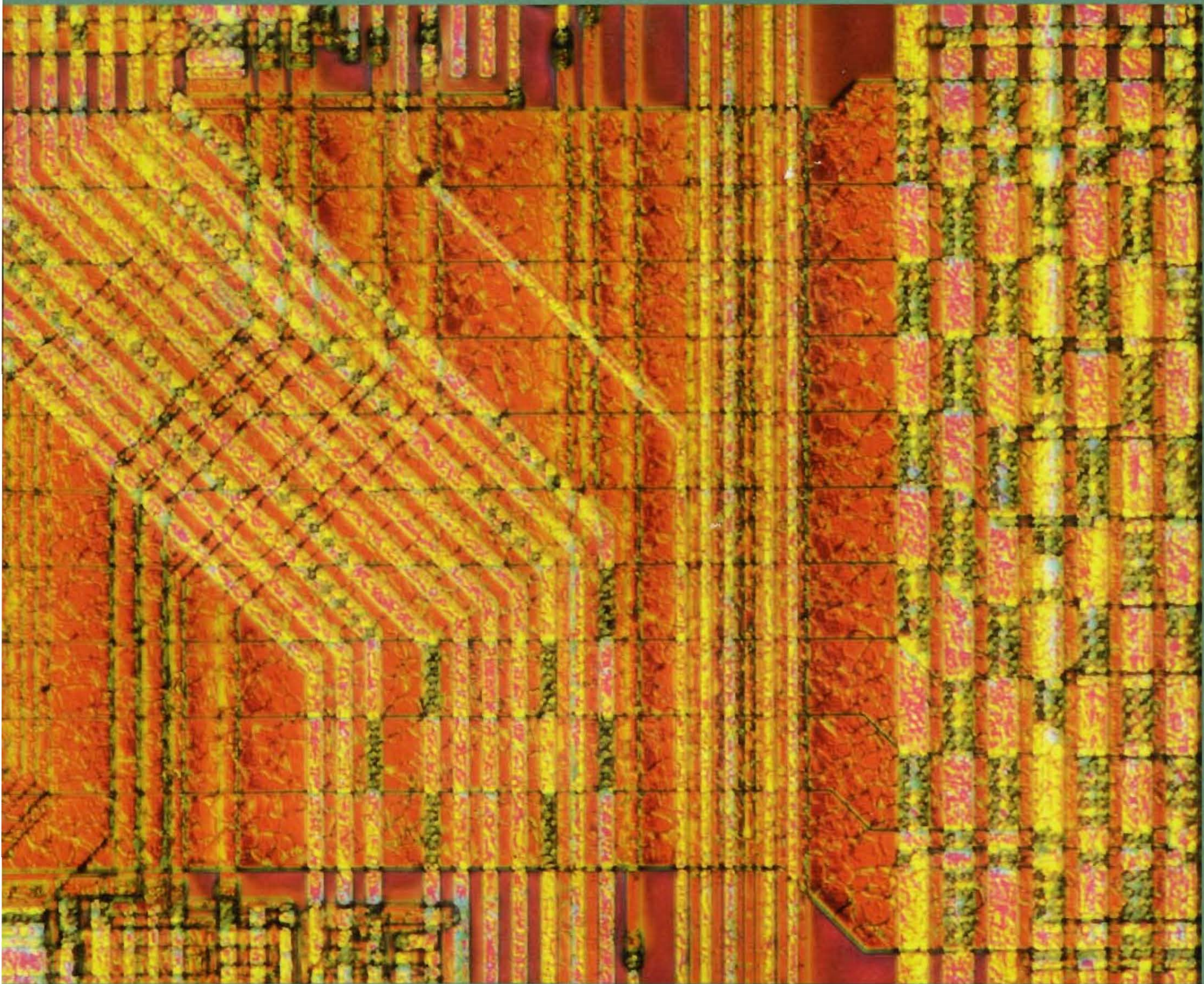
Semiconductor Technologies

Digital Technical Journal

Digital Equipment Corporation



Volume 4 Number 2
Spring 1992

Editorial
Jane C. Blake, Editor
Helen L. Patterson, Associate Editor
Kathleen M. Stetson, Associate Editor

Circulation
Catherine M. Phillips, Administrator
Sherry L. Gonzalez

Production
Mildred R. Rosenzweig, Production Editor
Margaret L. Burdine, Typographer
Peter R. Woodbury, Illustrator

Advisory Board
Samuel H. Fuller, Chairman
Richard W. Beane
Robert M. Glorioso
Richard J. Hollingsworth
Alan G. Nemeth
Victor A. Vyssotsky
Gayn B. Winters

The *Digital Technical Journal* is published quarterly by Digital Equipment Corporation, 146 Main Street MLO1-3/B68, Maynard, Massachusetts 01754-2571. Subscriptions to the *Journal* are $40.00 for four issues and must be prepaid in U.S. funds. University and college professors and Ph.D. students in the electrical engineering and computer science fields receive complimentary subscriptions upon request. Orders, inquiries, and address changes should be sent to the *Digital Technical Journal* at the published-by address. Inquiries can also be sent electronically to DTJ@CRL.DEC.COM. Single copies and back issues are available for $16.00 each from Digital Press of Digital Equipment Corporation, 1 Burlington Woods Drive, Burlington, MA 01830-4597.

Digital employees may send subscription orders on the ENET to RDVAX::JOURNAL or by interoffice mail to mailstop MLO1-3/B68. Orders should include badge number, site location code, and address. All employees must advise of changes of address.

Comments on the content of any paper are welcomed and may be sent to the editor at the published-by or network address.

Copyright © 1992 Digital Equipment Corporation. Copying without fee is permitted provided that such copies are made for use in educational institutions by faculty members and are not distributed for commercial advantage. Abstracting with credit of Digital Equipment Corporation's authorship is permitted. All rights reserved.

The information in the *Journal* is subject to change without notice and should not be construed as a commitment by Digital Equipment Corporation. Digital Equipment Corporation assumes no responsibility for any errors that may appear in the *Journal*.

ISSN 0898-901X

Documentation Number EY-L521E-DP

The following are trademarks of Digital Equipment Corporation: DEC, Digital, the Digital logo, VAX, VAX 6000, VAXcluster, VAX FORTRAN, and VMS.

DEC AVS is a trademark of Stardent Computer Systems.

MIPS is a trademark of MIPS Computer Systems, Inc.

PISCES and SUPREM3 are trademarks of the Board of Trustees of Leland Stanford Junior University.

SPICE is a trademark of the University of California at Berkeley.

Book production was done by the Design Group of Atlantic Graphic Services, Inc.

Cover Design

The CMOS-4 processes described in this issue are used to build Digital's high-performance NVAX and Alpha 21064 chips. The photomicrograph on our cover focuses on the high-speed clock driver and power pads of the Alpha 21064 microprocessor, which operates at a clock rate of 200 MHz.

The photomicrography is by Peter Catinella and Brian Edwards of Digital's Semiconductor Analysis Laboratory, Hudson, MA. The cover design is by Mike Call of Quantic Communications, Inc.

Contents

10 ***Foreword***
R. J. Hollingsworth

Semiconductor Technologies

12 ***Microprocessor Performance and Process Complexity in CMOS Technologies***
Bjorn Zetterlund, James A. Farrell, and Thomas F. Fox

25 ***Numerical Device and Process Simulation Tools in Transistor Design***
Marden H. Seavey, John V. Faricelli, Nadim A. Khalil, Gerd Nanz, Llanda M. Richardson, Christian O. Schiebl, Hamid R. Soleimani, and Martin Thurner

39 ***CMOS-4 Technology for Fast Logic and Dense On-chip Memory***
Andre I. Nasr, Gregory J. Grula, Antonio C. Berti, and Richard D. Jones

51 ***CMOS-4 Back-end Process Development for a VLSI 0.75 μm Triple-level Interconnection Technology***
Marion M. Garver, Joseph M. Bulger, Thomas E. Clark, Jamshed H. Dubash, Lorain M. Ross, and Daniel J. Welch

73 ***Implementation of Defect Reduction Strategies into VLSI Manufacturing***
Mary Beth Nasr and Ellen J. Mager

83 ***A Yield Enhancement Methodology for Custom VLSI Manufacturing***
Randall S. Collica, X. Joseph Dietrich, Rudolf Lambracht Jr., and David G. Lau

100 ***Transistor Hot Carrier Reliability Assurance in CMOS Technologies***
Daniel B. Jackson, David A. Bell, Brian S. Doyle, Bruce J. Fishbein, and David B. Krakauer

114 ***Electromigration Reliability of VLSI Interconnect***
J. Joseph Clement, Eugenia M. Atakov, and James R. Lloyd

Editor's Introduction

Jane C. Blake
Editor

The design of semiconductor chips has been the topic of several past *Digital Technical Journal* issues. With the introduction of Alpha 21064, the world's fastest microprocessor, this issue focuses for the first time on the development of semiconductor technologies that make possible the high-performance of Digital's VLSI chips. Engineers in Advanced Semiconductor Development present in-depth views into CMOS-4 technologies, which produce microprocessors with up to 1.7 million transistors and operating frequencies as high as 200 MHz.

The significant increase in performance achieved with each generation of CMOS technology is in part the result of a synergistic relationship between microprocessor design and process engineers. In their paper on process technology contributions to microprocessor performance, Bjorn Zetterlund, Jim Farrell, and Frank Fox describe the scaling theory that has led to a doubling of gate density and an increase of 30 percent in gate speed in four successive CMOS generations. They discuss process features implemented in CMOS-4, and close with a discussion of models that predict process variations.

Models and tools, essential in providing designers early insight into the characteristics of the transistors to be fabricated, are the focus of a paper by Marden Seavey, John Faricelli, Nadim Khalil, Gerd Nanz, Llanda Richardson, Christian Schiebl, Hamid Soleimani, and Martin Thurner. The authors describe several physical models that accurately simulate transistor behavior, and present numerical mathematical methods used to enhance existing simulators. An overview of Digital's and others' efforts to integrate simulation tools concludes the paper.

The need for both high-density logic gates and on-chip cache memory in microprocessors presents special challenges to process engineers. Andre Nasr, Greg Grula, Antonio Berti, and Rich Jones review the front-end process (formation of device and local interconnect) for the CMOS-4 0.75-μm technology and the steps taken to meet design requirements. They also describe the effects on submicron devices related to the scaling of feature sizes and examine some solutions.

Goals for the back-end process (formation of global metal interconnect) were also driven by the logic design requirements for higher circuit density. In addition, back-end development goals included the continued use of equipment developed for the 1.0-μm CMOS-3 technology. Marion Garver, Joe Bulger, Tom Clark, Jamshed Dubash, Lorain Ross, and Dan Welch relate how tools were modified for CMOS-4 and describe new blanket tungsten and planarization processes for submicron devices.

To produce a specified yield of CMOS devices, defect reduction and yield enhancements, like other processes, must be initiated concurrently with the design stage. Mary Beth Nasr and Ellen Mager review the principles of microcontamination control and outline defect reduction techniques to increase product yield in the areas of p-gate leakage and metal 2 short circuits. The paper that follows addresses the methodology of yield enhancement, including processing, process equipment, manufacturing, and design and test. Randy Collica, Joe Dietrich, Rudy Lambracht, and Dave Lau describe the use of test chip data, yield models, and the selected approach to yield analysis and forecasting.

An advanced method that helps designers predict circuit hot carrier lifetime and thus maximize transistor performance at the required reliability level is the topic of a paper by Dan Jackson, David Bell, Brian Doyle, Bruce Fishbein, and David Krakauer. The authors describe a physically based method for determining the acceptability of hot carrier-induced degradation in transistor characteristics.

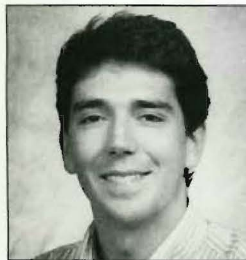
Another predictor of chip lifetime is the reliability of the interconnects. In their paper on electromigration reliability, Joe Clement, Eugenia Atakov, and Jim Lloyd provide a helpful overview of the potential erosion in metal interconnect due to electron conduction. They then present a scaling model developed to characterize the reliability of CMOS-4 chip interconnects.

The editors thank Rich Hollingsworth and Arlene Delvy of the Advanced Semiconductor Development Group for their guidance and unfailing support in developing this issue.

Jane Blake

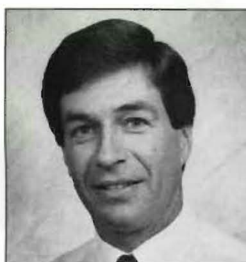
Eugenia M. Atakov As a principal engineer in the Advanced Semiconductor Development (ASD) Group, Eugenia Atakov is responsible for evaluating on-chip interconnect reliability. Since joining Digital in 1988, she has contributed to several CMOS VLSI projects, including the Alpha 21064 chip. Previously, Eugenia worked at the Moscow Institute of Semiconductor Electronics. She received a B.S.E.E. and an M.S.E.E. from the Moscow Energy Institute, and a Ph.D. from the Moscow Institute of Steel and Alloys. She holds two patents and has published 17 papers on semiconductor device and interconnect reliability.

David A. Bell David Bell attended Bates College and received a B.S. degree in physics and mathematics in 1980. In 1984, he completed his Ph.D. in physics at the Massachusetts Institute of Technology. David worked at Texas Instruments on submicron CMOS and BiCMOS development prior to joining Digital's Advanced Semiconductor Development Group in 1989. His current work includes submicron CMOS transistor design, characterization, and modeling. David is a member of IEEE and Phi Beta Kappa.



Antonio C. Berti Antonio Berti is currently contributing to the work of the Fab 4 engineering organization as a senior engineer. He has developed and implemented CMOS-4 processes in manufacturing, including metal deposition processes for cobalt, aluminum interconnections, TiN adhesion layers, and TiN local interconnects. Antonio also developed a cobalt silicide process to decrease defects and implemented TiN deposition hardware and processes to reduce particles. He graduated from Rensselaer Polytechnic Institute in 1988 with a B.S. in materials engineering and joined Digital in 1989.

Joseph M. Bulger Joseph Bulger is a principal engineer with the Fab 4 manufacturing engineering organization. His main areas of involvement are thin film manufacturing, chemical vapor deposition of tungsten, and metallization. Joe joined Digital in 1988 after 10 years with Analog Devices, Inc., working on thin film development and manufacturing. He received his B.S. degree in electrical engineering from Northeastern University in 1982.



Thomas E. Clark A principal engineer, Tom Clark is currently a member of the CMOS-6 technology team within the Advanced Semiconductor Development Group. Tom has developed blanket tungsten plug modules for CMOS-4, CMOS-5, and CMOS-6 technologies. He also performed feasibility studies of Co and Ti salicidation processes for contact, source, drain, and gate applications in the CMOS-4 technology. Tom holds a Ph.D. in inorganic chemistry from the University of Massachusetts and has published numerous papers on the subject of CVD tungsten metallization. He has received one patent.

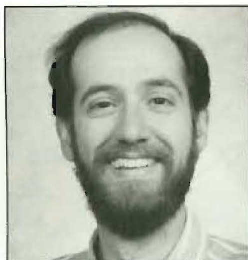
J. Joseph Clement A principal engineer on the interconnect physics project in the Advanced Semiconductor Development Group, Joe Clement works primarily on metal reliability and performance issues for VLSI. Before joining Digital in 1987, he held positions at Fairchild Semiconductor and Sandia National Laboratories. Joe received a Ph.D. in electronic materials and devices from Princeton University and a B.S.E.E. from the University of Notre Dame. He has published four papers on interconnect reliability and serves as a mentor for SRC research in electromigration at the Massachusetts Institute of Technology.

Randall S. Collica Randall Collica, a principal yield engineer, joined Digital in 1987. His primary areas of interest are SRAM fault analysis, yield modeling, fault tolerance, and experimental design and analysis. For the CMOS-4 technology, he established the use of yield models for redundancy on advanced processor chips containing on-board cache RAMs. These yield models are used for line control and productivity optimization for VLSI products. Prior to this work, Randall was engaged in failure analysis, test chip design, and process experimentation. He holds a B.S.E.E. from Northern Arizona University and is a member of IEEE.

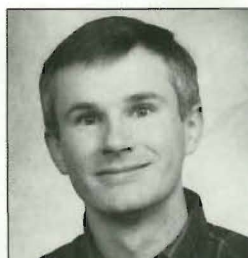
X. Joseph Dietrich II Joe Dietrich is a senior engineer whose work involves yield enhancement for CMOS technologies. His responsibilities for the CMOS-4 process included work on the design for manufacturability and yield modeling. He is currently performing the same yield analyses for the CMOS-5 process. Joe has also developed photolithography processes for earlier generations of CMOS technologies. Joe joined Digital in 1983 after graduating from Rensselaer Polytechnic Institute. He was awarded a B.S. degree in physics in 1983.

Brian S. Doyle Born in Dublin, Ireland, Brian Doyle graduated from Trinity College, Dublin, in 1975, and received an M.Sc. and a Ph.D. from the University of London. He held a postdoctorate position at the University of London from 1980 to 1982. For the next six years, Brian was employed by Bull, France, and worked on cryogenic MOS devices and hot carrier effects. He joined Digital in 1988 and is a principal engineer in the Gate Insulator and Transistor Reliability Group. Brian is a member of IEEE and is on the technical committee for the 1992 International Electron Devices Meeting.

Jamshed H. Dubash Jamshed Dubash is a senior process engineer with the Fab 4 manufacturing organization. For the CMOS-4 technology, he developed corrosion-free metal etch processes and a sloped via process for metal step coverage. Prior to this work, he contributed to the etch process development of CMOS-3 generation semiconductors. Jamshed is currently responsible for the process development of the integration of SOG planarization processes into manufacturing. He joined Digital after receiving his B.S. degree from the Rochester Institute of Technology in 1988.

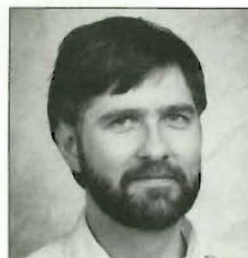


John V. Faricelli John Faricelli is a principal manufacturing engineer in the Advanced Semiconductor Development Group. Since joining Digital in 1983, he has worked on projects related to device physics and modeling, including cryogenic operation of MOS devices, hot electron effects, device isolation, and CMOS latchup. John has contributed to several device and process simulation programs, including MINIMOS, PISCES-II, and PROMIS. He holds a B.S.E.E. from Rensselaer Polytechnic Institute, and M.S.E.E. and Ph.D.E.E. degrees from Cornell University. He is a member of Tau Beta Pi and Eta Kappa Nu.



James A. Farrell Jim Farrell is a principal engineer in the Semiconductor Engineering Group (SEG). He was most recently engaged in the development of high-performance switching circuits, with particular emphasis on low-temperature operation. Prior to this work, Jim was a static RAM designer in SEG, with responsibility for developing the CMOS-3 process with the Advanced Semiconductor Development Group and for designing several memory chips. He joined Digital in 1985, after receiving B.S.E.E. and M.E.E.E. degrees from Rensselaer Polytechnic Institute in 1984 and 1985, respectively.

Bruce J. Fishbein Bruce Fishbein attended Cornell University, where he received a B.S. degree in materials science and engineering in 1983. He also received M.S. (1985) and Ph.D. (1988) degrees in electrical engineering from Stanford University. Since joining the Advanced Semiconductor Development Group at Digital in 1988, Bruce has worked on gate insulator and transistor reliability for CMOS microprocessors. Specific areas of interest include time-dependent dielectric breakdown, plasma-induced dielectric damage, thermal instabilities in gate dielectrics, and the use of high dielectric constant materials.



Thomas F. Fox Frank Fox, a consulting engineer in the Semiconductor Engineering Group, co-led the implementation of the NVAX microprocessor and consulted with the Advanced Semiconductor Development Group on the design of the CMOS-4 technology. He joined Digital in 1984 and worked on the implementation of the CVAX microprocessor. Frank received a B.E. degree from University College Cork, National University of Ireland (1974), and a Ph.D. degree from Trinity College, Dublin University (1978), both in electrical engineering. He has published papers on ultrasonic instrumentation and VLSI design; he has two patents.

Marion M. Garver Marion Garver was the project leader and supervisor of the CMOS-4 back-end process development group and is currently performing the same function for the CMOS-6 technology. Marion joined Digital in 1980 soon after the opening of the Hudson site. She initially worked in bipolar manufacturing fabrication and joined the Advanced Semiconductor Development Group in 1982 to develop some of the original plasma etching systems for NMOS and CMOS applications. Marion holds a B.A. degree (1973) from Oberlin College and was previously employed by Alpha Industries and Ohio Medical Products.

Gregory J. Grula A consultant engineer in the Advanced Semiconductor Development Group, Greg Grula has worked on process development and integration for CMOS-1, CMOS-2, and CMOS-4 technologies. Currently he is the supervisor of the CMOS-5 front-end process technology and has performed significant work in the isolation of transistors for CMOS-5. Greg joined Digital in 1978 after five years with RCA Corporation. He received a B.S.E.E. from Wilkes University and belongs to the Electrochemical Society and IEEE. Greg was awarded one U.S. patent and has three patents pending.

Daniel B. Jackson Daniel Jackson received a B.S.E.E. (1978) from Worcester Polytechnic Institute and was awarded M.S. (1980) and Ph.D. (1985) degrees from the University of Illinois for work on semiconductor device physics. He then joined Digital, working on advanced MOS devices and process development. Daniel has managed the Gate Insulator and Transistor Reliability Group in Advanced Semiconductor Development since 1989. He is interested in advanced process development and in MOS device physics and transistor reliability. A member of IEEE, Daniel has served on the technical committee for the IEDM.

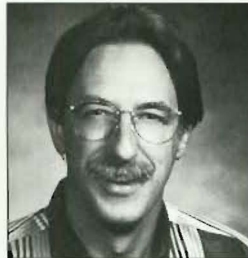
Richard D. Jones A principal engineer in the Advanced Semiconductor Development Group, Richard Jones developed etch processes such as local interconnect etch, polysilicon etch, and metal etch for the CMOS-4 technology. He is currently developing polysilicon etch and spacer etch processes for CMOS-6. Richard joined Digital in 1984. He received a Ph.D. (1986) and an M.S. (1982) in physics from Rensselaer Polytechnic Institute and is a member of the American Physical Society and the American Vacuum Society. He has published two papers on the properties of semiconductor materials.

Nadim A. Khalil Principal engineer Nadim Khalil is a member of Digital's Advanced Semiconductor Development Submicron Simulation Group. Currently developing optimization-based tools for process and device design, Nadim was involved in graphics and numerical software development and analytical modeling for CMOS technologies. He joined Digital in 1985, after receiving a B.S.E.E. (with distinction) from the American University of Beirut and an M.S.E.E. from Louisiana State University. Nadim is presently pursuing a Ph.D. degree at the Technical University of Vienna.

David B. Krakauer Since joining Digital's Advanced Semiconductor Development Group in 1989, David Krakauer has worked on topics in the area of transistor reliability, including hot carriers and electrostatic discharge (ESD). In particular, he has been involved in the design and characterization of ESD protection for Digital's advanced semiconductor processes. David graduated from the Massachusetts Institute of Technology in 1989 with S.B. and S.M. degrees in electrical engineering.

Rudolf Lambracht Jr. As a principal yield engineer, Rudolf Lambracht performs analysis to identify yield-limiting process steps and root causes for yield improvement. He is also interested in tools for failure analysis. Rudy worked on the CMOS-4 technology and is currently involved in the CMOS-5 technology. In his previous position as senior ion implant engineer, he was responsible for the implant process for MOS, CMOS, and advanced bipolar processes. Rudy joined Digital in 1978. He holds an A.S.E.E. degree from Waterbury State College and is coauthor of a paper concerning ion implantation.

David G. Lau As a senior manufacturing engineer for the Advanced Semiconductor Development Group, Dave Lau was responsible for driving yield issues during CMOS-4 development and analyzing yields for the CMOS-4 process in conjunction with the manufacturing yield group. Currently, he is developing yield strategies for CMOS-6. In a prior position, he developed and implemented device reliability characterization techniques and designed a CMOS process for liquid nitrogen temperature operation. Dave holds B.S. and M.S. degrees from MIT. He has coauthored three papers for IEEE International Electron Devices Meetings.



James R. Lloyd Principal engineer James Lloyd of the Design and Reliability Assurance Group specializes in metal reliability problems, primarily electromigration and stress voiding. He is also studying physical mechanisms related to thin film reliability in integrated circuits at the Max Planck Institut für Metallforschung in Stuttgart, Germany. Before joining Digital in 1988, Jim worked for IBM. He holds four patents, has written more than 50 papers for refereed journals, and is an invited speaker in his field. Jim received a Ph.D. (1978) from Stevens Institute of Technology in materials and metallurgical engineering.

Ellen J. Mager Ellen Mager is a member of the Fab 4 process engineering organization within Semiconductor Manufacturing. An engineering supervisor, Ellen is currently involved with the ongoing defect reduction of CMOS-4 material, as well as preparing for the transfer of the next technology level (CMOS-5). Ellen joined Digital in 1987 after working as a process engineer for National Semiconductor Corporation. She received her B.S. degree in chemical engineering (with high distinction) from Colorado State University.

Gerd Nanz Gerd Nanz received his Diplommathematiker degree from the Technical University of Munich, Germany, in 1985. From 1985 to 1986, he remained at the University as an assistant teacher in the Department of Civil Engineering, where he was involved in the development of self-adaptive finite element programs. Gerd then joined Professor Selberherr's group at the Technical University of Vienna, Austria, and completed his Ph.D. thesis on numerical methods in device simulation in 1989. Since 1990, he has worked on device and process simulation at Digital's Campus-based Engineering Center in Vienna.

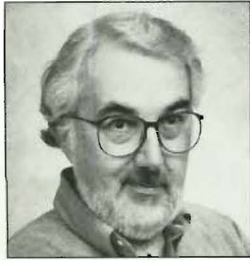
Andre I. Nasr A consultant engineer with the Advanced Semiconductor Development Group, Andre Nasr is the project leader in the 0.35-μm CMOS technology for front-end process development and on-chip memory implementation. He was the project leader for the CMOS-4 device group, responsible for the 0.75-μm process and device architecture. Andre introduced and implemented the drain engineering concept in Digital's CMOS processes. He also pioneered the high-performance BiCMOS process development. Andre has one patent issued and two pending. He holds an M.S. in physics from the University of Massachusetts.

Mary Beth Nasr Mary Beth Nasr is a member of the Advanced Semiconductor Development Group. She is currently the engineering supervisor of the Metrology and Microcontamination Group, with direct responsibility for defect inspection tools. Prior to joining Digital in 1987, Mary Beth was a microcontamination control engineer for Mitsubishi Semiconductor Corporation in Japan for one and one-half years. In 1985 Mary Beth received a B.S. in chemical engineering from Columbia University and a B.S. in mathematics from Providence College.

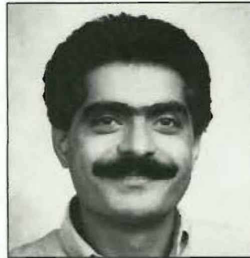
Llanda M. Richardson Consultant engineer Llanda Richardson manages the Submicron Physics and Chemistry Group within the Advanced Semiconductor Development Group. She established the Device and Process Simulation Group and is responsible for ASD's participation in the Campus-based Engineering Center in Vienna, Austria. Llanda is a member of the CMOS Development Committee and Digital's alternate representative on the SRC Board of Directors. She joined Digital in 1979 after receiving a Ph.D. in physics from the University of Vermont.

Lorain M. Ross Lorain Ross joined Digital in 1984 after two years with NCR Corporation. She has worked primarily toward the photolithography process enhancement of several generations of bipolar and CMOS devices. Lorain participated as a member of the original task force to introduce SPC methods for controlling critical parameters of semiconductor manufacturing at the Hudson site. As a principal process engineer for the Fab 4 organization, she is responsible for the integration of newly developed photo processes into manufacturing. Lorain holds a B.S. degree in chemical engineering from Yale University.

Christian O. Schiebl Christian Schiebl received a Diplomingenieur degree in physics from the Technical University of Vienna, Austria, in 1986. The subject of his master's thesis was quantitative electron probe microanalysis (EPMA) using characteristic M-lines. He studied at the Institut für Angewandte und Technische Physik at the Technical University of Vienna and received a Ph.D. in 1989. His dissertation was on characteristic fluorescence correction in quantitative EPMA. Christian then joined the Campus-based Engineering Center in Vienna, where he currently works in the field of semiconductor process and device simulation.

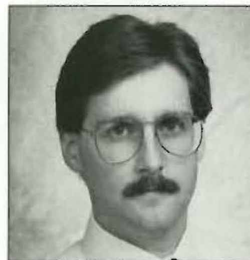


Marden H. Seavey Marden Seavey, a consultant engineer, came to Digital in 1981 from Raytheon Company. He was the supervisor of the device and process simulation projects within the Advanced Semiconductor Development Group for five years. Marden pioneered the adaptation of the MINIMOS program to the ASD CMOS process and device design. The mobility model that he developed also made it possible to apply MINIMOS in worst-case circuit design. Marden received a Ph.D. in applied solid-state physics from Harvard University. He retired from Digital in June 1992.



Hamid R. Soleimani Senior manufacturing engineer Hamid Soleimani joined Digital in 1987 and is a member of the Submicron Simulation Group within Digital's Advanced Semiconductor Development Group. He developed the transient diffusion model that has been incorporated in the SUPREM-III process simulation program and is used for CMOS technology development work. Hamid holds B.S.E.E. (1984) and M.S.E.E. (1986) degrees from Louisiana State University, Baton Rouge, Louisiana. He is a member of the IEEE CAD and Electron Device Societies.

Martin Thurner Martin Thurner is a principal engineer in the Submicron Physics and Chemistry Group within the Advanced Semiconductor Development Group. He is responsible for the three-dimensional device simulations investigating MOS device width effects. Martin has improved the efficiency and accuracy of the mathematical and physical models used in the MINIMOS program. Prior to joining Digital in 1988, he worked at the Federal Research and Test Institute of Austria. Martin received Diplomingenieur and Ph.D. degrees from the Technical University of Vienna, Austria.



Daniel J. Welch As a principal engineer in the Fab 4 process engineering thin films group, Dan Welch is responsible for the development and optimization of interlevel dielectric and planarization processes. He was project leader for the development and implementation of the SOG planarization process for CMOS-4. Dan joined Digital in 1991 after affiliation with Intel Corporation as a senior process engineer for the high-volume manufacture of advanced microprocessors. He received a B.S. in chemical engineering in 1983 from Clarkson University.

Bjorn Zetterlund Bjorn Zetterlund, a manufacturing consultant engineer, came to Digital in 1985 after 12 years with Raytheon Company. He is the device engineering supervisor for CMOS-5 and is responsible for CMOS-5 transistor design and the CMOS-5 technology file. Prior to this work, Bjorn was responsible for transistor design in CMOS-3 technology. He received B.S., M.E., and Ph.D. degrees in electrical engineering from Rensselaer Polytechnic Institute. Bjorn is a member of IEEE, Tau Beta Pi, and Eta Kappa Nu.

