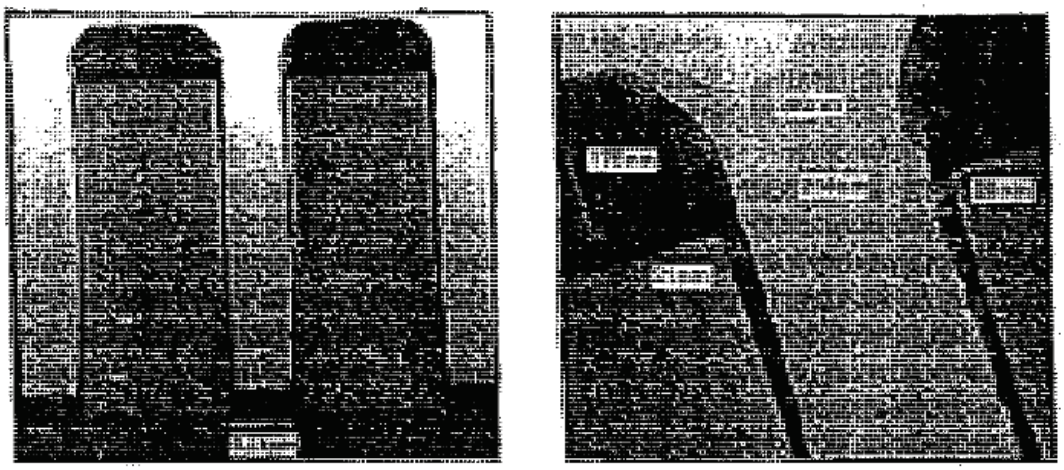


Fig. 16-24 SEM showing how thin the sidewall coverage is in high-aspect-ratio-features with I-PVD compared to the deposited layer thickness on the top surfaces of a wafer.

sidewall and bottom surface areas. For a via with an aspect-ratio of 5:1 the surface area of the sidewalls and bottom is 21x the surface area of the top opening. Thus, the material-thickness on all surfaces can only be 1/21 times as thick as the film deposited on the flat areas adjacent to the recess. Even at the 0.18-μm technology node, where the aspect ratio may be 5:1, a film-thickness of 2000-Å must be deposited in the flat regions to get a 200-Å-thick film on the via sidewalls (Fig. 16-24). Thus, at some time soon, it is predicted that PVD will no longer be capable of depositing an adequate seed-layer. Some have called the PVD-seed-layer the "weak-link" in the Cu-electroplating process.

At the point when the aspect-ratios for the lower levels of the interconnect system get to be too high to be adequately covered with any of the available PVD-processes (or if such a thick PVD-layer is needed that it causes too great an overhang), it may be necessary to switch to a CVD-Cu seed-layer. Some predict that this will occur for either the 100-nm or 70-nm CMOS nodes. Figure 16-25a shows a CVD Cu seed-layer on a CVD barrier layer deposited in a high-aspect-ratio recess that is only 0.14-μm wide. The CVD-Cu seed-layer is totally conformal and 500-Å-thick. The remaining 300-Å-wide slot was filled by electroplating without voids.[32] Figure 16-25b shows another example of a CVD-Cu seed-layer that is conformally deposited over high-aspect-ratio features.[34] One concern of using CVD-Cu seed-layers is the strength of their adhesion to the barrier-layer materials. Many freshly deposited barrier-layers (especially Ta-based), have highly reactive surfaces. These surfaces can degrade organic compounds resulting in interfacial contamination that prevents good adhesion.

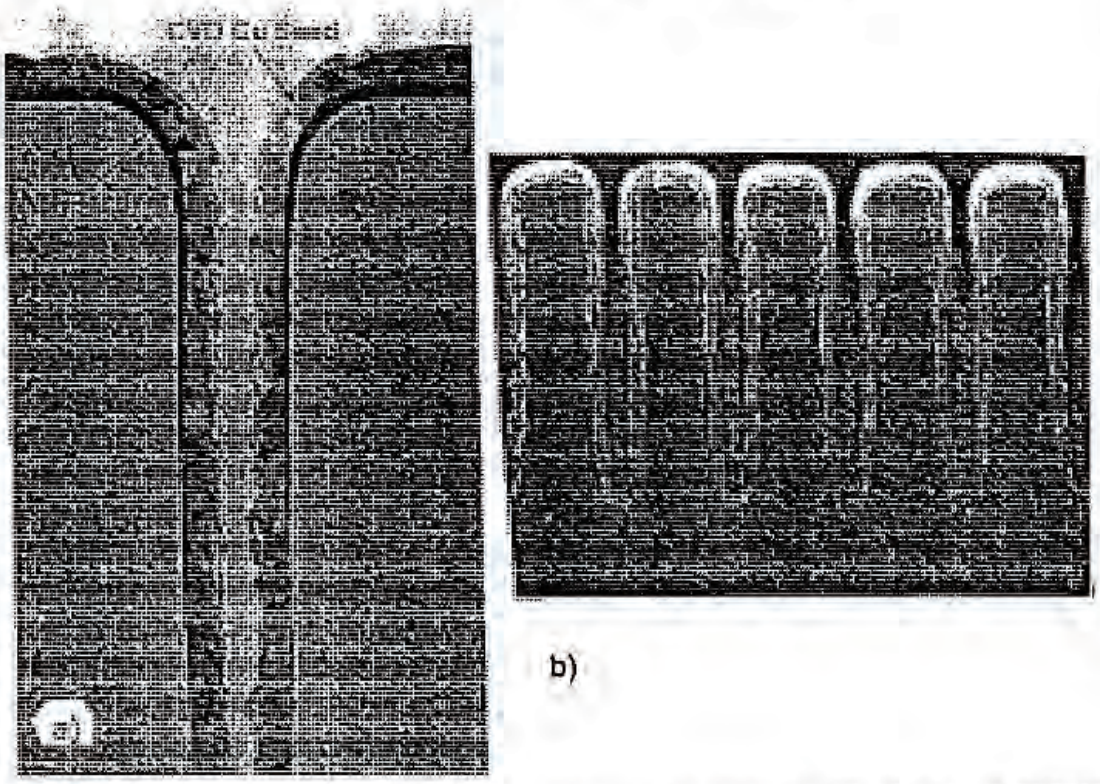


Fig. 16-25 a) SEM showing a CVD-Cu seed-layer integrated with clean and PECVD tungsten nitride (WN) barrier. (Courtesy Genus) b) SEM showing CVD-Cu seed-layer. (Courtesy Applied Materials)

This problem will need to be resolved by optimizing the barrier-layer choice, and improving the CVD deposition-process and precursor chemistry.

Two other alternative replacements for the PVD-Cu seed-layer have also been reported. The first is an electrochemical deposition process developed by Semitool. In their method, an ultra-thin Cu-layer (e.g., 200-Å-thick) is first deposited by PVD for adhesion purposes. This ultra-thin Cu-layer is then conformally thickened to provide a full-thickness seed-layer with an electrochemical process. Finally, the electrochemical process is switched over to a conventional electroplating chemistry to complete the feature fill.[35]

The other alternative replacement technology for PVD-Cu seed-layers is atomic-layer deposition (ALD - see Chap. 4, Sect. 4.6.1 for more background information on ALD).[36] ALD is a monolayer, stepwise growth-process that exploits the binding-energy-difference between chemisorption and physisorption. The growth proceeds by exposing the substrate-surface alternately to the precursors, and following each reaction, the excess species and by-products are removed with a nitrogen-pulse. The reactant-source pulses are thus separated by

a pulse of an inert gas. As a result, ALD is capable of depositing conformal-films with thicknesses that can be determined accurately by the number of deposition-cycles. The Cu precursor chosen for the work in Ref. 36 was $Cu(hfac)_2 \cdot XH_2O$ (see Sect. 16.5.3), and methanol, ethanol, and formalin were examined as the reducing agents. Smooth, conformal, near-bulk-resistivity Cu seed-layer films of 20-nm thickness with good adhesion were deposited by this ALD method.

As noted in Sect.16.3, another reported problem involving the PVD-seed-layer is its selective adhesion to barrier-layer materials. That is, the PVD-Cu seed-layers adhere well to Ta, but less well to TaN. Cu-films used with TaN barrier-layers can peel away during the later Cu-CMP process. Unfortunately Ta is not as good a diffusion barrier as TaN. To overcome this problem, use of a Ta/TaN bilayer has been reported.[37] Even better adhesion by the PVD Cu seed-layer was obtained by using a TaSiN barrier-layer.[26]

16.5 COPPER INTERCONNECT FILM DEPOSITION

Copper films for IC interconnect applications can be deposited in a number of ways, including: electrolytic-plating; electroless-plating; chemical-vapor-deposition (CVD); and physical-vapor deposition (PVD). Electrolytic-plating of Cu is the dominant technique used to fill recessed damascene features in deep-submicron ICs. Therefore, it is discussed in the greatest detail here (i.e., in Sect. 16.5.1). Electroless-plating of Cu and CVD-Cu are described more briefly in Sects. 16.8.2 and 16.5.3, respectively. PVD technology of Cu-films (used primarily to deposit Cu-seed-layers prior to the electrolytic-plating filling process), was covered in Sect. 16.4.

16.5.1 Electrolytic-Plating of Copper

As described in Chap. 15, dual-damascene processes have been developed for the multi-level Cu-interconnect structures of deep-submicron ICs. The damascene approach avoids the problems associated with having to etch Cu-films, but a challenge of a different sort arises from the need to fill the high-aspect-ratio damascene recesses without voids in the metal.

Line-of-sight PVD techniques are inadequate for filling such features because accumulation of deposits at the upper-corners of trenches closes-off the opening of the recess and this leads to void-formation. Chemical-vapor-deposition (CVD) has the potential for conformal coverage, but numerous chemical and hardware issues have hampered development of a manufacturable fill process for Cu. Thus, to the surprise of many in both the semiconductor and plating industries, electrolytic-plating has emerged as the deposition method of choice for

damascene-Cu interconnects (although advanced sputtering techniques are still needed for depositing thin barrier- and seed-layers, as described in Sects. 16.3 and 16.4) before the filling of the recesses with a Cu-plating process.

Electrolytic-plating (also often called *electroplating*) and *electroless-plating* of metals are older technologies still used in a variety of industries including hardware, auto, and electronics. In the electronics industry these processes have been (and still are) widely used in the printed-circuit-board (PCB) industry to plate copper layers on both sides of plastic boards for PCB interconnects. By the late 1990's, work was begun to develop electrolytic-plating processes for IC fabrication technologies (especially for depositing copper films for IC interconnect structures).

The key reasons why development of electroplating processes for on-chip metallization was pursued stemmed from the future requirements of IC interconnects. That is, it became clear that in deep-submicron ULSI technologies, low-temperature, low-cost, metal-deposition-processes compatible with thermally fragile low-k dielectrics would be needed. Electroplating of metals (especially Cu) represented a very attractive process for such applications. This is because deposition of Cu by metal-plating offers a high deposition-rate process (0.5-1.0-μm/min), which in turn permits high-throughput (50 wafers/hr), and thereby also a potentially much lower cost-of-ownership than PVD or CVD. In addition, such processes can be carried out at temperatures $\leq 40°C$, and a well-tuned electroplating-process can fill high-aspect-ratio structures without defects (i.e., trenches with aspect ratios of 5:1 can be completely filled). The details of electroplating copper-films for deep-submicron ICs are described here. As noted earlier, since electrolytic-electroplating requires a closed-conducting-path for continuous electric current, it is not possible to plate directly onto an insulator such as SiO_2. In this case, an electrically-conducting contact-layer must first be deposited (i.e., in Cu electroplating, a low-resistivity Cu seed-layer is deposited on the insulator).

Electroplating processes utilize conductive solutions containing ions of the metal to be deposited.[13,38] In the case of copper, the solution typically contains copper sulfate ($CuSO_4$ in the concentration range of 17.5-60g/L), sulfuric acid (H_2SO_4 in the concentration range of 45-325 g/L), and water. The $CuSO_4$ in this acidified solution breaks up into Cu^{2+} (cupric-ions) and SO_4^{2-} (sulfate-ions). The sulfuric acid added to the solution plays several roles: a) it dissolves any native copper-oxide on the copper seed-layer, so that the copper underneath is exposed, allowing the plating process to be reliably initiated across the entire wafer;* b) it

* Both CuO and Cu_2O (cuprous oxide) are unstable in a strongly acidic medium (pH < 3).

increases the conductivity of the electrolytic solution; and c) it improves the wetting capability of this solution. In commercial processes, proprietary organic additives are also added to the plating solution in order to improve the filling of recessed features by the deposited Cu film (as is described in more detail in Sect. 16.5.2.3). As with other types of deposition, the success of the process is measured by factors such as: deposition-rate, film-morphology, trench- and via-filling-capability, deposited-film-thickness-uniformity across-the-wafer and from wafer-to-wafer.** Plating chemistries were evolved from the plating bath formulations used in the PCB industry, and these were made to work for semiconductor applications.

In the Cu-electroplating-process for on-chip interconnects, wafers with a conductive top surface (which in this case is a PVD-Cu seed layer, as mentioned earlier) are immersed face-down in the electrolytic solution. The wafer-surface is also electrically connected to the negative-side of an external dc power-supply (by a conducting sealing-ring at the outer edge of the wafer, see Fig. 16-26a[50]), making it the *cathode* in this circuit. Electric-current is passed so that a chemical reaction occurs at the wafer surface which causes Cu plating to occur (with the cupric-ions that carry electric charge constituting the current in the electrolytic solution, see Fig. 16-26b). That is, as the positive cupric-ions arrive at the negative-biased-cathode they adsorb on its surface. They also acquire two electrons and are reduced to Cu metal (which plates out), with the reaction for Cu being:

$$Cu^{2+} + 2e^- \rightarrow Cu^0 \qquad (16.6)$$

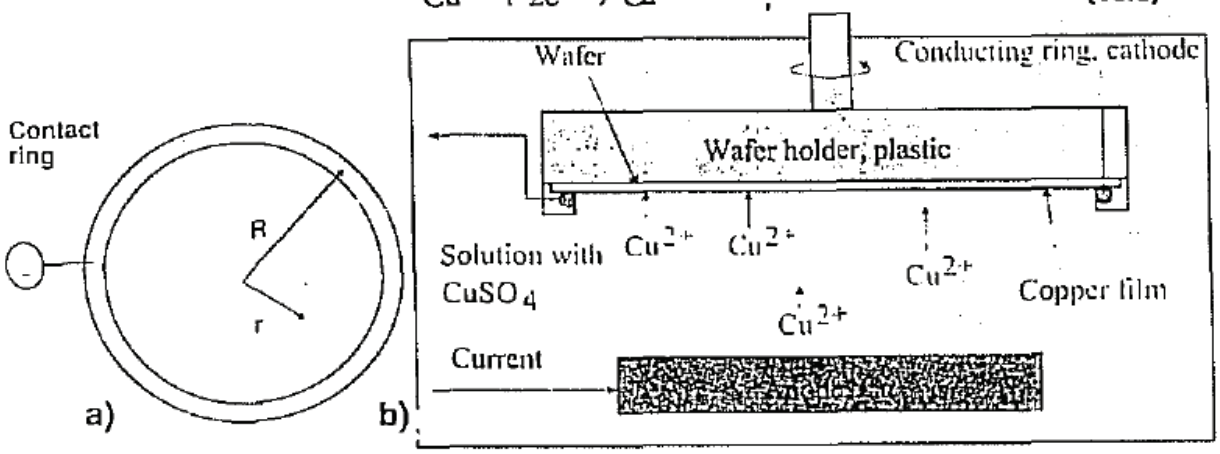


Fig. 16-26 a) Contact ring of a copper plating cell. b) Copper electroplating process.

** While the semiconductor industry has predominantly deposited Cu films electrochemically using acid baths, alkaline baths have also been considered. This is because alkaline baths can plate directly onto TiN diffusion barriers without a copper seed-layer and without the need for organic additives. On such bath studied by Semitool used an Cu-EDTA-complexed bath and the deposited Cu exhibited superior adhesion to TiN.[49]

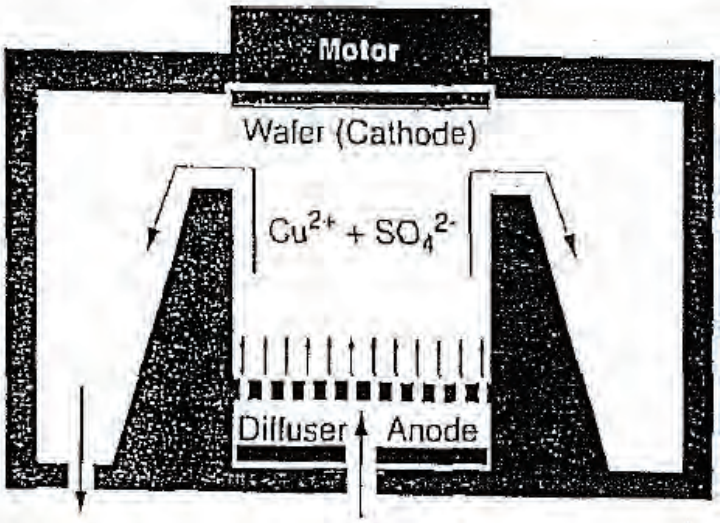


Fig. 16-27 a) Schematic drawing of a cross-section of a Cu electroplating cell.

Another electrically active surface (a solid-copper electrode which serves as the *anode*), is present in the conductive solution to complete the electric circuit (i.e., it is connected to the positive side of the power-supply). At the anode, the oxidation reaction occurs that balances the current flow at the cathode, thus maintaining the electrical neutrality in the solution. That is, all of the cupric-ions removed from solution by the plating of Cu onto the cathode are replaced by dissolution from a solid-copper anode immersed in the solution. (The copper atoms of the anode are oxidized to form the Cu^{2+} ions. These ions are released from the anode and dissolve into the solution.) The metal-ion formation-reaction is thus sustained by passing an electrical current through the bath.

If the applied potential is reversed, the cathode and anode are also reversed, and so are the locations where Cu deposition and dissolution occur. Alternating between deposition and dissolution (a process known as *pulse-plating*) may be useful in order to improve the filling capability of a plated film (see Sect. 16.5.2.5 for more details on pulse-plating). To enhance the uniformity of deposition across the wafer, the electrolyte is usually pumped against the wafer during the plating process, and the wafer may also be rotated at the same time (see Fig. 16-27).

The electroplating rate is a direct function of the current-density. To obtain uniform deposition across a wafer, the plating cell must be designed to give uniform potentials across the wafer surface. (Because of concerns about Cu contamination, it is also critical that the plating cell be designed so that no copper is deposited on the backside of the wafer.) The current delivered to a conductive surface is directly proportional to the quantity of metal deposited (according to

Faraday's Law of Electrolysis). By using this relationship, the mass deposited can be readily controlled through variations of plating current and time.

If no potential is applied across the interface between a metal and the electrolytic solution, an equilibrium potential exists between the two (see Fig. 16-28). Once the potential is shifted by an external power supply away from the equilibrium potential, a current will be driven across the interface. Under conditions typical of most plating processes, this current-flow is approximated by an exponential relationship known as the *Tafel equation*.

Figure 16-28 shows a current-vs.-potential curve of a copper electroplating process (also referred to as a *Tafel plot*). As the potential is increased from equilibrium to more negative values, the current increases in an exponential manner (Tafel region). In this region of operation the overall deposition-rate is determined largely by reaction-rate kinetics at the cathode. Ordinarily, conversion of Cu^{2+} ions to plated Cu-atoms occurs with nearly 100% efficiency, so that the electric current-density is a direct measure of the plating-rate. This also implies that ohmic effects (primarily through the uniformity of current distribution) are one of the important parameters that impact plating-thickness uniformity across a wafer. The strong dependence of current upon potential implies that the plating cells must be designed to enhance the uniformity of the potential across the wafer surface. The resistance of the barrier/seed-layer also impacts such ohmic effects, as will be described in later paragraphs.

As the potential continues to increase, mass-transfer effects become predominant and a limiting current-plateau is reached. This occurs because the Cu^{2+} species reacting at the cathode no longer reach the interface at a rate

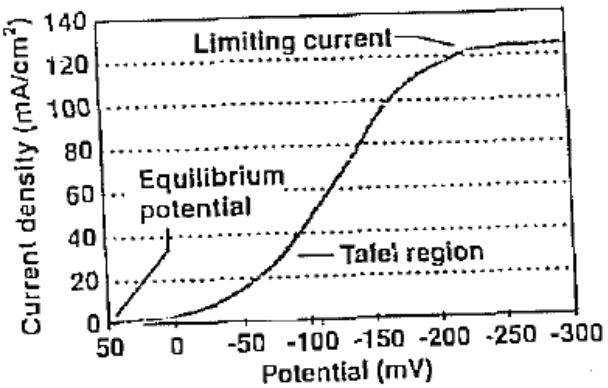


Fig. 16-28 A current-potential curve (Tafel plot) typical of a copper electroplating process.

sufficient to sustain the high reaction-rate. As a general guideline, Cu plating processes are operated at currents no greater than 30-50% of the limiting current to avoid undesirable film-deposition-characteristics.

To ensure that the rate of mass-transfer of Cu species to the interface is large compared to the reaction-rate (and that the mass-transfer rate is uniform across the wafer), the rates of migration, diffusion, and convection must be well understood and controlled. Convection is the most important mode of mass transfer in a liquid, and can vary from stagnant to laminar (or turbulent) flow. (See Vol. 1, p. 72 [Chap. 3] and p. 156 [Chap. 6] for a discussion of turbulent and laminar flow, and p. 157-8 [Chap. 6] for a discussion of mass-transfer in regions where movement by convection is zero [i.e., in stagnant regions].) In some electroplating systems, the wafer is also rotated while it is immersed in the electrolyte to get a uniform boundary-layer and to ensure that the fluid mechanics and material balance for the copper-diffusion match well everywhere on the wafer.

Generally, electroplating-processes have high deposition-rates (e.g., 400-1000 nm/min) and form desirable metallurgy. However, as noted in Sect. 16.4, electroplating requires a thin seed-layer of Cu (about 50-nm-thick) which must first be deposited by some other method (e.g., CVD or PVD) onto the barrier/adhesion-layer. As noted in Sect. 16.4, the characteristics of the seed-layer are another of the critical factors that impact the ability to completely-fill recesses by electroplating without voids or other defects. It would be ideal if a single layer could serve as both a barrier- and seed-layer, but currently two materials are used. An adhesion/barrier layer like Ta, TaN, Ta/TaN or TaSiN is first deposited, with a Cu seed-layer applied in a second step.

Note that to obtain uniform deposition across the entire wafer, the current-density should be equal at all locations. Since the conducting layer has some resistance, there will be a voltage-drop from the edge-to-center of the wafer. This means that the edges of the wafer will plate faster than the wafer center (and the deposition-rate non-uniformity is greatest at the initiation of the electroplating process). The problem also gets more severe as the wafer diameter is increased, and thus is worse for 300-mm- than for 200-mm-wafers. To minimize such deposition-rate non-uniformity-effects, the voltage-drop from edge-to-center must be kept tolerably small. Since the barrier-layer is very thin and typically has a relatively high resistivity, a Cu seed-layer must be formed on top of the barrier-layer to increase the conductivity to a sufficiently high degree. (It has also been reported that reverse pulse-plating [as discussed later], will help reduce the non-uniformity of deposition, as the copper also dissolves more rapidly during the part of reverse-plating cycle in places where it is most heavily plated.[25])

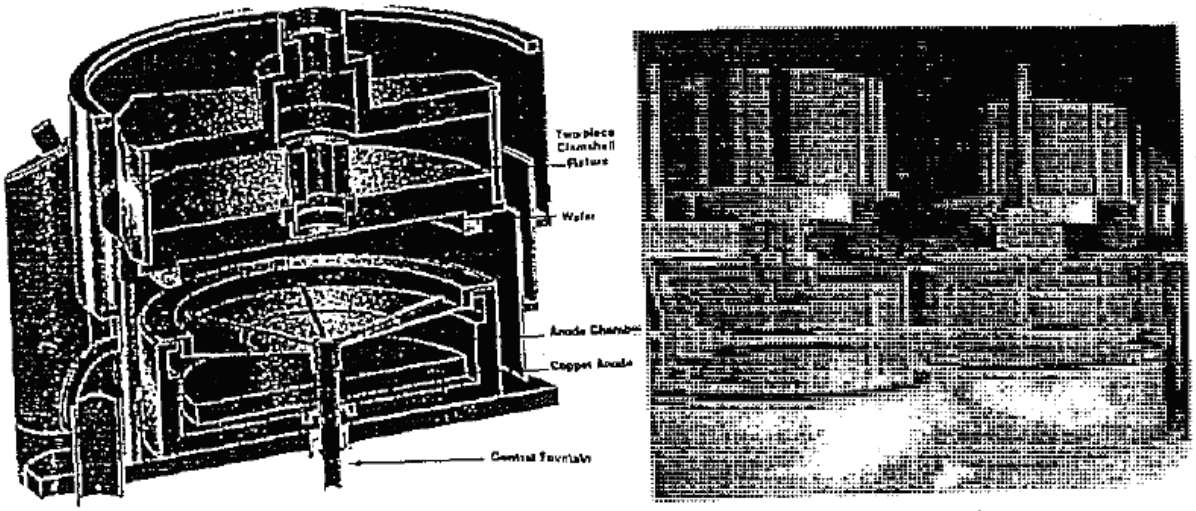


Fig. 16-29 Schematic drawings of plating cell of the Novellus SABRE™ electroplating tool. Courtesy of Novellus Systems.

Copper electroplating tools are commercially available from a number of companies. In 2000 the three companies with the largest market share were Novellus Systems, Semitool, and Applied Materials. The market for copper tools is expected to continue to grow significantly in the decade starting at 2001. Figure 16-29 shows illustrations of the Cu electroplating cell in the Novellus Systems SABRE™ elecroplating tool. Figure 16-30 is a drawing of the overall SABRE tool layout. This electrofill system enables void-free copper in extremely narrow high-aspect-ratio structures (up to 10:1) without backside contamination. Throughputs of up to 75 wafers are possible, with manufacturing availability above 90%.

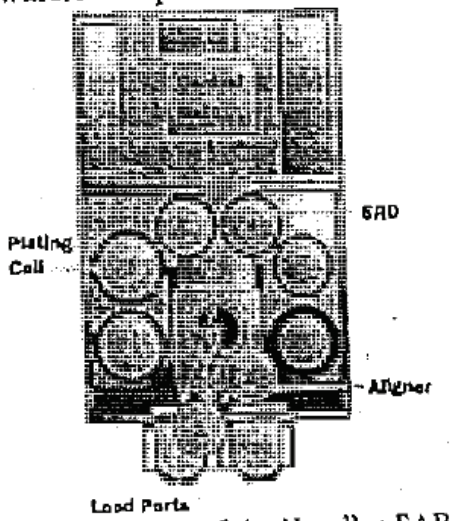


Fig. 16-30 Schematic drawing of the layout of the Novellus SABRE™ electroplating tool. Courtesy of Novellus Systems.

Figure 16-31 depicts the Applied Materials Electra™ Cu electroplating system. It has dual process cells and dual wafer handling blades for high throughput and reliability. Wafers are loaded into the twin plating cells. After plating they are processed in an edge-cleaning station followed by a spin-rinse-dry. The wafers are then moved to the integrated anneal chambers before being delivered back to the cassettes.

16.5.2 Filling High-Aspect-Ratio Recesses Without Voids Using a Cu Electrolytic Electroplating Process

The deposition process of Cu films for dual-damascene IC interconnects must not only have a high and uniform deposition-rate across the wafer, but it must also be capable of producing *void-free fill* of high-aspect-ratio damascene features. There a number of factors that impact the ability of the Cu electroplating process to fill such high-aspect-ratio recessed structures without the formation of voids or seams, including: 1) the incoming feature profile; 2) barrier-layer and seed-layer attributes; and 3) key electroplating process parameters that are optimized to encourage acceleration of deposition *near* the base of damascene features ("bottom-up" fill). That is, for deep-submicron IC applications an accelerated plating rate is desired within submicron recessed features to achieve void-free fill. The attributes of the barrier- and seed-layers needed for permitting void-free fill were discussed in Sect. 16.3 and 16.4, respectively. The void-free filling factors involving the electroplating process are considered in the following subsections. They include: 1) the electroplating solution must provide *rapid uniform wetting of the wafer surface*, especially within the area of high-aspect-ratio features; 2) the deposition process must be *properly initiated*; 3) appropriate

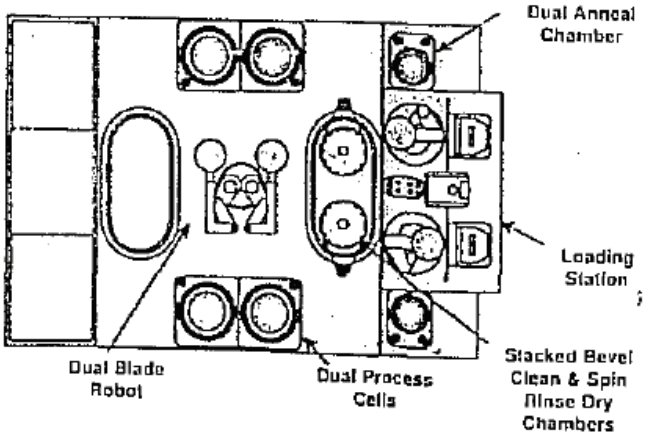


Fig. 16-31 Schematic drawing of the layout of the Applied Materials Electra™ Integrated ECP system. Courtesy of Applied Materials.

organic additives are added to the electroplating baths to facilitate void-free filling; and 4) the *electric-current* waveform used for the deposition process can impact the filling characteristics. Topics 1) and 2) involving the electroplating process itself are described in Sect 16.5.2.1, while the effect of additives in the plating bath is covered in Sects. 16.5.2.2 and 16.5.2.3. The topic of plating-bath chemistry-management is discussed in Sect. 16.5.2.4 and the subject of pulsed plating in Sect. 16.5.2.5. The issue of topography in plated Cu-layers is addressed in Sect. 16.5.2.6.

16.5.2.1 The Electroplating Process Sequence: The electroplating process sequence typically consists of the following events. The wafers are immersed face down in the plating cell containing the electrolytic bath. Adequate surface wetting prior to initial deposition of the copper must occur. That is, rapid uniform wetting of the wafer surface and the inside of the high-aspect-ratio features is another critical requirement for good filling performance. For a solution with good wetting properties, capillary action should force solution into submicron features. Wetting is highly dependent on plating chemistry, seed-layer structure, composition, surface oxidation, and contaminants, as well as how the wafer is immersed into the plating bath and how the plating is initiated.[39]

The time the wafer is being wetted is referred to as the "zero-current-induction- time," or "induction-time" (see Fig. 16-32). During this time no external potential is applied, and thus no cathodic current flows. It represents the time period over which the wafer is exposed to the plating electrolyte at an uncontrolled (floating) potential. If the wafer is not polarized during this time, copper corrosion and cuprous oxide dissolution can occur.[40] To avoid the possibility of corrosion during the initial immersion period, the wafer may instead be introduced into the bath with an immediate small current flowing (or with a very-small cathodic potential being applied). In this way these corrosion reactions can be reduced or reversed. Immersing the wafers while applying a small cathodic potential is applied is thus referred to as *cathodic protection*.[33]

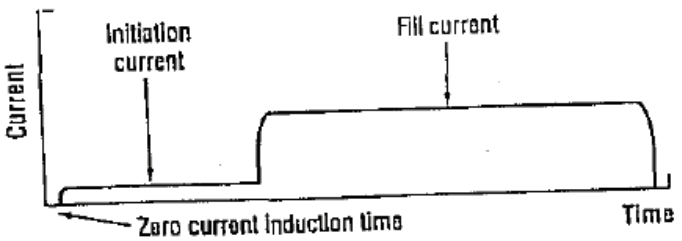


Fig. 16-32 Example of a two-step dc plating-waveform with three key features identified.[33]

After the induction time has passed, an "initiation current" is applied to the wafer (see again Fig. 16-32). During this time the current level is small, the deposition rate is slow, and the deposition occurs conformally. Thus, the seed-layer thickens uniformly during this period (allowing it to become continuous if a discontinuity in its coverage had occurred). In some cases, a thicker Cu seed-layer is needed to support the nucleation and high current density required in the subsequent fill step.

At the conclusion of the initiation-current stage, the "fill-step" stage is performed (Fig. 16-32). Since the damascene recesses are typically about 0.5-μm deep, and the Cu deposition-rate during the electrofilling stage of the plating process is ~1.0-μm/min, the fill-step lasts 20-40-sec, depending on bath conditions and feature size. Smaller features tend to fill earlier because they develop the conditions necessary for bottom-up fill (as will be described in the next sections). Since the larger features (i.e., with aspect-ratios much smaller than unity) fill more slowly, the copper layer must typically be deposited to a thickness somewhat greater than that of the damascene dielectric film stack. This issue will be considered in more detail later.

16.5.2.2 Additives in Plating Baths for Damascene Cu Interconnects: As noted earlier, the most important challenge for depositing Cu-films by plating for damascene applications is to completely fill vias, holes, trenches, and their combinations without voids or seams. For such defect-free filling, a higher deposition-rate along the sides and the bottom of the feature is needed than on the flat, top-wafer surface. As early as 1990 at IBM, it was discovered that plating from certain electrolytic solutions that contain additives leads to this kind of "bottom-up" filling. Since that time, work has continued to understand the role of such additives in plating solutions that will give the best filling results.

Additives to electroplating baths were originally developed to enhance the brightness and hardness of metallic deposits. Both of these attributes can be related to decreased grain-size in the deposits and reduction (or leveling) of the topography over rough surfaces. When the Cu-plating for submicron IC interconnects was being developed it was also observed that without additives the grain-size of the Cu was large (roughly 1-μm) and the film surface was rough. Additives led to a much finer grain-size (<1000-Å) and a smooth, highly reflective surface that had significantly less topography than the underlying substrate. From that work it was surmised that perhaps additional additives could be found that would also permit the damascene recesses to be filled without voids. It was discovered that a number of different additives (i.e., four or more)

need to be added to the plating solution to give the best filling results. (It also appears that the effects of these multiple additives are interactive.) Here such additives and their characteristics are described. In the next section a model is given for how these additives allow recesses to be filled from the "bottom-up."