Foreword

R. J. Hollingsworth
Manager, Advanced Semiconductor Development

Digital has developed and manufactures the world's fastest production complex instruction set (CISC) and reduced instruction set (RISC) microprocessors. The speed of these microprocessors is due, in large part, to a complementary metal oxide semiconductor (CMOS) technology having faster transistors, as dense on-chip wiring, and innovative performance-enhancing materials and structures. This advanced CMOS technology allows Digital the unique capability to design and produce microprocessors that operate twice as fast as common leading-edge devices produced by the world's premier semiconductor manufacturers.

In 1980, Digital recognized the strategic role that microprocessor chips played as core elements in reshaping and advancing the computer industry. A key observation was that the unrelenting advances in very large-scale integrated (VLSI) circuits would continue to allow vast amounts of logic and memory to be economically produced in a single silicon chip, thereby yielding dramatic improvements in performance, cost, and reliability. VLSI devices were demonstrating yearly improvements of 10 to 15 percent in gate switching speed and 25 to 35 percent in density. Microcontamination and process control methods coupled with increasing wafer size allowed larger chips to be fabricated at lower cost. It was clear that the ability to integrate more and more function into a piece of silicon was fundamentally changing the computer industry. The era of entire computing systems on a single chip was rapidly approaching. Chips were not just components in a system, they were becoming the system. To fully exploit this and lead Digital into what C. Gordon Bell termed a "semi-computer" company in the 1990s, the decision was made to develop and subsequently manufacture semiconductor technologies.

Digital's semiconductor operations group set a goal in the early 1980s to achieve leadership in the development and manufacture of the world's highest performance microprocessors. To meet this, a number of strategic positions were taken:

- Develop distinct generations of CMOS that would produce a wide range of VLSI devices, not only microprocessors.
- Be at the leading edge in density; be ahead of the industry in high-speed switching devices and system-level features.
- Make CMOS technology decisions by optimizing a wide range of requirements necessary to meet the goal: the world's fastest microprocessors. The approach would be a rigorous engineering optimization from computer architecture through chip manufacturing processes.
- Develop CMOS with a single, multi-disciplined technical team dedicated to the project from initial conception through manufacturing qualification—a four- to five-year endeavor.
- Use the microprocessor product, targeted to world-leading performance, as the specific focal point for CMOS development. Tie together the efforts of full-time chip architecture, design, test, reliability, packaging, and manufacturing people for the full project duration, i.e., four to five years.
- Develop CMOS technology in conjunction with the microprocessor architecture and design—an essential ingredient in delivering leading VLSI chips to the market first. This "concurrent" approach has been a mainstay in Digital's CMOS development since the early 1980s.
- Participate in, contribute to, and draw upon the best semiconductor research in the world.

Many leading semiconductor companies follow these strategies. Digital, however, is unique in practicing all of them.

To help guide the technical direction, a CMOS technology roadmap was created in the early 1980s. It defined the key pacing elements that delineate each distinct CMOS generation: minimum feature size, switching speed, manufacturable chip and wafer size, and other attributes necessary to deliver leading-edge microprocessors. This roadmap set the goals for the organization and allowed easy

reference to competitive trends. In its simplest form, the roadmap defined each CMOS generation by making logic gates switch 30 percent faster while occupying half the silicon area and by integrating these elements on chips that were growing 40 percent in area compared with the previous generation. The roadmap today defines eight generations of CMOS; four have been introduced to manufacturing (CMOS-1 through CMOS-4), and two are now under development (CMOS-5 and CMOS-6).

The papers in this issue of the *Digital Technical Journal* are focused on Digital's fourth-generation CMOS (CMOS-4) that is used to build a wide variety of VLSI chips, most notably the NVAX and Alpha 21064 microprocessors. CMOS-4, presently being manufactured in Digital's semiconductor facilities in Hudson, MA, and South Queensferry, Scotland, delivers microprocessors with up to 1.7 million transistors operating at clock rates up to 200 MHz. A variety of materials and structures have been crafted into CMOS-4, allowing advanced system-level capabilities not available in other state-of-the-art CMOS processes.

This issue spans the breadth of technical areas necessary to make CMOS-4 a successful element in establishing Digital's preeminent position in microprocessor technology. The discussions herein include how manufacturing process and chip design trade-offs are made; how a large number of complex manufacturing steps are integrated; how leading-edge speed, density, and materials are achieved; and what modeling, simulation, and measurements are critical to ensure reliability and produceability. The papers are a sample of the range of technological achievements in Digital's semiconductor operations.

Bjorn Zetterlund
James A. Farrell
Thomas F. Fox

Microprocessor Performance and Process Complexity in CMOS Technologies

Digital's CMOS technology is characterized by a scaling methodology that doubles the gate density and improves the gate speed by approximately 30 percent with each new generation. Decreasing feature size from one generation of CMOS technology to the next is fundamental to improving the performance of VLSI chips. Each of Digital's successive CMOS generations has added new technology features to improve performance further. Digital's latest, qualified CMOS technology incorporates features such as low voltage operation, low-resistance topside substrate contacts, low-resistance transistor gate material, local interconnects in SRAMs, three levels of metal interconnect, and fuses for redundancy.

The goal of Digital's semiconductor organization is to provide leadership in product performance and functionality, as most recently evidenced by the Alpha 21064 and NVAX microprocessors.[1,2] Internal development of complementary metal-oxide semiconductor (CMOS) processes has been crucial to the success of these chips because it allowed us to design the process to meet very specific needs. The identification and fulfillment of these needs has been a multigenerational, ongoing task that closely links the chip design effort with the process development.

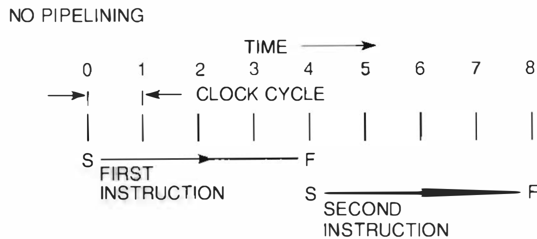
Each new generation of CMOS technology is scaled to double the gate density and improve the gate speed by 30 percent. In addition to the generation-to-generation improvements that are derived from scaling, a CMOS process that is designed specifically for high-performance microprocessor applications requires a number of features beyond those normally implemented in CMOS processes. As a new process is being developed, proposed new features are critically evaluated in order to arrive at the optimum trade-off between chip performance and process complexity.

This paper describes Digital's CMOS processes from the perspective of those process features that contribute to the performance and functionality of high-speed microprocessors. It begins with a short discussion on microprocessor architecture, which strongly influences the direction of process development.

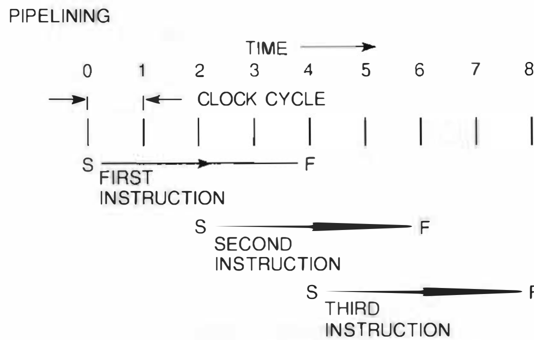
CMOS Microprocessors—General Considerations

Several factors determine the performance of the fastest microprocessor that can be built in a given CMOS technology. The performance of a microprocessor is inversely proportional to the product of clock cycles per instruction (CPI)[3] and the machine cycle time. From one generation of microprocessors to the next, improvements in both CPI and machine cycle time are required in order to meet the performance goal. CPI depends on the microarchitecture as well as the mix of instructions executed. The minimum machine cycle time depends only on circuit and microarchitectural issues.

Improvements in CPI are achieved in part by pipelining and in part by adding cache memory to the die. As shown in Figure 1, pipelining is defined as the simultaneous execution of two or more instructions; for example, the first part of a two-part instruction is executed at the same time as the second part of the previous instruction. Simultaneous execution reduces the number of CPI; the result is higher performance at the expense of more circuitry.[4] Adding on-chip cache memory also reduces the CPI. The microprocessor can quickly access instructions or data resident in its caches, rather than wait for this information to be transmitted from on-board random-access memory (RAM) chips.



NO PIPELINING: Second instruction does not start until the first instruction finishes. One instruction is executed every four cycles, so CPI = 4.



PIPELINING: Second instruction starts halfway through the execution of the first instruction. Third instruction starts halfway through the execution of the second. One instruction is executed every two cycles, so CPI = 2.

Note: It takes more circuitry to implement a pipeline since instructions are executed simultaneously.

KEY:

S START
F FINISH

Figure 1 Relationship between CPI and Pipelining

As the minimum feature size decreases, transistor density increases. Thus, for a given die size, a 0.75-micrometer (μm) minimum feature size technology (CMOS-4) can support four times as many transistors as a 1.5-μm technology (CMOS-2). In addition, advances in process technology lead to increases in the size of the largest die that can be built with an acceptable manufacturing yield.

Microprocessor designers take advantage of the extra transistors to increase the degree of pipelining and the size of caches. This reduces the CPI, which boosts the performance of the machine. Table 1 shows the difference between two generations of scaled CMOS processes. The halving of the feature size has been augmented by a larger die size and microarchitectural changes that increase the SPECmark[5] performance by a factor of 4.8.

Generation-to-Generation Scaling

The pace of improvement in microprocessor performance indicated by the REX520/NVAX comparison is ahead of the industry.[6] Microprocessor performance has been doubling approximately every two years, and there is no evidence that this pace of change is slackening. The shrinking of feature sizes with each new generation of CMOS process technology and the increase in yieldable die sizes have enabled this rapid improvement in the performance of very large-scale integration (VLSI) chips.

At the center of the reduction in the feature sizes of CMOS processes is the miniaturization of the MOS transistor. Over the past 15 years, a set of rules, known as scaling theory, has been developed to guide this process.[7,8,9]

In the fundamental form of scaling, called constant field scaling, the transistor's physical parameters and the power supply voltage are kept proportional to the feature sizes to maintain the magnitude and the contours of the electric fields within the transistor. All the dimensions of the transistor, e.g., length, width, gate dielectric thickness, and source/drain junction depths, and the power supply and threshold voltages are reduced by the scaling factor, [$1/\kappa$ (where κ is greater than 1)], while the doping concentrations are increased by κ. These rules are also extended, with some exceptions, to guide the miniaturization of the interconnect. In practice, scaling theory is not followed exactly, for reasons of both performance and standardization that are discussed below. Digital's implementation of scaling through four CMOS generations is shown in Table 2.

Improved Performance through Scaling

The reduced feature sizes made possible by scaling have a major impact on node capacitance and, hence, the speed of the chip. The minimum cycle time of a microprocessor is inversely proportional to the capacitances of the gates, sources, drains, and interconnect. The gate capacitance is inversely proportional to the thickness of the gate dielectric, and transistors with thinner gate dielectrics have higher drive current. Since the minimum cycle time is a stronger function of transistor drive current than gate dielectric capacitance, the trade-off should be made in favor of a thinner gate dielectric.

In Digital's family of CMOS processes, gate capacitances have scaled with minimum feature size.

Table 1 Comparison of Single Chip VAX Microprocessors

Process	Minimum Feature Size	Chip	Tape Out Date	Performance SPECmarks*	Cycles per Instruction†	Cycle Time (Nano-seconds)	Chip Size (Mils)	No. of Transistors
CMOS-2	1.5 μm	REX520	Sep 87	8.5	11.95	28	460×460	320,000
CMOS-4	0.75 μm	NVAX	Nov 90	40.5	5.85	12	636×574	1,300,000

Notes:

*These are combined integer and floating point SPECmarks run on a VAX 6000 Model 410 (REX520) and a VAX 6000 Model 610 (NVAX).

†CPI depends not only on the CPU chip, but also on the memory subsystem and the particular program being executed. The CPI values quoted here are a composite for the ten benchmark programs in the SPECmarks suite.

Table 2 Comparison of Feature Sizes in CMOS Generations

	CMOS-1	CMOS-2	CMOS-3	CMOS-4
Gate Dielectric Thickness (Å)	300	225	150	105
Minimum Feature/Space (μm)				
Active area	4/2	3/1.5	2/1	1.5/0.75
Polysilicon/polycide	2/2	1.5/1.5	1/1	0.75/0.75
Metal 1 with contact or via	4/2	3/1.5	2/1	1.5/0.75
Metal 2 with via	5/2	3.75/1.5	2.5/1	1.875/0.75
Metal 3 with via			4/6	3/4.5
Minimum Feature Size (μm)				
Metal 1 contact	2	1.5	1	0.75
Metal 2 contact	2	1.5	1	0.75
Metal 3 contact			4	3
Minimum Spacing (μm)				
Metal 1 contact/polysilicon (in active area)	2	1.5	1	0.75
P+/N+ active area	8	6	4	3

From the 1.5-μm minimum feature size of CMOS-2 technology to the 0.75-μm size of CMOS-4, the area of the gates was scaled by a factor of four and the gate dielectric thickness was halved. The result is a twofold reduction in gate capacitance ($C = \epsilon_0 A/t$). The typical gate dielectric thickness in the 0.75-μm CMOS-4 process is 105 angstroms (Å). Manufacturability and reliability considerations have been the major factors determining the minimum gate dielectric thickness used for each generation.

As shown in Figure 2, the sources and drains of n-channel metal-oxide semiconductor (NMOS) transistors form N+/P diodes to the substrate. Since the p-type doped substrate is held at a potential of V_{ss} (ground) and (during normal operation) the sources and drains are always at V_{ss} or higher, these diodes are always reverse biased and act as voltage-dependent capacitors. The sources and drains of p-channel metal-oxide semiconductor (PMOS) transistors form P+/N diodes to the n-wells; the n-well is held at V_{dd} (power supply voltage).

The capacitance of a reverse-biased diode is a function of its shape and size: there is both an area component and a perimeter component.

$$C_{total} = C_{area} \times \text{Area} + C_{perimeter} \times \text{Length_of_Perimeter}$$

Since the area scales with the square of the minimum feature size and the perimeter scales directly with the feature size, C_{total} scales by somewhat more than the minimum feature size: the exact amount depends on the shape of the source or drain. For the NVAX microprocessor, which was designed in CMOS-4, the area and perimeter components contribute about equally to C_{total}. In future technology generations, the perimeter component will tend to dominate.

In Digital's CMOS processes, metal interconnect widths and spaces are scaled with the minimum fea-

ture size, but metal thicknesses and dielectric thicknesses are held constant to avoid three undesirable effects. Scaling the metal thickness would increase the sheet resistance (leading to larger power supply voltage drops and RC time constant delays) and decrease the current-carrying capability of the conductor lines (from an electromigration viewpoint). Scaling the interconnect dielectric thickness would increase the capacitance per unit area. Because the thicknesses of conductor lines and dielectric layers are not scaled, the aspect ratios of the spaces between the conductors and the contacts or vias between interconnect layers increase. This makes fabrication more difficult.

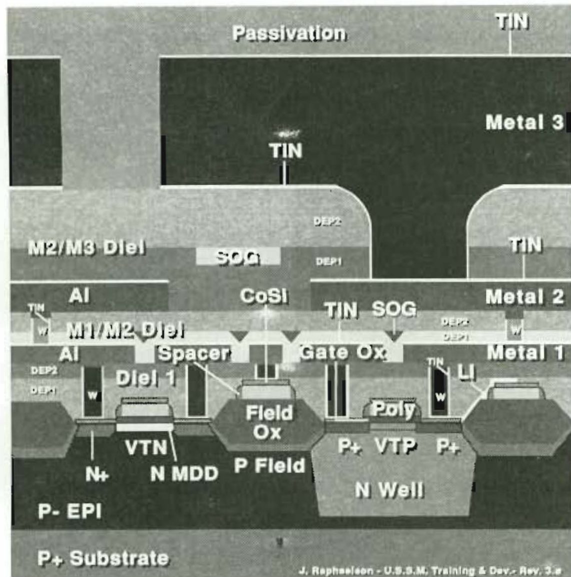


Figure 2 Diagram of NMOS and PMOS Transistors Showing Gate, Source, and Drain for CMOS-4

As with sources and drains, interconnect capacitance has both an area component, which scales quadratically, and a perimeter component, which scales linearly. Consequently, the total capacitance of interconnect scales by somewhat more than the minimum feature size. Because neither the interconnect thickness nor the dielectric thickness is scaled, the capacitance between adjacent conductor lines increases. The result is an increased susceptibility to cross-talk between adjacent bus signals, which is shown in Figure 3. In the NVAX microprocessor, greater-than-minimum spaces were used on some critical buses to reduce cross-talk.

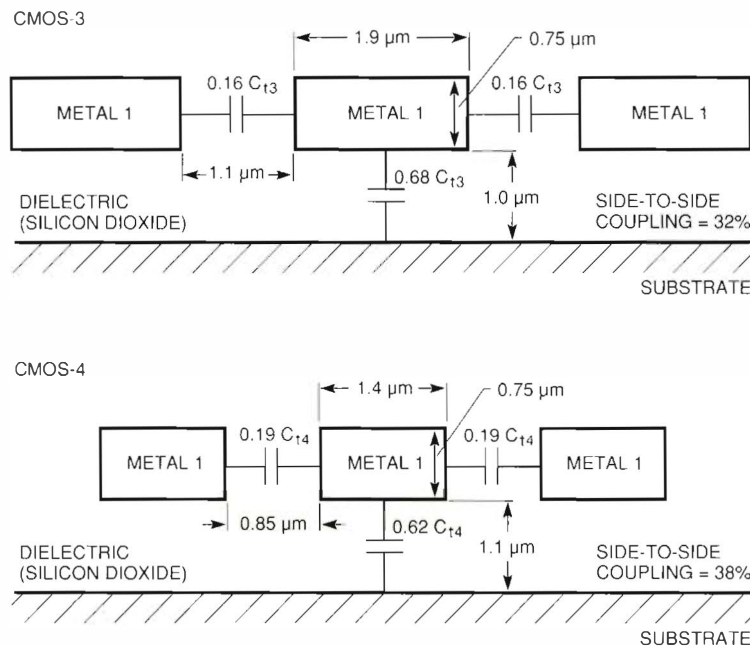


Figure 3 Cross Section of Three Minimum-spaced Metal 1 Lines Drawn to Scale for CMOS-3 and CMOS-4

The use of industry-standard power supply voltages results in a significant violation of constant field scaling rules. Nevertheless, power supply voltage is generally held constant across two or more process generations [5.0 volts (V) in CMOS-1 and CMOS-2 and 3.3 V in CMOS-3 and CMOS-4] in order to maintain voltage compatibility with industry-standard chips such as RAMs. However, a nonscaled power supply voltage presents formidable challenges for the design of reliable transistors.

Developing CMOS for Microprocessors

The particular implementation of transistors, interconnect, and special circuit elements in a CMOS process depends on the application. For Digital's high-speed microprocessors, performance, as measured in SPECmarks, is crucial.[5] In addition to optimizing the transistors for maximum drive current, performance in this application can be improved by adding process features to provide denser on-chip cache static RAM (SRAM), interconnect with high current capability, and precision resistors for impedance matching. As discussed above, performance can also be improved by increasing the die size. By contrast, a major part of the effort in designing a process for dynamic RAMs (DRAM) is directed toward developing a very small, high-capacitance memory element.

In addition to performance, the planned production volume is an important factor in defining a CMOS process. A process for low- to moderate-volume, high-performance microprocessors differs from a process optimized for fast turnaround gate arrays or high-volume RAMs. In a high-volume product, a great deal of effort is devoted to reducing the total number of process steps. For example, compensating blanket implants are often used to set the thresholds of the transistors to decrease the number of photolithographic masking steps. This approach couples the parameters for the NMOS and PMOS transistors, making parameter adjustments more difficult and requiring more development effort.

To produce high-performance microprocessors, Digital has developed many unique features for its CMOS technologies. Table 3 lists the new technology features that have been developed for each process generation to meet increasingly demanding performance requirements. We begin our discussion of the implementation of these features and the requirements for reliable circuit operation by addressing the issue of power dissipation in high-speed microprocessors.

Power Supply Voltage

It is well known that CMOS power dissipation is dominated by $C \times V_{dd}^{2} \times f$, where C is the switched capacitance, V_{dd} is the power supply voltage, and f is the clock frequency. A reduction in V_{dd} is an excellent way to counteract the increase in power due to

Table 3 Features Added by Generation for CMOS-1 to CMOS-4

	CMOS-1	CMOS-2	CMOS-3	CMOS-4
Masks	12	13	20	21
Minimum Dimension	2.0 μm	1.5 μm	1.0 μm	0.75 μm
New Features	1X photolithography n-well Epitaxial layer Borophosphosilicate glass planarization	Lightly doped drain junction/spacer Tungsten silicide gate Deep P+ ring Photoresist etch-back planarization	5X g-line photolithography n, p-polysilicon Cobalt silicide TiN barrier Al:1%Cu Metal 3 Fuse	Local interconnect Tungsten plug for M1C, M2C Spin-on-glass etch-back planarization

the higher frequencies and larger switched capacitance (which results from the increase in die size). When developing the CMOS-3 process, we chose to reduce the power supply voltage from the industry norm of 5 V to the Joint Electronic Device Engineering Council (JEDEC) 3.3 V standard.[10,11]

With the CMOS-4 process specified for 3.3-V power supply voltage, the NVAX and the Alpha 21064 microprocessors consume 16 watts (W) at 100 megahertz (MHz) and 27 W at 200 MHz, respectively. If the supply voltage were 5.0 V, the power would scale by $(5\ V)^2/(3.3\ V)^2 = 2.3$. This increase in power dissipation would have greatly increased the complexity and cost of the chip packages.

Significant changes to the NMOS and PMOS transistors were necessary to optimize the process for operation at 3.3 V. The most visible parameter change was a lowering of the target threshold voltages for the NMOS and PMOS transistors by about |0.4| V to 0.5 V and −0.5 V, respectively. To explain why this is necessary, we must consider the dependence of both the nodal transition time (which is a good measure of circuit performance) and the transistor currents on V_{dd}. The time required to transition a signal node between the power supply rails is proportional to the charge (Q) on the node and inversely proportional to the drain-to-source current (I_{ds}) of the driving transistor. Since $Q \alpha V_{dd}$ and, to first order, $I_{ds} \alpha V_{dd}^2$, the time required to transition a node is inversely proportional to V_{dd}. However, when second-order effects are considered, the 3.3-V technology is of about the same performance as the corresponding 5-V technology. The second-order effects include the benefits from lowering the threshold voltages of the transistors in the 3.3-V process and the compromises that would have to be made to the transistors in the 5-V process to make them reliable.

Hot Carrier Degradation

A CMOS transistor that is subjected to excessive voltages becomes damaged over time by hot carriers. Hot carriers are highly energetic current carriers that result from the high electric fields in the transistor. To date, the NMOS transistor has been more susceptible to hot carrier degradation than the PMOS transistor. Hot carrier damage gradually reduces the saturation current I_{DSAT} of the NMOS transistor as the damage increases over time. On chips with a nominal 3.3-V power supply, some transistors are subjected to source-drain voltage transients as high as 4.3 V. Table 4 gives details of the origins of these high voltage transients.

Hot carrier rules for the CMOS-4 process are illustrated in Figure 4, which shows the three legal regions of device operation on a plot of V_{gs} (gate-to-source voltage) versus V_{ds} (drain-to-source voltage). Devices may operate in any, or all, of three regions: (1) unconditionally safe region, (2) region subject to turn-on transient rule, and (3) extended safe region for "off" devices. Devices can spend up to

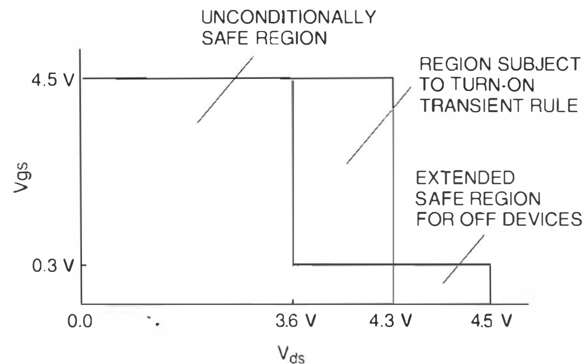


Figure 4 CMOS-4 Hot Carrier Rules

Table 4 Origins of High Voltage Transients

	V_{ds} (Volts)	Subtotal (Volts)	Comments
Power supply	3.30	3.30	Nominal voltage
Power supply tolerance	0.165	3.465	5% tolerance, includes ripple
On-chip power supply ringing due to package inductance	0.175	3.64	From NVAX SPICE simulations. Half of peak-to-peak noise (V_{dd}_internal with respect to V_{ss}_internal)
Booting above V_{dd} due to capacitive coupling	0.66	4.30	Capacitive coupling to susceptible nodes is limited to <20% by designers
Total	4.30		

100 percent of the time in the safe region, no more than 5 percent in the region subject to turn-on transient rule, and no more than 10 percent in the "off" devices safety region.

A wide variety of NVAX circuits were simulated to determine what constraints should be placed upon circuit design style in order to ensure that the CMOS-4 hot carrier rules were not violated. A set of general circuit design constraints was developed, and a computer-aided design (CAD) tool was written to ensure that all the circuits on the NVAX chip observed these constraints. The hot carrier CAD checks were run prior to fabrication, and circuits that violated the constraints were redesigned.

Electromigration Considerations

If the average current density ($J_{average}$) through an aluminum conductor line is too high, the conductor line is susceptible to metal migration. This phenomenon occurs over time as the electron current forms voids at one site and deposits downstream. Eventually a short circuit or an open circuit develops, which results in a circuit failure. Chip designers guard against electromigration failure by ensuring that $J_{average}$ for every conductor line on the chip is lower than the maximum allowed value.

For a conductor line that is switched every cycle, the relationship between average current density, microprocessor cycle time (T_{cycle}), V_{dd}, and cross-sectional area is given by $J_{average} = (C \times V_{dd}) / (T_{cycle} \times$ Cross-Sectional Area).

It is interesting to note the changes to $J_{average}$ for a conductor line as a chip is shrunk from one generation to the next. The node capacitance, C, decreases by slightly more than the scaling factor; V_{dd} remains constant; T_{cycle} reduces by the scaling factor since the chip can now run faster; and the cross-sectional area decreases by the scaling factor. Consequently, $J_{average}$ increases by slightly less than the scaling factor as the width of the conductor line shrinks. If $J_{average}$ now exceeds the maximum allowed value, the circuit must be redesigned. If there is enough space, $J_{average}$ can be reduced by widening the conductor line so that the cross-sectional area is increased. From this brief analysis, it is clear that it becomes more difficult to observe the electromigration limits as the technology scales—even when the metal thickness is not scaled. As can be seen from the $J_{average}$ equation, reducing V_{dd} from 5.0 V in CMOS-2 to 3.3 V in CMOS-3 helped to counteract the effect of scaling on $J_{average}$.

Scaling the interconnect and dealing with electromigration issues are some of the most formidable challenges that must be faced as feature sizes continue to decrease in the next decade.

Substrate Contact

As mentioned earlier, the substrate must be connected to the V_{ss} of the chip through a low-impedance path to prevent any rise in voltage. If the substrate voltage rise is severe, NMOS source/drain diodes will conduct, and if sufficient charge is injected, the chip may latch-up. Latch-up is a destructive mechanism involving the parasitic bipolar transistors formed by the CMOS process. The process, circuits, and V_{ss} substrate contact are designed to prevent latch-up from occurring.

The usual industry substrate connection method depends on a path through bond wires and the package to connect between internal V_{ss} and the substrate. To ensure a good substrate contact, Digital's CMOS technologies incorporate a deep P+ implant (DPI) around the edge of the die to connect the V_{ss} metal on the die surface to the low-resistance substrate. The implant creates a low-resistance path through the P-epitaxial layer in which the NMOS and PMOS transistors are formed.

The DPI is a low-inductance path when compared to the standard method that connects the substrate through the package. The additional area enclosed by the path through the bond wires and package implies greater inductance, which is undesirable for high-frequency signals. The DPI path between V_{ss} and substrate has low inductance because it is made directly on-chip.

Technology-limited Gate Dielectric Thickness

As stated above, maximizing the current that a transistor can supply at a given V_{dd} is of uppermost importance for circuit performance. Scaling the transistor gate length, gate dielectric thickness, and threshold voltage improves the drive current. The transistor gate length is constrained by the minimum polysilicon line-width feature; the minimum threshold voltage is set by the leakage current allowed when the transistor is turned off. However, scaling does not establish a fixed relationship between feature size and gate dielectric thickness; scaling only determines the change from one generation to the next.

Figure 5 shows how the saturation currents for both NMOS and PMOS transistors in CMOS-4 depend on the gate dielectric thickness. The dielectric thickness range plotted spans applications from microprocessors to SRAMs. Both curves are only slightly sublinear; thinning the gate dielectric provides almost a one-to-one return in transistor saturation current. In high-performance microprocessor applications, reliability and manufacturability considerations determine the extent to which the gate dielectric thickness can be reduced. Digital's CMOS technologies have consistently used thinner gate dielectrics than industry norms.

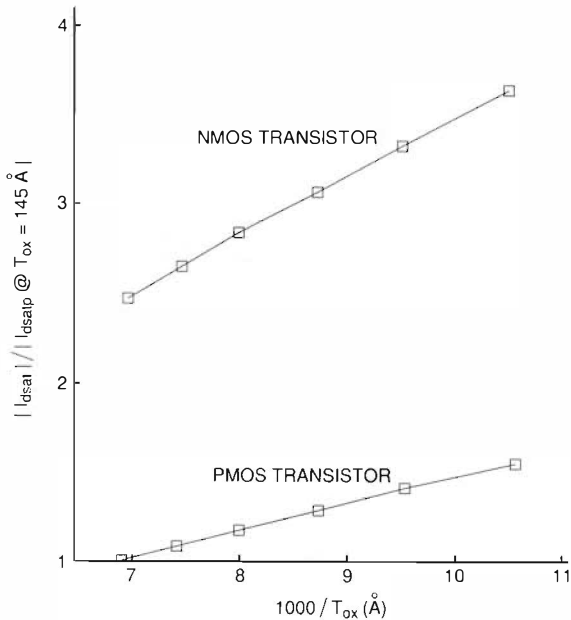


Figure 5 Normalized Drive Current as a Function of Gate Dielectric Thickness for CMOS-4

Silicided Source/drain and Gate

The basic gate material for an MOS transistor is highly doped polysilicon. The sheet resistivity of this material in the CMOS-1 process was 40 ohms per square. For CMOS-2, the RC time constant delay associated with this sheet resistance would have created nonuniform turn-on of wide, fast-switching output transistors. A tungsten silicide layer was added to the polysilicon to form a polycide. The sandwich structure reduced the sheet resistance of the gate material to 3 ohms per square.