The plating baths, as formulated by a number of specialty suppliers, include low concentrations (in the millimolar or micromolar range) of a variety of organic agents (generically referred to as *additives*) which serve to modify the properties of the deposited films as compared to results otherwise obtained from additive-free copper-sulfate/acid solutions. The manufacturing processes generally employ Cu plating-solutions purchased commercially, and the identity of the additives is proprietary. The examples of the various additives listed in this section come from the patent literature, as well as from various research articles that describe the additives being investigated. At the outset of the discussion the various classes of plating-bath additives are described. These classes include: suppressors, accelerators, and levelers (which, as will be noted, are also referred to by other names).

Accelerators are additives that speed or enhance deposition of the Cu during electroplating. Accelerators adsorb strongly on seed surfaces, where they also compete with suppressors for surface adsorption sites. Accelerators are usually unsaturated small organic molecules containing a polar sulfur, oxygen, or nitrogen functional group. In addition to increasing the Cu deposition-rate, they promote dense nucleation (and consequently the growth of films with fine grains). As a result, accelerators are also referred to as *brighteners* (i.e., because they cause the Cu-film to have a smooth, well-textured [i.e., bright] finish). Accelerators are added to plating baths in low concentration-levels (1-25 ppm). Examples of accelerators reported in the literature include: bis(3-sulfopropyl)disulfide (SPS);[41] sulfonated acetylthiourea;[42] 3-mercapto-1-propanesulfonate (MPSA);[33] and dibenzyl-dithio-carbammate.[43]

Suppressors (also called *inhibitors*, *carriers*, *surfactants or wetting-agents*) are plating-bath additives that reduce the Cu plating rate. Suppressors are usually present in the plating bath at higher concentrations (200-2000-ppm), so that their concentration at the interface is not strongly dependent on their rate of mass-transfer or diffusion to the wafer surface.

The suppressors are generally polymeric surfactants (with 1500 molecular-weight polyethylene glycol [PEG] reportedly being the most widely-used material for this purpose).[33] The suppressor molecules slow down the deposition-rate of Cu by adsorbing on the wafer surface, where they form a monolayer which offers a diffusion barrier to the cupric ions. In addition, they enhance the

cathodic polarization needed for fine-grain structure. In order for the polyethylene glycol to act as an effective suppressor, however, chlorine ions must also be added to the bath (in the form of HCl), and thus the chloride ion Cl^- can be considered a co-suppressor. In fact, HCl is the only non-organic additive in plating solutions.[33]

The role of chloride ions should be discussed in more detail. Chloride ions are believed to adsorb on the wafer surface, where they attract PEG molecules, and cause them to be more strongly adsorbed there as well. In fact, it has been observed that low chloride concentration leads to weak or negligible PEG adsorption.[44] This effect results in poor suppression of current at the wafer surface, and therefore the deposition on the wafer surface is not suppressed (which impairs the desirable condition of a differential deposition-rate between the top surface and inside the recesses that assists the complete filling of recesses). On the other hand, as the Cl^- concentration is allowed to increase to the appropriate level, good suppression of Cu deposition on the wafer surface is obtained. The levels of Cl^- within the recesses, however, might not be sufficiently high to maintain a high level of PEG adsorption, allowing accelerator molecules to be adsorbed on those surface instead. Thus, the conditions for good "bottom-up" filling are not disrupted. But too high a concentration of Cl^- concentration can interfere with "bottom-up" filling, because Cl^- can compete successfully with the accelerator molecules for surface sites in the recesses, and subsequently attract suppressing PEG molecules within the recess.

Levelers are plating-bath additives used to reduce surface roughness and to smooth the topography of the plated-film. They perform this role by locally reducing the Cu deposition-rate (i.e., by adsorbing on surface sites, blocking them from serving as locations where cupric ions can plate). Hence, levelers can also be considered as kin to suppressor additives. However, unlike suppressors, levelers are added to plating baths only in low concentrations (i.e., 1-25 ppm, comparable to the accelerator concentration). Hence, unlike suppressors, the concentration of levelers at the interface is mass-transport dependent and the large size of the leveler molecules also impedes their migration into trenches. In these ways, isolated locations (such as the insides of recesses, where mass-transport is limited) have fewer leveler species adsorbed on their surfaces. Thus, these regions experience less deposition-reduction effects caused by leveler adsorption than do the top, flat-surface regions of the wafer.

On the other hand, protruding-surfaces on wafer flat areas, or trench corners (to which mass-transport by diffusion or surface-migration is more efficient), experience a greater suppression of the deposition-rate. That is, the number of

leveler molecules adsorbed on wafer surfaces is continually being depleted during the plating process – by incorporation into the film or by electrolytic decomposition. The suppression of plating at protrusions on the wafer flat-areas and the upper corners of the trenches is maintained by constant replenishment of the leveler species that block surface sites (provided the leveler is present at the correct concentration). The suppression of the growth of protrusions on the wafer surface results in a more *level* film surface on the top-surface of the wafer (hence the name *leveler*). The suppression of deposition on the corners of the trenches allows the tops of the recesses to remain open while the "bottom-up" filling proceeds to completion. Examples of leveler additives listed in the literature are Janus Green B (JGB) a kind of dye,[45] and quatenary ammonium salt (QAS).[43]

For Cu-plating in dual-damascene applications, the composition and concentration of accelerators and suppressors are selected such that the accelerator surface-concentration dominates that of suppressors on the interior surfaces of trenches and vias. Local deposition-rates are thus suppressed on the flat surfaces on top of the wafer relative to the insides of recessed features, leading to the desired "bottom-up" deposition and void-free metal filling.

As a result, when these various additives are correctly balanced in the plating bath solution, it is possible to get a correct fill without voids and seeds. However, high precision is required to balance the roles of these materials. This topic is addressed in Sect. 16.5.2.4, where the issues of managing plating-bath chemistries are covered.

It should also be noted that in addition to void-free filling of recesses by Cu layers, it is also desirable to end up with a smooth (not rough) and topography-free (i.e., level) film after plating. This will allow a CMP process without an extensive overpolish time (i.e., that would be needed to get rid of the high spots that accompany rough and non-planar films) to be used. If an extensive overpolish time is required, this will lead to more severe dishing & erosion problems. Adding levelers to the plating bath and employing reverse-pulse plating waveforms are measures used to try and planarize Cu-layer topography.

16.5.2.3 Models of How Additives in Plating Baths Impact "Bottom-Up" Filling:

A model that describes the effects of the plating-additives on the "bottom-up" filling of damascene recesses was published by Reid in 2000.[33] Reid's model is based on a large volume of experimental work, and it proposes that the mechanism for "bottom-up" fill is primarily due to the accumulation of *accelerator species* within damascene features. It supercedes a model by

Andricacos in 1998 that held that the "bottom-up" filling was driven primarily by the establishment of a gradient of *suppressor polymers* in the recesses.[46]

Before presenting the mechanisms that Reid proposes are responsible for "bottom-up" recess-filling, a few preliminary remarks are in order. First it has been observed by others that if the co-suppressors PEG and Cl^- are the only additives present in a plating bath, feature filling is poor. If PEG and Cl^- and an accelerator (e.g., SPS) are the additives present, better filling is seen to occur. If all three additive types (i.e., suppressor, accelerator and leveler) are present, the best filling results are obtained.[47] Furthermore, the concentrations in the bath must be correctly balanced to obtain the best filling effects. This implies that the effects of the three types of additive are interactive, as is also proposed in Reid's model.

Figure 16-33 shows the possible ways for a profile of plated copper to evolve during plating. In conformal plating, a deposit of equal thickness at all points of a

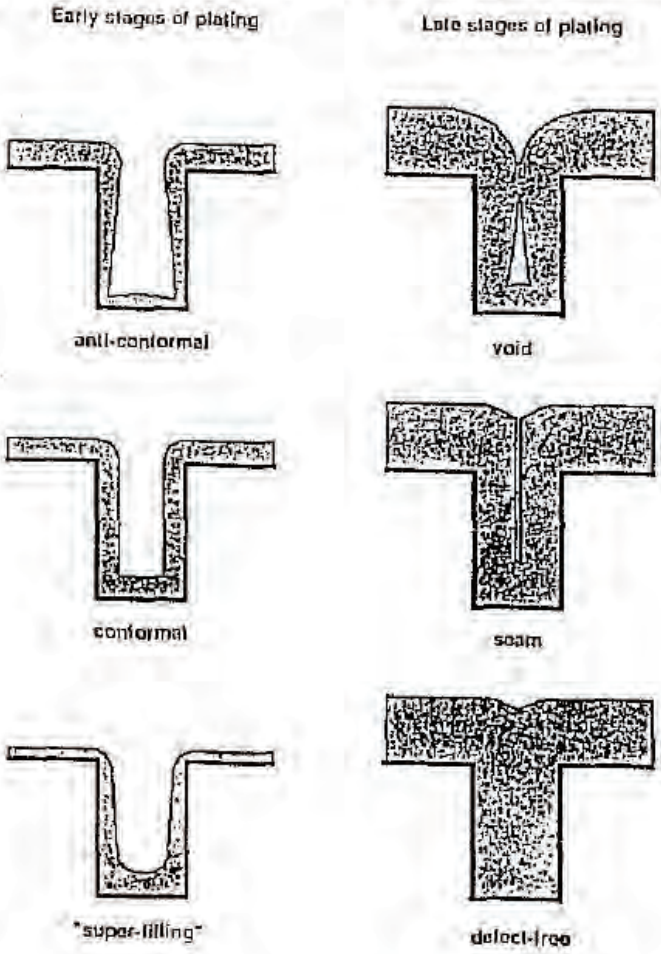


Fig. 16-33 Types of profile evolution in damascene plating.

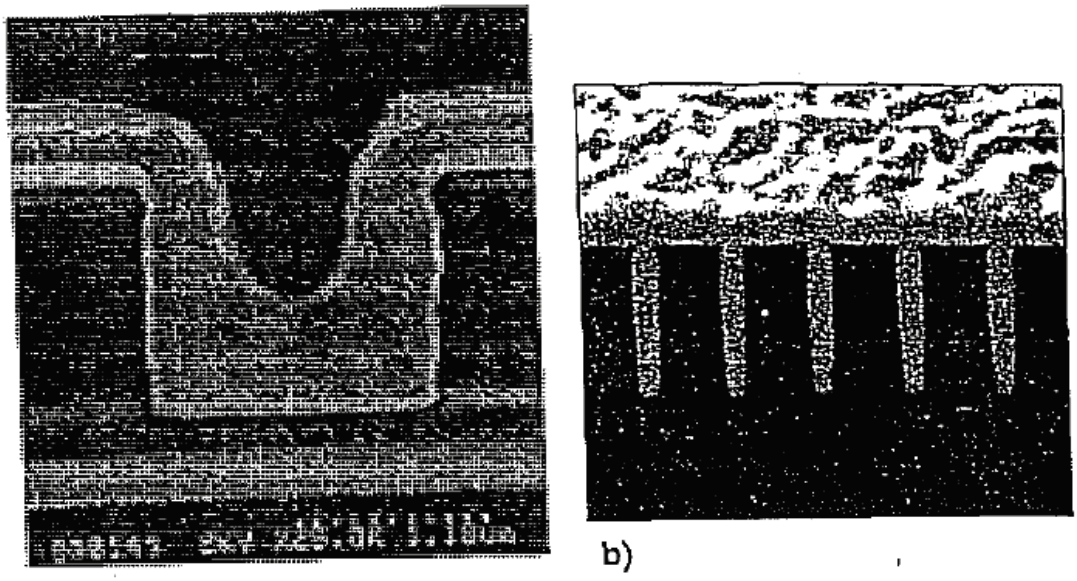


Fig. 16-34 a) Cross-section of a partially-filled line showing the profile of electroplated copper that exhibits the "bottom-up" filling phenomena.[48] b) Vias completely filled with Cu.

feature leads to the creation of a seam. If the shape of the feature is re-entrant, a void is formed. Subconformal plating also leads to void formation, even in straight-walled features. Subconformal plating occurs when substantial depletion of the cupric ions in the plating solution leads to significant concentration overpotentials, which in turn, cause the current to flow preferentially to more accessible locations outside of the feature. Thus, for defect-free filling, an increasing deposition-rate along the sides and bottom of the feature is desired ("bottom-up" filling). Fig. 16-34 shows a SEM of partially-filled line showing the profile of electroplated copper that exhibits such a "bottom-up" filling phenomenon.[48] Figure 16-35 shows SEMs of the time-evolution of filling a trench with a Cu-layer in a "bottom-up" manner.[49]

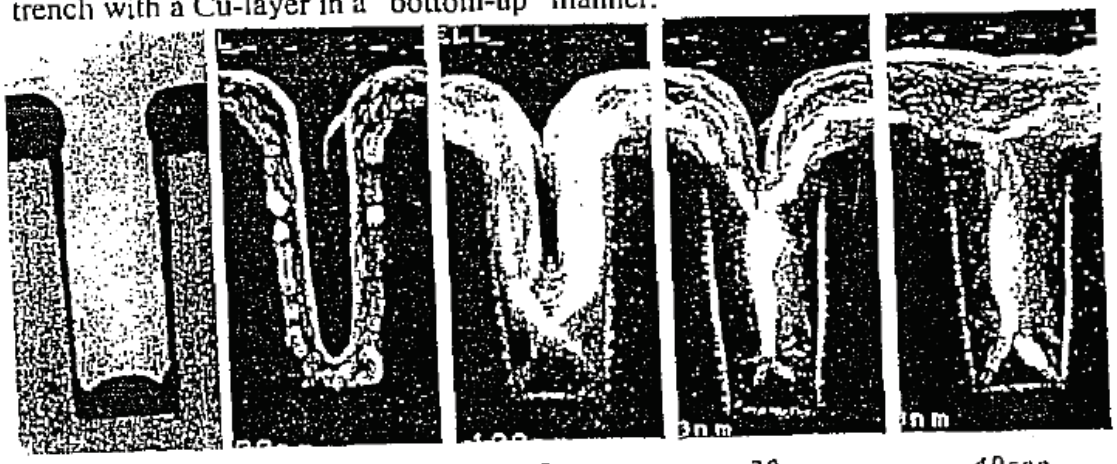


Fig. 16-35 SEMs of how "bottom-up" filling of Cu damascene recesses occurs versus time.[49]

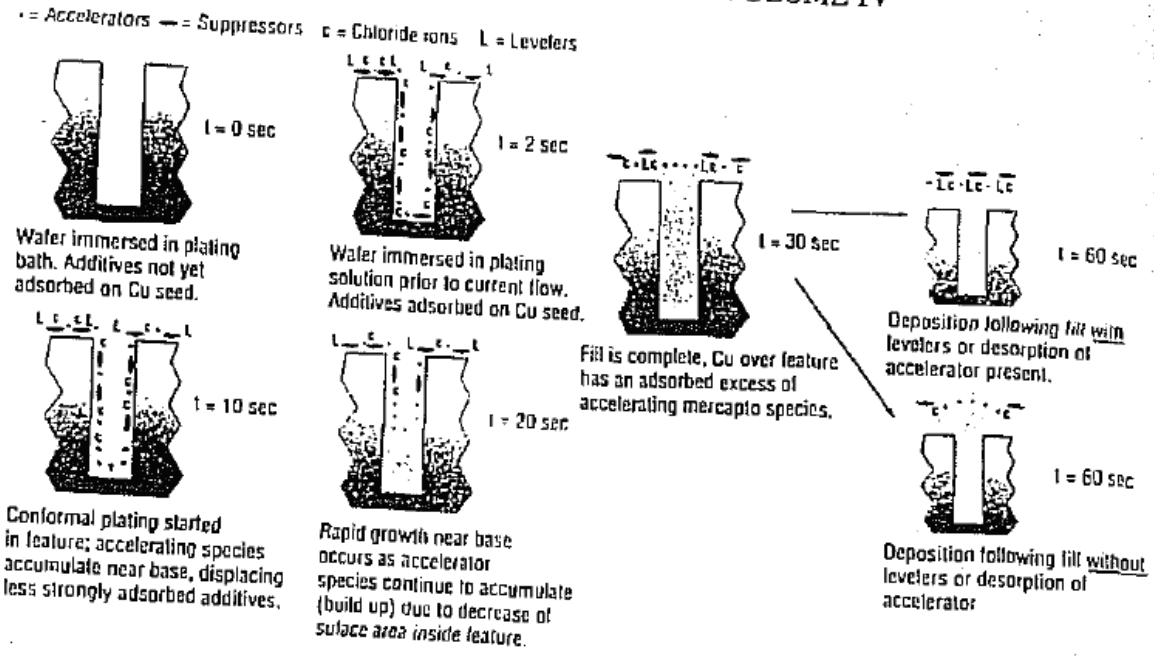


Fig. 16-36 Illustration of additive adsorption behavior during plating. A mechanism for the establishment and propagation of "bottom-up" fill is suggested. By permission of Solid State Technology. Published by PennWell.[33]

Reid's model of the sequence of events that accompanies "bottom-up" filling during Cu-plating is shown using Fig. 16-36. It begins by assuming that when a wafer is initially immersed in the plating bath ($t = 0$ in Fig. 16-36), that no additives are adsorbed on the Cu seed. However, within a few seconds (i.e., $t = 2$ sec in Fig. 16-36) both the suppressor and accelerator additives have been adsorbed on all areas of the wafer (including the insides of recesses) at their equilibrium levels. Thus, when the plating process is initiated by applying an external voltage to the cathode, the *initial* currents on all surfaces should be approximately equivalent on all surfaces. This leads to a deposition-rate that is constant everywhere. This perception is supported by the observed conformal deposition seen in the first 5-10 sec of a filling process (see also Fig. 16-35b - 10 sec). As was shown in Fig. 16-32, the plating current is kept low during this time (initiation current), to keep this conformal deposition thin.

After some time (which depends on current density, chemical concentrations, and feature size - e.g., $t = 10$ sec in Fig. 16-36), two phenomena are thought to begin to contribute to turn the fill process from a *conformal* filling-mode to a "*bottom-up*" filling-mode. First, the accelerator molecules start to accumulate at higher concentrations at locations near the bottom of narrow recesses. The increase in accelerator concentration near a recess bottom occurs because the surface area within the recess decreases as it fills (even if the fill is conformal)

and the accelerator molecules are neither incorporated in the depositing Cu nor are they desorbed into solution. (Recall from Sect. 16.5.2.2 that accelerator species adsorb strongly on seed-layer surfaces.) Meanwhile the PEG and Cl^- species are held less tightly on the seed-layer surface, and so they are displaced by the accelerator molecules and desorb into the solution. Thus, with a higher concentration of accelerator molecules in the areas of geometric concentration, the current-density (and the Cu-deposition-rate) is increased (t = 20 sec in Fig. 16-36).

The concentration of accelerator species in the solution must remain within an appropriate range. Too much accelerator in a bath disrupts fill because the accumulation of accelerating species also begins to take place on the top wafer surface between the recesses, and the differential deposition-rate with respect to the bottom of recesses is lost.

16.5.2.4 Electroplating Bath Control: Copper plating-baths are complex mixtures of chemicals, whose compositions contain various components that must be maintained at correct concentration levels. This is a challenging requirement because the composition of the bath changes as the plating process proceeds. The most dynamic ingredients of electroplating solutions are the organic additives. During normal electroplating the organic chemicals may be consumed in stoichiometric reactions (or they may break down in a variety of oxidative side reactions, some of which are non-stoichiometric). In any event, for consistent performance of the plating-bath, the original concentrations of these organics must be maintained, and their breakdown products must be removed. Even a small imbalance between components of the additives can cause defects in the filling of the trenches and/or degradation of the properties of the electroplated copper. As higher aspect-ratio-recesses are encountered, the ability to fill them properly will thus increase the importance of electrolyte management systems. (It should be noted that the bath solutions themselves are supplied by specialty-chemical vendors, such as Shipley Ronal.)

To enable the plating tools to be used in a high-volume production environment, the complex plating-bath chemistries have to be continuously monitored and controlled. Rigorous analysis of the plating bath chemistry has to be performed to determine how the bath composition changes with time as a consequence of plating, and then appropriate measures to counteract these changes must be taken. Such real-time analysis and feedback control of electrolyte are key to operating plating processes in a manufacturing mode. A self-adjusting, self-monitoring electrolyte is necessary to ensure precise control

of the plating chemistry to guarantee repeatable lot-to-lot fill. This has to be done on a large number of plating tools in a wafer fab that may be producing 20,000 wafers per month (with six or more levels of copper interconnect).

In some cases a separate, standalone, closed-loop analysis-and-dosing-system for multiple plating tools is used. In other cases the tool is designed with an on-board closed-loop electrolyte management system (e.g., as is done in the Applied Materials Electra ECP Deposition System™).[50] In the Electra tool, the bath management system measures (and then doses) the bath on-board at the system level, enabling precise control of suppressor, accelerator, leveler, copper, and acid concentrations. The system has a 200-liter tank with filtration on-board and six dosing channels. If the monitoring unit detects that, for example, suppressor x is under-concentrated, it immediately calculates how much is needed to dose the tank, and then does it. Feedback is on the order of minutes, rather than hours.

One method for monitoring the electroplating bath is a technique called *cyclic voltametric stripping* (CVS). The CVS principle has been widely used in the PCB industry, and was subsequently implemented for monitoring the electrolyte baths for IC Cu-plating. In the CVS method, a Cu film is deposited on an inert electrode and then this film is stripped by anodic dissolution. During stripping, the current is integrated over time to quantify the electric charge required for complete film dissolution. This charge is directly related to the molar quantity of the metal stripped (and thus to the amount initially deposited). From this information, calibration curves can be generated that provide reasonably accurate quantification of the total additive concentration in a plating bath. The stripping charge is monitored instead of the deposition charge because it is less sensitive to changing electrode surface states and impurity currents. Monitoring a process that proceeds to a well-defined endpoint also has inherent advantages.

The small plating cells used for plating Cu onto silicon wafers create very dynamic changes in chemical component activity, and this has generated a demand for a continuous, on-line, closed-loop analytical system. ECI Technology (Rutherford, N.J.) has developed such a CVS system for monitoring Cu-plating processes for silicon fabrication,[51] the Quali-Line QLC-5000.™ The CVS bath controller samples multiple plating tanks as well as a standard solution and then outputs the results.

The main limitation of CVS is its sensitivity to "matrix effects." Many bath components and their breakdown products display convoluted electrochemical interactions, so that the stripping-charge response is ambiguous if several components simultaneously change concentration. Thus, it is difficult to precisely correlate observed changes of electrochemical activity to variations in

individual bath components over extended times. These attributes make CVS less than ideal if the user's goal is individual monitoring of each bath component. However, under conditions where the principal requirement is to detect and correct changes in one or two electrochemically active components that have a rapid rate of consumption compared to others, the CVS method may be adequate for a closed-loop control system. In addition, CVS can provide rapid measurements with low-hardware investment, and little added waste-disposal burden.[52]

Another more elaborate plating-bath monitoring technology involves metal-free and ion liquid chromatography (also termed *high performance liquid chromatography* [HPLC] and *ion chromatography* [IC]). This method works on the principle of separating each of the individual components in the bath prior to detecting and quantifying them. It can provide information on the concentration and proportional balances of components in the plating solutions, as well as information on contaminants and degradation products.[53] A bath monitoring HPLC system for IC Cu-plating applications is being developed by Dionex Corp.[63] A short-coming of the chromatography methods include waste generation and the fact that the time required to perform the analyses can be long. It appears that a workable approach may be one in which both CVS and chromatography are integrated into a hierarchical method that exploits their relative strengths and minimizes their shortcomings.

When it has been determined that a plating bath needs to be adjusted (or replaced), several strategies can be employed. The simplest (and crudest) approach is to dump the bath when the plating results no longer meet the required specifications (and this is associated with a condition where the organic contaminant level exceeds a specified value). This technique is referred to as *batch dump*, or *bleed-and-dump-to-drain*. The level of contaminants in the bath is determined by correlating the Cu deposition tool throughput (and its related current-density) to the concentration of organic contaminants produced by breakdown of the organic additives. The increase in contaminant concentration is correlated to the amount of charge passed through the plating bath, and this relationship is expressed in units of *mg/amp-hr* (see Fig. 16-37). Thus, if this quantity is known (and the maximum acceptable contaminant-level is specified), the length time that a bath can be used before dumping is known. However, this method generates large quantities of waste that must be treated due to the copper and acid content. There is also no control (in either the general, or statistical sense), over the bath chemistry between dumps.

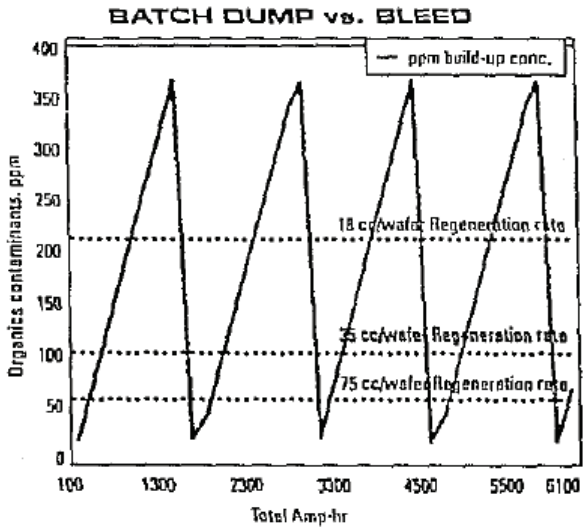


Fig. 16-37 "Batch dump" control versus "Bleed" control (0.3 amp-hr/wafer used to plate 1-μm thickness).[54] By permission of Solid State Technology. Published by PennWell.[11]

In a more refined technique, some fraction (typically about 10%) of the bath is removed daily, and it is replaced with new electrolyte solution. This is called the *bleed-and-feed* method. By analyzing a sample of the removed bath solution (e.g., with CVS), the concentration of additives in the new solution added to the bath can be adjusted to keep the total additive concentration within specified limits. However, as shown in Fig. 16-37, only a higher bleed-and-feed flow-rate can actively control the bath contaminant build-up that adversely impacts the bottom-up filling mechanisms. In addition, the bleed-and-feed method generates about the same amount of waste over time as the batch-dump method. Finally, while it does remove some of the contaminants that arise from breakdown of the organic additives, it does not completely remove them. It only dilutes them somewhat to a steady-state concentration. These drawbacks have led to a new approach to managing bath chemistry, referred to as a *remove-and-reconstitute* technique.

In this closed-loop method, the bath chemistry is continuously regenerated by stripping all organics (active and contaminant) from the recirculating bath. Then new organic additives (with the original concentration), are added back in. The bath solution is removed with flow-rates that are high enough to stabilize the concentration of contaminants at low levels. This flow rate is adjusted automatically, based on such inputs from the process tool as: amp-hours; wafer throughput; tool status, etc. The organic components are removed from the electrolyte by filtration, followed by their destruction using oxidizing agents

combined with heat and UV-light exposure. The organics are photochemically decomposed to carbon dioxide by exposure to the UV light - a clean process that involves no direct contact with the process fluid.

Process stabilization of both by-product and active organics are the benefits of this remove-and-reconstitute approach. The cost of consumables is also greatly reduced as only the organic additives are being removed (so only the organic additives need to be replaced). By using a closed-loop system, costly treatment of waste copper and acid is eliminated. A system in which this process is carried out is built by BOC Edwards. They call it the CuBIS™ (Copper Bath Integrated System), and it is shown in Fig. 16-38.[54] Its centralized monitoring system can simultaneously track five plating tools (from different equipment manufacturers).

16.5.2.5 Pulsed Plating: Utilizing pulsed-current waveforms (see Fig. 16-39b) during plating (instead of a dc-current) is a technique that complements the use of additives to obtain complete Cu filling of recessed features without voids. By periodically ceasing the current-flow, the plating reaction at the surface is halted in that time interval. The Cu filling process is improved for the following reason:

During electroplating a diffusion boundary layer is established between the bulk of the electrolyte solution and the wafer surface. In this boundary layer the Cu concentration trails off towards the wafer surface (since cupric ions are plated

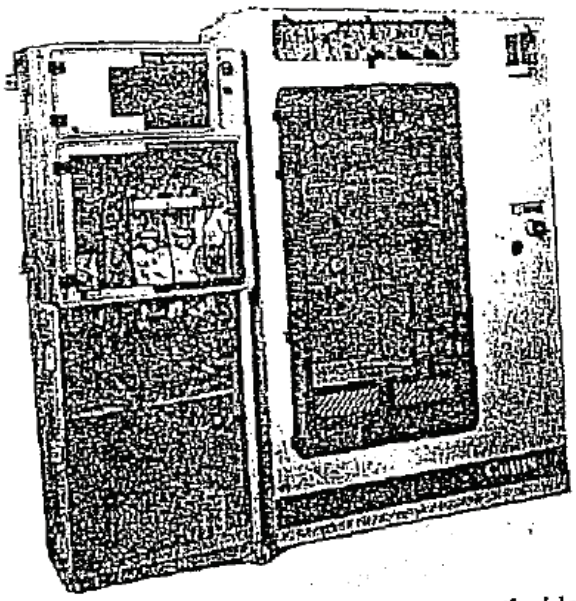


Fig. 16-38 Chemistry management is moving away from costly bleed-and-dump-to-drain approaches to bleed-and-feed. BOC's CuBIS Autorator uses a "remove-and-reconstitute" technique, which recycles the bleed stream, replenishing the bath at a rate automatically determined by plating conditions and by-product formantion. Courtesy of BOC Edwards.[54]

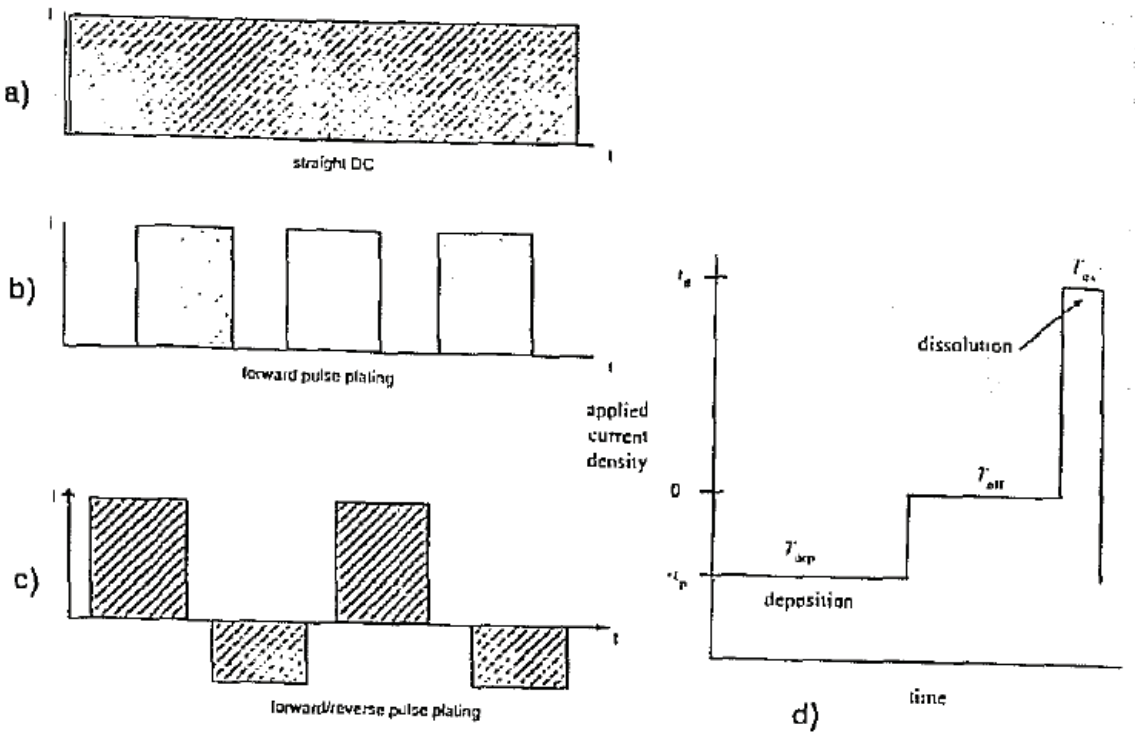


Fig. 16-39 a) dc current plating waveform; b) pulsed plating waveform; c) reversed-pulse-plating waveform; d) Example of reverse-pulse-plating waveform with optimized current amplitudes and pulse durations.[57]

out of the solution onto wafer surface). This boundary-layer is typically several microns thick. For a boundary-layer thickness that is larger than the pattern dimension, when the plating process reaches a steady-state condition, the copper-ion density is higher at the top-edges of trenches than at the bottom, Thus, for a dc-current-plating-process, a higher deposition rate is expected at the top of the trenches than at the bottom - and plating in this manner can lead to void formation.

However, by using pulsed-mode current, the concentration of cupric-ions depleted deep in the damascene recesses by the plating reaction can be restored to a higher level by Cu-ion diffusion to the bottom of the recess during such "rest-periods." This allows the deposition-rate at the bottom of the recess to equal the deposition-rate at the top when the current is restored during the next pulse of the waveform. This type of plating process is thus called *pulse-plating*. With an optimally shaped voltage pulse (both in amplitude and temporal shape), Cu filling of the recesses can be significantly improved.

Not only can the dc plating-current waveform be replaced with a pulse-waveform, but the polarity of the pulses can also be periodically reversed (Fig.

16-39c). The effect of reversing the polarity is to cause the Cu-film to undergo dissolution during the time polarity is reversed. Under the proper waveform conditions, the dissolution occurs more selectively at the top-corners of the trenches, thus allowing the trenches to remain open during the deposition-part of the plating waveform. Figure 16-39d shows an example of pulse-reverse waveform.[57] In one report, the application of periodic polarity reversal, together with adequately formulated and dosed surface-active additives allowed high aspect-ration via/trench filling capability. [58] One report observed that using small pulse periods gave better filling characteristics inside the trench. Using a small duty cycle and current density along with small pulse period appeared to give the best results.[59]

16.5.2.6 Copper Thickness Variation Within a Wafer After Electroplating: The copper topography across a wafer after the completion of a conventional electroplating process is not planar. Such within-wafer non-planarity is due to several effects. First, as discussed in Sect. 16.5.1, the higher current-density at the edge of the wafer next to the ring electrodes causes the Cu deposition-rate to be higher there than at the wafer-center. Thus, there is a buildup of copper thicker than at the center. One of the significant issues of achieving tight dishing and erosion specs after Cu-CMP is that these thicker Cu-regions require extra polishing time to get rid of them. This tends to increase dishing and erosion on other areas of the wafer.