Changes to the transistor process for CMOS-3 technology, which were continued into CMOS-4, required development of a new silicided gate process. The new process, known as salicide for self-aligned silicide, forms the silicide on the gate and on the source/drain regions after all the required transistor implants have been completed. This reduces the sheet resistance of the source/drain regions by more than an order of magnitude and allows them to be considered for use in local signal routing. The reduced sheet resistance, however, does little to improve the current drive of the transistor; for a typical CMOS-4 NMOS transistor, MINIMOS simulations show that the use of silicided source/drain regions improves the saturation current by only 0.6 percent.[12]

Precision Resistor

Although transistors are the dominant element in logic design, a resistor is sometimes needed, for example, to match the impedance of an output driver with the impedance of a board-level transmission line that it drives. MOS transistors make poor controlled impedance drivers because they change impedance as a function of drain voltage. One method of controlling the impedance of a driver is to use a diffusion resistor as the dominant element, as shown in Figure 6. The MOS transistors are sized such that their on-state resistance is much lower than that of the resistor. Therefore, if the transmission line impedance is 50 ohms, the resistor plus transistor impedance can be sized to match that value with little influence from the variations in transistor impedance.

Resistors are constructed from nonsilicided diffusion to meet tolerance requirements that would not be possible with a silicided version. Because a silicided resistor has lower sheet resistance, it is much longer and narrower than a nonsilicided

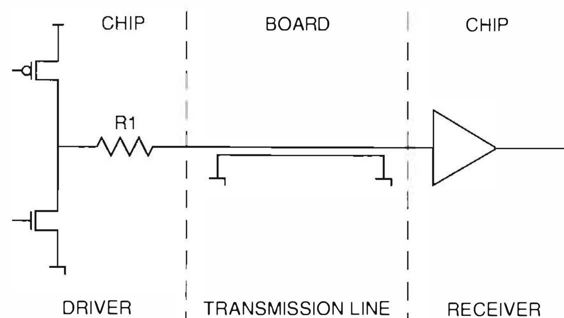


Figure 6 Precision Resistor Use in Impedance Matching

resistor of the same value. A narrow resistor is more susceptible to variations in field dielectric encroachment, which lowers the tolerance of the resistor. The tolerance is lowered further by process variation; it is more difficult to control the sheet resistance of the silicided diffusion than that of the nonsilicided diffusion.

The precision resistor is also an important element of our electrostatic discharge (ESD) protection strategy. Figure 7 shows a simple schematic of an I/O driver with the ESD protection components. Clamp 1 is the main path for shunting current during an ESD event. The position and construction of resistor (R) 1 and clamp 1 ensure that the clamp impedance is lower than that of a shunting path through R1 and the driver. R1 is also the impedance matching resistor for the output driver. R2 provides an additional level of protection for the gates of the input driver in conjunction with the smaller clamp 2 placed near the input driver. Both R1 and R2 are fabricated using the precision resistor mask layer.

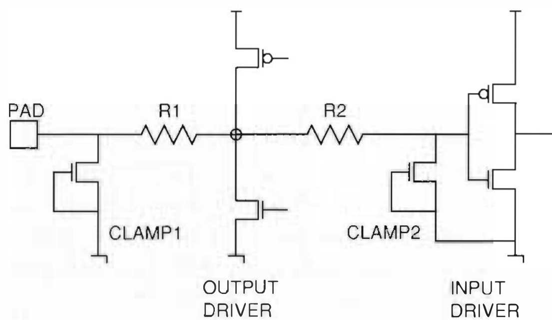


Figure 7 Precision Resistor Use in Electrostatic Discharge Protection

Local Interconnect in the SRAM

As stated earlier, one of the microarchitectural methods for increasing performance is to build as large a cache memory as possible on the die. Due to fabrication complexity, RAM-specific process enhancements are generally not implemented in a process tailored to microprocessors. However, the CMOS-4 technology does include one feature, called local interconnect, that significantly decreases the cell size. In the CMOS-4 implementation, local interconnect is a titanium-nitride (TiN) layer that provides direct contact between the polysilicon and diffusion layers.[13]

Figure 8 shows a comparison of layouts of memory cells with and without local interconnect. The six-transistor static cell layouts show an array of four cells with the PMOS load transistors at the top and bottom of the arrays and the NMOS pass transistors, which provide access to the bit lines, in the center of the array. The cell without local interconnect is 120 μm^2 in CMOS-4, compared to 98 μm^2 for the cell with local interconnect. The 18 percent improvement in area is important, but the cell also has features that increase yield. There are only 2.5 contacts between the first level of aluminum interconnect (M1) and diffusion or polysilicon for the local interconnect cell, compared to 8 in the non-local interconnect cell. None of the M1 in the local interconnect cell is at minimum pitch (where pitch equals M1 width plus space), while all the M1 in the other cell is at minimum pitch. Finally, the M2 pitch is smaller in the local interconnect cell, but at 3.19 μm, it is still greater than the minimum of 2.63 μm allowed by the technology. All of these factors add up to a significantly more yieldable cell; this is important since 15 percent of the area and about two-thirds of the transistors on the Alpha 21064 chip are SRAM cells.

Thick Metal 3 Interconnect

High-speed operation of the dense circuitry in large microprocessors results in a level of power dissipation not encountered in gate arrays or RAMs. The interconnect of a microprocessor has to carry tens of amperes of instantaneous current into and out of the chip in addition to routing signals. Furthermore, since high-performance microprocessor clock frequencies are of the order of hundreds of megahertz, the on-chip clocks must be distributed with very low RC time delay constants. These requirements lead to a number of differences between the interconnects used for microprocessors, gate arrays, and RAMs. RAMs at the 1-μm feature size are usually designed with two levels of aluminum-based conductors (M1 and M2) that are very similar in thickness (approximately 1 μm), minimum width, and spacing. Gate arrays generally add a third level of aluminum-based interconnect (M3) to improve signal routing and gate utilization in the array. This level has very similar characteristics to M1 and M2. Since the transistor density in a gate array is low and not all the gates are used in a design, the power dissipation is usually moderate by microprocessor standards.

The first two layers of interconnect on a high-performance microprocessor are very similar to the corresponding layers on an SRAM or gate array.

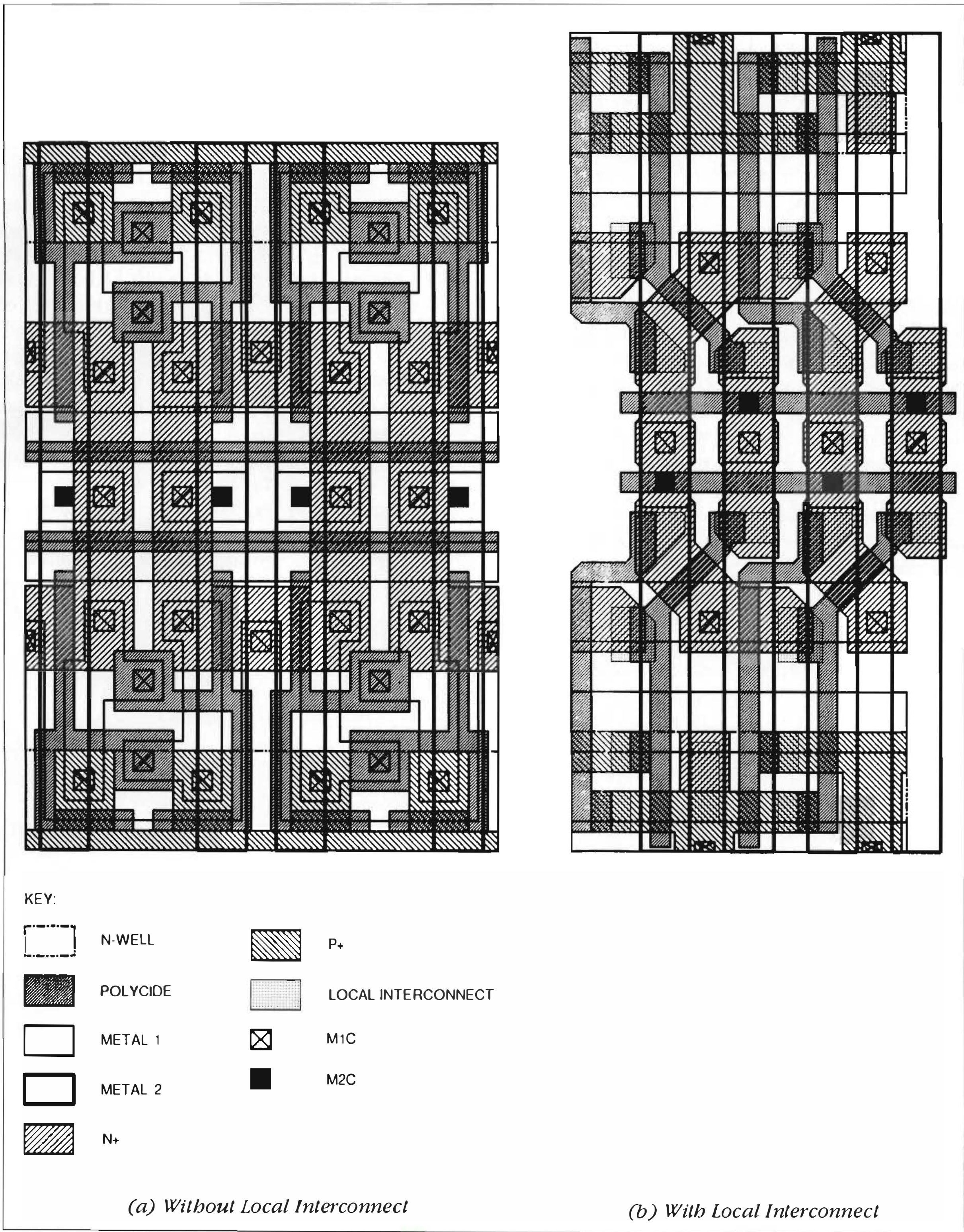


Figure 8 Six-transistor SRAM Cell Layout

However, because of the high power supply currents and low skew clock distribution required for high-speed operation, M3 in CMOS-3 and CMOS-4 processes is approximately 2.5 times thicker. To avoid an impact on yield, the pitch of M3 is chosen to be approximately three times larger than that for M2. To reduce capacitance to M1 and M2, the dielectric under M3 is thicker than that between M1 and M2.

Fuses

As the number of memory cells increases on a microprocessor, the impact of those cells on the yield of the die increases. CMOS-3 and CMOS-4 technologies implement redundancy by a standard technique of laser-fusible links to remove bad cells and incorporate new ones into the array. Digital's process differs from others in the implementation of the fuses, however. Standard RAM processes use polysilicon fuses for their small size and ease of ablation. Because of the thickness of dielectric layers that would need to be etched to uncover the fuse, the CMOS-3 and CMOS-4 technologies could not use polysilicon fuses. Sufficient control of the dielectric etch rate and the selectivity of the dielectric etch to polysilicon cannot be achieved in a manufacturing environment for the thin polysilicon fuse layer to be left intact. Instead, the 80-nanometer (nm) film of TiN that forms the bottom layer of M3 is used. The upper layer of aluminum copper (AlCu) is selectively etched through a special mask to leave a 3-μm wide strip of TiN that can be ablated by industry-standard lasers.[14]

Fusible links can be placed on a chip to form an identification register. A laser can then be used to program a unique code into this register. The contents of the identification register on the NVAX chip can be read by system software so that individual die can be uniquely identified not only during system manufacture, assembly, and test, but also in the field.

Manufacturing Process Variations and Chip Design Strategy

As mentioned earlier, the performance of a CMOS process is a function of the current driving capability (I_{ds}) of the PMOS and NMOS transistors, as well as the capacitances of the gates, sources, drains, and interconnect that the PMOS and NMOS transistors must charge and discharge. All of these parameters vary in manufacturing. In order to ensure that chips will function correctly and at the planned speed, chip designers must account for these manufacturing process variations when the chips are being designed.

The lot-to-lot variation in characteristics for CMOS-4 transistors is significantly larger than for bipolar transistors. Figure 9 shows how the saturation current (I_{DSAT}) varies in manufacturing for CMOS-4. The five points on the plot of PMOS I_{DSAT} versus NMOS I_{DSAT} represent the process extremes and are often referred to as the process corners. Each process corner has a two-letter label: FF, TT, SS, FS, and SF. The first letter of the pair is used to refer to the PMOS device: F indicates a fast (i.e., high current) device; S indicates a slow (i.e., low current) device; and T indicates a typical (i.e., manufacturing target current) device. The second letter of the pair refers to the NMOS device. Thus, the FF point in Figure 9 represents the fastest PMOS device paired with the fastest NMOS device; the SS point represents the slowest PMOS device paired with the slowest NMOS device; and the TT point represents a typical PMOS device paired with a typical NMOS device.

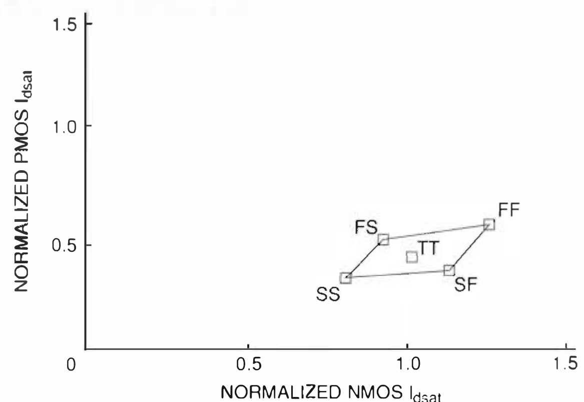


Figure 9 Comparison of Saturation Currents of NMOS I_{DSAT} and PMOS I_{DSAT} for CMOS-4 Process Corners

The FF SPICE[15] models are used to predict the speed of the fastest chips, the maximum power dissipation, the transient current demands on the power supply, the maximum voltage drops in the on-chip power and ground routing, the worst-case current density (checked to ensure that the electromigration limits are not violated) in power supply and signal lines, and the maximum rate at which the chip's signal pins will transition. Power supply current transients and signal-pin transition rates

have important implications for the electrical design of chip packages. Both place limits on how much inductance can be tolerated in the leads. The maximum power dissipation has obvious thermal implications for package and heat sink design.

The SS SPICE models are used to predict the speed at which the slowest chips will run. The TT models are used to check that the circuits on the chip will run at the desired speed when the manufacturing process is at the center of its range. The bulk of the circuit design work for the NVAX and Alpha 21064 chips was done using the TT models.

The FS and SF SPICE models are used to determine the noise margin for circuits (DC circuit analysis) rather than to predict the speeds at which circuits will run (AC analysis). The FS model has a semi-fast PMOS transistor paired with a semi-slow NMOS transistor; the SF model is just the opposite. The parameters that determine I_{DSAT} for PMOS and NMOS transistors are correlated. These correlations are taken into account in the FS and SF models. The correct operation of some CMOS circuits is particularly sensitive to the ratio of the PMOS to NMOS currents. By using the FS and SF models to simulate these circuits, designers can verify that the circuits will function correctly in spite of variations in the manufacturing process. A simple example is given in Figure 10, which shows how the switching point of a CMOS inverter changes from FS to TT to SF. Circuit designers use DC simulations like these to determine the safe bounds for the sizes of transistors in a variety of common circuit structures. These bounds are incorporated into the design methodology for the project, and CAD tools are used to search the circuit schematic database for structures that violate the methodology.

When creating schematics, circuit designers use technology-specific rules of thumb to estimate the interconnect capacitance on signal lines, etc. Layout for the schematics is then generated, exact capacitances are extracted from the layout using CAD tools, the capacitance estimates are replaced with the extracted values, and the circuits are resimulated to ensure that they still meet the specifications. The capacitance extraction tool can be rerun for a different process corner (dielectric thicknesses, etc.) by changing its parameter file.

Although not discussed here, environmental effects such as operating temperature and power supply variations must also be taken in account.

If CMOS chips are to be manufacturable, designers must account for process variations during the design phase by following procedures such as those just outlined. In order to get to market quickly, the NVAX microprocessor was being designed while the CMOS-4 process was being developed. Process

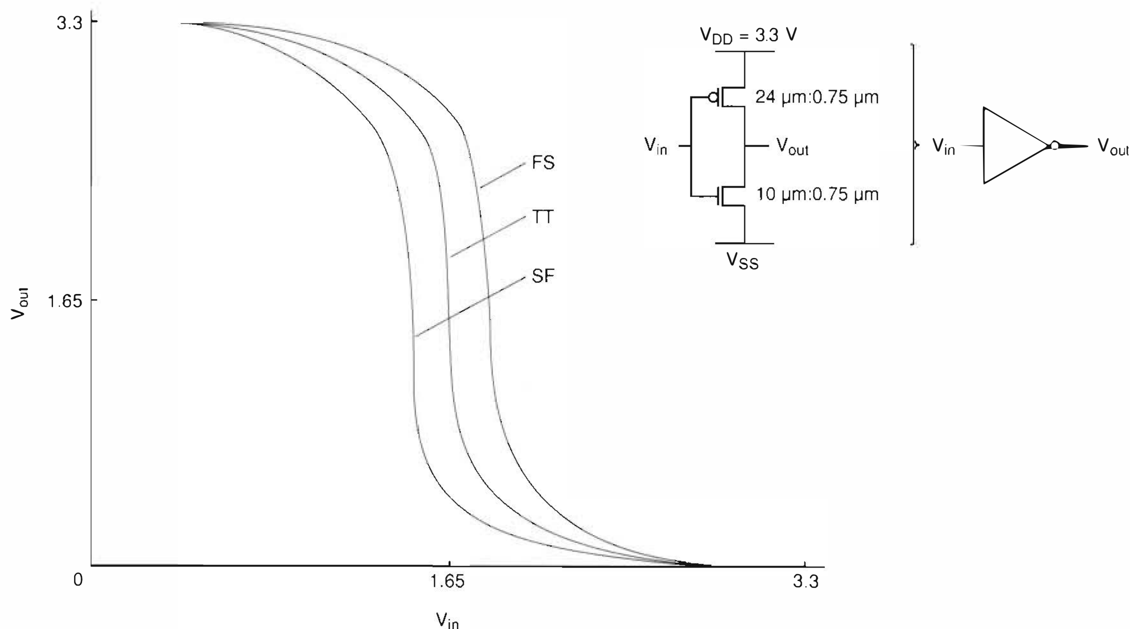


Figure 10 CMOS-4 Inverter Switching at Process Corners

simulators and test chips were used to generate the CMOS-4 worst-case and typical electrical models for transistors and interconnect during the early phase of process development. The accuracy of these models was critical to the successful and timely completion of the NVAX design: it was never necessary to redesign circuits due to process or model changes during the course of the project.

Conclusion

Digital's CMOS processes have been developed specifically for high-performance microprocessors. Generation-to-generation improvements derived from scaling, increased die area, and new technology features have allowed increased performance every two years. The Alpha 21064 and NVAX chips, implemented in CMOS-4, are the highest performing reduced instruction set (RISC) and complex instruction set (CISC) microprocessors reported in the industry to date.

References and Notes

1. D. Dobberpuhl et al., "A 200MHz 64b Dual-Issue CMOS Microprocessor," *IEEE International Solid-State Circuits Conference Digest of Technical Papers* (February 1992): 106-107.

2. R. Badeau et al., "100MHz Macropipelined CISC CMOS Microprocessor," *IEEE International Solid-State Circuits Conference Digest of Technical Papers* (February 1992): 104-105.

3. To compute clock cycles per instruction (CPI), typical application programs and benchmarks are first run. Then the number of clock cycles required to execute these programs is divided by the total number of instructions executed. Note: strictly speaking, CPI is also a function of the system configuration (i.e., the size and speed of the memory subsystem, etc.) and the compiler.

4. J. Hennessy and D. Patterson, *Computer Architecture: A Quantitative Approach* (San Mateo: Morgan Kaufmann, 1990).

5. SPECmark is a quantitative measure of performance determined by running a suite of 10 benchmark programs. For example, a VAX 11/780 system has a SPECmark value of 1.0.

6. R. Allmon et al., "CMOS Implementation of a 32b Computer," *IEEE International Solid-State Circuits Conference Digest of Technical Papers*, vol. 32 (February 1989): 80-81.

7. R. Dennard, F. Gaensslen, H. Yu, V. Rideout, E. Bassous, and A. LeBlanc, "Design of Ion-implanted MOSFETs with Very Small Physical Dimensions," *IEEE Journal of Solid-State Circuits*, vol. SC-9, (1974): 256-268.

8. R. Dennard, F. Gaensslen, E. Walker, and P. Cook, "1 μm MOSFET VLSI Technology: Part II—Device Designs and Characteristics for High-performance Logic Applications," *IEEE Transactions on Electron Devices*, vol. ED-26 (April 1979): 325-333.

9. G. Baccarani, M. Wordeman, and R. Dennard, "Generalized Scaling Theory and Its Application to a ¼ Micron MOSFET Design," *IEEE Transactions on Electron Devices*, vol. ED-31 (April 1984): 452-462.

10. R. Allmon et al., "System, Process and Design Implications of a Reduced Supply Voltage Microprocessor," *IEEE International Solid-State Circuits Conference Digest of Technical Papers* (February 1990): 48-49.

11. Joint Electronic Device Engineering Council (JEDEC) JEDEC Standard No. 8 *Standard for Reduced Operating Voltages and Interface Levels for Integrated Circuits* (Washington D.C.: Electronics Industries Association, March 1984).

12. S. Selberherr, A. Schültz, and H. Pötzl, "MINIMOS—A Two-Dimensional MOS Transistor Analyzer," *IEEE Transactions on Electron Devices*, ED-27 (1980): 1540-1550.

13. A. Nasr, G. Grula, A. Berti, and R. Jones, "CMOS-4 Technology for Fast Logic and Dense On-chip Memory," *Digital Technical Journal* vol. 4, no. 2 (Spring 1992, this issue): 39-50.

14. M. Coffey and R. Hollingsworth, "Integrated Circuit Having Laser-Alterable Metallization Layer," U.S. Patent 4,849,363, Issued July 1989.

15. SPICE is a general-purpose circuit simulator program developed by Lawrence Nagel and Ellis Cohen of the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley.

Marden H. Seavey
John V. Faricelli
Nadim A. Khalil
Gerd Nanz
Llanda M. Richardson
Christian O. Schiebl
Hamid R. Soleimani
Martin Thurner

Numerical Device and Process Simulation Tools in Transistor Design

Numerical device and process simulation programs are fundamental tools in the design and characterization of silicon transistors. These tools employ numerical mathematical methods to simulate the operation of the elemental transistor structures that are the building blocks of CMOS VLSI circuitry. When designing these basic structures, CMOS process and device design teams require efficient, high-performance simulators that use accurate physical models. Digital has developed thermal annealing, mobility, and avalanche models, and has improved the numerical methods used in its process and device simulation programs. Also, supporting software was developed to help integrate the various simulation tools.

With the increasing number of transistor functions implemented on each chip in complementary metal-oxide semiconductor (CMOS) very large-scale integration (VLSI) circuitry, computer simulation tools have become essential at all levels of the design process. This is particularly true at the level of metal-oxide semiconductor (MOS) device fabrication and elemental transistor design. The continuing reduction in the geometrical scale of the transistor structures means that careful control and design of these structures is necessary to maintain the required switching properties and current drive capabilities.

Process and device simulators contribute to this design and control by employing numerical mathematical methods to simulate the ion flux that occurs in the device fabrication process and the current flow that occurs during transistor operation. Process simulation requires physical models of ion implantation, diffusion of ions, and thermal oxidation, for example. In device simulation, the basic semiconductor equations of drift and diffusion are solved using microscopic physical models, such as models of the mobility of electrons and holes and models of electron-hole pair generation, also referred to as avalanche generation.

For practical application, the simulators must be capable of high-speed performance and must provide accurate predictions. Digital's Advanced Semiconductor Development (ASD) Submicron Simulation Group modifies and extends the capabilities of simulators that have already been developed. These modifications and extensions have resulted in accelerated performance, higher accuracy, and thus improved application of the simulators to Digital's CMOS technologies. The process simulators SUPREM3 from Stanford University and PROMIS from the Technical University of Vienna (TUV), Austria, and the device simulators MINIMOS from the TUV and PISCES from Stanford are the main simulators that Digital modified and applied. In addition, the ASD Group has developed software to support the simulators. This software consists of programs to interface the process simulators with the device simulators, and programs to automatically perform simulator calibration and sensitivity analysis.

The modified simulators and supporting software are vital elements in the development of Digital's CMOS technologies. These tools provide invaluable insight into transistor fabrication and operation. Using these tools in the design of lot splits considerably decreases fabrication time and thus reduces cost. To predict circuit performance limits, designers use calibrated simulation results as input to circuit simulators. The ability to predict these limits has made possible concurrent

technology and circuit design of Digital's CMOS-2, CMOS-3, CMOS-4, and CMOS-5 technologies.

This paper describes the nature of Digital's process and device simulation tools. Examples of the important physical models, numerical mathematical methods, and supporting software for the simulators are discussed. The paper closes with a summary of how these tools benefit Digital's semiconductor process development teams.

Physical Models

The key to accurate simulations of transistor characteristics is in the physical models employed by the programs. This section describes examples of models developed by the ASD Group. First, a rapid thermal annealing model for process simulation is presented. Next follow discussions on ion implantation through nonplanar surfaces in two dimensions and on the mobility and avalanche models used in the MINIMOS program. The section concludes with information about the use of simulators to predict transistor capacitance values.

Rapid Thermal Annealing Model

To alter the electrical properties of the semiconductor substrate material, i.e., the silicon wafer, atoms from groups III and V of the periodic table are used for doping. A known amount of these atoms, also called impurities or dopant, must be placed in the silicon lattice. Ion implantation is the main technique for incorporating the impurity atoms. However, the implanted atoms do not move into the proper sites upon implantation. A high-temperature treatment, known as thermal annealing, is required to achieve this.

There are two types of thermal annealing used in semiconductor processing: conventional furnace annealing (CFA) and rapid thermal annealing (RTA). CFA is a long-term (minutes to hours) annealing step carried out at moderate temperatures (below 1000 degrees Celsius). RTA is a recent technique conducted at higher temperatures (usually above 1050 degrees Celsius) for extremely short times (seconds).

Ion implantation is a defect-producing process that creates lattice disorder and point defects, such as silicon vacancies and interstitials, in an otherwise perfect lattice. (A silicon vacancy occurs when a silicon atom is missing from a perfect silicon lattice, whereas a silicon interstitial occurs when an extra silicon atom is squeezed into a perfect lattice.) Since ion implantation is performed at room temperature, the collection of implantation-induced defects is retained in a stable state in the lattice. However, at high temperatures the defects become highly mobile and influence the migration of impurity atoms.

The migration of impurities under annealing is governed by the diffusion phenomena, which are mediated by point defects. Ion-implantation-induced point defects can cause anomalous diffusion of the dopant. This effect is highly transient in nature because either the ion-implantation-induced defects disappear to the surface or to the depth of the silicon (referred to as the silicon "bulk"), or the defects may recombine while interacting with the dopant. An accurate understanding of point defect behavior is particularly important for small geometry transistors requiring ultra-shallow junctions with high doping levels.

Although research in point defect physics in silicon is extensive, point defect behavior is still not well understood. One contributing factor is the lack of a reliable technique to study point defects quantitatively, i.e., knowledge of point defects is obtained only through indirect observations of their effect on dopant diffusion. There remains significant controversy surrounding the experimental observations and the corresponding interpretations. At the macroscopic level, there are models that solve the diffusion equation using an average diffusivity for dopant. However, the conventional diffusion models do not take into account the details of defect-dopant interactions resulting from high-dose ion implantation. Recently, several models, both physically based and empirical, have been proposed to simulate transient-enhanced diffusion under high-dose ion implantation. The physically based models require an accurate account of ion-implantation-induced defect concentration and often use time-consuming Monte Carlo methods. Empirical models, on the other hand, are fast, but must be based on detailed physics and well-calibrated parameters.

The ASD Group developed a new empirical model called implantation-enhanced transient diffusion (IETD).[1] This model has been used successfully in the CMOS technology development in Digital's semiconductor manufacturing facility in Hudson, MA. The IETD model is phenomenological and is based on a dual vacancy-interstitial mechanism, with model parameters determined empirically. The model uses a relationship that links the

amount of ion-implantation-induced defects to ion implantation conditions, such as dose and energy.

The overall transient diffusion process, which depends on the annealing temperature, takes from several seconds to a few minutes to complete. Consequently, a relatively short time interval is available to limit the role of point defects and to control the transient diffusion, particularly when fabricating shallow junctions below 0.2 micron (μm) for devices less than 0.5 μm in size. The IETD model provides good predictability of transient diffusion, based on point defect behavior. This feature allows designers to study the effect of various processing conditions on transient diffusion and thus to optimize the process. Figure 1 shows a comparison of secondary ion mass spectroscopy (SIMS) data with arsenic diffusion, both with and without the use of the IETD model in the SUPREM3 process simulator. Clearly, using the IETD model gives more accurate results.

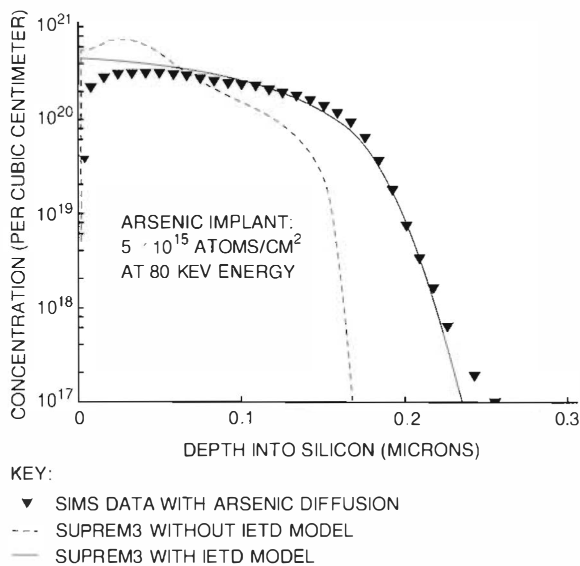


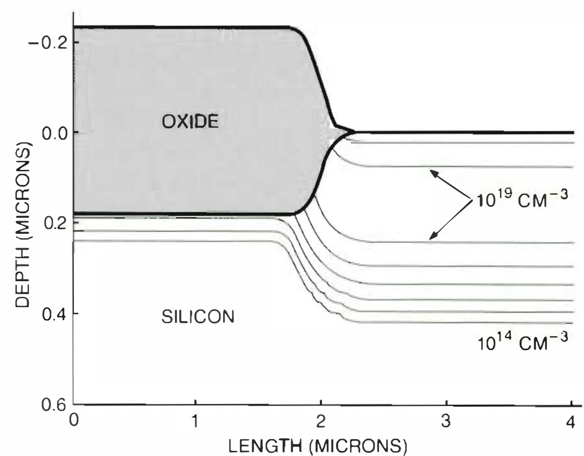
Figure 1 Comparison of SIMS Data with Results from Annealing Models

Implantation through Nonplanar Surfaces

A problem frequently encountered in semiconductor technology is the implantation of a dopant through thin dielectric layers on semiconductors or, in general, the implantation in multilayer structures. Calculating the exact depth distribution of implanted ions in a multilayer structure requires the solution of Boltzmann transport equations or Monte Carlo simulation. This process, however, is very complicated and requires a great deal of computing time.

To circumvent these disadvantages, several analytical models have been developed that give reasonably good results under certain conditions. One of these models, the numerical range-scaling model, is applicable in the limit of both very thin films and very thick films and leads to the most realistic profiles of all known analytical models.[2,3,4] Therefore, this model is implemented in the two-dimensional (2-D) process simulator, PROMIS. Using this numerical range-scaling model makes it easy to extend the 2-D model for arbitrarily shaped, nonplanar multilayer structures. In addition, the tilt angle of the implantation can be varied. Thus, it is possible to see the dependence of the doping profile on the tilt angle. The change in the analytical distribution function caused by changes in channeling by different implantation angles has not yet been considered.

Figure 2 shows an application of the newly implemented ion implantation feature in PROMIS. The simulation starts from a rectangular semiconductor region. The first process step is a local wet oxidation for 30 minutes at 1100 degrees Celsius. Boron ions are implanted into this structure with an energy of 50 kilo electron volts (keV) and at a dose of $5.0 \times 10^{11}/cm^2$. The iso-concentration lines of the 2-D as-implanted boron profile can be seen in Figure 2. Each contour line represents an order of magnitude change in boron concentration.



Note: Each contour line represents an order of magnitude change in boron concentration.

Figure 2 Two-dimensional Boron Implant into Silicon, Simulated Using PROMIS

Although the numerical range-scaling model is the most accurate way to analytically describe implanted dopant distributions for multilayer structures, the model has some shortcomings and restrictions. For example, the analytical distribution function does not depend on the tilt angle and the silicon orientation. This shortcoming could probably be solved by introducing a term that depends on these two parameters and could be fitted using measured profiles. Another shortcoming, as compared to a Monte Carlo simulation, is that the lateral distribution function does not consider interfaces. This deficiency could be especially important for structures with very steep interfaces, such as trenches. However, this neglect of the interfaces in the case of the lateral distribution function is only significant if the mean atomic numbers of the materials (on both sides of the interface) differ considerably. In the case of Si and silicon dioxide (SiO_2), for example, it is not a major shortcoming.

The model described above is implemented in PROMIS in such a way that there are, in principle, no restrictions in the simulation geometry. This means that the implantation module can handle arbitrary multilayer structures. The implementation of this ion implantation model is thus a significant step towards extending PROMIS to a fully multilayer simulation tool.

Mobility Model

Carrier mobilities in semiconductor material are determined by a large variety of physical mechanisms. Electrons and holes are scattered by thermal lattice vibrations, ionized impurities, neutral impurities, vacancies, interstitials, dislocations, surfaces, and the carriers themselves. The saturation of the drift velocity caused by interactions with lattice vibrations results in a further mobility reduction. For MOS transistors, however, the effect of the surface, i.e., the silicon-silicon dioxide (Si-SiO_2) interface, is of overriding importance. The sheet of conducting charge (either electrons or holes), called the inversion layer, is forced by the applied electric fields to flow close to this interface and interact with it.

The surface scattering is a complex and poorly understood process consisting of a combination of roughness, interface charge, and surface phonon-scattering mechanisms. Nevertheless, there is a universal empirical model, first demonstrated in 1979 and 1980, that can be calibrated against measured results.[5,6] An effective mobility exists that depends only on an effective field perpendicular to the silicon surface and is independent of the doping level near the surface.

The ASD Group modified the microscopic mobility model in the MINIMOS simulator to reflect this universal mobility model.[7] The microscopic model contains three adjustable dimensionless parameters called MR, MT, and MX, which are generally close to unity. These parameters scale the magnitude and the two field dependencies of the microscopic mobility. The model was originally calibrated against CMOS-2 electrical data in 1986. Note particularly that recent CMOS-4 data compare well to simulated data with only minor adjustments made to these three parameters. Figure 3 is a comparison of MINIMOS simulation for a device having a long-channel (51-micron) length with the linear region drain current graphed as a function of the gate voltage (V_g). In this section and the following two sections, the width of all the simulated and measured MOS devices is 50.5 μm. The scaled mobility parameters for this excellent fit are an MR of 1.02, an MT of 1.10, and an MX of 1.00, as compared to the default 1986 values of 1.00 for each parameter. Figure 4 shows the fit for a CMOS-4 short-channel device, i.e., 0.62 μm in length (an effective length of 0.5 μm), using the same mobility parameters as for the long-channel device. Adjustments were made to allow for the effects of interface charge and contact resistance. The agreement

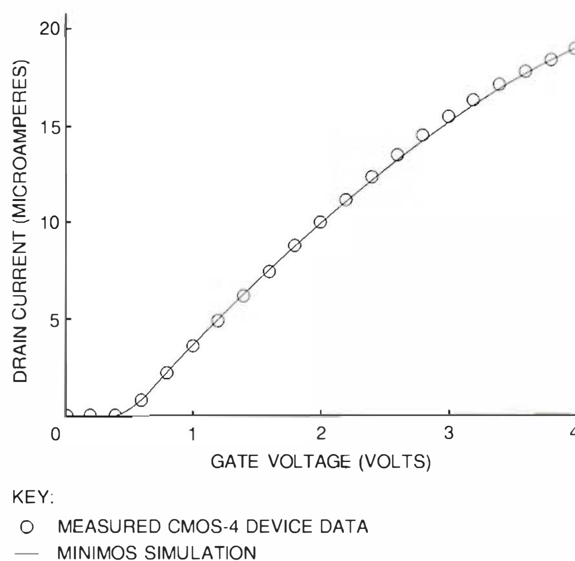


Figure 3 Comparison of MINIMOS Simulation to Long-channel CMOS-4 Device Data

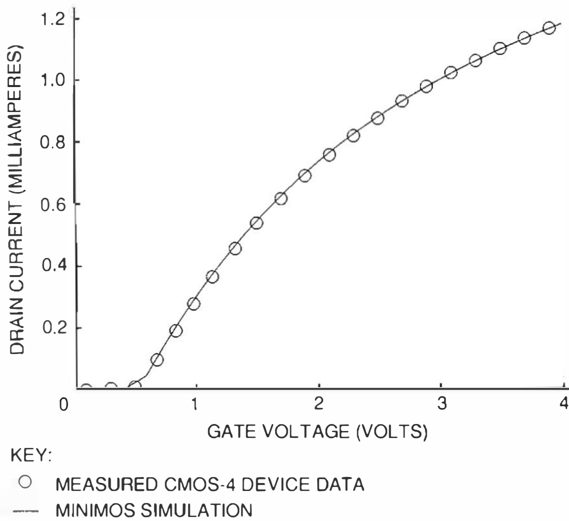


Figure 4 Comparison of MINIMOS Simulation to Short-channel CMOS-4 Device Data

of the simulated results with the measured data is again excellent and demonstrates the validity of the MINIMOS mobility model.

Avalanche Model

An important part of submicron MOS device design is device reliability. Aggressively scaled devices contain high electric fields. These fields are unavoidable, even with the reduced power supply of 3.3 volts for the CMOS-4 device. In addition, circuit effects, such as ringing, can cause voltages on devices to rise well above the magnitude of the power supply. Electrons in these high electric fields accelerate and begin to acquire energy faster than they can dissipate it to the underlying silicon lattice. A small fraction of the electrons, namely energetic or "hot" electrons, gain enough energy to generate an electron-hole pair in a process called impact ionization. The newly created electrons are swept to the drain of the transistor, and the holes are collected as a current called substrate current.

Other energetic electrons surmount the barrier at the Si–SiO_2 interface and inject themselves into the oxide, appearing as gate current, which flows out of the gate terminal. Carrier injection into the oxide can damage the oxide and cause a shift in the device threshold voltage or degrade the ability of the device to conduct current. This injection is the physical cause of device degradation.

Unfortunately, gate current is difficult to measure directly, because the currents involved are extremely small. As a fallback, it is common practice to use the substrate current as a monitor of hot electron damage, since both currents arise from the same high field conditions in the device. Using an accurate substrate current model can assist device designers in optimizing the CMOS transistor for reliability.

In late 1988, at the time of early CMOS-4 development, the avalanche generation model in the MINIMOS simulator was reexamined. The design team found that the peak substrate current for a given drain voltage (V_d) occurred at gate voltages lower than predicted by simulation. They resolved this problem by modifying the microscopic model for impact ionization in the MINIMOS simulator to include a depth-dependent term, similar to the one used by Slotboom et al.[8,9] The modification sharply reduced impact ionization near the Si–SiO_2 interface. Impact ionization model parameters were derived from CMOS-3 and CMOS-4 measured data.

Figure 5 shows measured and simulated CMOS-4 substrate current data for drawn gate lengths of 0.62 and 2.00 μm at a V_d equal to 3.3 volts. Figure 6 shows the substrate current as a function of the V_g for the 0.62-micron gate length device for three drain voltages around the design center of 3.3 volts. The agreement shown using one set of model parameters is quite remarkable, given the limited

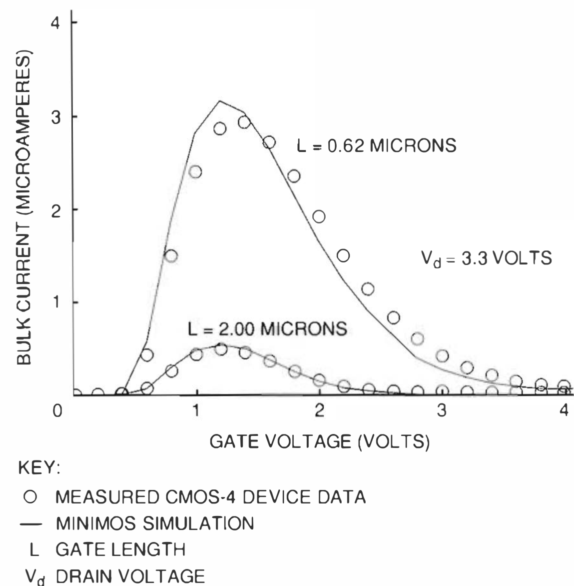


Figure 5 Comparison of MINIMOS Simulation to CMOS-4 Substrate Current Data for Two Gate Lengths

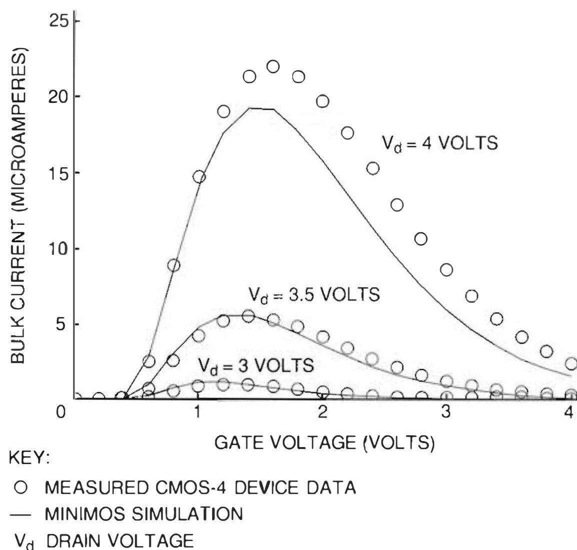


Figure 6 Comparison of MINIMOS Simulation to CMOS-4 Substrate Current Data for a Short-channel Device and Three Drain Voltages

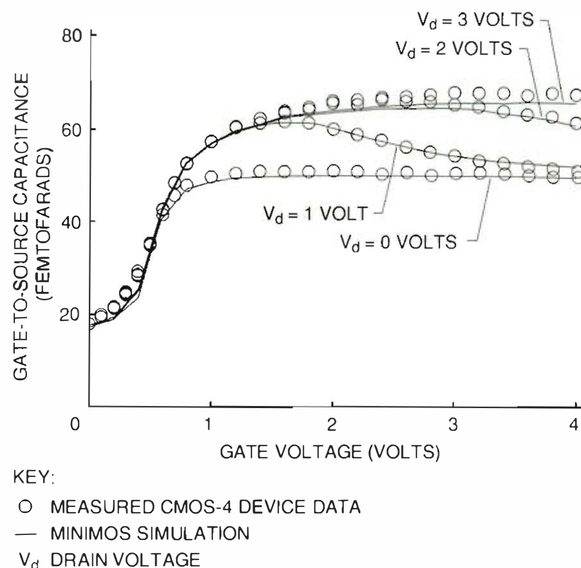


Figure 7 Comparison of MINIMOS Simulation Data to CMOS-4 Gate-to-source Capacitance Data for Four Drain Voltages

physics and empirical nature of the model. Although the modified model has proven satisfactory for CMOS-4 devices, we are examining other, more physically based, models of substrate and gate current for our future generations of CMOS devices.

Capacitance Simulation

The speed of MOS circuits partially depends on the amount of device capacitance that must be charged during circuit state transitions. Device simulators can tell device designers how large the device capacitances will be and what effect changes in the manufacturing process will have on capacitance size. The gate capacitance of the MOS transistor is one of the two key types of device capacitances; the other consists of the so-called "diode" capacitances of the source and drain junctions. The gate capacitance is split into three parts: gate-to-source (C_{gs}), gate-to-drain (C_{gd}), and gate-to-bulk (C_{gb}). Figures 7 and 8 show simulated and measured C_{gs} and C_{gd} values versus gate and drain voltages for a device with a width of 50.50 μm and a length of 0.62 μm. C_{gb} is not shown because it is a negligibly small quantity for voltages above the transistor threshold voltage.

Achieving the good agreement between simulation and measurement shown in Figures 7 and 8 is a difficult undertaking. Accurate capacitance mea-

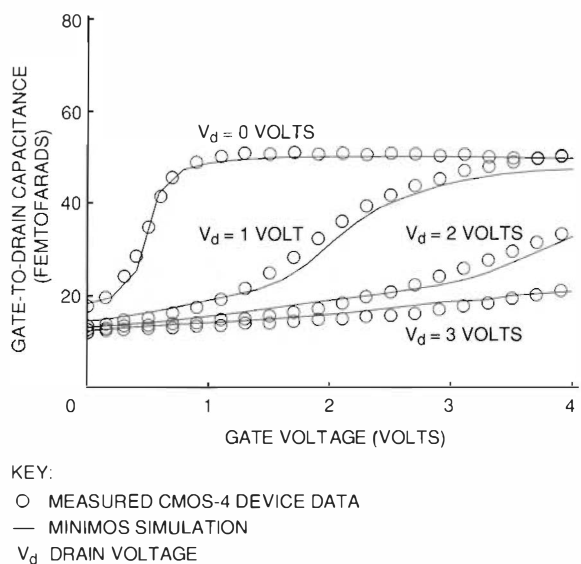


Figure 8 Comparison of MINIMOS Simulation Data to CMOS-4 Gate-to-drain Capacitance Data for Four Drain Voltages

surements on such short-channel devices require careful experimental techniques to reduce the effects of parasitic capacitance and to ensure that

the DC power supply can sink sufficient current. For example, with the aid of simulation, it was shown that small parasitic resistances, on the order of 10 ohms, can significantly shift the capacitance curves of the device. Such resistances must be minimized to yield accurate results and to verify that simulation and measurement can be correlated.

On the simulation side, uncertainties in the concentration of dopant in the channel region and the source-drain regions have a significant impact on the capacitance characteristics. The ASD Group used this fact to advantage in a "reverse engineering" of the doping profiles. They compared the measured data with the simulated capacitance using doping profiles generated by the process simulators. Then, they revised the profiles until the simulated capacitances matched the measured ones. The ASD team applied this reverse engineering method to both channel doping into depth and lateral source-drain doping. In the former case, the deep depletion capacitance-to-voltage measurements of doping were verified using simulation. In the latter, the doping was extracted by examining the relationship between the C_{gd} and the V_d when there is no V_g. Using this method, it was possible to extract doping profiles that were unattainable by other analytic means, such as SIMS.

The CMOS-4 devices exhibit a slight polysilicon depletion effect, which decreases the measured capacitance of the MOS transistor gate. Work performed in parallel by the ASD Group and at the TUV has resulted in an implementation of a polysilicon depletion model in the MINIMOS simulator.[10] The model is unique in that it analytically solves the Poisson equation in the polysilicon gate and thus can be implemented in MINIMOS as a simple modification of the gate boundary condition. Figure 9 shows a comparison of the model with simulation. The measured total gate capacitance of a large-area MOS diode (50.5 by 50.5 μm) as a function of the V_g is presented, along with MINIMOS simulation results. For reference, Figure 9 also shows the simulated gate capacitance without the polysilicon depletion effect, computed using the PISCES device simulation program. The PISCES results were used as an independent check on the MINIMOS simulator results. The curves indicate that the modeled results are a good approximation of the measured data. At CMOS-4 dimensions, the polysilicon depletion effect is rather small, i.e., on the order of 5 percent. The effect becomes more pronounced at thinner gate oxides.

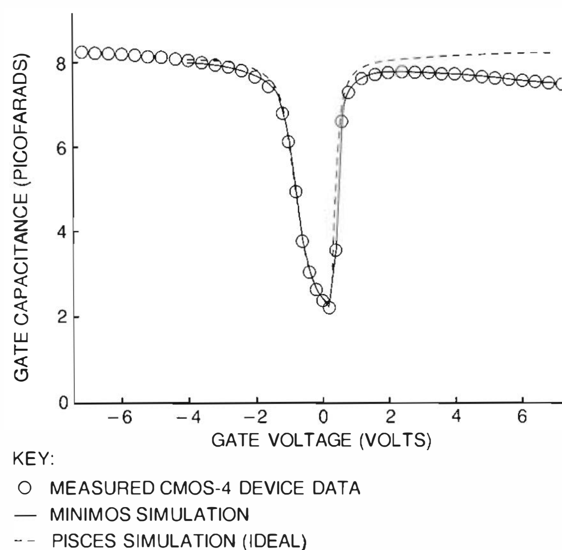


Figure 9 Comparison of MINIMOS and PISCES Simulation Data with Capacitance-voltage Data in the Presence of the Polysilicon Depletion Effect

Numerical Mathematical Methods

The efficient use of CPU time and memory in the device and process simulators is directly dependent on the efficient implementation of state-of-the-art numerical mathematical algorithms and physical models. Because of the limitations on representing numbers in computers, these algorithms and models must incorporate many detailed enhancements in order to maintain stability of the simulation programs in the numerical and physical sense. This is especially true in the case of simulation in three space dimensions. This section first describes the improvements to the MINIMOS device simulator achieved during the last few years. An application of the improved MINIMOS to a parasitic device in shallow trench isolation is then presented, followed by a discussion of automatic space grid design.

Performance Improvements Achieved for the MINIMOS Device Simulator

A three-dimensional (3-D) simulation program requires the solution of very large matrices and is thus time-consuming. Therefore, an accurate use of resources is obligatory. This goal was achieved for the MINIMOS device simulator by means of software techniques and hardware options.[11,12]

Various algorithms and numerical methods were thoroughly tested with practical examples, because not all algorithms presented in the literature are applicable. Most of the proposed algorithms, and most of the improvements related to overall performance gain, are found among the linear solvers. These solvers are applied to the solution of the large sparse matrix systems that arise from the discretization of the partial differential equations. The best-known linear solver algorithm is called Gaussian elimination; this algorithm is the most stable method of solving a system of linear equations. Unfortunately, Gaussian elimination is generally not applicable to the 3-D simulations because enormous amounts of memory and CPU time are required. Instead, iterative solvers have been tested for stability, using several techniques, which are required because of the variation over a range of many orders of magnitude in the coefficients and in the dependent variables.

After an accurate comparison of the convergence behavior of the different linear iterative solvers, a final set of algorithms was identified, one which is adequate for the present. These algorithms make the 3-D simulation program more efficient by taking into account the different properties of the Poisson equations (solutions for electric potential) and the continuity equations (solutions for current flows).

Strong emphasis has been placed on reducing the amount of duplicate code in the simulation program. MINIMOS is a hierarchically ordered program which first solves an approximate 2-D problem in three steps and then proceeds to solve the actual 3-D problem with the 3-D algorithms. This method guarantees an efficient use of the computer resources. However, sometimes the incompatibility of the 2-D and 3-D codes means that separately functioning code must be used. Additional development work has been done to create common code in all important areas, including physical models (e.g., mobility models) and the numerical algorithms. Future upgrades and modifications for 2-D and 3-D codes can be done simultaneously. This improvement also contributes significantly to the performance improvement.

An investigation using the VAX Performance and Coverage Analyzer (PCA) was carried out to identify so-called "hot spots," which are code sequences that consume high amounts of CPU time when executed. Eliminating or improving the performance of these code segments significantly reduces simulation time.

The VAX FORTRAN high-performance option (HPO) compiler was also used to improve the performance of the MINIMOS 3-D simulation program. The ASD Group investigated both parallelization and vectorization on a multiprocessor VAX system with vector hardware. The performance improvement resulting from parallelization (using two CPUs) and vectorization (using two vector units) was as much as a factor of five. During the last few years, software enhancements, including the removal of hot spots, the improvement of algorithms, and the creation of common code, have resulted in a thirtyfold performance improvement of the program. Thus, the combined software and hardware enhancements improved the performance of the MINIMOS 3-D simulation program by approximately a factor of 150.

Many of these improvements also enhance performance on scalar processors. The MINIMOS program is run on a variety of VAX and MIPS processors at several sites in the United States and in Europe. Engineers can submit MINIMOS (or other) simulation programs from their VAXcluster systems running under the VMS operating system to various fast processors on the local network. This improves the overall turnaround time of computationally intensive jobs and provides a painless way for engineers to better use available computational resources.

Application to a Parasitic Device

The improved performance of the 3-D MINIMOS simulation program made it possible to analyze the behavior of shallow trench isolated devices. Electrical data on these devices in early development work showed a "bump" in the subthreshold drain current versus gate voltage characteristic under certain conditions. The ASD Group simulated this bump using the 3-D MINIMOS program and demonstrated that the origin of the bump was in a parasitic current along the trench sidewall. The 2-D version of the MINIMOS program was unable to explain the phenomenon. Only when the 3-D version was applied to the problem (taking into account a slight overlap of the gate polysilicon into the trench) was it apparent that a parasitic device at the edge of the trench was turning on at a lower gate voltage than was the main device, thus causing the bump.

Figure 10 shows a comparison of the MINIMOS simulations with the electrical data. The bump is resolved by the 3-D MINIMOS simulation but is not present in the case of the 2-D simulation. The para-

sitic device turns on before the main device, but its effective width is much less than that of the main device. Thus at higher gate voltages, the main device, which can be simulated with the 2-D version of the MINIMOS program, is dominant. The simulations have been adjusted for threshold and back bias. Further simulations indicated that increased back bias on the device greatly enhances the bump; the parasitic device at the corner of the trench is partially shielded from the effects of the back bias. The improved 3-D MINIMOS program is currently used in the design of the CMOS-5 technology, for continued electrical analysis of shallow trench isolation.

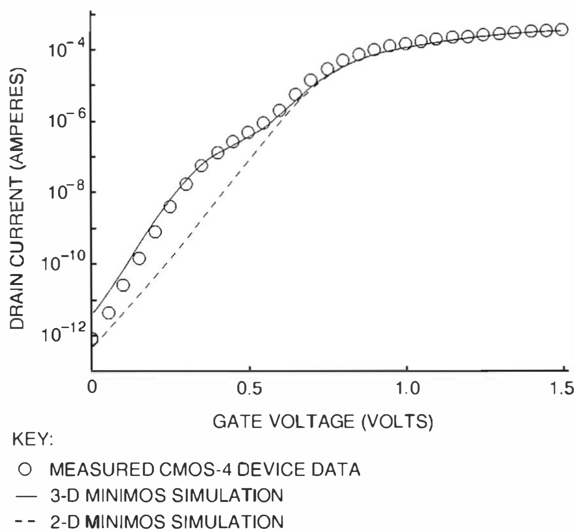


Figure 10 Comparison of MINIMOS Simulation with Data from a Shallow Trench Isolated Device with Parasitic Sidewall Current

Automatic Grid Design

In process and device simulation, the numerical mathematics requires the construction of a discretization scheme, an underlying space grid and, for time-dependent problems, a time grid, which should be controlled automatically. This basic gridding determines not only the accuracy of the solution but also the run time of the program, and in some cases even the success or failure of the simulation.

The design of an automatic space grid requires both mathematically based and physically based strategies.[13,14] Mathematical criteria for self-adaptive gridding involve the remainders in series expansions, equidistribution of the discretization error, the degree of coupling of the differential equations to be solved, and finally, the control of the ratio of adjacent mesh distances. Attention to these details results in the achievement of theoretical convergence rates and reasonable CPU times.

Figures 11 and 12 show respectively an implanted boron profile and the error indicator for the masked implantation of boron. The mask ends at the origin, and the error criterion is based on the difference between the second mixed derivatives of the solution function. Figure 12 clearly shows that the error is concentrated at the curvature of the profile around the mask edge.

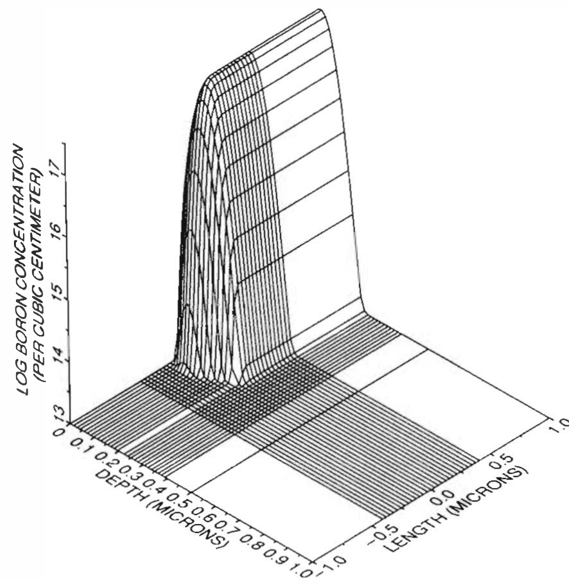


Figure 11 Masked Ion Implantation of Boron Showing the Space Grid

Almost all physical criteria are heuristic, i.e., they use quantities like the current densities or the net generation/recombination of carriers to design the grid properly in regions of physical interest. For example, numerical experiments indicate that in device simulation, mesh refinement is essential in regions where the net generation/recombination of carriers becomes higher than $1.0 \times 10^{22}/cm^3 \cdot s$.

Conservation laws play an important role in the definition of these strategies. For instance, coarse grids at curved pn-junctions introduce a significant integration error, thus leading to insufficient charge conservation, which is essential for accurate capacitance calculations.[15] For the automatic calculation

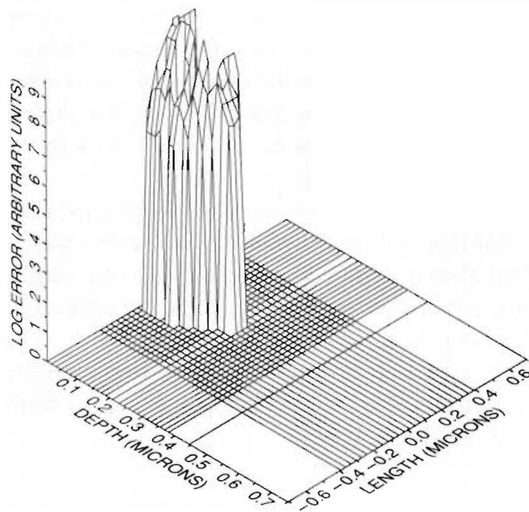


Figure 12 Error Indicator for the Masked Ion Implantation of Boron

of time steps, heuristic criteria are used also. These criteria take into consideration the change in boundary conditions and try to estimate the time discretization error by analyzing the solution for two successive time steps. More mathematically based criteria, such as utilizing the change of the potential energy, are under investigation.

Grid design strongly affects the run times of a simulation program. Usually the numerical effort of a two-dimensional simulation program depends quadratically on the number of grid points. Thus a 50 percent reduction in the number of grid points reduces the run time by a factor of four. Recent work at the Campus-based Engineering Center (CEC) in Vienna on gridding strategies in the PROMIS simulator has resulted in improving performance by a factor of 2.5 for the same accuracy. Automatic grid design is absolutely necessary to optimize the ratio between the number of grid points and the desired accuracy for each specific boundary condition. For time-dependent simulations, in particular, moving grids is the only way to guarantee the same accuracy for all time steps. Work is continuing in this area at both the ASD Group in Hudson and the CEC.

Technology Computer-aided Design

Historically, process and device simulation tools have been developed individually as standalone tools. No single tool covers the whole range of processing steps and device simulation needs that can arise in CMOS transistor design. Consequently, users must combine different tools to perform required simulations. The ASD Group has developed interface programs to assist in this effort. On a larger scale, research is going on in industry and in academia to develop the so-called technology computer-aided design (technology CAD or TCAD) frameworks for integrating the various simulation tools.[16,17,18] The ASD Group has been involved in this effort by actively participating in the United States-based TCAD group of the CAD Framework Initiative (CFI) and by interacting with the researchers at the TUV, who are developing the Viennese Integrated System for Technology CAD Applications (VISTA) framework discussed later in this section. These efforts hold much promise for the future, but the ASD Group has required intermediate-term solutions to speed up and automate the various procedures used by its transistor designers. This section presents two such solutions followed by a brief description of the VISTA system currently under development.

TCAD Command Language

The simulation needs and demands of Digital's CMOS technology development have increased beyond isolated simulation runs to include more complex tasks such as sensitivity analysis, optimization, and macromodeling. To support these higher-level needs, the ASD Group developed the TCAD Command Language (TCL). TCL is a specialized programming language for TCAD task management and execution, an instance of a command extension language tailored to TCAD users' needs. The following list highlights features of TCL that are necessary in the ASD TCAD environment:

- General programming constructs and control mechanisms
- Specialized subroutines suitable for optimization and sensitivity analysis, with arguments divided into input parameters, and input and output variables
- Tool control and manipulation that support environment customization, journaling, distributed execution, and parameterized tool invocation
- Special analysis and optimization commands bundled in a callable run-time library

TCL has been used in a variety of design and analysis tasks that include exploration of the design space, parameter sensitivity, design through optimization, model parameter extraction and charac-

terization, and statistical analysis. For example, Figure 13 shows TCL code that illustrates the application of the TCL OPTIMIZE command for simulator calibration. Using this code, the linear region mobility parameters used for the MINIMOS program and discussed in the Mobility Model section are automatically adjusted so that the MINIMOS calculated current values fit the experimental data. This automates a routine and time-consuming task formerly done manually.