Second, topography variations are caused by the feature-size-dependence of the filling process that is a characteristic of conventional electroplating. For example, the additives present in the electroplating baths used to enhance "bottom-up" filling of recesses with Cu can cause copper mounds to grow above regions containing dense arrays of narrow features (Fig. 16-40a). This occurs because the concentration of additives which accelerate the Cu-deposition rates in the recesses remain adsorbed on the top of the copper layer after the recesses

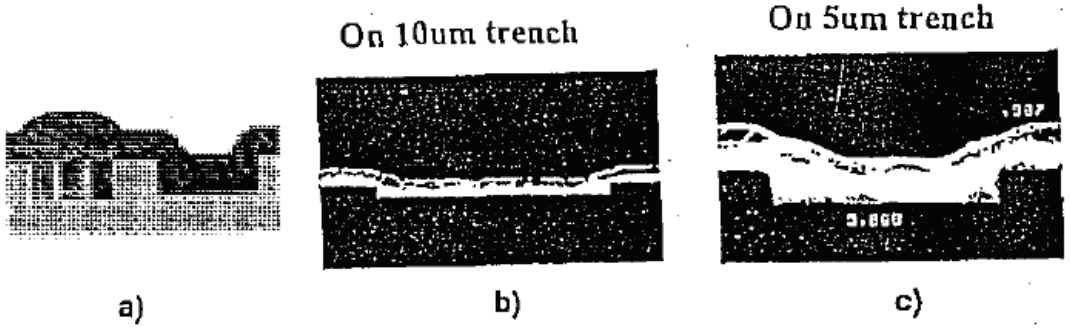


Fig. 16-40 a) Topography of a Cu-layer after plating into damascene recesses of varying widths; b) SEM of Cu plating into 10-μm-wide trench; c) into a 5-μm-wide trench.[61]

filling has been completed in those areas. This accelerates copper growth with respect to other areas of the wafer. If plating is done using electroltye solutions in which no additives are present to inhibit this effect, the copper layer over dense features can be twice as thick as the copper layer on wafer regions with no recesses. This reflects a continuation of copper growth beyond the time of filling completion, and the phenomenon is sometimes referred to as *overplating* (or *overburdening*, see also Fig. 16-36, t = 30 sec, and t = 60 sec).[60] The presence of mounds of excess copper in such regions again requires additional CMP over-polish. As a result, levelers are added to the plating bath (and reverse-pulsed waveforms are also used) to suppress the degree of such overburdening.

On the other hand, wide features do not fill as rapidly with Cu as narrow ones. Thus, there are steps in the copper layer at the edge of wide features. Figure 16-40a also shows such a step at the edge of a wide recess. Figures 16-40b and 16-40c show SEMs of Cu layers plated onto wide recesses, again showing the topography that results from a slower filling of wide recesses by plating relative that at which narrower recesses are filled.[61]

In summary, the topography that results from conventional electroplating processes exacerbates the problems of dishing and erosion that occur in a Cu-CMP process. Modifications to the plating process (such as those described in Sect. 16.5.4) have been proposed to alleviate this difficulty.

16.5.3 Post-Electroplating Cu Annealing and Recrystallization

When work began on copper-electroplating for IC interconnects, it was discovered that Cu-films are unstable after being deposited by electroplating, in that they self-anneal at room temperature (with the transient film-changes occurring over a period of hours to days). Furthermore, the transformation time is a strong function of film thickness. Thinner films take substantially longer to transform than do thicker films. In addition, the Cu seed-layer thickness also appears to play a role in the time and the degree to which annealing reduces the resistivity of the plated Cu-films. Greater resistivity reductions were observed as the plated films self-annealed on thinner Cu seed-layers than on thicker ones.[62]

In any case, if these variables are not controlled, each plated wafer headed for CMP may be different, depending on how much time has elapsed before it reaches the CMP step (a period that can vary depending on load size or tool selection). This can lead to wide distribution in electromigration behavior due to grain-size differences. To prevent such lack of control in plated-Cu-film characteristics, a post-ECP anneal at an elevated temperature was instituted. Such

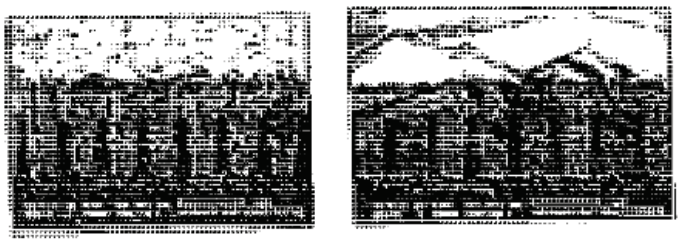


Fig. 16-41 Comparison of grain size in damascene trenches: a) without; and b) with anneal.

post-plating anneals have since become a standard part of the Cu-deposition sequence.

The annealing process causes several changes in the properties of the as-plated Cu-films. First, grain-growth and strong (111)-texture is observed after 400°C annealing (Fig. 16-41). The as-plated grain-size of 1000 Å increases to several microns.[77] The increase in grain-size is significant because the film resistivity is also a strong function of grain-size. (Thin films with larger grains generally have a smaller resistivity than do those with smaller grains. This is because in films with large grains there are fewer grain boundaries – which cause carrier scattering, and thus degrade conductivity.) Thus, annealing of plated-Cu films also reduces their resistivity by up to 20%.[50]

Furthermore, annealing before CMP stabilizes both the Cu-grains and the Cu film-density (i.e., they no longer undergo changes with time at room-temperature). The result of such stabilization is not only a more controllable (and repeatable) Cu-CMP polishing-rate, but this rate also tends to be somewhat higher (by about 20-30%).[64] However, the temperature and time-duration of the annealing process can have an impact on the reliability of the Cu lines, and this must be considered when specifying the annealing conditions (as is discussed in more detail in Sect. 16.8.2).

To avoid having to raise the temperature of the Cu annealing process beyond the point where it makes no further improvements, studies of changes in the sheet resistance versus anneal-temperature and time were conducted. In one study it was found that a temperature of 200°C for 30 seconds would adequately reduce the resistivity and improve the Cu CMP removal-rate (Fig. 16-42). Thus, by designing a resistively-heated-pedestal-based annealing chamber within an electroplating tool enclosure, an annealing process can be integrated into the plating system. A chill-plate is also used in this arrangement to quickly cool the wafer to 50°C after annealing before exposing it to atmosphere (to keep the hot copper from oxidizing, as it would do if it were allowed to cool in air). The wafer

filling has been completed in those areas. This accelerates copper growth with respect to other areas of the wafer. If plating is done using electroltye solutions in which no additives are present to inhibit this effect, the copper layer over dense features can be twice as thick as the copper layer on wafer regions with no recesses. This reflects a continuation of copper growth beyond the time of filling completion, and the phenomenon is sometimes referred to as *overplating* (or *overburdening*, see also Fig. 16-36, t = 30 sec, and t = 60 sec).[60] The presence of mounds of excess copper in such regions again requires additional CMP over-polish. As a result, levelers are added to the plating bath (and reverse-pulsed waveforms are also used) to suppress the degree of such overburdening.

On the other hand, wide features do not fill as rapidly with Cu as narrow ones. Thus, there are steps in the copper layer at the edge of wide features. Figure 16-40a also shows such a step at the edge of a wide recess. Figures 16-40b and 16-40c show SEMs of Cu layers plated onto wide recesses, again showing the topography that results from a slower filling of wide recesses by plating relative that at which narrower recesses are filled.[61]

In summary, the topography that results from conventional electroplating processes exacerbates the problems of dishing and erosion that occur in a Cu-CMP process. Modifications to the plating process (such as those described in Sect. 16.5.4) have been proposed to alleviate this difficulty.

16.5.3 Post-Electroplating Cu Annealing and Recrystallization

When work began on copper-electroplating for IC interconnects, it was discovered that Cu-films are unstable after being deposited by electroplating, in that they self-anneal at room temperature (with the transient film-changes occurring over a period of hours to days). Furthermore, the transformation time is a strong function of film thickness. Thinner films take substantially longer to transform than do thicker films. In addition, the Cu seed-layer thickness also appears to play a role in the time and the degree to which annealing reduces the resistivity of the plated Cu-films. Greater resistivity reductions were observed as the plated films self-annealed on thinner Cu seed-layers than on thicker ones.[62]

In any case, if these variables are not controlled, each plated wafer headed for CMP may be different, depending on how much time has elapsed before it reaches the CMP step (a period that can vary depending on load size or tool selection). This can lead to wide distribution in electromigration behavior due to grain-size differences. To prevent such lack of control in plated-Cu-film characteristics, a post-ECP anneal at an elevated temperature was instituted. Such



Fig. 16-41 Comparison of grain size in damascene trenches: a) without; and b) with anneal.

post-plating anneals have since become a standard part of the Cu-deposition sequence.

The annealing process causes several changes in the properties of the as-plated Cu-films. First, grain-growth and strong (111)-texture is observed after 400°C annealing (Fig. 16-41). The as-plated grain-size of 1000 Å increases to several microns.[77] The increase in grain-size is significant because the film resistivity is also a strong function of grain-size. (Thin films with larger grains generally have a smaller resistivity than do those with smaller grains. This is because in films with large grains there are fewer grain boundaries – which cause carrier scattering, and thus degrade conductivity.) Thus, annealing of plated-Cu films also reduces their resistivity by up to 20%.[50]

Furthermore, annealing before CMP stabilizes both the Cu-grains and the Cu film-density (i.e., they no longer undergo changes with time at room-temperature). The result of such stabilization is not only a more controllable (and repeatable) Cu-CMP polishing-rate, but this rate also tends to be somewhat higher (by about 20-30%).[64] However, the temperature and time-duration of the annealing process can have an impact on the reliability of the Cu lines, and this must be considered when specifying the annealing conditions (as is discussed in more detail in Sect. 16.8.2).

To avoid having to raise the temperature of the Cu annealing process beyond the point where it makes no further improvements, studies of changes in the sheet resistance versus anneal-temperature and time were conducted. In one study it was found that a temperature of 200°C for 30 seconds would adequately reduce the resistivity and improve the Cu CMP removal-rate (Fig. 16-42). Thus, by designing a resistively-heated-pedestal-based annealing chamber within an electroplating tool enclosure, an annealing process can be integrated into the plating system. A chill-plate is also used in this arrangement to quickly cool the wafer to 50°C after annealing before exposing it to atmosphere (to keep the hot copper from oxidizing, as it would do if it were allowed to cool in air). The wafer

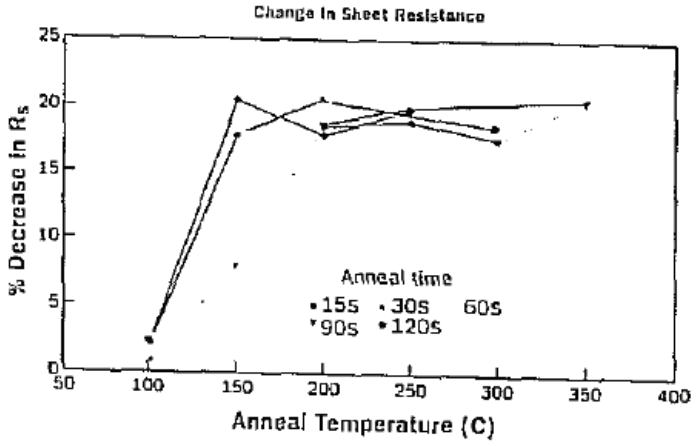


Fig. 16-42 Sheet resistance of Cu film as a function of anneal time and temperature after electroplating.[50]

is actually moved close to the chill plate (i.e., proximity cooling is used). To prevent the copper film from oxidizing during annealing, the ambient in the annealing chamber is maintained with < 5-ppm oxygen.

The Cu self-annealing effect can be explained as a recrystallization and purification process.[65] That is, an as-deposited plated-Cu-film is highly defective and the resistivity of the as-deposited film is high. The defects in the film are likely to be impurity inclusions caused by additives in the baths. (The large-grained deposits from additive-free baths do not undergo this recrystallization.) The inclusion of impurities at grain boundaries from the additives or their electrolyte by-products would be expected to cause pinning that would suppress grain growth. The as-deposited electroplated Cu film is not stable because of this high defect density. The film tends to reduce the defect level by recrystallization. During recrystallization, new grains are formed with fewer defects. These grains grow larger by consuming defective grains, leaving defects mainly in the grain boundaries and film interfaces. The stress in electroplated copper also changes from compressive-stress to tensile-stress after self-annealing. This is the macro evidence of film contraction caused by defect reduction.

16.5.4 ElectroChemical Mechanical Deposition (ECMD™)

As noted in Sect. 16.5.2.2, the Cu electroplating deposition-rate within recessed regions of a wafer exhibits a feature-size-dependence. That is, the Cu-deposition-rate is higher on areas of the wafer where there is a high-density of narrow, recessed damascene-features (and the plating process continues beyond the point

where the recess is filled). This results in mounding of the copper to a local thickness above those regions that is greater than in the flat regions devoid of recessed features (overburdening). In wide recessed-features, the deposition-rate is close to that on the flat-areas devoid of features, and hence a downward step can occur in the copper film when going from a flat area adjacent to a wide-recessed area. This has several detrimental consequences, especially with respect to the CMP process. That is, the thickness of the plated film must be increased to ensure that the wide recesses are completely filled. Thus, the process time of the CMP to polish the thicker Cu-film and the overburdened areas must be longer (which reduces the CMP throughput). There is also apt to be more Cu dishing in the wider-recessed areas. This problem would be overcome if the Cu could be deposited so that its surface after plating was globally planar. While techniques such as the use of leveler additives and pulsed plating can reportedly reduce the non-planarity to some degree, further improvement in the global planarity is still being pursued.

A novel process and deposition tool has been introduced by NuTool Inc. which seeks to address this issue directly. That is, in a process they call *electrochemical mechanical deposition* (ECMD^{TM}), as a copper layer is deposited in their tool (NuTool 2000^{TM}), a polishing pad is in physical contact with the wafer. This mechanical polishing coupled with the NuTool process chemistry simultaneously deposits copper and removes it. A differential growth rate between Cu deposition on the wafer flat areas and Cu deposition in the features is obtained. The speed, pressure, and mechanical path of the pad can all be controlled with software, eliminating the common problems of polishing copper, such as dishing and erosion.[66]

The ECMD process solves the problem of *excess-overburdening* (i.e., the buildup of excess copper over small densely arrayed features) and the downward steps that occur at the edges of wide recessed-features, by providing *aspect-ratio-independent filling* of features. Because the copper surface is not only globally planar but it is thinner in the field area than in a conventional Cu electroplating deposition process, the ECMD technique permits the CMP polish time to be reduced.

The tool and process were introduced in 2000, and thus information about its use in production is not available at the time of this book's publication. Some observers have commented that the ECMD process combines electroplating with some form of CMP on the wafer. Whether this combination of phenomena can be made to work simultaneously in a manufacturing environment will have to be demonstrated.

16.5.5 Electroless-Plating of Copper

Copper films can also be deposited by an electroless plating process, and this technology is already used in electronic packaging. Some work has also been done to develop Cu electroless plating processes for IC interconnect films. However, while this method has not gained the acceptance of electrolytic electroplating for forming Cu films, it is also briefly described here.

An *electroless* process involves the formation of a thin-film from an electrolytic solution without externally applied voltage. The deposition is due to an electrochemical reaction among the metal ions, reducing-agent, complexing-agents, and pH-adjusters on catalytic surfaces. The chemical process involves a CuO_4 and formaldehyde reduction:[67]

$$Cu^{2+} + 2HCHO + 4OH^- \rightarrow Cu^0(\text{solid}) + 2H_2O(\textit{liquid}) + H_2(\textit{gas}) \qquad (16.7)$$

An electroless Cu plating process for ICs was described in 1993 to give films with resistivities of 1.9 $\mu\Omega$-cm, close to that of bulk-Cu (1.7 $\mu\Omega$-cm).[68] Later work demonstrated filling of vias with aspect ratios of greater than 4:1 onto diffusion-barrier films of TiN.[69]

Although potentially a low-cost process because of its batch nature, electroless plating lost ground to electrolytic plating because of concerns due to hydrogen evolution, complexity of process control, low deposition rates, and potentially adverse environmental impact. However, it is possible that electroless plating will offer significant advantages for future wiring structures with improvements in sophistication and understanding. For example, some think that electroless plating may be used to deposit the Cu seed-layer at the 0.7-μm technology node.

16.5.6 CVD of Copper

While electroplating is the dominant method used to deposit Cu films for IC interconnects, work has been done on CVD-Cu deposition methods as well. One reason why interest remains in Cu-CVD processes is that it becomes increasingly difficult to deposit a uniform, conformal Cu seed-layer in the vias and trenches as their aspect-ratio continues to increase. At some point, it may not be possible to deposit a Cu-PVD-layer with sufficient thickness and/or continuity into such narrow features. A more conformal CVD-process could potentially produce a better Cu seed-layer in such high-aspect-ratio features. Thus, CVD-Cu technology approaches are briefly described here. (It should be noted however, that a copper PVD-process which deposits a very-thin nucleation layer will still be still needed before the Cu-CVD process.)