Vienna Interactive Data Editor

The Vienna Interactive Data Editor (VIDE) provides a set of tools for data manipulation and data transformation. The program, i.e., toolbox, consists of three parts: input, manipulation, and output. Separating the parts in this way makes it easier to implement additional data formats from a variety of programs. For this toolbox, there are five classes of operations that apply to either one, two, or three dimensions:

- Geometry handling (e.g., stretching, scaling, and shifting)
- Quantity handling (e.g., stretching, scaling, shifting, and least squares fit)
- Quantity arithmetic (e.g., multiplication, square root, and logarithm)
- Grid handling (e.g., creation of new grid, interpolation, and merging of grids)
- Tools (e.g., plotting and rotation)

Additionally, powerful macros can be defined as procedures in command files. The expansion of 1-D doping profiles to 2-D by elliptic rotations is an example of such a macro. To illustrate the expansion, this section describes the combination of the 2-D source and drain profiles (arsenic and phosphorus) of a MOS transistor simulated by PROMIS, and the 1-D channel profile (boron) simulated by SUPREM3. First, the 1-D boron profile from the SUPREM3 simulator is expanded to 2-D and stretched to the appropriate geometry. Then, the profile is interpolated on the grid defined by PROMIS. The boron profile is the acceptor profile, whereas the sum of the arsenic and the phosphorus profiles gives the donor concentration. The result, consisting of the net 2-D doping, is shown in Figure 14.

The applications of VIDE are not restricted to process and device simulation. The toolbox allows the manipulation of data of arbitrary origin. The program has been successfully used to analyze the accuracy of simulation results for the calculation of the absolute and relative error, as well as for least squares fits to measurement data. Furthermore,

```
BEGIN
!
! First the block definition.
!
DEFINE BLOCK TUNE
 PARAMETERS MR, MX, MT
 INPUTS  L, VGS
 OUTPUTS IDS
 BEGIN
   READ COMMANDS T1.MMI TEMPLATE
   REPLACE $MR-VALUE$  /BY=MR   /IN=TEMPLATE
   REPLACE $MX-VALUE$  /BY=MX   /IN=TEMPLATE
   REPLACE $MT-VALUE$  /BY=MT   /IN=TEMPLATE
   REPLACE $VGS-VALUE$ /BY=VGS  /IN=TEMPLATE
   REPLACE $L-VALUE$   /BY=L    /IN=TEMPLATE
   FWRITE MINIRUN.MMI TEMPLATE
   MINIMOS MINIRUN.MMI /OUT=MINIRUN.MMO /DOPING=T1.2DOP
   READ MINIRUN.MMO IDS /INTO=IDS
 END
! Initial parameters values in INIT.PAR, Final values in T1.SAV
! Experimental data in T1.DAT.
OPTIMIZE TUNE /DATA=T1.DAT /INIT=T1.PAR /SAVE=T1.SAV
END
```

Figure 13 TCAD Command Language Code Illustrating the Application of the OPTIMIZE Command

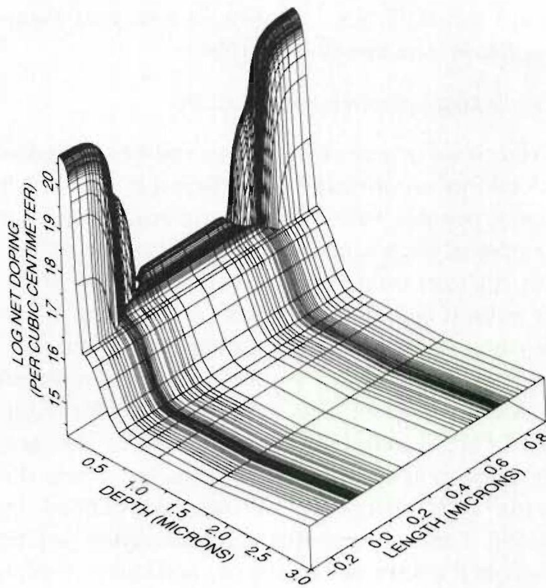


Figure 14 Net Doping Profile Combining the 2-D Source and Drain Profiles from PROMIS with the 1-D Channel Profile from the SUPREM3 Simulator

VIDE serves as a transformation program between different data formats of simulators.

VISTA Framework

VISTA, developed by Professor Selberherr's group at the TUV, is the first working TCAD framework based on a standardized data format.[19,20] A brief description of the VISTA system follows.

The Back End A multilanguage programming interface, e.g., FORTRAN, C, or XLISP, permits access to the simulation database. Such an interface has a well-defined data format, which is a basic requirement for tool-to-tool communication. The Program Interchange Format (PIF) data format in the VISTA environment provides a standardized way to transport simulation data, while at the same time remains open to future demands and extensions through its highly flexible structure.

The Front End The point-and-click interface of the TCAD shell, together with a visual programming interface, allows easy interaction for inexperienced users. The user interaction surface can be customized to accommodate user and program requirements. Interactive development of complex process flow simulations is supported. The TCAD shell automatically performs implicit parallelization and job control; the shell is also capable of quick standard visualization. Interfaces to advanced visualization tools, such as DEC AVS, provide state-of-the-art graphics. An on-line documentation system with automatic documentation generation from the source code always guarantees up-to-date information for users and programmers.

The Tool Aspect Within the VISTA package, a generic toolbox enables common data manipulations such as interpolation, gradient calculation, and arithmetical operations. The toolbox covers most of the standard data manipulations occurring in process and device simulation and thus allows program developers to focus on the main issues of their tasks. A tool abstraction concept helps with the automatic generation of front and back interfaces for new simulation tools. Strict rules ensure a consistent extension of the VISTA system to new and complex simulation tools.

General Aspects To face future challenges, a TCAD environment requires a consistent architecture together with high-level concepts. Modern software development techniques, such as automatic code generation and documentation, layered structures, and a high abstraction level of the underlying concepts, are the basis on which VISTA is built.

Present Status of VISTA within Digital's Development Work The most recent version of VISTA has been installed at the CEC in Vienna. Because of the close proximity of the CEC and the TUV, feedback on the concepts and the implementation details will directly influence the future progress of VISTA. This test version permits the application of the common data interface, i.e., the PIF application layer, and the simple coupling of the VISTA-PROMIS and VISTA-MINIMOS simulators, both of which are based on PROMIS and MINIMOS but have been further developed for use within a TCAD framework. It is expected that during 1993, VISTA will replace parts of the intermediate solutions for Digital's needs. Integration of the existing tools such as VIDE into VISTA is in progress.

Conclusions

The use of process and device simulation tools provides Digital's semiconductor process development teams with the following benefits:

- Decreases the number of experiments required to optimize the fabrication process. Experimental lots may take many months to process, whereas simulators can give results in minutes or hours. Simulation can never totally replace experimentation, however, because simulation is only a model of what we know about process and device physics, not reality itself. Smaller dimensions and new manufacturing processes require constant revision of our physical models and simulation tool capabilities.
- Allows the design teams to better estimate the spread of device performance, i.e., the so-called "worst case" conditions, before the process is well established. Thus, circuit designers can begin design earlier and with more confidence.
- Gives insight into the internal behavior of processes and devices to back up engineering judgment. For example, a layout design rule had been violated in an obscure way in the design of an I/O driver circuit. Redesign would have cost both time and space on the chip. Device simulation verified the device engineer's opinion that the violation would not cause any problem in actual practice.

References

1. H. Soleimani, "Modeling of High-dose I/I-induced Dopant Transient Diffusion and Dopant Transient Activation," *Journal of the Electrochemical Society* (Forthcoming, 1992).
2. H. Ryssel, J. Lorenz, and K. Hoffmann, "Models for Implantation into Multilayer Targets," *Applied Physics,* vol. 41 (1986): 201-207.
3. H. Ryssel, W. Krüger, and J. Lorenz, "Comparison of Monte Carlo Simulations and Analytical Models for the Calculation of Implantation Profiles in Multilayer Targets," *Nuclear Instruments and Methods in Physics Research B,* vol. 19/20 (1987): 40-44.
4. H. Ishiwara, S. Furukawa, J. Yamada, and M. Kawamura, *Ion Implantation in Semiconductors,* edited by S. Namba (New York: Plenum, 1975): 423.
5. A. Sabnis and J. Clemens, "Characterization of the Electron Mobility in the Inverted <100> Si-surface," *Proceedings of the IEEE International Electron Devices Meeting (IEDM)* (1979): 18-21.
6. S. Sun and J. Plummer, "Electron Mobility in Inversion and Accumulation Layers on Thermally Oxidized Silicon Surfaces," *IEEE Transactions,* ED-27 (1980): 1497-1508.
7. S. Selberherr, W. Hänsch, M. Seavey, and J. Slotboom, "The Evolution of the MINIMOS Mobility Model," *Electronics and Communication* (1990): 44, 161.
8. J. Faricelli, "Improved Substrate Current Prediction Using Depth Dependent Impact Ionization Parameters," *MINIMOS User's Symposium* (June 1989).
9. J. Slotboom, G. Streutker, G. Davids, and P. Hartog, "Surface Impact Ionization in Silicon Devices," *Proceedings of the IEEE International Electron Devices Meeting (IEDM)* (1987): 494-497.
10. P. Habas and J. Faricelli, "Investigation of the Physical Modeling of the Gate-depletion Effect," *IEEE Transactions* (Forthcoming, 1992).
11. M. Thurner, P. Lindorfer, and S. Selberherr, "Numerical Treatment of Nonrectangular Field Oxide for 3D MOSFET Simulation," *Simulation of Semiconductor Devices and Processes (SISDEP) Proceedings* (Bologna, 1988): 375-381.
12. K. Traar, M. Stiftinger, O. Heinreichsberger, and S. Selberherr, "3D Simulation of Semiconductor Devices on Supercomputers," *Proceedings of the ACM International Conference on Supercomputing* (1991).
13. G. Nanz, W. Kausel, and S. Selberherr, "Self-adaptive Space and Time Grids in Device Simulation," *International Journal for Numerical Methods in Engineering,* vol. 31, no. 7 (1991): 1357-1374.
14. G. Nanz, "Self-adaptive Space Grids in Process Simulation," *Microelectronics Journal* (Forthcoming, 1992).
15. G. Nanz and C. Schiebl, "Remarks on Numerical Methods in Capacitance Calculations," *IEEE Transactions on Computer-aided Design* (Forthcoming, 1992).
16. TCAD Architecture Committee, "Technology CAD Framework Architecture," edited by D. Boning (Research Triangle Park, NC: Semiconductor Research Corporation, May 1990).

17. CFI TCAD Framework Group, Semiconductor Wafer Representation Working Group, "The SWR Programming Interface, Version 0.9c," edited by A. Wong and G. Chin (Austin, TX: CAD Framework Initiative, Inc., February 1992).

18. S. Selberherr et al., "The Viennese TCAD System," *Proceedings of the International Workshop on VLSI Process and Device Modeling* (1991).

19. H. Pimingstorfer, S. Halama, S. Selberherr, K. Wimmer, and P. Verhas, "A Technology CAD Shell," *Proceedings of the International Conference on Simulation of Semiconductor Devices and Processes,* vol. 4, edited by W. Fichtner and D. Aemmer (Konstanz: Hartung-Gorre, 1991).

20. F. Fasching, H. Read, C. Fischer, S. Selberherr, H. Stippel, and W. Tuppa, "A PIF Implementation for TCAD Purposes," *Proceedings of the International Conference on Simulation of Semiconductor Devices and Processes,* vol. 4, edited by W. Fichtner and D. Aemmer (Konstanz: Hartung-Gorre, 1991).

Andre I. Nasr
Gregory J. Grula
Antonio C. Berti
Richard D. Jones

CMOS-4 Technology for Fast Logic and Dense On-chip Memory

Digital's fourth-generation CMOS technology has produced the industry's highest performance microprocessors. The NVAX and Alpha 21064 chips are based on 0.75-μm, 3.3-V CMOS technology capable of producing operating frequencies of up to 100 MHz and 200 MHz respectively. The high-performance CMOS transistors consist of a 105-Å gate oxide, symmetric n+ and p+ doped polysilicon for surface channel conduction, low threshold voltage, and good turn-off characteristics. The transistor has an on-wafer electrical gate length of 0.5 μm, a shallow medium doped drain junction for hot electron immunity, a $CoSi_2$ salicided gate, and source and drain regions for low interconnect sheet resistance. A $TiN/CoSi_2$ local interconnect scheme was used to strap the drain and gate regions to form a six-transistor memory cell with an area equivalent to 100 μm^2.

High-performance complementary metal-oxide semiconductor (CMOS) microprocessor design began in Digital's Hudson, Massachusetts, site in the mid-1980s. Digital's strategy calls for the scaling of feature sizes with each new generation of CMOS technology, coupled with larger die size to achieve higher circuit density and higher system performance. CMOS CPU operating frequency has doubled every two years since 1986, from 10 megahertz (MHz) for the CVAX chip fabricated with the CMOS-1 technology, to 100 MHz for complex instruction set computer (CISC)-based architecture as demonstrated by the NVAX chip.[1] It reached 200 MHz in 1991 for the Alpha 21064 reduced instruction set computer (RISC) architecture.[2] On-chip caches increased from 1 kilobyte (KB) for the CVAX chip in 1986 to a total of 16KB for the Alpha 21064 chip fabricated with CMOS-4 technology in 1991.

The transistor count implemented on a single microprocessor chip beginning with CMOS-1 technology swelled from 100,000 in 1986 to 1.7 million devices in 1991. During the same period of time, the polysilicon line that forms the transistor channel length was scaled from 2.0 microns (μm) to 0.75 μm, and the gate oxide thickness was decreased from 300 angstroms (Å) to 105 Å for the CMOS-4 process. Concurrently, the power supply was scaled from 5 volts (V) to 3.3 V to reduce power dissipation and to provide extra protection against gate oxide wear-out mechanisms.

High-performance microprocessors have posed a number of technological challenges in the design of transistors and on-chip caches. A number of process and device innovations have been introduced. For example, symmetric submicron n+ and p+ doped polysilicon transistors must provide low threshold voltage and good turn-off characteristics. Also, graded drain junctions must balance the driving current of the high-performance transistor with adequate hot carrier resistance. Self-aligned cobalt silicide ($CoSi_2$) polysilicon gate, and source and drain regions must provide low interconnect sheet resistance for improved circuit density and performance.

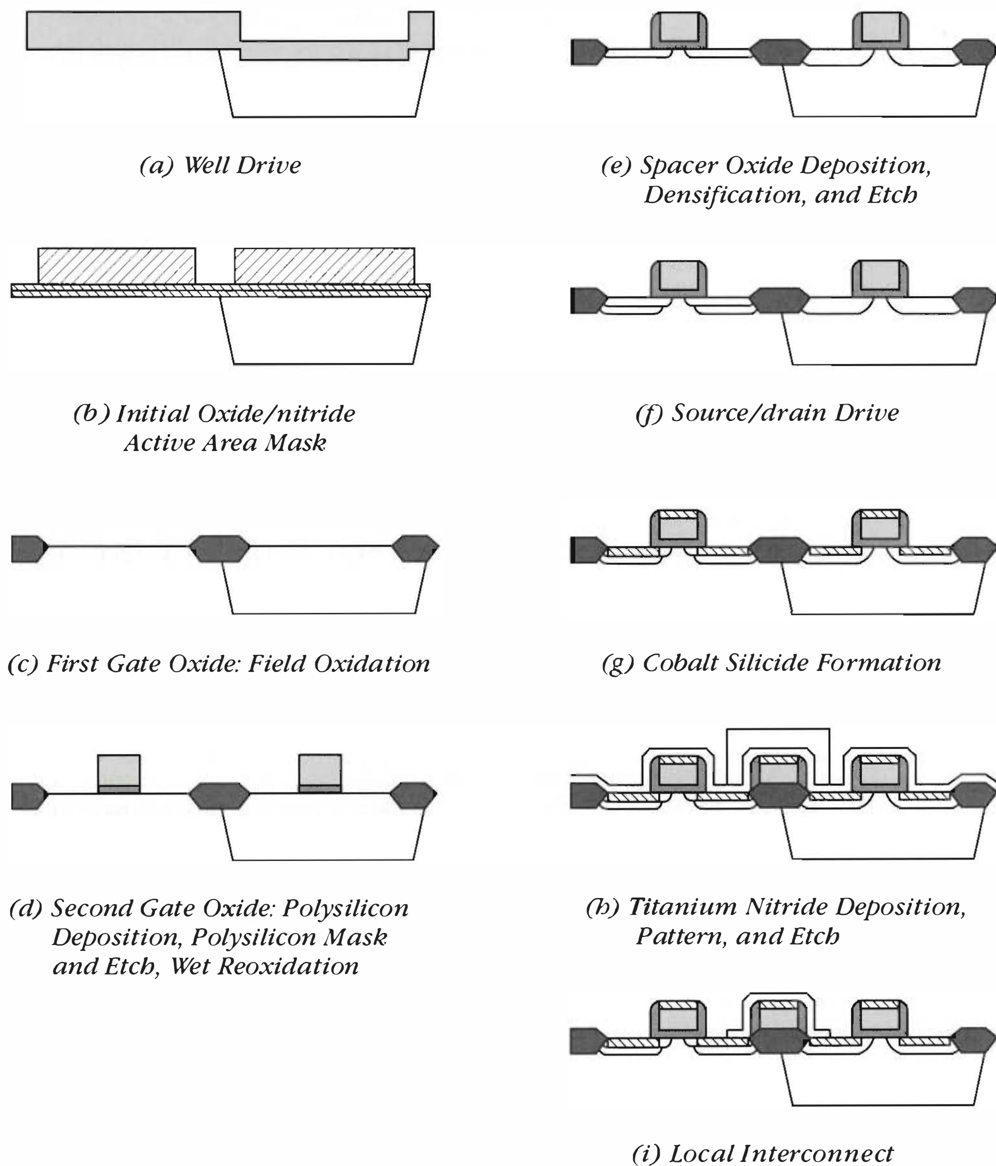
On-chip, high-density, static random-access memory (SRAM) is required for high-performance CMOS microprocessors. Previous technology scaling restricted the size of the on-chip cache due to the relatively large cell area. A variety of techniques were proposed to improve the RAM density, including the four-transistor (4T) cell with a second-layer polysilicon resistor and the six-transistor (6T) cell with buried contact (BC) or local interconnect (LI). The most attractive technique was the local interconnect scheme that was used to fabricate the 100-μm^2 cell, the cell being used successfully in the Alpha 21064 and NVAX microprocessors.

This paper describes the front-end process flow for Digital's fourth-generation 0.75-μm CMOS technology. It emphasizes the methods used to form

the low-resistance $CoSi_2$ and the titanium-nitride-based local interconnect used for the dense on-chip cache. The paper also discusses the CMOS transistor design and device characteristics. It concludes with a description of the SRAM cell.

Process Description

The front end of the CMOS-4 process is divided into six process modules: well formation, device isolation, gate formation, medium doped drain junction formation, $CoSi_2$ formation, and the local interconnect process. CMOS-4 technology was built upon the previous CMOS generations. Additional steps were incorporated into the process flow to meet new circuit design and layout requirements. Process steps were modified to accommodate scaling of device dimensions. A schematic process flow that depicts the various process modules is shown in Figure 1. The CMOS-4 layout design rules are given in brief in Table 1.



(a) Well Drive

(b) Initial Oxide/nitride Active Area Mask

(c) First Gate Oxide: Field Oxidation

(d) Second Gate Oxide: Polysilicon Deposition, Polysilicon Mask and Etch, Wet Reoxidation

(e) Spacer Oxide Deposition, Densification, and Etch

(f) Source/drain Drive

(g) Cobalt Silicide Formation

(h) Titanium Nitride Deposition, Pattern, and Etch

(i) Local Interconnect

Figure 1 CMOS-4 Front-end Process Flow

Table 1 Layout Design Rules for CMOS-4 Process

Design Rule	Dimension
Minimum active area width	1.5 μm
p+ to n+ spacing	3.0 μm
n+/n+ or p+/p+ spacing	0.75 μm
Polysilicon width/space	0.75/0.75 μm
Metal contact to polysilicon or active area	0.75 μm

Well Formation

Digital's CMOS technology uses an n-well formation process. Starting silicon wafers, or substrates, are doped p-type. The starting wafer is composed of a thin, high-resistivity epitaxial layer on a low-resistivity substrate (p on p+). A thick oxide is thermally grown on the wafer. Well regions are opened in photoresist using a photomasking process. The oxide is removed in the well regions. The n-well is formed by phosphorous ion implantation (approximately 10^{13} atoms per square centimeter [cm^2]) into the open regions. A high-temperature diffusion step is performed to drive the n-well implant to a specified depth in the epitaxial layer. For the CMOS-4 process, the diffusion step was carefully adjusted to account for the closer spacing between p and n transistors and the thin epitaxial layer thickness (see Figure 1a).

Epitaxial and well process requirements for Digital's CMOS-1 and CMOS-4 technologies are given in Table 2. The time required for well diffusion was shortened to 2.5 hours from 9 hours. Epitaxial thickness was reduced to 6.5 μm to improve latch-up immunity. After the well diffusion step, all oxide is removed from the wafer before formation of the device isolation regions.

Device Isolation

Isolation regions between devices are formed using a conventional, semirecessed local oxidation of silicon scheme referred to as LOCOS isolation.[3] After the pad oxide is grown and the silicon nitride is deposited, device areas (or active areas) are imaged in photoresist. Silicon nitride is removed from the isolation regions (or field regions) and retained on the active areas. In the substrate field regions, it is necessary to increase the concentration of p-type dopant to prevent unwanted current flow between devices. Boron is selectively implanted (approximately 10^{13} atoms per cm^2) into the substrate field regions and is blocked from entering the well regions by the photoresist layer. Note also that the nitride not covered with photoresist must be thick enough to block the implant (see Figure 1b). Next a thick thermal oxide approximately 0.45 μm is grown in the field regions. The nitride, however, prevents oxidation of the active area.

The LOCOS isolation has been tailored to CMOS-4 dimensions. Narrow-width transistors, routinely used in dense SRAM layouts, are particularly sensitive to isolation process conditions. Pad oxide and silicon nitride thicknesses have been selected to minimize excessive lateral oxidation of the active area, which could cause a reduction in the physical transistor width and thus lower the saturation current. The oxidation temperature and time must be optimized to grow the desired field oxide thickness without negative impact on other device parameters. For CMOS-4, the lateral oxide encroachment was reduced to 0.25 μm (per side). To minimize undesirable lateral diffusion of the field dopant, a relatively low temperature (950 degrees Celsius) is used for field oxidation (see Figure 1c).

After field oxidation, the nitride is chemically removed from the active areas, and the pad oxide is chemically stripped. The field oxide regions are semirecessed. Approximately 50 percent of the field oxide is below the silicon surface due to silicon consumption during the oxide growth.

Gate Formation

To achieve the desired electrical and reliability behavior, microcontamination must be controlled during gate region formation. The gate oxide material must be free of any defects and contain minimal amounts of ionic and metallic impurities. Rigorous

Table 2 Epitaxial and Well Diffusion Process for Two CMOS Technologies

Technology	Epitaxial Thickness	Well Diffusion Temperature	Well Diffusion Time
CMOS-1	11 μm	1130 degrees Celsius	9 hours
CMOS-4	6.5 μm	1100 degrees Celsius	2.5 hours

chemical cleaning of the wafer is required before gate oxide growth. Gettering (collecting) of mobile ions is performed during gate oxidation by the use of chlorine species. The wafer is transferred from gate oxide to polysilicon deposition immediately to minimize exposure of the dielectric film to airborne contamination.

Channel Doping A 450-Å thermal oxide is grown on the active areas to condition or remove impurities from the silicon surface. Often referred to as a "sacrificial" oxide, the thermal oxide is etched from the silicon surface before the real gate oxide is grown. Before the growth of the gate oxide, channel dopant for p- and n-type transistors is implanted through the sacrificial oxide. Photomasking steps are done to selectively implant boron into n-channel regions and phosphorus into p-channel regions. For n-channel regions, two separate boron implants are done: a shallow implant (approximately 10^{12} atoms per cm^2 at 20 kilo electron volts [keV]) for threshold voltage adjustment and a deep implant (approximately 10^{12} atoms per cm^2, 110 keV) to guard against various punch-through mechanisms.

Being a surface channel device, the p-channel region receives only a phosphorous threshold adjust implant (approximately 10^{12} atoms per cm^2 at 100 keV) since the well concentration is sufficient for proper subthreshold operation.

Gate Oxide and Gate Electrode Next the sacrificial oxide is removed, and 105-Å thin gate oxide is grown at 900 degrees Celsius. Then polycrystalline silicon of 3500 Å thickness is deposited, patterned, and etched to define the CMOS n-channel and p-channel transistor gate electrodes (see Figure 1d). The requirement for appropriate symmetric device design for the n- and p-channel devices is discussed in greater detail in the Symmetric Device Requirement section of this paper.

Medium Doped Drain Junction Formation

Next, a thin, 170-Å silicon dioxide layer is grown on the polysilicon and the active area regions. This layer is followed by a photoresist step that defines the n-channel device. Phosphorus is implanted (approximately 10^{13} atoms per cm^2 at 25 keV) to form a shallow, 0.1-μm, medium doped drain (MDD) junction for hot carrier protection[4] and good turn-off characteristics. The p-channel junction is formed next. A photomasking step defines the p-channel region, followed by a shallow boron difluoride (BF_2) implant (approximately 10^{15} atoms per cm^2 at 50 keV).

Spacer Formation A 2000-Å silicon dioxide layer is deposited and densified at 850 degrees Celsius. The oxide is reactive ion etched. This particular etch is anisotropic and leaves a vertical oxide sidewall 2000-Å wide along the 3500-Å thick polysilicon lines. The MDD junction is located under the spacer oxide wall and extends 2000 Å toward the drain contact region (see Figure 1e). The oxide spacer protects the MDD junction from receiving the heavy dose source and drain implants, and it protects the gate oxide during the $CoSi_2$ formation.

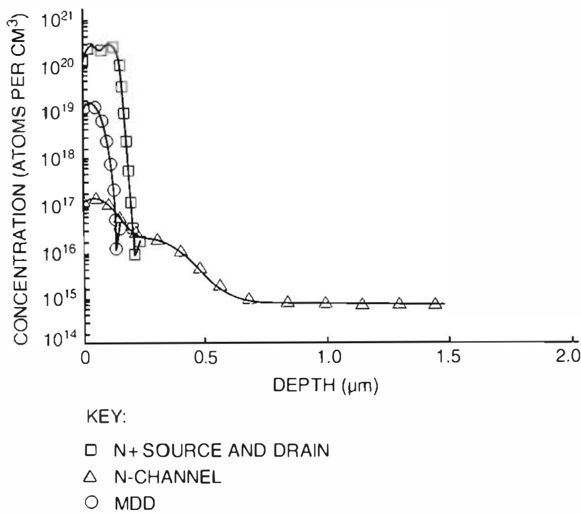
Source and Drain Junction Formation After the spacer formation, the photoresist step that defines the n-channel transistor is repeated. An arsenic implant (5×10^{15} atoms per cm^2, 100 keV) forms the n+ source/drain junction. A final high-temperature drive is performed to anneal the implant damage and to drive the junctions for the n-channel and p-channel devices to a final depth of 0.22 μm (see Figure 1f).

SUPREM-based simulation[5] of the CMOS-4 channel, MDD, and source and drain doping profiles are shown in Figure 2a for the n-channel device and in Figure 2b for the p-channel device. The channel surface concentration for both devices is approximately 1×10^{17} atoms per cm^2; the MDD junction depth is 0.1 μm; and the n+ and p+ junction depths are both 0.22 μm.

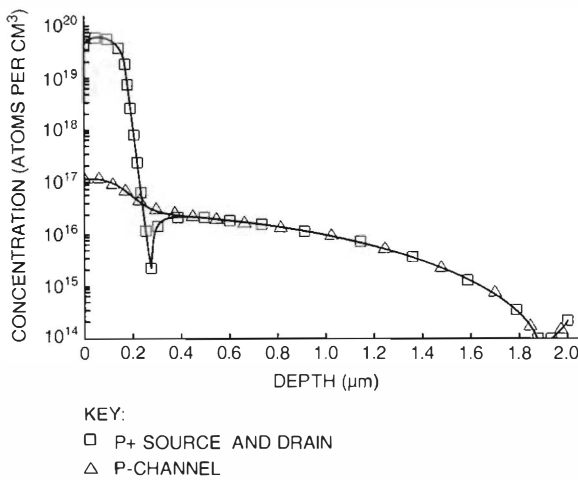
Cobalt Silicide Module

The CMOS-4 process technology has relied on the self-aligned $CoSi_2$ (salicide) to improve device performance by reducing the parasitic RC delay time associated with the gate and active area interconnects. $CoSi_2$ also provides good ohmic characteristics for metal contacts to gates and active areas, and acts as an etch stop during metal contact formation.[6] The term "salicide" refers to the formation of silicide on the gate and the source and drain region without the use of a masking layer. Because the silicide forms only in areas where the deposited metal can react with the exposed silicon surface, no silicide forms over silicon dioxide areas.

Prior to $CoSi_2$ formation, a wet chemical clean removes any surface contamination, and a hydrofluoric acid dip removes any residual silicon dioxide (approximately 100 Å). The wafers are then



(a) N-channel Device



(b) P-channel Device

Figure 2 SUPREM Process Simulations for CMOS-4 Channel Devices

introduced into a multistation, high-vacuum (down to 5×10^{-8} torr) sputtering system where another 100 Å of silicon dioxide is etched. This is followed by a sputter deposition of approximately 200 Å of pure cobalt film on the surface of the wafer.

The initial high-resistivity phase of CoSi is formed using a rapid thermal annealer. Each wafer is annealed at approximately 475 degrees Celsius for 90 seconds in nitrogen gas. Next the wafers are immersed in a selective etch, which is based on phosphoric acid to remove all the unreacted cobalt on the silicon dioxide without attacking the already formed CoSi or silicon dioxide. This 30-minute etch is self-limiting (once all the unreacted cobalt is removed, the reaction stops as the acid etches cobalt and nothing else). A one-minute 700 degrees Celsius anneal in nitrogen is then performed to form 700 Å of $CoSi_2$ with sheet resistance at approximately 5 ohms per square. (See Figure 1g.)

Figure 3 shows a cross-sectional photomicrograph of $CoSi_2$ film formed simultaneously over the polysilicon gate and the active area region. The $CoSi_2$ film thickness is approximately 700 Å. The spacer oxide separates the silicide in the active area from the silicide on the polysilicon gate.

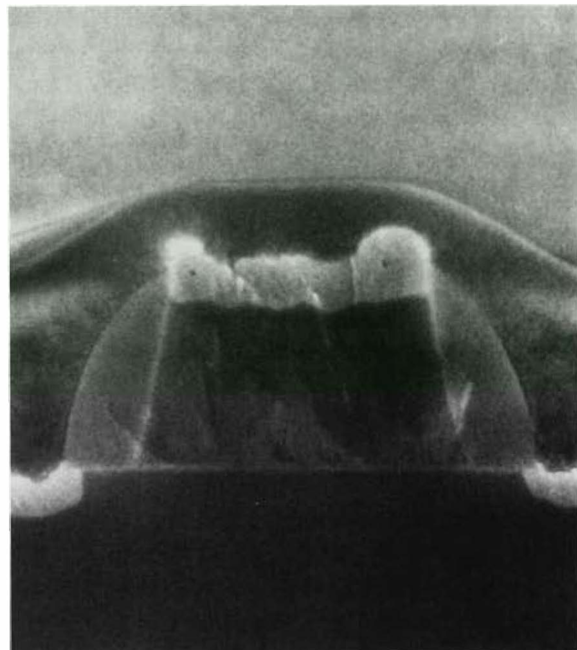


Figure 3 Cross Section of CMOS-4 Process Salicide $CoSi_2$ over Polysilicon and Active Area

Local Interconnect Process

The requirements for local interconnect material[7] include low-resistivity film with good etch selectivity to the underlying silicide film. Also the film must have good electrical contact resistance to both the silicided gates and silicided source and drain regions. The material selected was a laminate of 200 Å of titanium and 600 Å of low-resistivity titanium nitride (TiN). Titanium reduces the contact resistance through a chemical reaction with any surface oxides over the $CoSi_2$ to form a good ohmic contact. The titanium nitride has a higher resistivity, but is more chemically stable and does

not oxidize as readily as titanium during subsequent processing steps, such as oxygen plasma strips. Titanium nitride is also much easier to pattern with photoresist for the subsequent reactive ion etch step.

Local Interconnect Deposition The local interconnect is deposited using the same high-vacuum sputtering system used for cobalt deposition. The titanium and the titanium nitride are sequentially deposited in the same chamber. Nitrogen is reacted with the titanium on the target surface to form titanium nitride. Subsequently, the titanium nitride is sputter deposited onto the wafer. Varying the nitrogen flow can control the chemical and electrical properties of the titanium nitride film. The deposition process is adjusted to produce TiN with a resistivity of 50 μohm $\times$ cm.

Local Interconnect Etch Local interconnect etch consists of patterning TiN interconnects between source/drain and unrelated polysilicon (see Figure 1h). At this layer, there are two main concerns: (1) complete removal of TiN residue (stringers), and (2) etch selectivity to the other exposed materials on the wafer.

At this point in the fabrication process, the TiN is 4.5 times thicker at the side of an oxide spacer than over a flat region on the wafer. Continued etching to remove the TiN residue by the spacer would result in unacceptable overetching in the flat regions. Due to the conformality of the TiN to the spacer, the width of the TiN at the side of the spacer is the same as the thickness of the TiN in the flat regions. By controlling the ratio of the anisotropic etch component to the isotropic etch component, a portion of the TiN residue can be removed laterally, which significantly reduces the amount of overetch required. This combination of vertical and lateral etching is accomplished by using a chlorine/trifluoromethane (Cl_2/CHF_3) plasma chemistry at low pressure.

During the TiN etch, four other materials are exposed to the reactive plasma: silicon oxide, $CoSi_2$, photoresist, and silicon. To minimize material loss, the TiN etch rate must be optimized to significantly exceed the etch rates of these other materials. This is accomplished by using response surface methodology.

The optimized etch process consists of a two-step etch. The first step has an anisotropic characteristic. The pressure is 25 millitorr (mtorr); bias voltage is -220 V; Cl_2 is 20 standard cubic centimeters per minute (sccm); CHF_3 is 30 sccm; and boron trichloride (BCl_3) is 120 sccm. The second step is a more isotropic overetch step to remove residual TiN. It consists of a 25-mtorr pressure and a -94-V bias voltage. The Cl_2 is 60 sccm; CHF_3 is 40 sccm; and BCl_3 is 90 sccm. The temperature of the cathode is maintained at 50 degrees Celsius. The photoresist is stripped using a three-step process. The first step is a wet solvent strip that removes any soluble residues and plasma-hardened resist from the wafer. The second step is an oxygen plasma strip using a single-wafer stripper. The final step is another wet solvent strip (see Figure 1i).

Figure 4 is a photomicrograph of the top view of the active area and polysilicon regions. It shows the thin layer of TiN local interconnect strap that short-circuits the two layers together.

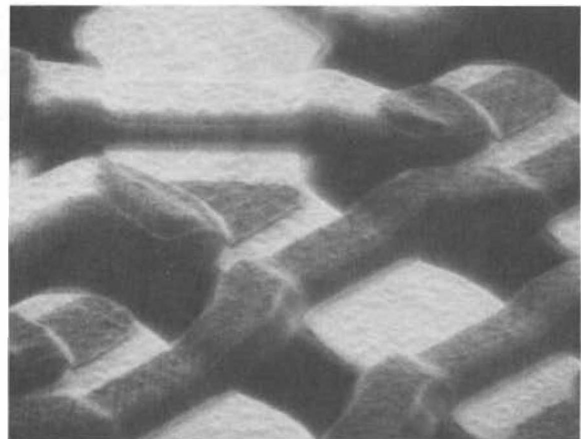


Figure 4 Photomicrograph Showing the Top View of the Active Area, Polysilicon, and TiN Local Interconnect

Transistor Design Considerations

Figure 5 shows a cross section of a CMOS-4 transistor. It highlights the polysilicon gate region, spacer region, $CoSi_2$ region, and the metal contact regions filled with TiN and tungsten films.

Unless adequately designed, submicron CMOS devices suffer many undesirable electrical effects. These effects are related to the scaling of the transistor channel length and the gate oxide thickness. Scaling the effective channel length for both n- and p-channel devices requires a reduced junction depth and an increased channel surface concentra-

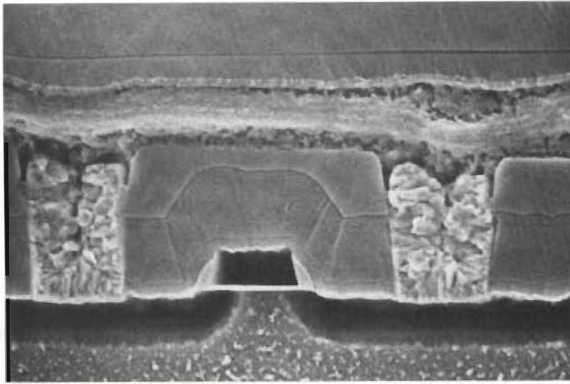


Figure 5 Cross Section of CMOS-4 0.75-μm Transistor

tion aimed at improving the short channel effects. These solutions may also contribute to an increase in the built-in electric field, an increase in the impact ionization, an increase in the substrate current, and a decrease in the punch-through voltage.

Arsenic doping profiles are usually used to fabricate shallow junctions in n-channel devices. They produce a high electric field that can cause a high rate of impact ionization between the electrons injected from the source and the fixed ions in the depletion region of the drain junction. On impact, electron-hole pairs are generated. The holes are swept to the source or substrate region and are known as the substrate current. The electrons subjected to the high drain electric field can gain enough energy to inject in the gate oxide region above the drain junction. These "hot electrons" may create interface states or may lose enough energy and be trapped in some defect location. Long-term effects of the trapping mechanism give rise to transconductance and saturation current degradation which eventually lead to circuit failure.

N-channel Junction Formation

To protect against the reliability hazards associated with arsenic profiles, a graded junction was implemented. Heavy emphasis was placed on the SUPREM process simulator and the MINIMOS device simulator.[8] The use of these simulation tools allowed us to accurately predict the device behavior under certain electrical conditions, and ensured that the device characteristics were optimized for high performance and superior reliability. Use of these simulators also reduced the dependence on wafer usage for process optimization.

The two-dimensional device optimization resulted in the following graded junction process parameters: a phosphorous dose of 7×10^{13} atoms per cm^2 at 25 keV, diffused to a 0.1-μm junction depth with a spacer width of 0.15 to 0.2 μm. The phosphorous surface concentration was set to approximately 10^{19} atoms per cm^3 to reduce the source and drain series resistance, R_t, and to accomplish the highest possible saturation current, I_{dsat}, while maintaining low substrate current for improved hot carrier reliability. The graded phosphorous junction was called medium doped drain (MDD) to refer to the relatively high doping concentration (1×10^{19} atoms per cm^3). In contrast, the lightly doped drain (LDD) junction, with doping in the 10^{17} to 10^{18} atoms per cm^3 range, suffers from large R_t and low driving current capability.

By implementing the MDD process, we were able to accomplish the following n-channel device characteristics: (1) minimum device lifetime of 20 years at a drain voltage bias (V_{ds}) of 4.3 V, (2) source and drain series resistance of 0.05 ohm $\times$ cm, (3) driving current capability I_{dsat} of 0.385 milliampere (mA) per μm, and (4) punch-through voltage (BVDSS) above 7 V. See Table 3.

Table 3 Electrical Parameters for CMOS-4 Transistors

	N-channel	P-channel
MDD X_j	0.1 μm	NA
P+ X_j	NA	0.18 μm
X_{well}	NA	1.75 μm
T_{OX}	105 Å	105 Å
V_{TX}	0.5 V	−0.5 V
L_{eff}	0.5 μm	0.5 μm
Delta L	0.25 μm	0.25 μm
Delta W	0.55 μm	0.65 μm
I_{dsat}	0.385 mA per μm	−0.167 mA per μm
BVDSS	>7 V	>−7 V

Figure 6 shows the results of two-dimensional MINIMOS simulation of the lateral electric field distribution in the CMOS-4 n-channel device with MDD junction profiles. The electric field is plotted along the transistor channel region starting from the source toward the drain region. The device has a drawn polysilicon length of 0.75 μm and a gate oxide thickness of 105 Å. The drain voltage biases, V_{ds}, were set to 4.3 V and 3.3 V, and the gate voltage,

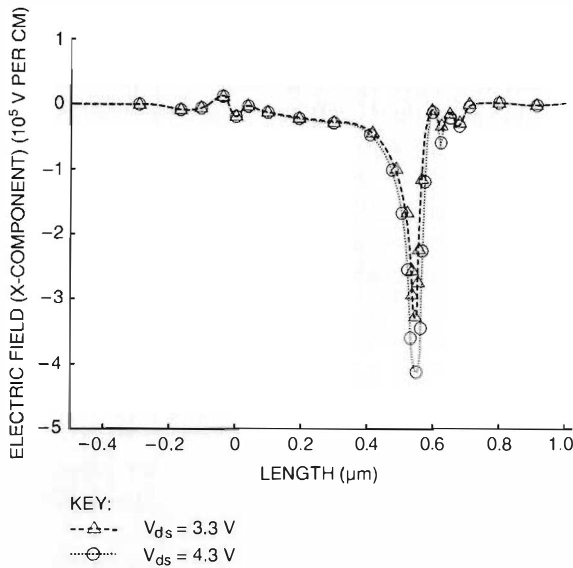


Figure 6 N-channel MINIMOS Simulation of Lateral Field for Medium Doped Drain Junctions

V_{gs}, was set to 2.2 V. Notice that the peak field for 3.3-V operation is approximately 0.3×10^6 V per cm compared to the worst-case field condition of 0.4×10^6 V per cm, where V_{ds} equals 4.3 V.

P-channel Junction Formation

The p-channel device requires special care in the design of the p+ source and drain junction depth to ensure that boron does not penetrate the thin gate oxide and enter the n-well region. The p+ junction dose, junction depth, and temperature cycle were optimized to accomplish a low R_t, by using BF_2 (1×10^{15} atoms per cm^2 and an energy of 50 keV). In addition, the diffusion time was minimized, and good threshold voltage control and punch-through protection were maintained. The boron distribution in polysilicon obtained with extensive secondary ion mass spectroscopy (SIMS) analysis, as well as negative bias and temperature instability (NBTI) tests, ruled out any boron penetration.

Symmetric Device Requirement

High-performance submicron CMOS technologies require the simultaneous optimization of the n-channel and p-channel devices for high driving current capability and excellent short-channel device characteristics. This is best accomplished with symmetric design where channel doping, junction depth, and threshold voltage of both transistors are equivalent.[9] Furthermore, the n-channel has an n-type doped polysilicon formed during the n+ source/drain junction and the p-channel has a p-type doped polysilicon formed during the p+ source/drain junction. The technique of symmetrically designing the devices allows the threshold voltages to be equal in magnitude, and suitably low for high driving current capability while maintaining good punch-through characteristics.

Figure 7 shows the results of the MINIMOS simulation of the potential distribution in CMOS-4 transistors with p-type doped polysilicon gate. The bias points were set for V_{gs} equals 0 V and V_{ds} equals −3.3 V. The channel length was 0.75 μm, and the gate oxide was 105 Å. Superior punch-through characteristics are observed since the potential contour lines do not spread significantly toward the source.

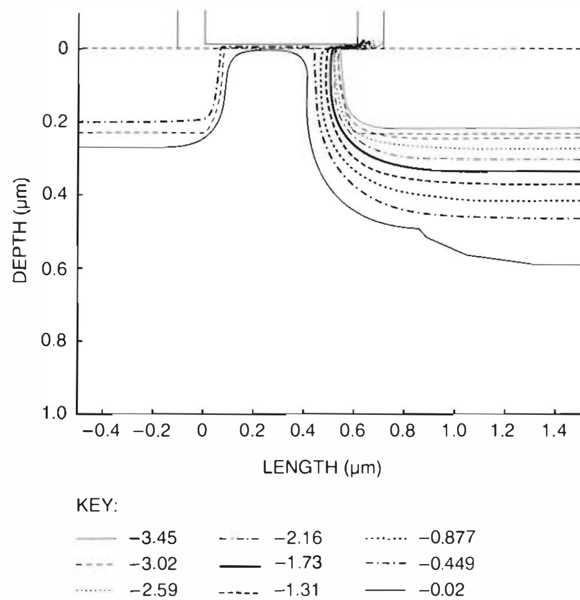


Figure 7 P-channel MINIMOS Simulation of Potential Distribution

Device Characteristics

The extrapolated threshold voltage for an n-channel device plotted as a function of gate length is shown in Figure 8. Excellent threshold voltage control is shown for channel length down to 0.5 μm. The n-channel drain current, I_{ds}, is plotted in Figure 9 as a function of drain voltage, V_{ds}, while V_{gs} is varied from 0 to 5 V with 0.5-V steps. In Figure 10, the drain current is plotted on a logarithmic scale as a function of gate voltage to highlight the subthreshold

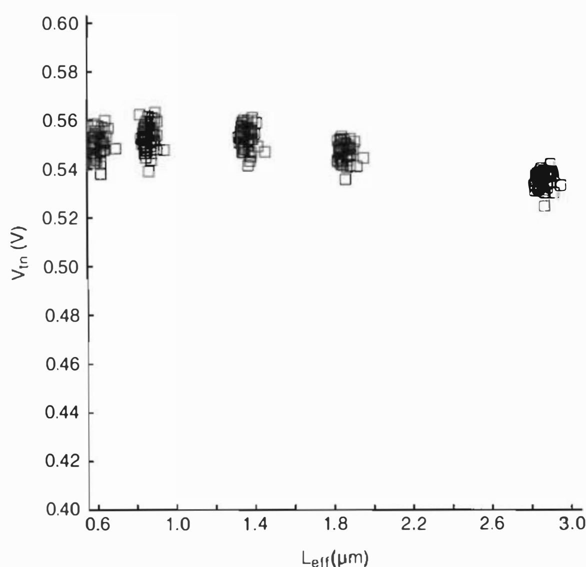


Figure 8 N-channel Threshold Voltage Plotted as a Function of Effective Channel Length

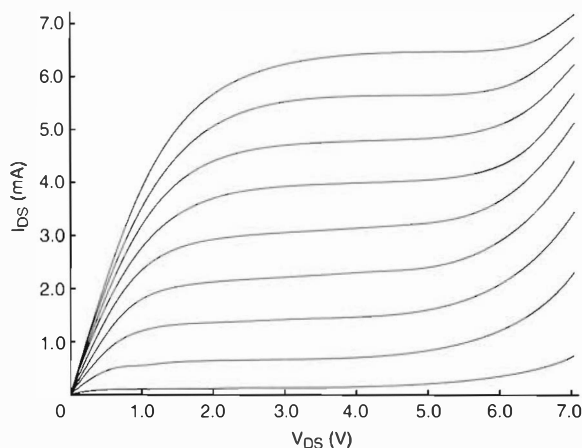
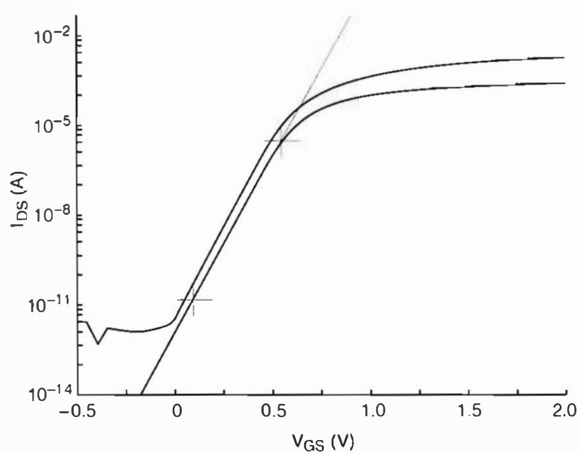


Figure 9 N-channel Drain Current Plotted as a Function of Drain Voltage

slope behavior for V_{ds} of 0.1 V and 3.6 V. The subthreshold slope was measured to be 86 mV per decade and is characterized by good drain-induced, barrier-lowering characteristics. The drawn dimensions of the transistor are 12.5 μm wide by 0.75 μm long.

Similar characteristics are observed for the p-channel device and are shown in Figures 11, 12, and 13. The subthreshold current conduction and punch-through characteristics are very similar to those of the n-channel device.



	GRADIENT	1/GRADIENT	X INTERCEPT	Y INTERCEPT
LINE 1	11.6	86.1×10^{-3}	1.02	1.27×10^{-12}
LINE 2				

Figure 10 N-channel Drain Current Plotted as a Function of Gate Voltage

Table 3 shows typical CMOS-4 transistor process and device parameters. The junction depths, X_j, and the n-well depth, X_{well}, are simulated with the SUPREM process simulator and verified with SIMS analysis. T_{OX} is the physical gate oxide thickness; V_{TX} is the extracted threshold voltage; L_{eff} is the nominal final channel length, and delta L and delta W are electrically extracted using the Terada method, which accounts for the parasitic series resistance. I_{dsat} is the saturation current measured with the drain and gate voltage at 3.3 V. BVDSS is the punch-through voltage measured with V_{gs} set at 0 V.

Silicided Interconnects Characteristics

Table 4 shows the effects of the salicide process on the parasitic resistance in four consecutive technologies. The CMOS-1 process uses no silicided gate or drain and therefore is expected to have a high interconnect sheet and contact resistance. CMOS-2, on the other hand, uses a low sheet tungsten silicided (WSi_2) polysilicon gate with a sheet resistance of 3 ohms per square. The CMOS-3 and CMOS-4 technologies use salicided low sheet resistance $CoSi_2$ for both the polysilicon gate and the source/drain region with a sheet resistance of 5 ohms per square.

SRAM Implementation

A six-transistor (6T) cell was selected for its process simplicity and cell stability. To provide a dense, cost-effective SRAM capability, the 6T cell was

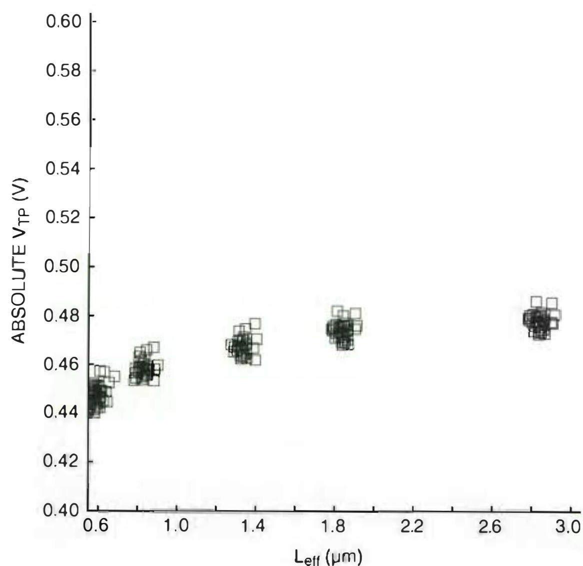


Figure 11 P-channel Threshold Voltage Plotted as a Function of Effective Channel Length

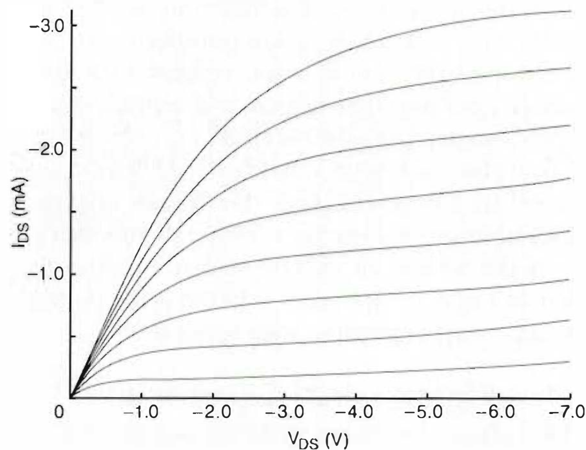
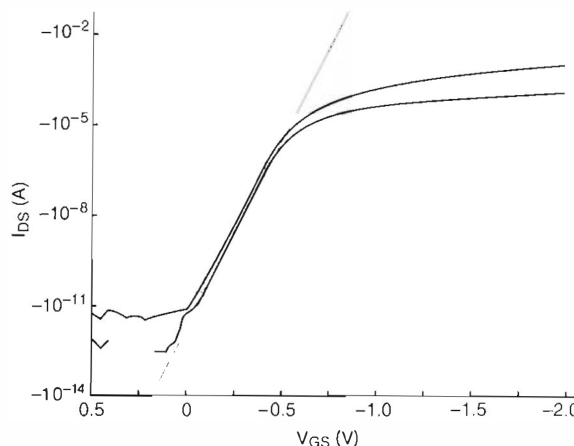


Figure 12 P-channel Drain Current Plotted as a Function of Drain Voltage



	GRADIENT	1/GRADIENT	X INTERCEPT	Y INTERCEPT
LINE 1	-12.1	-0.082	-0.97	-1.65×10^{-12}
LINE 2				

Figure 13 P-channel Drain Current Plotted as a Function of Gate Voltage

chosen over the 4T cell, which requires complex, two-level polysilicon films.

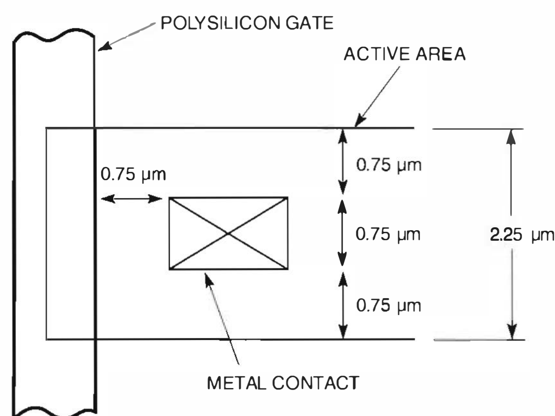
During the initial 6T cell process development, the TiN local interconnect scheme was considered advantageous to the buried contact scheme. In the buried contact procedure, the gate oxide is patterned and etched, and then a polysilicon gate is deposited to provide the contact between polysilicon and the active area. This technique allows the polysilicon film to access the source/drain region without the need for area-consuming metal contact. Unfortunately, this technique is not readily compatible with symmetric n+ and p+ doped polysilicon structures. In addition, silicon grooves might form during polysilicon etch, which could jeopardize the junction integrity and cause leakage or short circuits to the silicon substrate.

The preferred method to access the source/drain region was the use of TiN strap over $CoSi_2$. TiN local interconnect is a conductive material. When it is sputter deposited on the wafer, pattern and etch can be used to strap the node of one transistor to the gate or drain of another transistor. Also, the TiN local interconnect provides excellent etch selectivity to the underlying $CoSi_2$ material. The TiN local interconnect process proved superior to the buried contact scheme because the improved etch selectivity to $CoSi_2$ prevents junction leakage.

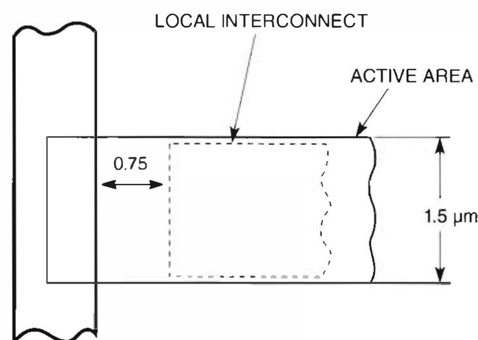
In standard layout techniques, the metal 1 contact is spaced 0.75 μm from the edge of the polysilicon gate and the isolation. This spacing results in a 2.25-μm wide active area, as shown in Figure 14a. In contrast, the local interconnect technique does not require a contact region; therefore the active area width can be scaled to 1.5 μm, as shown in Figure 14b. The use of local interconnect has reduced the 6T cell area from 120 μm^2 (no LI) to 100 μm^2 (with LI). In addition, the use of LI has improved yield due to a relaxed metal 1 contact requirement and metal spacing.

Table 4 Sheet Resistances for CMOS Technologies (Ohms per Square)

	CMOS-1 N+/P+	CMOS-2 N+/P+	CMOS-3 N+/P+	CMOS-4 N+/P+
Source/drain sheet resistance	40/75	40/75	5	5
Polysilicon sheet resistance	40/NA	3	5	5
Local interconnect	NA	NA	NA	6



(a) Standard Layout



(b) Local Interconnect Layout

Figure 14 Layout Schematics Comparing Metal Contact with Local Interconnect

Figure 15a is a photomicrograph of a CMOS-4 6T SRAM cell taken after LI etch and photoresist strip. It highlights the active area regions covered with $CoSi_2$ and TiN LI straps. Figure 15b shows a layout of the SRAM cell used in the Alpha 21064 microprocessor. Transistors T1, T2, etc., are highlighted in Figures 15 and 16 to simplify their identification. The cell area is 6.75 by 14.8 μm^2 (100 μm^2). The cell uses only three metal 1 contacts (V_{DD} and V_{SS}) to active area regions, compared to eight contacts in the cell with no LI. The minimum transistor width in the cell is 1.5 μm.

Summary

CMOS-4 technology for the Alpha 21064 and the NVAX microprocessors was discussed in detail. Process and device features for fast logic and dense on-chip SRAM were presented. The high-performance transistor requires the simultaneous optimization of the drain junction for hot carrier resistance and for high driving current capability. Low-resistance silicided interconnect uses a robust $CoSi_2$ process. On-chip SRAM based on a 6T cell with TiN local interconnect provides high-density and high-yield products.

Acknowledgment

The authors wish to acknowledge the contributions of all the members of the CMOS-4 group and the FAB-4 manufacturing group.

References

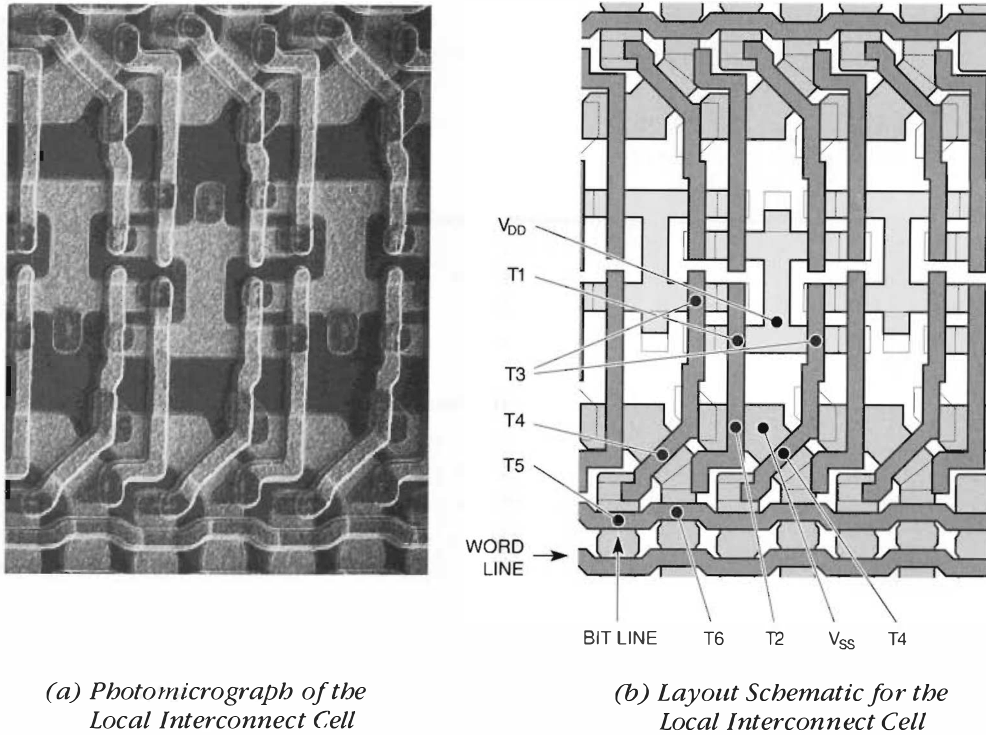
1. R. Badeau et al., "100 MHz Macropipelined CISC CMOS Microprocessor," *IEEE International Solid-State Circuits Conference Digest of Technical Papers* (February 1992): 104-105.

2. D. Dobberpuhl et al., "A 200MHz 64b Dual-Issue CMOS Microprocessor," *IEEE International Solid-State Circuits Conference Digest of Technical Papers* (February 1992): 106-107.

3. J. Appels et al., "Local Oxidation of Silicon and Its Application in Semiconductor Device Technology," *Philips Research Reports,* vol. 25 (1970): 118-132.

4. P. Tsang, S. Ogura, and W. Walker, "Fabrication of High Performance LDDFET's with Oxide Sidewall-Spacer Technology," *IEEE Transactions on Electron Devices,* vol. ED-30 (1988): 652-657.

5. C. Ho, J. Plummer, S. Hansen, and R. Dutton, "VLSI Process Modeling SUPREM-III," *IEEE Trans-*



(a) Photomicrograph of the Local Interconnect Cell

(b) Layout Schematic for the Local Interconnect Cell

Figure 15 Local Interconnect Cell Used in the Alpha 21064 Chip

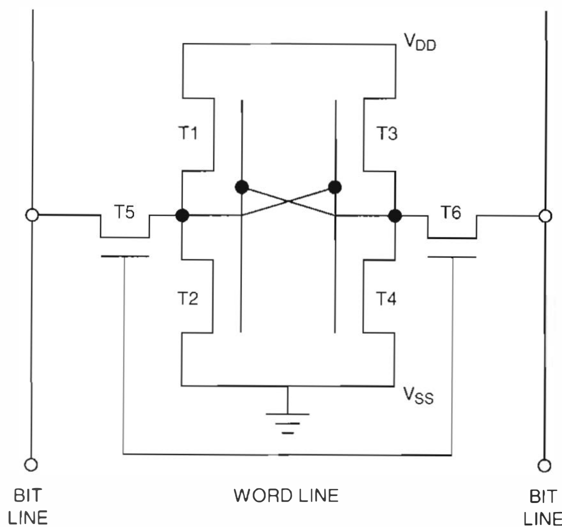


Figure 16 Six-transistor Cell Identifying the Various Transistors

actions on Electron Devices, vol. ED-30 (1983): 1438.