Organo-metallic sources are the most widely used precursors of CVD copper. These include Copper(II) {bis-hexa-fluoro-acetyl-acetonate copper} [$Cu^{II}(hfac)_2$] and Copper(I) {hexa-fluoro-acetyl-acetonatoL copper} or [$Cu^{I}(hfac)L$], where L is a neutral ligand weakly bonded to the Cu. One such ligand that has been used is vinyl-tri-methyl-silane (or *vtms*). $Cu^{II}(hfac)_2$ is a solid at room temperatures and must be dissolved in isopropyl alcohol in order to form a liquid precursor that can be delivered as a vapor (with hydrogen used as the carrier-gas). High quality Cu-films with low resistivity can be formed by the hydrogen reduction of $Cu^{II}(hfac)_2$. However, to get the lowest-resistivity Cu-films, the deposition temperature must be between 350–450°C. Since Cu is likely to be used with thermally-fragile low-*k* dielectrics, such high deposition-temperatures are undesirable. A low-pressure plasma process has reportedly been developed which allows the deposition temperature to be reduced to about 170°C, and which produces films with resistivities of 2.2 $\mu\Omega$-cm.

The $Cu^{I}(hfac)L$ precursor (where the L = vtms) is commercially available as a product called Cupra Select,™ developed by Schumacher.[70] It is a liquid at room temperature and is fed into the process chamber by direct-liquid-injection (i.e., the liquid is pumped to a vaporizer and then injected). This delivery method is needed for such precursors as $Cu^{I}(hfac)$-vtms, which decompose after long periods of heating in a bubbler. Low-resistivity Cu films can be formed at deposition temperatures of 150°C. The Cu is then deposited through a disproportionation mechanism:

$$2Cu^{I}(hfac)\text{-vtms}\ (gas) \rightarrow Cu^{0}(solid) + Cu^{II}(hfac)_2\ (gas) + 2\text{vtms}\ (gas) \qquad (16.8)$$

Both the $Cu^{II}(hfac)_2$ and the vtms ligand are stable compounds and readily desorb from the surface at the deposition temperature, leaving behind a Cu film relatively free of residues. Since the reaction is reversible, vtms can be used to dry clean the deposition chamber. By removing the Cu deposited on the parts inside the chamber, particle contamination by the film's cracking and flaking can be prevented. However, an issue that must also be addressed with CVD-Cu is wafer-backside Cu-deposition.

16.6 WET-ETCHING AND ELECTROPOLISHING OF CU FILMS

Work has also been carried out to develop processes that would replace (or shorten the process time) of Cu-CMP. Two basic approaches have been pursued: 1) chemical *spin-etch planarization* (SEP) - in which the Cu film is etched with wet-chemicals using a special spin-processing tool; and 2) electrochemically removing the Cu film with a process called *electropolishing*.

16.6.1 Spin-Etch Planarization of Cu Films

A SEP process was jointly developed by SEZ and Honeywell.[71] The SEP process removes layers of Cu material at a controlled rate by exposure to a reactive chemical solution while the substrate is spinning. Its goal is to remove the copper overfill material after a plating process with the same planarization results as CMP (but with no risk of damage to the low-k materials in the structure), and at much lower cost than CMP. Damage to the low-k is avoided and low-erosion of the dielectric is also achieved because there is no mechanical downforce as occurs in CMP.

During SEP the wafer is suspended horizontally on a nitrogen cushion above a rotating chuck (utilizing the Bernoulli principle). The substrate is held in place laterally with locating pins about the wafer perimeter. As the chuck and wafers are spun, wet-etchants are dispensed onto the wafer from a radially oscillating nozzle. An appropriate etching solution and spinning of the wafer result in a planar final surface (with the simultaneous radial and tangential etchant flow greatly enhancing planarization over simpler static etch processes).[72] DI water and nitrogen are then applied to the wafer to achieve rapid cleaning and drying. Etch rates of copper layers can range from 500-2000-nm/min, depending on the chemistry selected, and almost infinite selectivity exists with respect to etching of the underlying Ta and TaN barrier-layers.

By 2000, SEP was still in an early stage of development. However, it was demonstrated that the SEP process could provide a high degree of planarization for small features (0.35-0.5-μm-wide), and work was proceeding to improve the planarization of larger features in the 5-100-μm range. One way that SEP may become more practical is if a short CMP process (with a small downforce) could be used to knock off the large mounds of excess overburden present on parts of the wafer where high densities of narrow features are present. This would make the surface substantially more planar, making the SEP process more effective.

16.6.2 Electropolishing of Cu Films

Electropolishing of the overfill material after Cu electroplating is another process that has been pursued to eliminate or shorten the Cu CMP process. That is, after the electroplating step is completed, the current flow can be reversed, and the Cu film will then de-plate. The idea is to use such de-plating to remove the Cu overburden while leaving the Cu in the damascene recesses. Since reverse electroplating will remove Cu uniformly across the wafer, at first glance the concept appears feasible. However, it proves not to be practical as long as the

problem of aspect-ratio-dependent filling-effects occur during the initial electroplating process (see Sect. 16.5.2.6). That is, the non-uniform topography of conventional electroplating processes will result in overpolishing in the wide trench-regions and under-polishing in the circuit regions where there are dense-arrays of small-features. However, with the advent of a globally-planar Cu-deposition-process such as described in Sect. 16.5.4, electropolishing is once again being investigated.

In 2001 a Cu-electropolishing process was described in which the Cu-CMP step was either eliminated or significantly shortened.[61] A globally-planarized Cu-film was then deposited (such as was described in Sect. 16.5.4). With a globally planarized Cu film, the electropolishing can remove copper across the wafer without thinning the metal in wide trenches. In the Cu-CMP-free process that was described in Ref. 61, the Cu-overburden was electroplolished down to the barrier-layer. The barrier-layer was then removed by dry-etching with a fluorine-based plasma (that does not etch the exposed Cu remaining in the trenches). The wafer was then cleaned with brush scrubbing (Fig. 16-43). In the CMP-less process, the Cu-overburden was partially-removed by electropolishing and the remainder by conventional CMP. Less dishing could be achieved with the CMP-less process. A smooth Cu surface and highly (111) texture was also obtained by selecting the proper current density during the electropolish step.

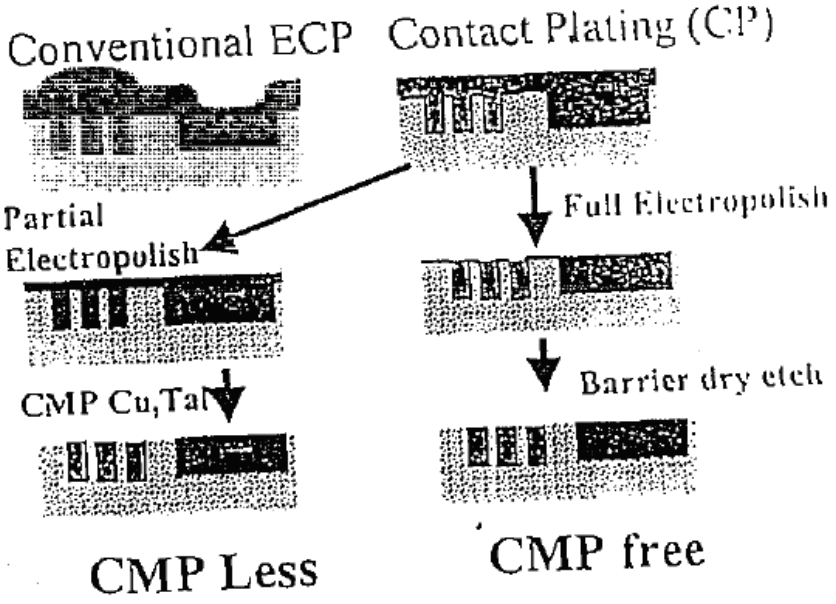


Fig. 16-43 Schematic diagram of CMP-less and CMP-free process flows by using a process that produced planar-Cu films after the plating process.[61]

A tool designed for precise electropolishing was introduced by ACM Research Inc. They refer to the process as *stress-free polishing* (SFP), and tool is called the ACM Ultra SFP.™ It is claimed that the SFP Ultra can polish Cu supported by ultra-low-*k* dielectrics (i.e., with a *k*-value less than 1.8).

16.7 CONTAMINATION CONTROL IN COPPER TECHNOLOGY

In Section 16.1 the issue of copper-contamination was introduced. Also briefly listed were the on-chip lines-of-defense and Cu-manufacturing protocols developed by wafer fabs to guard against this occurrence. Here additional information is provided about methods used to control copper-contamination in wafer manufacturing environments.

16.7.1 Sources of Copper Cross-Contamination in a Wafer Fab

In a completed device with a dual-damascene interconnect (and during many stages of the wafer manufacturing sequence), the embedded copper lines and vias are fully encapsulated with barrier-layers to prevent Cu out-diffusion into the surrounding dielectric and silicon areas. In those processes the likelihood of copper contaminating either other parts of the chip, or the wafer-fab environment, are small. However, in several process steps, Cu is exposed or is present outside of these sealed boundary areas (e.g., during those steps that occur from Cu-deposition to Cu-CMP). The most obvious risk of Cu-contamination occurs when large open areas of Cu are present on the wafer, or when Cu-containing chemicals or waste products - such as electroplating solutions or Cu-CMP slurry - are present. Another critical mode of Cu-contamination results from the high-energy bombardment of exposed Cu surfaces during plasma processing (e.g., via-open etch, and resist-strip). The following wafer processing steps (in roughly descending order of importance), can thus be considered to carry the highest risk of spreading Cu contamination - either into the deeper layers of an individual wafer or from wafer-to-wafer: a) Cu-CMP; b) Cu-deposition (including PVD seed-layer-deposition and electroplating); c) post-plating anneal: d) cap diffusion-barrier layer deposition; and e) via opening.[73]

During Cu-seed-layer deposition, electroplating, and post-plating anneal processes, the front side of the wafer surface is almost completely covered with a blanket Cu-film (and some Cu may find its way to the bevel and wafer backside during one or more of these steps). Therefore, cleaning of the wafer backside and bevel at certain critical steps during this stage may be necessary to prevent the spread of Cu particles from wafer-to-wafer by way of wafer-handling equipment (see Sect. 16.8.3). A barrier layer should also ideally extend over the whole wafer

surface, including around the bevel and on the backside to prevent Cu from diffusing into the wafer. This could be accomplished by using a wafer frontside barrier-layer (typically the conductive barrier-layer deposited before Cu), which overlaps a diffusion-barrier layer present on the wafer backside (e.g., CVD-SiN). Thus, the wafer bevel would also be completely covered by one or both of these barrier-layers. In many cases, the backside SiN diffusion-barrier-layer is not present. The backside and bevel cleaning is even more important in such cases. In some cases, if there is no backside diffusion-barrier layer, a silicon etch is performed after the backside and bevel cleaning step.[74] Its purpose is to remove a thin layer of silicon from the wafer backside that contains any copper that was deposited by the previous processes.

As noted above, contamination from copper can arise from tools and equipment involved in the deposition and handling of wafers. For example, a deposition tool that coats wafers with copper may cause it to be deposited on the bevel of the wafer. This wafer with copper on its bevel may be sent to a metrology tool equipped with a wafer handler that manipulates the edges of the wafers. (Such metrology tools are commonly shared among various process steps because manufacturers generally find that the acquisition of dedicated metrology equipment for copper processing is cost-prohibitive. These tools may include in-line SEMs for critical-dimension measurement, film-thickness monitors, and optical defect inspection devices.) The metrology-tool wafer-handler is now contaminated with copper, and can then cross-contaminate other wafers (for example, those that are on their way to an etch tool). The etcher thus eventually becomes contaminated with copper, and this contamination can spread throughout the fab (accumulating in plasma process chambers, wet benches, and lithography steppers). Contamination can also be spread by way of the wafer backside, as many wafer-handlers and robots hold and lift wafers by their backsides.

Wafer-handling by transport in cassettes is another likely source of copper cross-contamination. If copper is left on the wafer bevel as it exits a process tool, the carrier will invariably be contaminated with copper. Great care must thus be taken to either limit the use of wafer carriers to a specified area or to switch to a clean carrier once the copper is removed from the bevel. Wafer carriers may also need to be cleaned more frequently than otherwise required just for particle control. Furthermore, to prevent wafer-handlers and wafer-carriers from being contaminated by Cu particles on the wafer edge, it is preferable to maintain a sufficient edge exclusion of the Cu seed-layer or electroplated film. However, a full-faced seed-layer is typically required to allow electroplating with an edge-exclusion region of less than 4-mm. If dedicated cassettes are used to move the

wafers from the PVD-tool to the plating tool, and a cleaning step that removes the Cu from the bevel (and wafer backside) after electroplating is used, this problem can be circumvented.

Another source of possible copper cross-contamination is wafer breakage. Although the incidence of wafer breakage in processing tools has dramatically declined in recent years, it still occasionally happens. Thus, protocols for cleaning tools in which a wafer that contains copper has broken, are crucial to preventing factory-wide cross-contamination. A decontamination plan is required for each area and tool, which mandates appropriate cleaning methods and wipedowns in the immediate area in which the breakage occurred. However, there is no easy test for demonstrating the cleanliness of a tool after such breakage and cleanup, short of analyzing a test-wafer for copper-contamination.

Copper can reach the wafer bevel or backside during a deposition process. However, these surfaces can be protected by clamping and sealing the wafer hermetically during the plating operation so that only the face of the wafer is exposed during the process as required. The SABRE Electrofill™ tool from Novellus Systems uses a clamping method known as a *clamshell* to prevent the wafer backsides from being exposed to copper-containing solutions during the electroplating process. Surface analysis over a period of time during which 100,000 wafers were processed through a SABRE tool showed that a consistently low copper concentration of $< 10^{11}$ atoms/cm^2 were present (which was similar to the copper concentrations on control-wafers that did not undergo the plating process).[75]

The plating tool from Applied Materials (Electra™) uses a different approach. It contains a separate backside- and bevel-cleaning station integrated into the tool to treat each wafer after the electroplating step such that any copper that is deposited is removed (see Sect. 16.8.3 for more details).

16.7.2 Effects of Copper-Cross Contamination at the Back-End-of-Line (BEOL)

The concern for copper cross-contamination at the front-end-of-line (FEOL) stages of wafer processing is that the copper can diffuse into the active device regions and cause degradation of the gate-oxide-integrity or *pn*-junction leakage. The method of ensuring that copper-contamination is not occurring in the FEOL bays of a wafer fab is to monitor the MOS-C characteristics of wafers passing through the process line (much in the same way that is also done to safeguard against other forms of metal contamination - e.g., sodium).

However, the concerns at the BEOL phase of the wafer processing sequence are different from those feared at FEOL. That is, the solid solubility of Cu at

BEOL processing temperatures is relatively low, so not much Cu is expected to be incorporated into the silicon substrate from Cu on the wafer surface as long as the process temperature does not exceed 400°C. A test at SEMATECH took wafers with Cu plated on the backside, and put them into a furnace at 400°C.[76] Subsequent tests for leakage and threshold-voltage-shift showed no deviation from control wafers. Apparently the gettering mechanisms that protect devices from other contaminants can also handle levels of copper-contamination from such an event. As a result, worries about Cu entering the Si wafer at BEOL are probably somewhat exaggerated. On the other hand, the concern that copper will diffuse into the interlevel dielectric and cause leakage during BEOL processing is more real. The failure due to such a degradation of interconnect-leakage from copper contamination could also be one that develops slowly during the life of a part.

16.7.3 Cu Removal from the Wafer Bevel/Edge and Backside

Isolating Cu production from the rest of a wafer fab's manufacturing area is one approach to control copper contamination. But, many fabs share such critical and sophisticated tools as metrology and lithography equipment. Thus, such tools must accept Cu-processed wafers but it must be ensured that no contamination transfers to non-Cu wafers. While backside hermetic sealing and clamp rings are used to provide electrical contact and to mask the wafers exclusion zone from Cu deposition, they may not be 100% effective in preventing Cu deposition in the exclusion zone and on the wafer backside. This condition is further exacerbated by trends that minimize the exclusion zone to 3-mm (and even 2-mm), effectively putting the deposited Cu blanket onto the beveled edge of the wafer (Fig. 16-44a). Furthermore, since CMP cannot remove copper from the bevel and

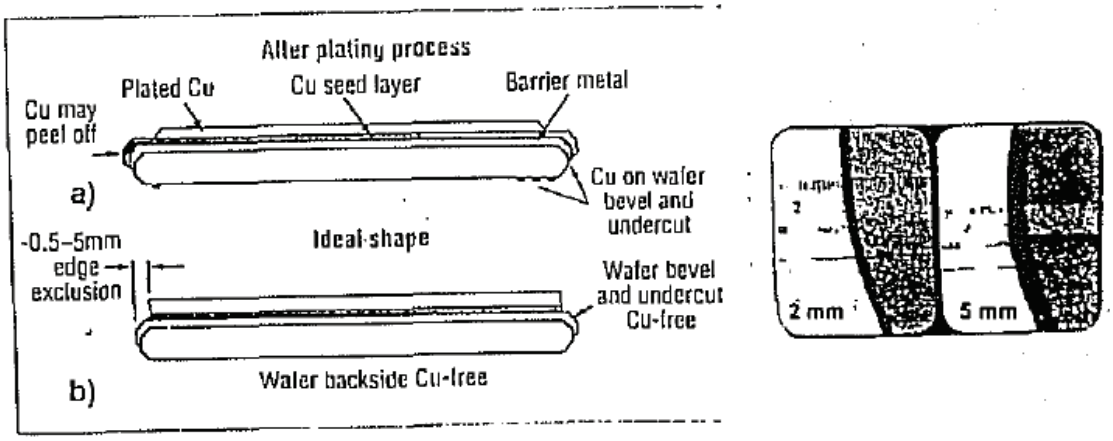


Fig. 16-44 a)Full coverage electrochemical Cu plating; b) The ideal Cu condition on the wafer after SPCE. By permission of Solid State Technology. Published by PennWell.[74]

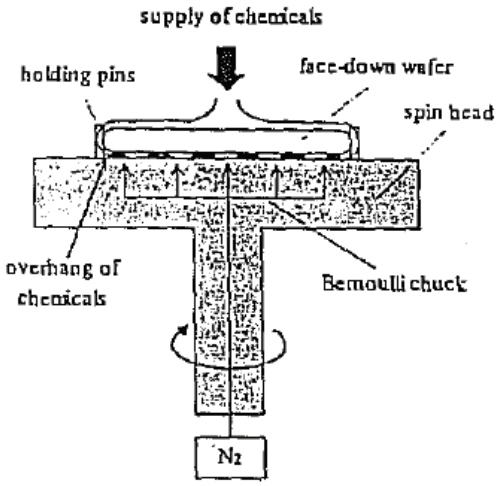


Fig. 16-45 Schematic diagram of the single-wafer cleaning system with a Bernoulli chuck.

backside, after a few layers, Cu-peeling (and resulting Cu cross-contamination) due to film buildup may occur.