6. S. Murarka, D. Fraser, A. Sinha, and S. Hillenius, "Self-Aligned Cobalt Disilicide for Gate and Interconnection and Contacts to Shallow Junctions," *IEEE Transactions on Electron Devices,* vol. ED-34 (1987): 2108-2114.

7. T. Tang, C. Chia-Wei, R. Haken, and T. Holloway, "VLSI Local Interconnect Level Using Titanium Nitride," *IEDM Technical Digest* (December 1985): 714-717.

8. S. Selberherr, A. Schültz, and H. Pötzl, "MINIMOS—A Two-Dimensional MOS Transistor Analyzer," *IEEE Transactions on Electron Devices,* vol. ED-27 (1980): 1540-1550.

9. K. Cham and D. Wenocur, "Submicrometer Thin Gate Oxide P-Channel Transistors with P+ Polysilicon Gates for VLSI Application," *IEEE Electron Device Letters,* vol. EDL-7, no. 1 (January 1986).

Marion M. Garver
Joseph M. Bulger
Thomas E. Clark
Jamshed H. Dubash
Lorain M. Ross
Daniel J. Welch

CMOS-4 Back-end Process Development for a VLSI 0.75-μm Triple-level Interconnection Technology

Digital's CMOS-4 on-chip interconnect technology, developed for and used in production of the NVAX and the Alpha 21064 microprocessor chips, is a three-level aluminum alloy metallization process, with planarized TEOS-based silicon dioxide dielectrics, tungsten-filled contacts and vias, and a minimum feature size of 0.75 μm. The process development effort was a twofold approach based on the maximum use of existing manufacturing capability and the introduction of required new process features. For photolithography, plasma etch, and PVD metallization, the 1.0-μm manufacturing equipment set and processes were modified and reoptimized for the submicron regime. In addition, two new process features, a blanket CVD tungsten process and a TEOS-based oxide planarization process, were developed and implemented in manufacturing to meet the CMOS-4 technology requirements.

Each generation of Digital's complementary metal-oxide semiconductor (CMOS) very large-scale integration (VLSI) microprocessor development has the goal of providing a 30 percent net incremental performance improvement and a twofold area density improvement from the previous technology. This logic design need for higher density and improved performance places a considerable demand on ultra-large-scale integration (ULSI) circuitry to provide processes that permit a scaling of horizontal geometries with vertical film thicknesses remaining constant.[1]

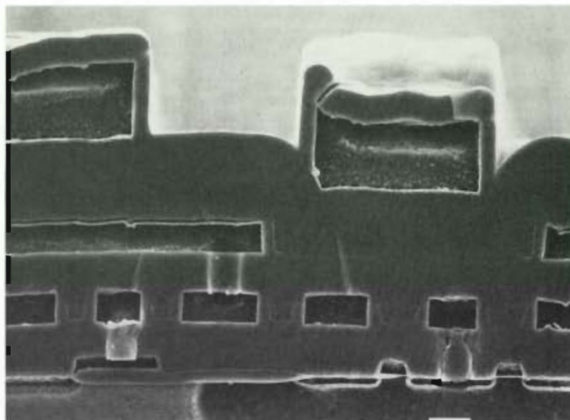
The technology goals for fourth-generation CMOS (CMOS-4) were met by providing a 25 percent algorithmic reduction of horizontal feature size from 1.0 micron (μm) to 0.75 μm, accompanied by minimal or no reduction in back-end interconnect or dielectric thicknesses. The process materials and critical parameters are described in Table 1. The small spatial resolution required vertical-walled vias to access smaller-pitched metal layers efficiently. Interconnect reliability was maintained through implementation of a tungsten-filled via-plug to improve current spreading and to maintain metal step coverage into contacts.[2] The CMOS-4 interconnect structure is shown in Figure 1.

The CMOS-4 process was developed in a manufacturing fabrication clean room originally configured for the preceding 1.0-μm CMOS-3 technology. The goal of the Advanced Semiconductor Development (ASD) and Manufacturing Engineering Groups was to introduce as few process changes and new pieces of equipment as possible. For two of the processes, joint development efforts at equipment vendor sites were conducted to develop hardware and assess process feasibility. The equipment was purchased and installed in the manufacturing clean room with final process characterization and integration performed at Digital's Hudson facility.

This paper discusses how the existing tools were modified for use in the CMOS-4 process in the areas of photolithography, plasma etch, and physical vapor deposition (PVD) metallization. It describes the addition of blanket tungsten and plasma-enhanced tetraethylorthosilicate (PE-TEOS) oxide processes that were developed in clustered, multichamber tools and optimized to meet the submicron-level requirements of CMOS-4 technology.

Table 1 CMOS-4 Film Types, Thicknesses, and Critical Dimensions

Process Level	Material	Final Thickness Target	Critical Dimensions
Dielectric 1	Phosphorus-doped PE-TEOS and boron oxide	7500 Å	—
Metal 1 contact	Ti/TiN and W plug size	Plug recess <3000 Å	0.75 by 0.75 μm contact
Metal 1	Al:1%Cu/TiN cap	7500 Å	1.50/0.75 μm line/space
Dielectric 2	PE-TEOS and SOG	7500 Å	—
Metal 2 contact	Ti/TiN and W plug	Plug recess <3000 Å	0.75 by 0.75 μm via size
Metal 2	Al:1%Cu/TiN cap	7500 Å	1.88/0.75 μm line/space
Dielectric 3	PE-TEOS and SOG	18,000 Å	—
Metal 3 contact	Al:1%Cu	>0.3 μm	3.0 by 3.0 μm via size
Metal 3	TiN/Al:1%Cu/TiN cap	20,000 Å	4.5/3.0 μm line/space
Passivation	PE-TEOS	7500 Å	—



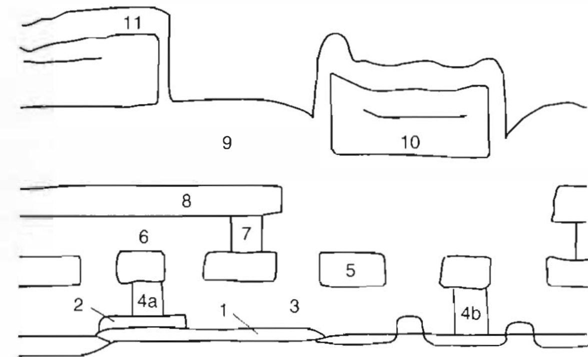


Figure 1 Cross-sectional Photomicrograph and Schematic of CMOS-4 Interconnect Structure

Modification of Back-end Processes

The details of development efforts and their resolution for photolithography, plasma etch, and PVD metallization are discussed in the following sections.

Photolithography

The CMOS-4 photolithography process uses single-layer photoresists, reduction steppers with an exposing wavelength of 436 nanometers (nm) and numeric aperture lenses of .45NA and .54NA, and spray/puddle develop. A photoresist thickness of 1.2 μm is used at contact levels to enhance resolution. The photoresist thickness used at metal 1 and metal 2 is 2.0 μm in order to prevent total photoresist erosion from the higher steps during the etch process. Metal 3 contact and metal 3 use a thicker, 3.5-μm photoresist to accommodate tapered oxide etching, extreme topography, and high loss of photoresist during etch. In the manufacturing line, multiple track and stepper combinations are allowed. Critical dimension control is maintained by running fixed exposures and monitoring process E_o once a shift for the possible coat-and-expose equipment sequences. Overlay control of ±.20 μm is achieved by running a lot pilot wafer and using

alignment offsets to center the lot distribution at zero.

The CMOS-4 photolithography process was transferred into the existing CMOS-3 production line without any modifications to existing equipment. However, the introduction of back-end processes did present resolution problems at the metal 2 and contact layers and poor overlay performance at all metal layers.

Resolution The challenge for many of the upper levels was to routinely resolve 0.75-μm geometries. For contact and via levels, part of the solution was to use thinner photoresist, overexposure, and only the higher NA steppers. Initially, material processed at these levels periodically exhibited distorted or missing contacts. Experimentation with focus/exposure matrices showed no further improvement with overexposure. Also, slight focus shifts were enough to considerably distort the contact pattern. A 0.5-μm focus offset provided the necessary latitude to eliminate resolution problems.

Metal 2 Processing Poor pattern definition caused both bridged photoresist lines over topography and rounding of the tips of lines. The photoresist could not be thinned below 2.0 μm due to the high loss of photoresist during etch. Elimination of dyed photoresist was an option because titanium nitride (TiN) antireflective coating (ARC) was used below the photoresist layer. Undyed photoresist eliminated the bridging problem. A focus shift similar to that used at the contact levels was necessary to best resolve the tips of the lines. Factorially designed experiments indicated that higher exposure and further defocus would help minimize metal short circuits. The increased exposure widened the gap between potentially short-circuited lines, and the defocus increased the degree of proximity effect,[3] leaving the dense lines significantly smaller than the isolated lines. The 0.5-μm focus offset was again selected as optimum. It decreased the photolithography contribution to metal short circuits yet still produced line widths that met design guide criteria.

Alignment Maximum misalignment tolerance is dictated by the reliability requirement to ensure 100 percent contact coverage. The alignment tolerance calculation was based on maximum allowable contact size and minimum metal line widths. For example:

Maximum metal 2 contact = 0.95 μm
Minimum metal 2 line width = 1.55 μm
Misalignment tolerance = (1.55 − 0.95) / 2 = 0.30 μm

Similar calculations performed at metal 1 contact and metal 3 contact indicated similar tolerances were needed. However, signal-to-noise ratios with the alignment systems were extremely low and overly sensitive to minor process fluctuations, such as changes in film thickness, film reflectivity, surface roughness, extent of planarization, and grain boundary highlighting. Initial attempts indicated alignment, when possible, had an average value above 0.40 μm for a lot. New alignment systems specifically designed for metal and planarized dielectric were considered. However, the decision was made to develop a new processing technique on the existing equipment.

A logical approach to the problem was to eliminate the metal from the alignment target areas. All efforts at optimizing target size still rendered only a marginally acceptable process. "Cutout" processing at metal 3 was used from the beginning of CMOS-4 development and was gradually introduced to the other metal layers as wafer lot volume increased and problems with alignment increased cycle times and scrap rates.

The cutout process involves running an extra masking step prior to metal alignment, which exposes the underlying alignment targets. Since the cutout openings are large, misalignment tolerance is on the order of microns and is achievable by aligning the cutout mask in a manual global alignment mode followed by a blind step sequence. Wafers are subsequently either dry or wet etched to remove the TiN and aluminum (Al), then returned to photolithography for standard metal alignment processing. Overlay results using the cutout at metal 2 and metal 3 average approximately 0.20 μm (±3 σ). Metal 1 performance is approximately 0.15 μm (±3 σ) by aligning to the still visible active area marks.

Plasma Etch/Strip Processes

The CMOS-4 back-end oxide etches, metal etches, and photoresist strip processes were developed on the existing manufacturing equipment. A straight-walled contact process for metal 1 contact (M1C) and metal 2 contact (M2C) needed to be developed for incorporation with the tungsten plug technology. Additionally, the 0.75-μm wide contacts with straight sidewalls (greater than 85 degrees) and

aggressive aspect ratios (height/width), required optimization of the photoresist strip processes. The metal 3 contact (M3C) tapered etch process was also redeveloped to meet CMOS-4 electromigration reliability requirements.

In addition to a straight-walled contact process for use with tungsten plugs, excellent selectivity to underlying materials was required to compensate for the increased planarity of the CMOS-4 dielectrics. For example, the planarity of the dielectric between polysilicon and metal 1 meant that the contacts to a polysilicon gate would be etched through the thinnest dielectric (approximately 0.75 μm), whereas the contacts to active area regions would need to be etched over a much thicker dielectric. The worst-case difference is approximately two times the thickness of the dielectric above the polysilicon layer, or approximately 1.5 μm, as shown in Figure 2. This difference in dielectric film thickness meant that the cobalt silicide ($CoSi_2$) film over the polysilicon contact would be overetched by approximately 100 percent during the M1C etch process. Therefore, the M1C etch process needed a very high differential etch rate or selectivity between the etch rate of the oxide as compared to the etch rate of the $CoSi_2$.[4]

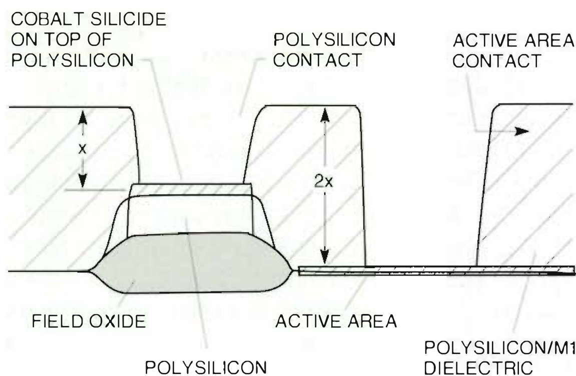


Figure 2 Schematic Drawing Showing the Difference in Step Height for Metal 1 Contact to Polysilicon and Source/Drain Regions

Straight-walled Contact Process The straight-walled contact process development was accomplished by experimenting with bias voltage, trifluoromethane:oxygen (CHF_3:O_2) gas ratio, and pressure.[5] The substrate bias (to control photoresist, oxide, and $CoSi_2$ etch rates) and the ratio of CHF_3:O_2 flows (to control sidewall profile and photoresist pullback) were determined to be the most critical parameters. The optimized process resulted in a high-throughput, uniform, 0.75-μm straight-walled contact process with an oxide-to-cobalt silicide selectivity of greater than 25:1. The underlying material at M2C etch is an aluminum (Al) alloy. Because of a very high selectivity of oxide-to-Al etch rates, an overetch of up to 100 percent at M2C etch did not impact the contact profile or the contact resistance.

Optimized Photoresist Strip Process The optimized M1C/M2C straight-walled etch process altered the amount of sidewall polymer formed in the contacts during the contact etch process.[6] In addition, the change in aspect ratio affected the ability of wet chemicals to remove all the photoresist and polymer during photoresist strip processing. It was shown empirically that residual polymer remaining in contacts after photoresist strip impacted contact resistance. The contact photoresist strip process was modified from a two-step dry/wet process to a three-step wet/dry/wet process. The first wet strip was required to pre-wet the polymer/photoresist in the contacts. The bulk of the photoresist was then removed in an oxygen downstream plasma stripper (dry) and was followed with a final wet strip to remove any residues. Beginning with a wet strip cycle also improves sidewall polymer removal for metal etch processes. Consequently, all the CMOS-4 back-end resist strip processes follow a wet/dry/wet strip process flow.

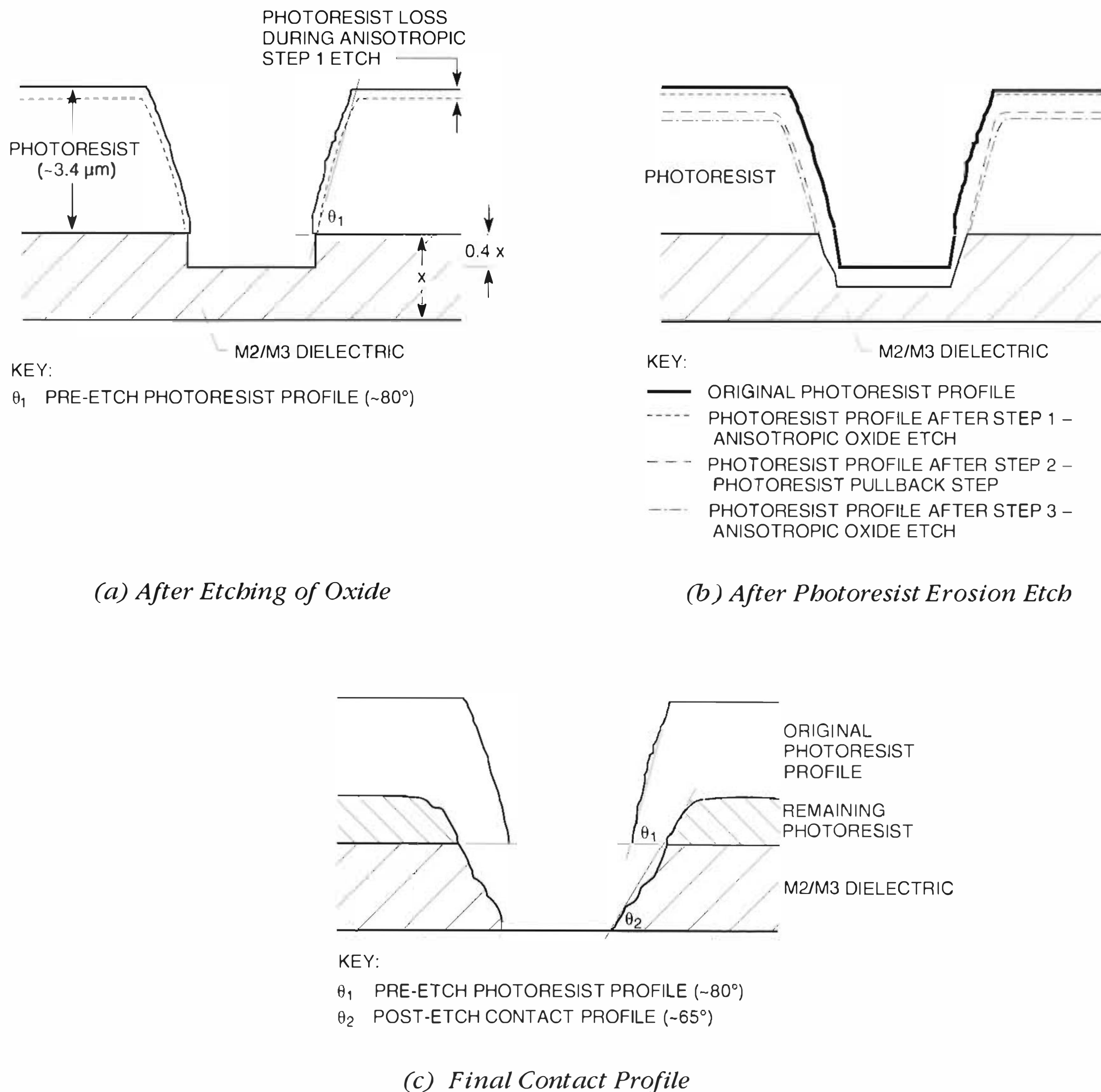
Metal 3 Contact Etch Process Initially, the M3C tapered etch process did not consistently meet the CMOS-4 electromigration minimum step-coverage requirement of 0.30 μm of metal on M3C sidewalls. A unique set of problems existed at M3C etch. The nonuniformity in the underlying topography caused the photoresist coat to be thinned over isolated metal lines. This thin coat led to early photoresist breakthrough and subsequent dielectric erosion during M3C etch. A thicker photoresist coat led to steeper sidewalls, which resulted in a degradation of metal step coverage.

An experimental design was used to optimize the photoresist and etch processes in unison.[7] The photoresist thickness was increased to 3.4 μm, and a focus offset was implemented during exposure to slope the photoresist profile prior to etch. The optimized photolithography process resulted in a pre-etch photoresist profile of 80 degrees $\pm$3 degrees. A multiple-step, tapered etch process was devel-

oped in which step 1 etched 40 percent of the contact depth anisotropically to maintain final contact critical dimension control. Steps 2 and 4 are photoresist pullback steps. These critical etch steps were optimized by running a series of designed experiments to characterize the responses of photoresist etch rate, uniformity, and lateral-to-vertical erosion rates. Etch rates were determined using patterned oxide test wafers which were cross-sectioned and analyzed using the scanning electron microscope (SEM). Substrate bias and oxygen flow were the primary parameters controlling lateral-to-vertical photoresist erosion rates. Steps 3 and 5 transfer the tapered photoresist profiles into the dielectric film. Figure 3 depicts this progression. Step 6 is a final clean-up step. The optimized M3C process demonstrates a post-etch contact slope of 65 degrees ±5 degrees, which consistently results in metal step coverage exceeding the 0.30-μm requirements. Measurements obtained by SEM are compared in Figure 4.

Physical Vapor Deposition Metallization

The PVD metallization processes are performed in a single-wafer, multichamber, high-vacuum sputter deposition system. The chambers include a radio frequency (RF) etch, titanium, cobalt, and aluminum-copper cathodes. The RF etch provides a sputter



(a) After Etching of Oxide

(b) After Photoresist Erosion Etch

(c) Final Contact Profile

Figure 3 Schematic Drawing Showing Metal 3 Contact Tapered Process

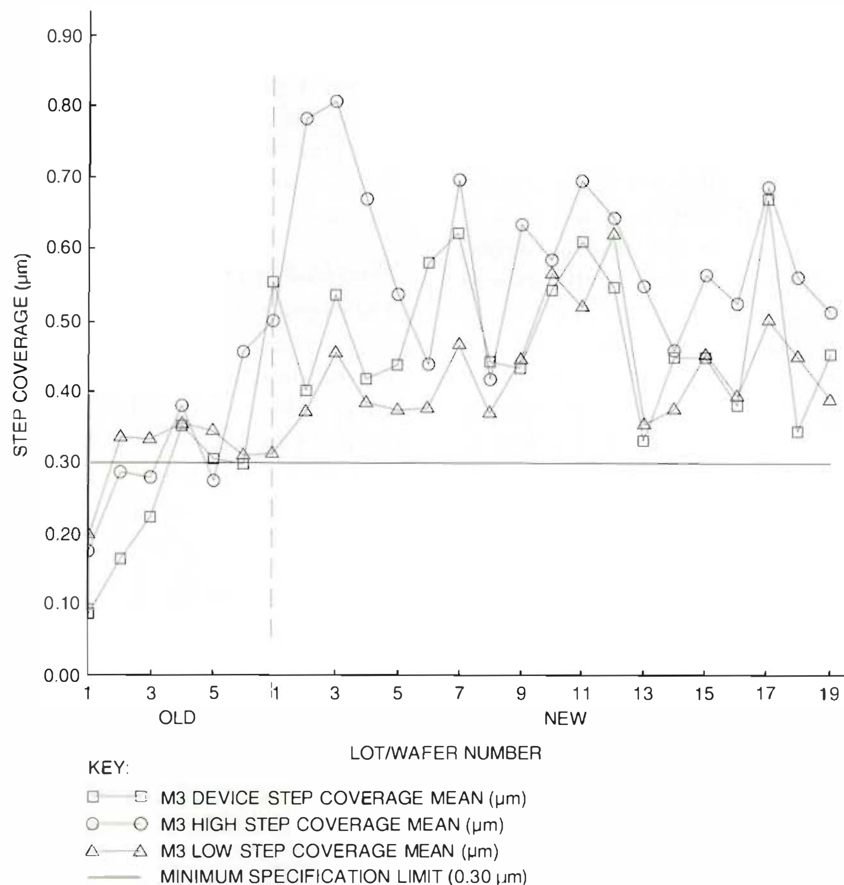


Figure 4 Comparison of Metal 3 Step Coverage with Standard and Optimized Contact Tapers

cleaning of the wafer to remove any native oxide from the wafer surface before the film is deposited. The titanium target is used to reactively form the TiN film that is used for local interconnect, tungsten adhesion layers, antireflective coatings, and a fuse link. The cobalt target is used for silicide formation on gate and source/drain regions. The Al:1%Cu alloy is used for three levels of on-chip low-resistance interconnect.

The reactive sputtering of the TiN film has provided a number of challenges in the area of particle control. Early modeling in the CMOS-4 development cycle using the CMOS-3 yield model predicted the impact of the TiN particle levels on the reduced metal pitch areas of the CMOS-4 process. The model indicated that the number of defects per 100 meters of interconnect would have to be reduced from the level of 20 defects added. To prevent the TiN film and associated interconnect from being among the top yield limiters, a level of less than 5 defects had to be attained.

To lower the number of defects, Digital process engineers worked with the equipment vendor to design and develop a new cathode. The CMOS-3 cathode used for both the TiN and Al films was a magnetron configuration, designed to enhance the deposition rate of the target material by creating additional bombarding ions. The magnetron has a fixed set of magnets oriented to confine the ionizing electrons and thus cause the target to erode in a racetrack pattern. This confinement results in re-deposition on areas of the target that are not sputtered. These areas become a major particle source. The new design is based on a set of rotating magnets that move the erosion pattern over the entire surface of the target and keep the surface free of any material build-up. Because of this enhanced sputter uniformity, the rotating configuration maintains a low particle count throughout the life of the target. The conventional magnetron and rotating magnetron defect density levels are compared in Figure 5.[8]

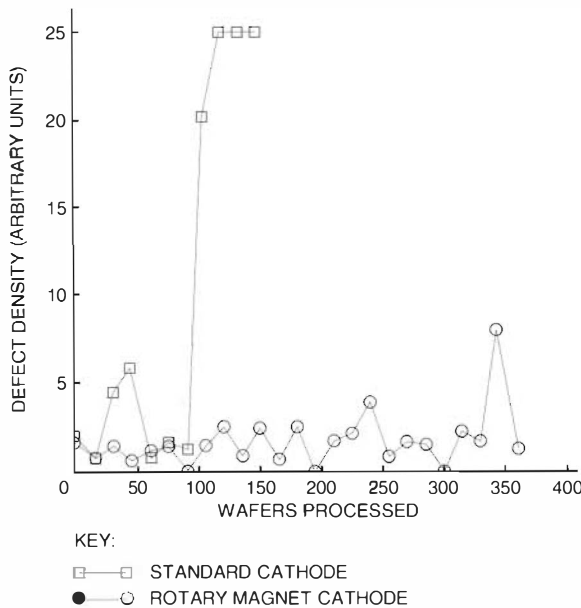


Figure 5 Cathode Configuration as Compared to Defect Density

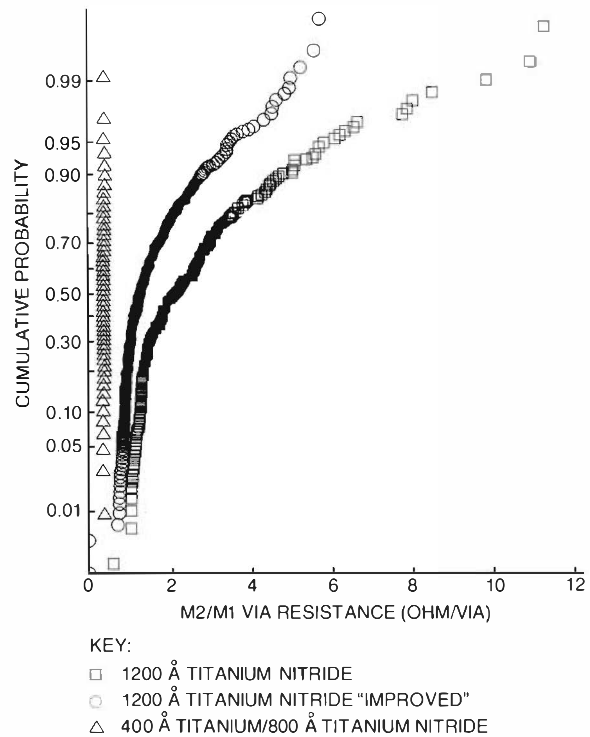


Figure 6 Titanium Nitride Adhesive Layer Optimization

Titanium-nitride Film The TiN process required re-optimization to characterize the rotating cathode and accommodate the CMOS-4 application of TiN as a tungsten plug adhesive layer. In addition to lowered particle levels, the process focused on achieving low contact resistance between the various metallurgical interfaces. Screening studies were performed to characterize adhesion layer properties for M1C and M2C via plugs. While minor effects on via resistance were observed for some of the factors, the single most important factor was the presence or absence of a titanium underlayer. With as little as 15 nm of titanium beneath the TiN, low via resistance was obtained. Without titanium, no conditions resulted in acceptably low resistance. These results are shown in Figure 6, which is a cumulative probability plot of via resistance for three adhesion layer processes: (1) 120 nm of TiN deposited using conditions acceptable at the M1C level, (2) 120 nm of TiN deposited using a modified deposition process, and (3) a film with 40 nm of titanium beneath 80 nm of TiN.

The importance of overall process integration became apparent when it was determined that the rotating cathode was damaging transistor characteristics during sputtering of the local interconnect TiN. The TiN deposition process was modified to decrease the substrate bias, and the new process recipe was characterized and retrofitted into the contact and via levels.[9] This modification resulted in a more manufacturable single recipe to maintain for all CMOS-4 TiN levels.

Deposition of the Al interconnect film for CMOS-4 required little alteration from the CMOS-3 recipes. Film thicknesses were reduced by 500 angstroms (Å) for metal 1 and metal 2 to maintain aspect ratios of no greater than 1. Metal 3 film thickness remained the same. The incorporation of the tungsten plug process into CMOS-4 technology simplified the metal step-coverage requirements because filling the contact with tungsten reduced the effect aspect ratio significantly. The metal 3 contacts used a tapered non-tungsten process and therefore required that the high step-coverage process developed for CMOS-3 technology be maintained and further improved for the CMOS-4 process.

The PVD Al step-coverage process is a three-step process, optimized for wafer temperature, aspect ratio, and the underlying material. The first step is a low-temperature deposition of a nucleation layer that provides a continuous coating over all surfaces. Step 2 uses a low-power temperature ramp to allow enhanced surface mobility. Controlling the time during this step increases the average distance

that material can move along the surface and into the contacts. A high-temperature, high-power third step is used to reach the final film thickness.[10]

Development efforts provided a manufacturable, high-throughput, and high-yielding metallization process. A modified TiN cathode made the existing equipment set serviceable for another technology generation. The use of the historical equipment database minimized development time for the new technology, and increased the level of understanding for a deposition technology with known benefits. The incremental nature of the changes from the CMOS-3 to the CMOS-4 process allowed an efficient technology transfer due to a reduced number of learning cycles and the use of a familiar equipment set.

Blanket Tungsten Plugs Process Development

As stated previously, new process technologies were developed to meet the CMOS-4 process criteria. Blanket tungsten plugs are an example of a new technology. As Figure 1 illustrates, blanket tungsten plugs are used in the CMOS-4 technology to vertically interconnect metal 1 to the silicon substrate or to polysilicon (i.e., contacts), as well as to vertically interconnect metal 1 and metal (i.e., vias). Blanket tungsten plugs were selected for use in the CMOS-4 technology because they minimize spatial requirements for contacts and vias by allowing the use of vertical-wall openings rather than the tapered openings used in earlier technologies.