The approach used to overcome this problem is to remove any copper that may have accumulated on the wafer-backside, bevel/edge, and exclusion-zone with a spin-etching process. Such a process is carried out immediately after the electroplating step is completed, and may be conducted either on a standalone spin-processor or on spin-processor that is integrated into the platform of the eletroplating tool.

A *standalone, single-wafer spin-cleaning-system* for backside Cu removal has been developed by SEZ Inc. They call their technology *spin-process contamination elimination* (SPCE™). During SPCE the wafer is suspended horizontally and frontside-down on a nitrogen cushion above a rotating chuck, utilizing the Bernoulli principle (Fig. 16-45). The substrate is held in place lateral with locating pins about the wafer perimeter. As the chuck and wafers are spun, the wafer is lifted about 1-mm. Wet-etchants are dispensed onto the wafer backside from a radially oscillating nozzle (e.g., modified HNO_3 chemicals are used for Cu removal), and the N_2 flow protects the entire front of the wafer (except the edge) from the chemicals. The degree of wraparound of chemicals onto the front of the wafer is a function of the N_2 flow-rate, the rotation speed of the chuck, and the viscosity and composition of the chemicals. This enables contamination and film removal from the back, edge/bevel, and front of the wafer without damaging the devices and without the use of masking materials on the frontside of the wafer. The width of the exclusion ring is tunable from 0.5-5-

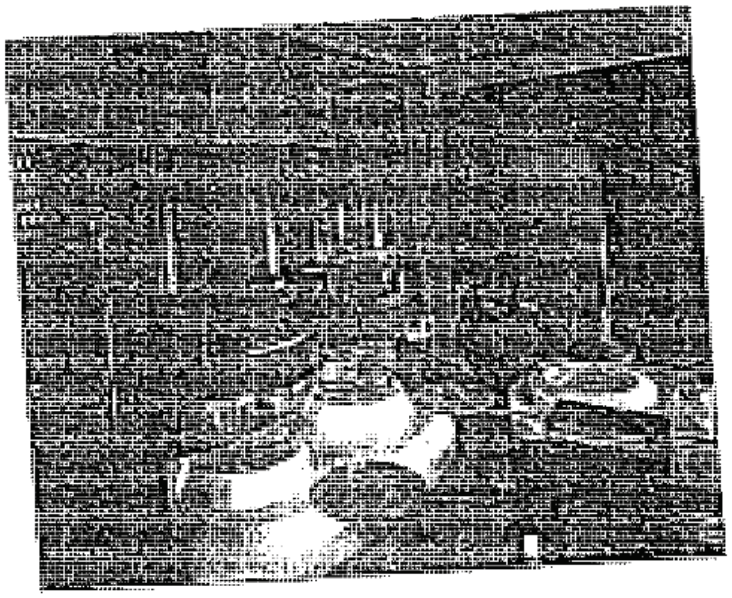


Fig. 16-46 Edge-clean chambers in the Applied Materials Electra electroplating tool. (Courtesy of Applied Materials)

mm.[74] DI-water and nitrogen are applied to the wafer for rapid rinsing and drying. Fig. 16-44b shows ideal Cu coverage on a wafer after an SPCE process.

A backside and edge *Cu-removal spin processor* has also been *integrated* into the Electra™ electroplating tool of Applied Materials. After the electroplating process, the wafer is moved to the cleaning chamber, where it is loaded face down by a robot onto the self-centering wafer hoop. A wafer chuck holds each wafer, which is then lowered into a cavity with three chemical channels below and three above the wafer (Fig. 16-46). While the wafer is rotated, a DI-water rinse first rinses the electrolyte from the wafer surface. As the wafer is spun dry at high rotation speeds, the chemical dispensing arm moves into position to release its etching/cleaning chemical precisely at the edge of the wafer. Three sets of nozzles (120 degrees apart) enable simultaneous chemical dispensing for very short edge clean times. While the chemicals remove any edge and backside copper, DI-water is sprayed on the front side. By rotating the wafer and balancing the flows, it is possible to control how much material from the backside overwraps, either on the side or the bevel, or just above the bevel, or whether its kept from coming to the front side at all. Figure 16-47 shows the programmable edge exclusion with this edge clean process.[50]

An integrated bevel and backside cleaning process is also offered in the Novellus SABRE™ electroplating system. In this approach, the Cu is removed

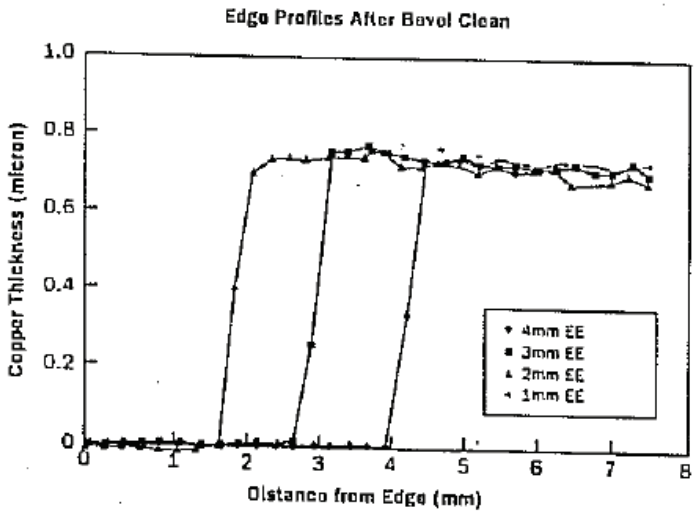


Fig. 16-47 Programmable edge exclusion with the applied Materials Electra™ edge-clean process. (Courtesy of Applied Materials)

from the bevel after the Cu seed-layer is deposited (i.e., prior to the electroplating process). The removal is carried out in the cleaning station of the tool.[78]

16.8 RELIABILITY OF COPPER INTERCONNECTS

As described in Sect. 16.1, copper lines offer the benefit of improved reliability over Al lines, primarily due to their superior electromigration lifetimes. However, the issue of copper interconnect reliability has other aspects as well, and some of these are described in this section.

16.8.1 Reliability of Copper Interconnects: "Missing-Metal" Defects

One Cu reliability issue involves post-CMP defects in narrow lines, which are referred to as "missing-metal" defects (Fig. 16-48). Such defects occur as a result

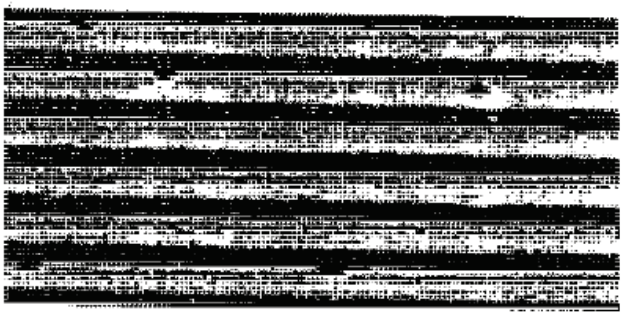


Fig. 16-48 Example of "missing-metal"-type defects. These defects arise from grain pull-out or corrosion and are exposed during the CMP process.[39] Reprinted with permission of Semiconductor International.

of corrosion or *grain-pullout* of lines exposed to a CMP-process. While the mechanism for these defects is poorly understood, it appears that their occurrence is influenced by a number of factors, including: plating-chemistry; plating-waveform; composition of the barrier-layer; and thermal-treatment. The latter effect appears to be most significant, and is thus discussed further here.[39]

Copper-lines are annealed after electrodeposition with the Cu overburden still in place. High thermal-excursions can cause large stress in the Cu films. That is, as the temperature is increased, the thermal expansion of the copper is greater than that of the surrounding dielectric material. This Cu expansion thus induces compressive-stress in the copper film that causes the film to re-crystallize and relax through dislocation motion. However, upon cooling the copper contracts and goes into tensile stress that is much higher under ambient temperature conditions than in the as-deposited film, and such stressed crystalline-films are prone to grain-pullout during CMP.

To reduce the incidence of defects related to thermal treatment it is necessary to minimize the temperature excursion of the post-deposition-anneal process (i.e., while the Cu overburden remains in place). Temperatures should be kept below the Cu plastic-flow limits (~200°C) so the copper can expand and contract elastically. Figure 16-49a shows the sensitivity of via-chain yield to the post-electroplating annealing temperature. Some have said that controlling the defects related to the copper process is perhaps the most important reliability issue facing the drive to achieve higher yields for copper in production.[39]

16.8.2 Reliability of Copper Interconnects: Plated vs. CVD Cu Films

Another reliability aspect observed in copper interconnects is that electroplated copper is more reliable than CVD-Cu with respect to electromigration failures.

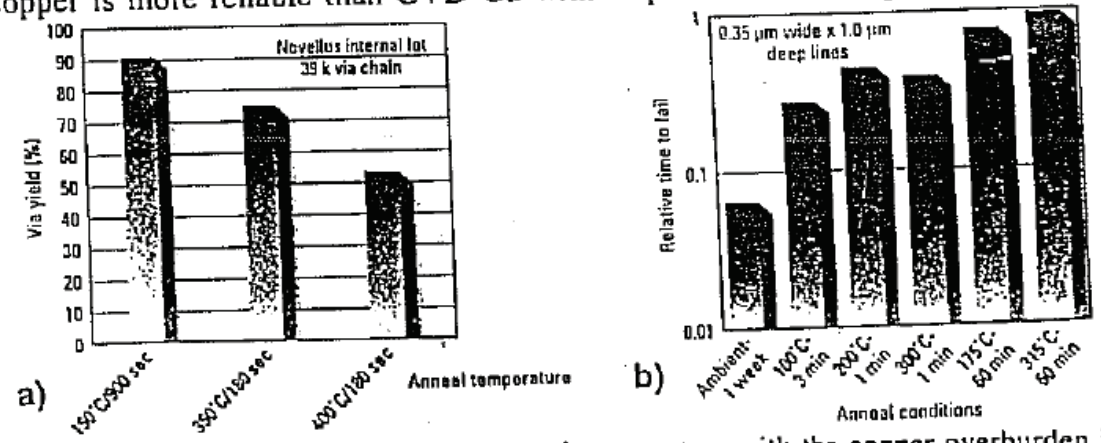


Fig. 16-49 a) The effect of post-plate anneal temperature with the copper overburden in place on via-chain yield.[39] b) Electromigration lifetimes for narrow lines (0.35-μm) with various pre-CMP anneals.[39] Reprinted with permission of Semiconductor International.

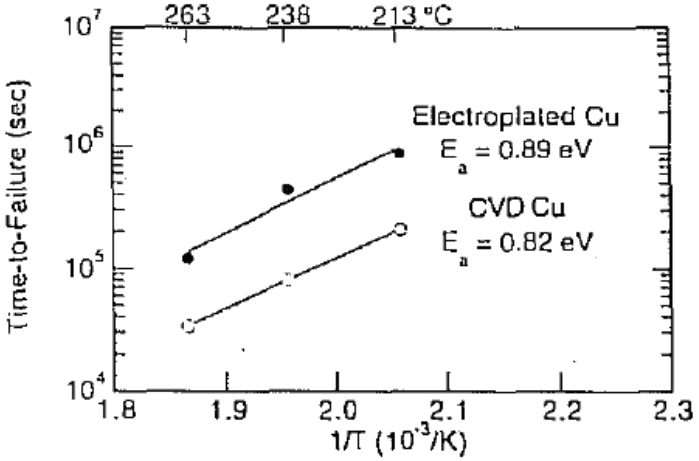


Fig. 16-50 Plot of time-to-failure of CVD-Cu and electroplated-Cu lines.[80] (© IEEE 1998)

As shown in Fig. 16-50, this is especially true for deep-submicron Cu linewidths. (The 0.4-μm-thick Cu-lines in Fig. 16-50 were tested for electromigration behavior after deposition, annealing [at 400°C for one hour], and CMP.[80]) Copper-lines deposited by electroplating exhibit longer lifetimes under electromigration-stressing than do CVD-Cu-lines because CVD-Cu has a smaller as-deposited grain size. In the narrow trenches of deep-submicron dual-damascene interconnects, electroplated-Cu has a bamboo grain-structure, which is conducive to superior electromigration behavior (whereas CVD-Cu has smaller grains after annealing than Cu deposited by electroplating, as can be seen in Fig. 16-51).[80,81]

As far as the electroplated-Cu-lines themselves, their electromigration reliability is impacted by the line-width and post-plating anneal-cycle. Wide lines anneal fully at the lowest anneal-temperatures, and when fully annealed they

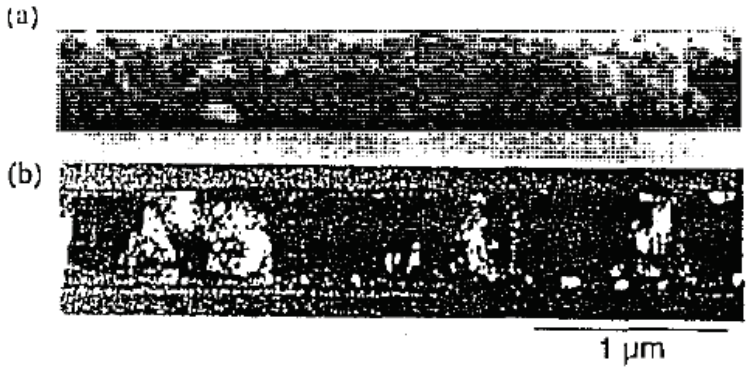


Fig. 16-51 Plan-view TEM micrographs of Cu lines in trenches after annealing: a) CVD-Cu; and b) electroplated Cu. (© IEEE 1998)

exhibit the longest electromigration lifetimes. Narrower linewidths require a higher-temperature anneal-cycle to achieve the same reliability behavior (see Fig. 16-49b, which shows the electromigration lifetimes of narrow copper-lines versus anneal-time and temperatures).[39] However, higher-temperature annealing cycles lead to greater "missing-metal" defect levels in narrow Cu-lines. Thus, a tradeoff must be made between deteriorating defect-performance and improved electromigration-performance as the anneal-temperature is increased.[39]

It was also reported that by incorporating a small amount of magnesium into the copper, various characteristics of the Cu interconnect were improved. That is, by forming a CuMg/Cu/CuMg stack, it was observed that the Mg diffuses through the copper and aggregates on the Cu surfaces. At the interface between the Cu seed-layer and the TiN barrier-layer this Mg improves the adhesion of copper to the TiN. At the top surface of the Cu-layers, the Mg forms a self-limiting oxide that hinders copper oxidation. Electromigration characteristics of such Cu/Mg stacked layers showed no difference relative to pure copper layers.[79]

16.9 ENVIRONMENT, HEALTH, AND SAFETY ISSUES OF COPPER TECHNOLOGY

The implementation of copper technology in IC fabrication introduces new chemistries, tools, and waste streams into the manufacturing arena. Before copper processes are put into production, it is important to address the *environment, health, and safety issues* (EHS) involved with them. Copper brings five new technologies to wafer fabrication, and their EHS impacts are discussed here: barrier-layer deposition; copper seed-layer deposition; Cu-fill by electroplating; copper CMP; and copper wafer reclaim.[82]

The first two of these processes, *barrier-layer deposition* and *Cu seed-layer deposition* are currently PVD processes which use solid sources, and hence do not involve any special EHS considerations. *Electrolytic plating for Cu fill*, however, is a new technology for the semiconductor industry, which also involves new chemistries, tools, and wastes. Nonetheless, commercially available electroplating chemistries used in the printed-circuit wiring board (PWB) industry have been evaluated and found to have no particular hazards or worker exposure issues. That is, all materials are TSCA listed (Toxic Substance Control Act). The primary concern is the two waste streams generated by the process: rinse water and a concentrated bath.

Copper CMP involves the use of a slurry that may or may not contain any chemicals that are environmentally hazardous. However, the process effluent is a potential environmental hazard, insofar as the it contains all the copper removed from the wafer by the CMP process. Development and manufacturing fabs

generate a large number of *test wafers that are reclaimed.* When these have a copper film, they cannot be reclaimed using existing methods. Nevertheless, relatively environmentally friendly techniques have been developed for this purpose (e.g., ammonium persulfate/sulfuric acid),[83] which is also compatible with electroplating waste. Nevertheless, the wafer reclaim process, like the electroplating process, generates the same two waste streams.

Thus, the two key steps in a copper dual-damascene structure – electroplating and CMP – are both wet-chemical processes. They use large amounts of water and create large amounts of acid and heavy-metal waste. They require large quantities of slurries, plating solution, and other liquid feedstocks. Bulk chemical distribution systems and waste-water treatment facilities for these materials can be built into new fabs. Accommodating them in an existing fab is more difficult. Typical fab delivery systems are designed around gases and DI-water, not slurries and plating chemicals.

16.9.1 Copper Waste Handling

The liquid chemical wastes from the various copper process steps described above must be evaluated for their toxic substance concentrations, and treated appropriately. It has been determined that the rinse water from the copper plating and wafer reclaim process ordinarily contains relatively low-levels of copper and contribute only minor increases in the total copper concentration of the industrial waste-water discharge. However, copper CMP contributes more than 80% of the discharged copper on a mass basis, and is likely to require treatment before being discharged to the industrial waste-water.

16.9.1.1 Handling Concentrated Waste from Plating and Wafer Reclaim: The liquid chemical wastes and slurries containing concentrated copper must be treated to remove the copper. A process similar to the electroplating process (called *electrowinning)* is one method for treating the concentrated plating baths and wafer reclaim wastes. That is, copper is reduced and plated out of solution at the cathode, and water is oxidized (split) to form oxygen at the anode. This process is capable of reducing the copper concentration in these waste liquids to levels below the current concentration based discharge limit for *publicly owned treatment works* (POTW). While, the time to reach such limits with the electrowinning process is 20-40 hours, it is still a cost-effective method and it does not generate other waste streams (e.g., sludge).

16.9.1.2 Handling Copper CMP Effluents: The treatment of the Cu-CMP effluent is done in another way. The primary goal of such treatment is to remove the

copper from the effluent. Recycling of the water is a secondary goal. In many CMP effluents the copper is entirely soluble and is not incorporated into the abrasive particles of the slurry residue. Thus, the treatment first involves microfiltering the effluent to remove the nonhazardous abrasive solids (which are then discharged into the fabs industrial wastewater and ultimately to the POTW). Next the water containing the slurry chemicals and copper is treated. The slurry oxidizer (e.g., H_2O_2) is removed by appropriate chemical addition. Then the copper can be removed by several different methods:

1. An ion exchange process can be employed, using cation-exchange canisters. These require periodic on-site or off-site regeneration (which again involves an electrowinning procedure). Such a treatment scheme can achieve very high levels of copper removal from the relatively dilute CMP effluent.

2. Coagulation and microfiltration is another method that can effectively reduce the copper content to less than 0.1 ppm.[83]

3. To insure copper removal to sub-0.1 ppm levels due to stringent POTW discharge requirements, it may be necessary to use water-soluble heavy-metal removal agents. These materials typically contain sulfide functionality, which bonds strongly to transition metals and forms insoluble precipitants.[84]

REFERENCES

1. P. Singer. "Tantalum, Copper, and Damascene: The Future of Interconnects," *Semiconductor International*, June 1998, p. 90.
2. R.L. Jackson *et al.*, "Processing and Integration of Copper Interconnects," *Solid State Technology*, March 1998, p. 49.
3. S.C. Sun, "Process Technologies for Advanced Metallization and Interconnect Systems," *Tech. Dig. .IEDM*, 1997, p. 765.
4. T. Takewaki *et al., Symp. on VLSI Technology, Dig. of Tech. Papers*, (1995), p. 31.
5. S. Venkatesan et al., "A High Performance 1.8 V, 0.20 μm CMOS Technology with Copper Metallization," *Tech. Dig. IEDM*, 1997, p. 769.
6. D. Edelstein *et al.*, "Full Cu Wiring in Sub-0.25-μm CMOS Technology," *Tech. Dig. IEDM*, 1997 p. 773.
7. M. Bohr, "Interconnect Scaling: The Real Limiter to High Performance ULSI," *Tech. Dig. IEDM*, 1995, p. 241.
8. C. Ryu *et al., International Reliability Physics Symp. Proc.*, 1997, p. 201.

9. L. Peters, "Leaders Describe 'Surmountable' Copper and Low-*k* Hurdles," *Semiconductor International*, September 2000, p. 25.
10. M. Igarashi *et al.*, "The Best Combination of Aluminum and Copper Interconnects for a High Performance 0.18-μm CMOS Logic Device," *Tech. Dig. IEDM*, 1998, p. 829.
11. A.E. Braun, "Aluminum Persists in a Copper Age," *Semiconductor International*, August 1999, p. 58.
12. P.K.K. Ho *et al.*, "Direct Integration of Damascene Copper onto Aluminum Interconnect," *Abs. Mtg. Electrochem. Soc.*, Fall 2000, Abs. 525.
13. P.C. Andricacos, "Copper On-Chip Interconnections,," *Electrochemical Society Interface*, Vol. 8, No. 1, Spring 1999, p. 32.
14. S. Rossnagel and J. Hopwood, *Applied Physics Letters*, **63**, 3285 (1993).
15. S. Rossnagel and J. Hopwood, "Metal Ion Deposition from Ionized Magnetron Sputtering Discharges," *J. Vac. Sci Technol.*, **B12**, 449 (1994).
16. S.M. Rossnagel, "Ionized Magnetron Sputtering for Lining and Filling Trenches and Vias," *Semiconductor International*, February 1996, p. 99.
17. G.A. Dixit *et al.*, "Ion Metal Plasma Deposited Titanium Liners for 0.25 μm/0.18-μm Multi-level Interconnections," *Tech Dig. IEDM*, 1996, p. 357.
18. B.L. Chin *et al.*, "Barrier and Seed Technologies for Sub-0.1 μm Copper Chips," *Semiconductor International*, May 2001, p. 107.
19. P. Gopalraja and J. Forster, *Appl. Phys. Letts.*, Vol 77, 2000, p. 3526.
20. C. Michaluk *et al.*, "Tantalum 101: Economics and Technology of Ta Materials," *Semiconductor International*, July 2000, p. 271.
21. H.E. Roosendaal, "Sputtering by Particle Bombardment I," in *Physical Sputtering of Single-Element Solids*, Behrisch (Ed.), Springer-Verlag, Berlin 1981, pp. 219-56.
22. D. Tracy, "The Emerging Copper Revolution Will Impact ICs with Critical Dimensions Smaller than 100-nm," *Chip Scale Review*, March 2001, p. 83.
23. C. Ryu *et al., Proc. Mtg. Material Res. Soc.*, Spring 1998, p. 75.
24. C. Ryu *et al.*, "Barriers for Copper Interconnections," *Solid State Technology*, April 1999, p. 53.
25. S.S. Wong *et al,. Proc. International Interconnect Conf.* (IEEE) 1998, p. 107.
26. T. Hara and K. Sakata, "Stress in Copper Seed Layer Employed in Copper Interconnects," *Electrochem. Soc. Letts.*, March 2001, G77.
27. T. Hara *et al.*, "Control of the (111) Orientation in Copper Interconnection Layer," *Electrochem. Soc. Letts.*, March 2002, C41.
28. A.A. Istratov and E.R. Weber, "The Importance of Understanding the Physics of Copper in Silicon for Development and Characteristics of Copper Diffusion Barriers," *Abs. Mtg. Electrochem. Soc.*, Fall 2000, Abs. 522.

29. M. Juppo *et al.*, "Atomic Layer Deposition of Titanium Nitride Thin Films Using *tert*-Butylamine and Allylamine as Reductive Nitrogen Sources," *Electrochem. Soc. Letts.*, January 2002, C4.
30. S. Hirao *et al.*, *Dig. Tech. Papers Symp. VLSI Technology*, 1997, p. 57.
31. J.-S. Min *et al.*, "The Mechanism of Si Incorporation with the Digital Control of Si Content during the Metallorganic Atomic Layer Deposition of Ti-Si-N Thin films," *J. Electrochem. Soc.* 147, October 2000. P. 3868.
32. A.E. Braun, "Copper Electroplating Enters Mainstream Processing," *Semiconductor International*, April 1999, p. 58.
33. J.D. Reid *et al.*, "Factors Influencing Damascene Feature Filling Using Copper PVD and Electroplating," *Solid State Technology*, July 2000, p. 89.
34. A.E. Braun, "ECP Technology Faces Chemical, Dielectric Hurdles," *Semiconductor International*, May 2000, p. 60.
35. I. Chen and T. Ritzdorf, "Seed-Layer Enhancement in Ultra-Thin Seed for Inlaid Copper Metallization," *Abs. Mtg. Electrochem. Soc.* Spring 1999, Abs. 275.
36. R. Solankey and B. Pathangey, "Atomic Layer Deposition of Copper Seed Layers," *Electrochem. Soc. Letts.*, October 2000, p. 479.
37. T. Nogami *et al.*, *Proceedings of the IITC*, p. 298 (1998).
38. J.D. Reid, "Fundamentals of Copper Electroplating," *Semiconductor International*, June 1998, p. 94.
39. V. Shannon and D.C. Smith, "Copper Interconnect for High-Volume Manufacturing," *Semiconductor International*, May 2001, p. 93.
40. S. Mayer *et al.*, "Integration of Copper PVD and Electroplating Processes for Damascene Feature Filling," *Abs. Mtg. Electrochem. Soc.*, Fall 1999, Abs. 732.
41. J. Horkans and J.O. Dukovic, "Monitoring of SPS-Based Additives in Cu Plating," *Abs. Mtg. Electrochem. Soc.*, Spring 2000, Abs. 360.
42. J.C. Hu *et al.* "Investigation of the Leveling Effect on Electrodeposited Cu Films for ULSI Applications," *Abs. Mtg. Electrochem. Soc.*, Spring 2000, Abs. 388.
43. P.L. Cavalotti *et al.* "Copper Electrodeposition for On-Chip Interconnections: Role of Different Additives," *Abs. Mtg. Electrochem. Soc.* Spring 2000 Abs. 359.
44. M. Yokio *et al.*, "Adsorption Behavior of Polyethylene Glycol on a Copper Surface in Acid Copper Sulfate Bath," *Denki Kagaku*, Vol. 52, p. 218, 1994.
45. J.L. Kelly *et al.*, "Leveling and Microstructural Effects of Additives for Copper Electrodeposition," *J. Electrochem. Soc.*, 146, July 1999, p. 2540.
46. P.C. Andricacos *et al.*, "Damascene Copper Electroplating for Chip Interconnections," *IBM J. Res. & Develop.* 42, 567 (1998).
47. J.J. Kelly and A.C. West, "Experimental and Numerical Study of Leveling of Submicron Features by Organic Additives," *Abs. Mtg. Electrochem. Soc.*, Spring 1999, Abs. 262.
48. P. Andricacos *et al.*, "Damascene Copper Electroplating for Chip Interconnects," *Abs. Mtg. Electrochem. Soc.*, Spring 1998, Abs. 254.

49. W.Y. Hsieh, "Interconnect Manufacturing Issues," *Advanced Interconnects: Design, Process, Integration, IEEE Short Course*, IEDM 2000.
50. S. Subrahmanyan *et al.*, "Integrated Electroplating – Key to Chip Production," *Vacuum Technology & Coating*, March 2000, p. 50.
51. G. Chalyt and P. Bratin, "On-Line Analysis of Organic Additives in Damascene Copper Processes, *Semiconductor International*, April 1999, p. 66.
52. T. Taylor *et al.*, "Electroplating Bath Control for Copper Interconnects," *Solid State Technology* November 1998, p. 56.
53. "Analysis of Copper Plating Baths," Dionex Corp. Website: www.dionex.com.
54. O. Blachier and J. Clark, "Taking a New Approach to Copper ECD Bath Management," *Semiconductor International*, July 2000, p. 229.
55. M. Biberger *et al.*, "Low Pressure Sputtering of Copper and Related Barriers for Seed Layers and Complete Planarization," Novellus Systems Website.
56. C. Marcadal *et al.*, "Metallorganic Chemical Vapor Deposited TiN Barrier Enhancement with SiH_4 Treatment," *J. Electrochem. Soc.* January 2002, p. C52.
57. A.C. West *et al.*, "Pulse-Reverse Copper Electrodeposition in High-Aspect-Ratio Trenches and Vias," *J. Electrochem. Soc.*, 145, September 1998, p. 3070.
58. S. Lopatin, "Extendability of Electrochemical Deposition for High-Aspect-Ratio Cu Interconnects," *Abs. Mtg. Electrochem. Soc.*, Spring 1999, Abs. 260.
59. D. Varadarajan *et al.*, "A Mass Transfer Model for the Pulse Plating of Copper Into High-Aspect-Ratio sub-0.25 Micron Trenches," *Abs. Mtg. Electrochem. Soc.*, Spring 1999, Abs. 268.
60. P. Singer, "Copper CMP: A Question of Tradeoffs," *Semiconductor International*, May 2000, p. 73.
61. M. Tsai *et al.*, "CMP-Free and CMP-Less Approach for Multilevel Cu/Low-*k* BEOL Integration," *Tech. Dig. IEDM*, 2001, p. 80.
62. B. Baker *et al.*, "The Interactions Between PVD Copper and ECD Copper as Observed in the Self-Annealing Process," *Abs. Mtg. Electrochem. Soc.* Spring 2000, Abs. 364.
63. E. Kaiser *et al.*, "Improved Determination of Additives in Acid Copper Plating Baths," *Abs. Mtg. Electrochem. Soc.*, Fall 2001, Abs. 1303.
64. T. Hara *et al.*, "Chemical Mechanical Planarization of Copper and Barrier Layers by Manganese (IV) Oxide Slurry," *Electrochem. Soc. Letts.*, December 2001, G109.
65. E. Shan *et al.*, "Effect of Deposition Temperature on ECD Copper Self-Annealing," *Abs. Mtg. Electrochem. Soc.* Spring 1999, Abs. 274.
66. J. Chappell, "NuTool Unveils ECMD Tool," *Electronic News*, 12/18/2000.
67. V.M. Dubin *et al.*, "Selective and Blanket Electroless Cu Platingby Contact Displacement for Deep Submicron Via and Contact Filling," *Proc. VLSI Multilevel Interconnect Conf.*, 1995, p. 315.

68. J. Van der Meerakker, "On the Mechanism of Electroless Plating - I - Oxidation of Formaldahyde at Different Electrode Surfaces," *J. Appl. Electrochem.*, 11, p. 387 (1981).
69. A. Contolini *et al.*, "Copper Electroplating Process for Sub-Half Micron ULSI Structures," *Proc. VLSI Multilevel Interconnect Conf.*, 1995, p. 322.
70. J.A. Norman *et al.*, "New OMCVD Precursors for Selective Copper Metallization," *Proc. VLSI Multilevel Interconnect Conf.*, 1991, p. 123.
71. D. DeBear *et al., Solid State Technology*, March 2000, p. 53.
72. J.A. Levert *et al.*, "A Novel Spin-Etch Planarization Process for Copper Dual-Damascene," *Abs. Mtg. Electrochem. Soc.* Fall 1999, Abs. 731.
73. D. Dornisch *et al.*, "Cu Contamination Control for Advanced Interconnect Manufacturing," *Solid State Technology*, May 2000, p. 137.
74. P.S. Lysaght and M. West, "Addressing Cu Contamination Via Spin-Etch Processing," *Solid State Technology*, November 1999, p. 63.
75. T. Cacouris, "Preventing Cross-Contamination Caused by Copper Diffusion and Other Sources," *Micro*, July 1999.
76. L. Peters," SEMATECH Explores Copper Contamination," *Semiconductor International*, May 2000, p. 89.
77. M.E. Gross *et al*, "Implications of Damascene Topography for Electroplated Copper Interconnects," *Solid State Technology*, August 1999, p. 47.
78. For more information on this process visit www.novellus.com.
79. L. Peters, "Following a Steep Learning Curve for Copper," *Semiconductor International*, October 2000, p. 25.
80. C. Ryu *et al.*, "Electromigration of Damascene Copper Interconnects," *Dig. Tech. Papers VLISI Symp.* 1998, p. 156.
81. H.S. Rathore *et al., 16th International VLSI Multilevel Interconnect Conf.* (1999), p. 89.
82. L. Mendocino and P.T. Brown, "The Environment, Health, and Safety Side of Copper Metallization," *Semiconductor International*, June 1998, p. 105.
83. R.D. Letterman (Ed.) *Water Quality and Treatment*, McGraw-Hill, New York, 1999, 5th Edition.
84. J.H. Golden *et al.*, "Evaluating and Treating CMP Wastewater," *Semiconductor International*, October 2000, p. 85.
85. T. Hara *et al.*, "Improved Barrier and Adhesion Properties in Sputtered TaSiN Layer for Copper Interconnects," *Electrochem. Soc. Letts.*, March 2002, G36.