At the outset of the CMOS-4 process development cycle, two tungsten plug technologies, selective and blanket, were evaluated. Although these plug technologies are very similar in their final structural form, their formation involves important differences. As shown in Figure 7a, selective tungsten plugs are formed by the deposition of tungsten only on conductive surfaces and not on the surrounding oxide surfaces. Such growth leads to a filling of openings from the bottom up. In comparison, blanket tungsten plugs, as shown in Figure 7b, are formed by a three-step process:

1. Sputter deposition of an adhesion layer
2. CVD of a conformal blanket tungsten film
3. Reactive ion etch back of the tungsten and adhesion layer to leave tungsten plugs surrounded by the adhesion layer

Blanket tungsten plug processing thus fills openings from the sides as well as from the bottom up. As a result, variable contact and via depths do not pose a problem for blanket plug technology (see Figure 7b).

Although selective tungsten is the simpler of the two tungsten plug formation schemes, it proved not to be manufacturable in the CMOS-4 development time frame. Our evaluations demonstrated a general inability to reproducibly generate low-resistance interconnections while simultaneously controlling selectivity loss. As a consequence, blanket tungsten plugs were chosen for use in the CMOS-4 process.

Blanket Tungsten Plug Requirements

Tungsten plugs and their processing have simple structural and electrical requirements. Structurally, the plugs must be free of voids and flush with the surrounding oxide surface after all undesired tungsten and adhesion layer residues have been etched from the top surface of the entire wafer. The void-free constraint ensures that potentially damaging process materials are not trapped in voids. The requirement that the plug be flush with its surrounding surface ensures good step coverage during the subsequent deposition of Al. The photomicrographs in Figure 8 illustrate these structural attributes. Electrically, the plug resistivity must be high enough to induce current spreading in the plug, but the interfacial resistance between the plug and other conducting materials must be low enough not to adversely affect circuit performance. In addition, the processing of the plugs must not induce device damage. The latter point refers to the results of early investigations of tungsten plugs, which indicated that device damage could occur as a result of the attack of underlying materials during the CVD process of tungsten.[11,12]

Finally, in addition to the structural and electrical objectives, the plug formation process had to be optimized from a cost perspective. Cost was of particular concern because the industry-wide blanket tungsten deposition method of choice at the time (low-pressure CVD) involved very low deposition rates and made very inefficient use of the expensive source gas, tungsten hexafluoride (WF_6).

Blanket Tungsten Plug Formation—Equipment and Process

An Applied Materials Precision 5000 tungsten system was used for tungsten plug processing. The Precision 5000 system performs both tungsten deposition and etch back without breaking vacuum. Processing involves first loading a wafer into

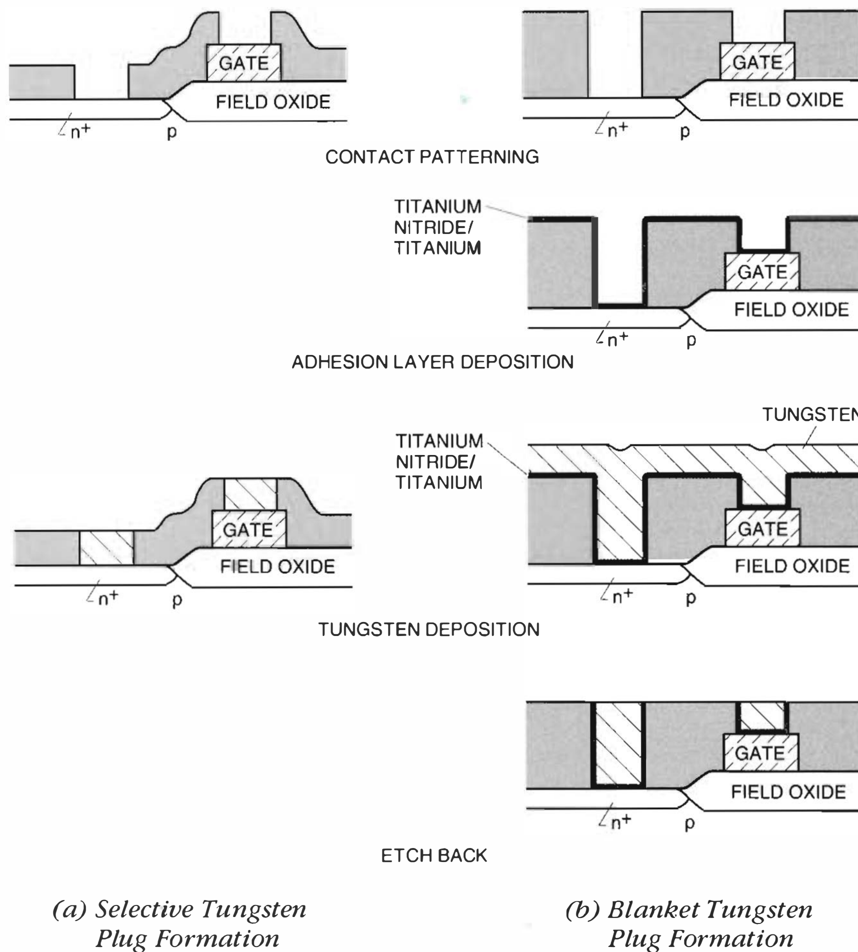
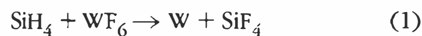


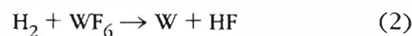
Figure 7 Comparison of Plug Technology

the deposition chamber. Then a nucleation layer approximately 50 nm thick is deposited at approximately 475 degrees Celsius by the silane reduction of WF_6:



$$SiH_4 + WF_6 \rightarrow W + SiF_4 \quad (1)$$

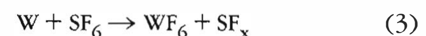
The bulk of the tungsten layer, approximately 800 nm, is then deposited using the hydrogen reduction of WF_6:



$$H_2 + WF_6 \rightarrow W + HF \quad (2)$$

Hydrogen reduction chemistry is used for the bulk of tungsten deposition because it yields good step coverage, whereas silane reduction does not.[13] However, silane reduction chemistry is used to initiate tungsten growth because hydrogen reduction chemistry involves an incubation period before film deposition begins on TiN[14], a step not required by silane reduction.[15]

Following tungsten deposition, the wafer is raised to expose its backside, and a short nitrogen trifluoride (NF_3) plasma etch is then performed in this same chamber to remove the small amount of tungsten that deposits on the backside edges of the wafer. The wafer is then transferred to a chamber in which a two-part etch back is performed. The bulk tungsten film is first etched in a sulfur hexafluoride/argon (SF_6/Ar) plasma according to equation (3):



$$W + SF_6 \rightarrow WF_6 + SF_x \quad (3)$$

until an optical emission from nitrogen, which has been liberated from the underlying adhesion layer, is detected. At this point, a chlorine/argon (Cl_2/Ar) plasma is used to remove any remaining adhesion layer according to equation (4):

$$TiN/Ti + Cl_2 \rightarrow TiCl_4 + N_2 \quad (4)$$

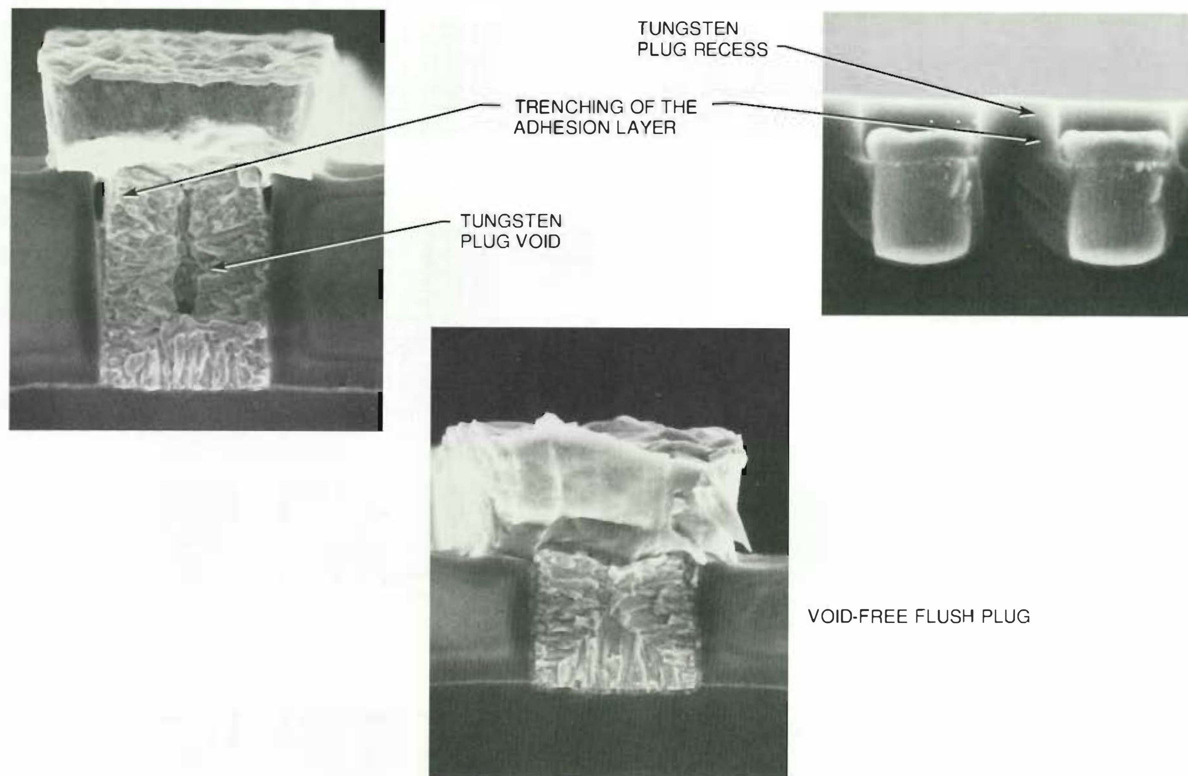


Figure 8 Photomicrographs Showing Several Possible Plug Structural Attributes

Both etching steps may employ a rotating magnetic field, which serves to improve the uniformity of the etching.

Blanket Tungsten Deposition

The more important properties associated with the tungsten deposition process include:

- Thickness uniformity
- Deposition rate
- Film resistivity
- Step coverage
- Film stress
- Surface smoothness
- Tungsten hexafluoride conversion

During development of the blanket tungsten deposition process, these properties were studied and "globally" optimized with respect to the plug objectives. A screening study was performed, followed by a response surface modeling (RSM) study. Table 2 shows the results of the screening study of these properties. Seven process factors were varied over the ranges given in Table 2. Four factors were identified to have an impact on the responses of interest (i.e., the properties listed above).

A detailed determination of the effect of these four factors was next performed in the context

Table 2 Process Factors and Ranges Used in the Blanket Tungsten Deposition Screening Study

Factor	Range
Spacing†	200–600 mils
Susceptor temperature†	400–475°C
Total pressure	60–80 torr
Partial pressure, H_2†	16–32 torr
Partial pressure, WF_6†	1–1.9 torr
Partial pressure, carrier gas	1.9–7.2 torr
Backside purge, N_2	300–700 sccm‡

Notes:
*Table 2 is adapted from data published in reference 16.
†Indicates factors found to have the greatest impact on responses of interest.
‡sccm = standard cubic centimeters per minute

of an RSM study. The process factors and ranges used for this study are given in Table 3. RSM studies produce mathematical models that relate factors and responses of interest. In this study, seven different models, one for each response, were generated. The models were then used to search for points of interest either computationally or graphically. An example of a set of contour plots used for a graphical search for tungsten step coverage and sheet-resistance uniformity (used to monitor thickness uniformity) is shown in Figure 9.

Table 3 Process Factors and Ranges Used in the Blanket Tungsten Deposition RSM Study

Factor	Range	Preferred Settings†
Spacing	200–600 mils	–400 mils
Susceptor temperature	430–490°C	475°C
Partial pressure, H_2	6–30 torr	–18 torr
Partial pressure, WF_6	1–2 torr	–1.75 torr

Notes:
*Table 3 is adapted from data published in reference 17.
†Also shown are the preferred deposition conditions that satisfy the optimization criteria shown in Table 4.

The specific criteria for the tungsten deposition optimization search are given in Table 4. The seven mathematical models/contour plots described above were used to find a region of the factor space where all of the optimization criteria could be met simultaneously. The preferred deposition conditions associated with this region are shown in Table 3. The models also indicated a relatively large process range within which the optimization objectives could be met. This range is important data for a production process.

Contour plots for deposition rate, resistivity, stress, WF_6 conversion, and reflectance (used to monitor surface smoothness) are not included in this paper. However, evaluation of these plots showed that the relevant optimization criteria could be met throughout the factor space studied. As a result, these responses did not impose constraints on the selection of the optimized process.

The criteria set for deposition rate and WF_6 conversion corresponded to values significantly higher (by a factor of approximately 3 to 10) than those typical of blanket tungsten processes at the time of our study.[15,16,17] The improvement resulted mainly from the use of higher deposition pressures compared to those previously used (80 torr versus less than 1 torr). Higher pressure also improved smoothness of the film, as seen in Figure 10. A smoother film is important because roughness on the tungsten film can be transferred into the underlying oxide during tungsten etch back.

The optimization criterion for film stress was set at a level corresponding to a mechanically stable film (i.e., one that would not peel spontaneously). Thus, although the stress values obtained are relatively high, they are below the critical level associated with delamination. For the tungsten resistivity, the observed values ranged from approximately 7.7 to 10.5 μohm per centimeter (cm) and were all acceptable.

Unlike the other optimization criteria, those for step coverage and sheet-resistance uniformity were not met throughout the factor space studied (see Figure 9). Tungsten step coverage can directly impact void formation, and tungsten thickness uniformity can impact plug recess control. Figure 11 illustrates how thickness variation across a wafer can lead to variations in plug recess following etch back. Because of the importance attached to meeting these two optimization criteria, the allowed process window was diminished in size. Figure 9 shows that a step coverage greater than or equal to 95 percent restricted the WF_6 partial pressure to approximately greater than or equal to 1.5 torr, hydrogen (H_2) partial pressure to approximately less than or equal to 18 torr, and gas-inlet-to-wafer spacing to less than or equal to 400 mils. A sheet-resistance uniformity less than or equal to 3 percent served to further restrict the spacing to values between approximately 400 and 300 mils.

In addition to the tungsten deposition process properties mentioned, the tungsten thickness had to be optimized. The upper limit on thickness was influenced by cost considerations and by the fact that a thinner tungsten deposit has less likelihood of being trapped in dielectric troughs. A thinner tungsten deposit also requires less overetch to remove tungsten spacers that may form on the trough sidewalls. As shown in Figure 12, the size of the spacer formed on a nonplanar dielectric for zero overetch with a nonisotropic etch depends on the absolute magnitude of the deposit thickness, T_d, and the worst-case dielectric sidewall slope, α, which together determine the local tungsten thickness range, $T_n - T_d$.

The lower limit on thickness was influenced by the need to fill contact openings with tungsten

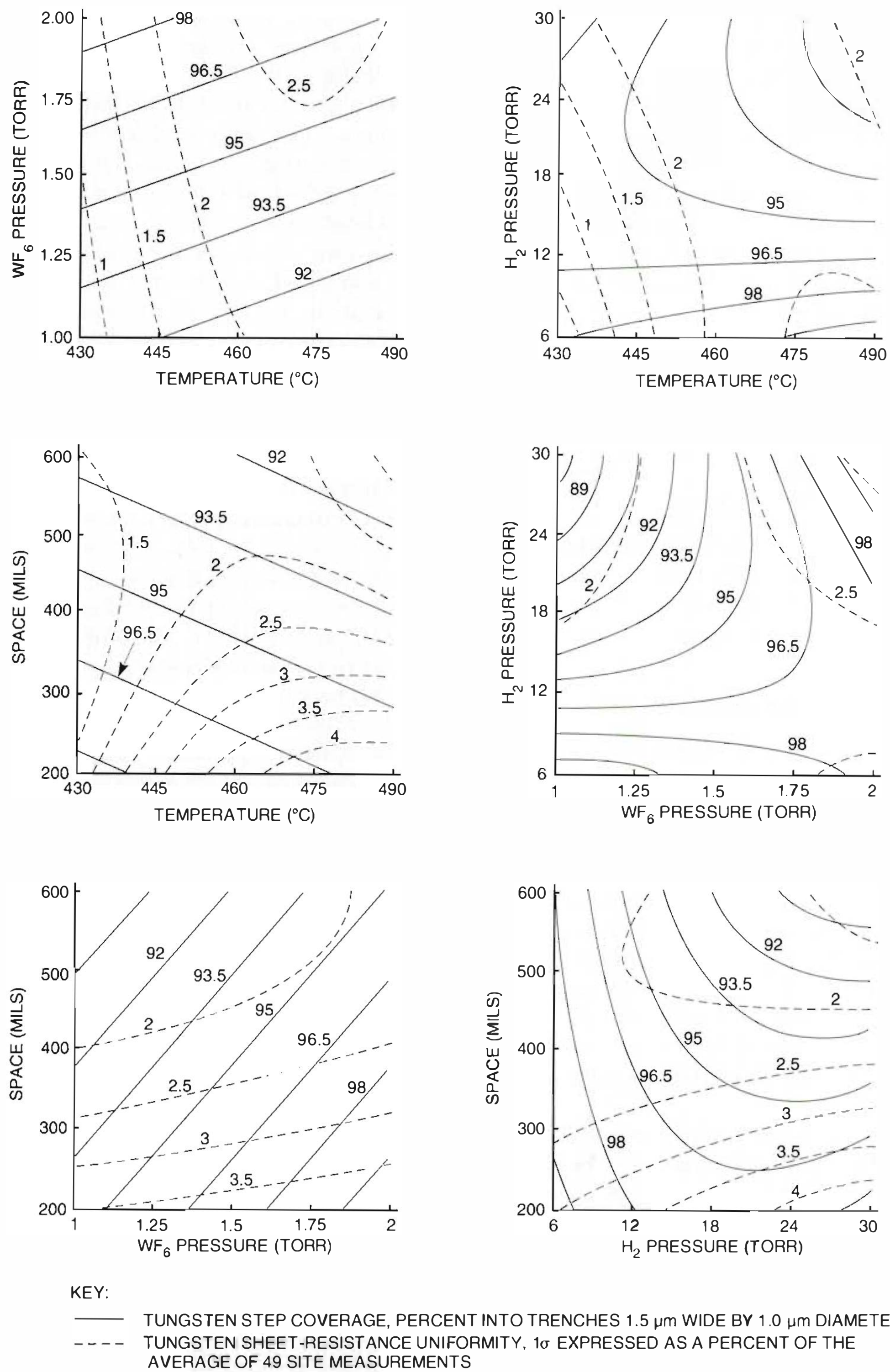


Figure 9 Contour Plots of Tungsten Step Coverage and Sheet-resistance Uniformity (Reprinted from the Journal of Vacuum Science Technology; *see reference 17.)*

Table 4 Tungsten Deposition-related Optimization Criteria for a Blanket Tungsten Plug Application

Parameter	Optimization Criteria
Growth rate	$\geq$300 nm per min
Resistivity	$\geq$ ~8 μohm-cm
Sheet-resistance uniformity	$\leq$3% (σ)
Tensile stress	$\leq 16 \times 10^9$ dyne per cm^2
Step coverage†	$\geq$95%
WF_6 conversion	$\geq$12%
Reflectance	$\geq$25% vs Si @436 nm

Notes:
*Table 4 is reprinted from the *Journal of Vacuum Science Technology;* see reference 17.
†Step coverage for a trench 1.5 μm deep and 1.0 μm wide.

prior to etch back, and further, to planarize the tungsten over the opening to minimize plug recess following etch back, as shown in Figure 13. Tungsten thickness studies showed that a tungsten film greater than approximately 650 nm is required for the submicron-level contacts and vias of CMOS-4 technology.

Tungsten Etch Back

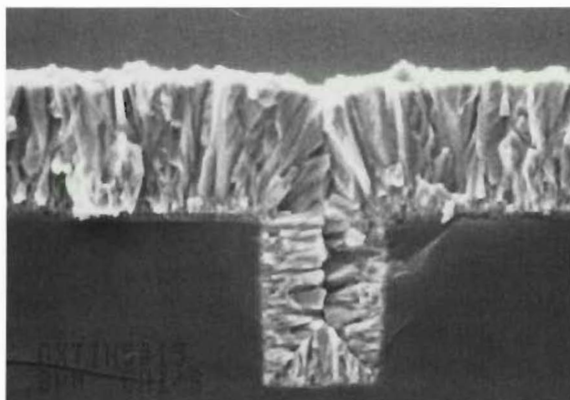
In this section, the tungsten and adhesion layer etch-back processes are discussed.

Bulk Tungsten Etch Chemistry (SF_6/Ar) The more important etch-back process properties for bulk tungsten etch include:

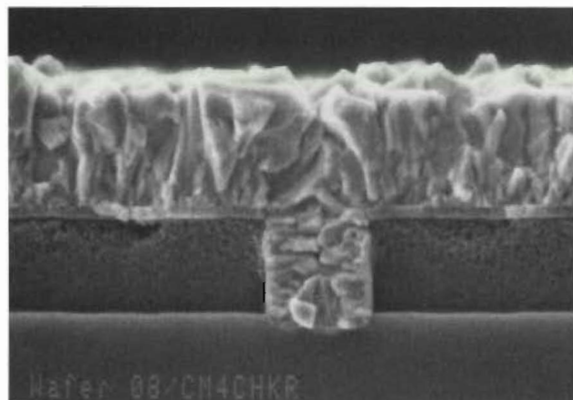
- Tungsten etch rate
- Tungsten etch-rate uniformity
- Tungsten/titanium nitride etch-rate ratio
- Isotropy (lateral etch rate/vertical etch rate)
- Microloading (tungsten plug/tungsten bulk etch-rate ratio)

For bulk tungsten etch, the tungsten etch rate impacts throughput and cost. The tungsten etch-rate uniformity, tungsten/titanium nitride etch-rate ratio, etch isotropy (lateral etch rate versus vertical etch rate), and microloading (tungsten plug etch rate/bulk tungsten etch rate) all impact the ability to produce a residue-free surface, while simultaneously maintaining flush plugs. A nonuniform etch rate affects plug recess because it leads to the clearing of one region of the wafer before the rest. This first-to-clear region becomes the site of the worst-case plug recess on a wafer. Figure 14 shows the plug recess that occurs in high etch-rate regions on a wafer when a tungsten film of uniform thickness has been etched to the proper end point for the lowest etch-rate regions.

The tungsten etch isotropy affects plug recess by increasing the etch time required to clear the final residues on a nonplanarized dielectric. Figure 15 illustrates that an etch isotropy that equals 1 (i.e., lateral etch rate equals vertical etch rate) can lead to uniform clearing of tungsten on nonplanar surfaces, and that an etch isotropy less than 1 tends to produce spacers that must be removed by overetching. The additional time required to clear tungsten spacer residues leads to increased plug

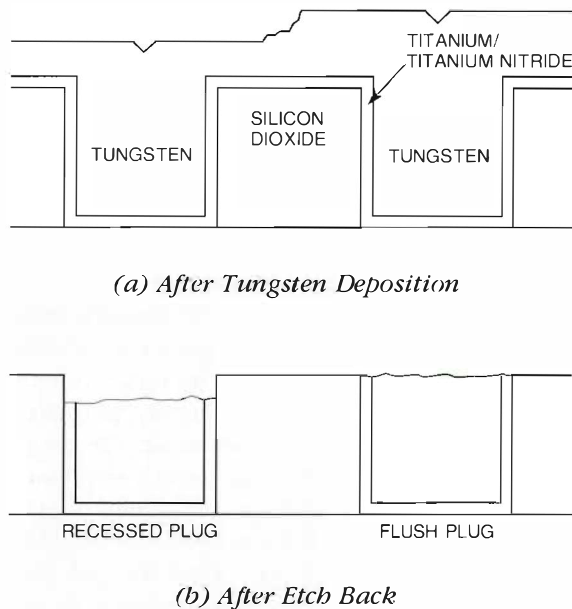


(a) Deposited at 80 Torr



(b) Deposited at Less Than 1 Torr

Figure 10 Photomicrographs Showing Surface Smoothness of Tungsten Films



(a) After Tungsten Deposition

(b) After Etch Back

Figure 11 Plug Recess as a Function of Thickness Variation

recess. Finally, a high tungsten/titanium nitride etch-rate ratio affects plug recess control indirectly by preventing the liberation of oxygen from the underlying dielectric. Oxygen liberation has been shown to greatly increase the tungsten micro-loading factor.[18]

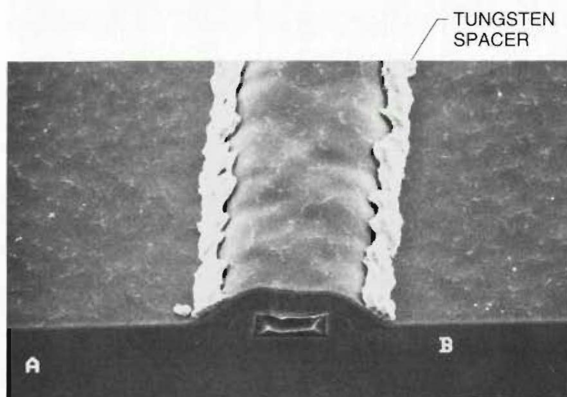
Development of the bulk tungsten etch in SF_6/Ar began with an RSM study of the tungsten etch rate and etch-rate uniformity.[19] Table 5 shows the factors and ranges used in the initial study, along with the preferred etch conditions identified. Attempts to improve the tungsten/titanium nitride etch-rate ratio in SF_6/Ar using only the factors in Table 5, while simultaneously maintaining high tungsten etch rates (greater than 500 nm per minute), were largely unsuccessful.[20] The desired improvement was eventually obtained through the incorporation of active wafer temperature control. Controlling the wafer temperature between 20 and 40 degrees Celsius improved the tungsten/titanium nitride etch-rate ratios from approximately 2:1 to between 10:1 and 50:1, respectively.[20]

Adhesion Layer Etch Chemistry (Cl_2/Ar) The more important etch-back process properties for adhesion layer etch include:

- TiN etch rate
- TiN etch-rate uniformity
- TiN/Tungsten (W) etch-rate ratio
- TiN/oxide etch-rate ratio

For adhesion layer etch chemistry in Cl_2/Ar, the titanium nitride etch rate impacts throughput and cost, while etch-rate uniformity affects the level of adhesion layer undercutting or trenching that occurs around a plug (see Figure 8c). The other two etch properties impact plug recess and oxide loss. Consequently, a process is derived that etches TiN at a high and uniform rate while it simultaneously and slowly etches tungsten and silicon oxide.

Process development for the adhesion layer etch in Cl_2/Ar began with an RSM study of the TiN etch rate and etch-rate uniformity.[19] Table 5 shows the factors and ranges used in the initial study, along with the preferred etch conditions identified.



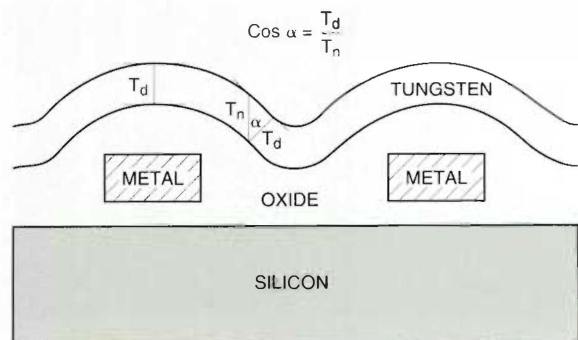


Figure 12 Photomicrograph and Schematic Drawing of a Tungsten Spacer

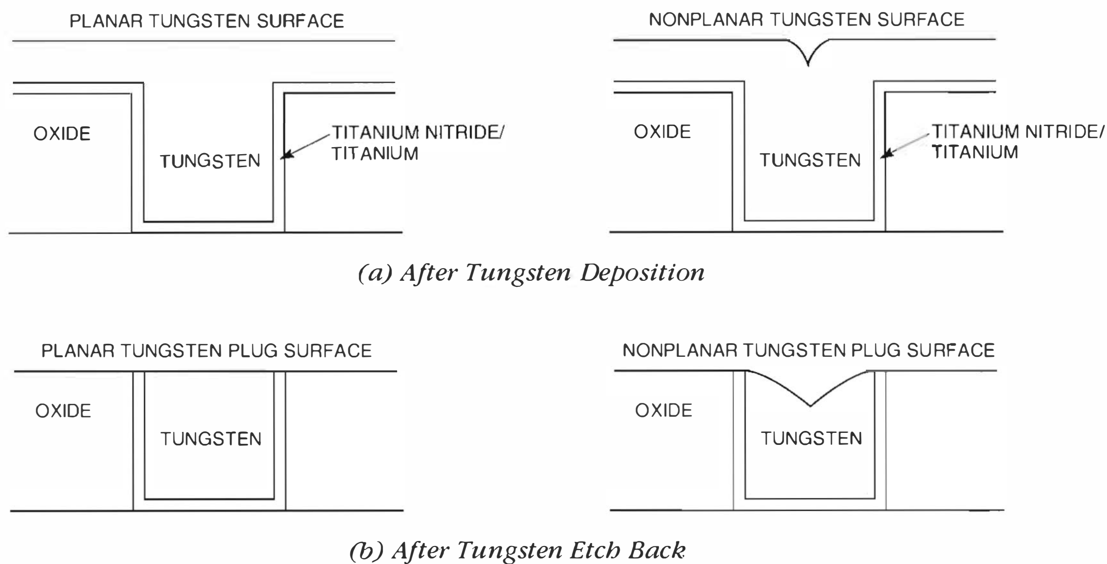


Figure 13 Nonplanarity in a Blanket Tungsten Deposit Transferred into Plug Recess (Reprinted from the Journal of Vacuum Science Technology; *see reference 17.)*

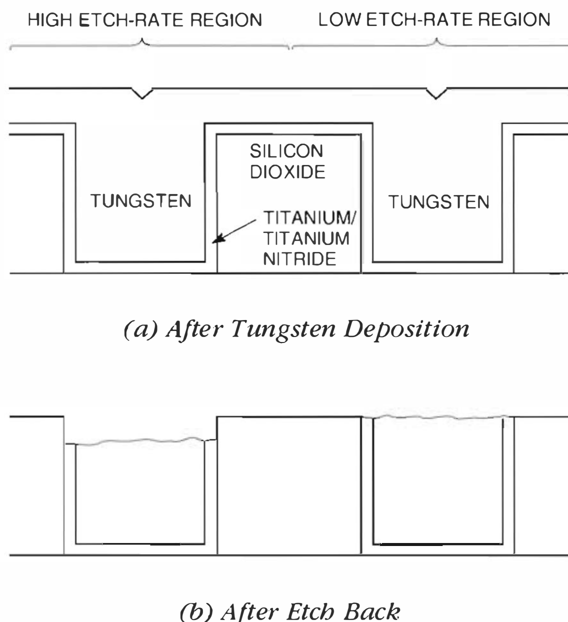


Figure 14 Plug Recess in High Etch-rate Regions on a Wafer

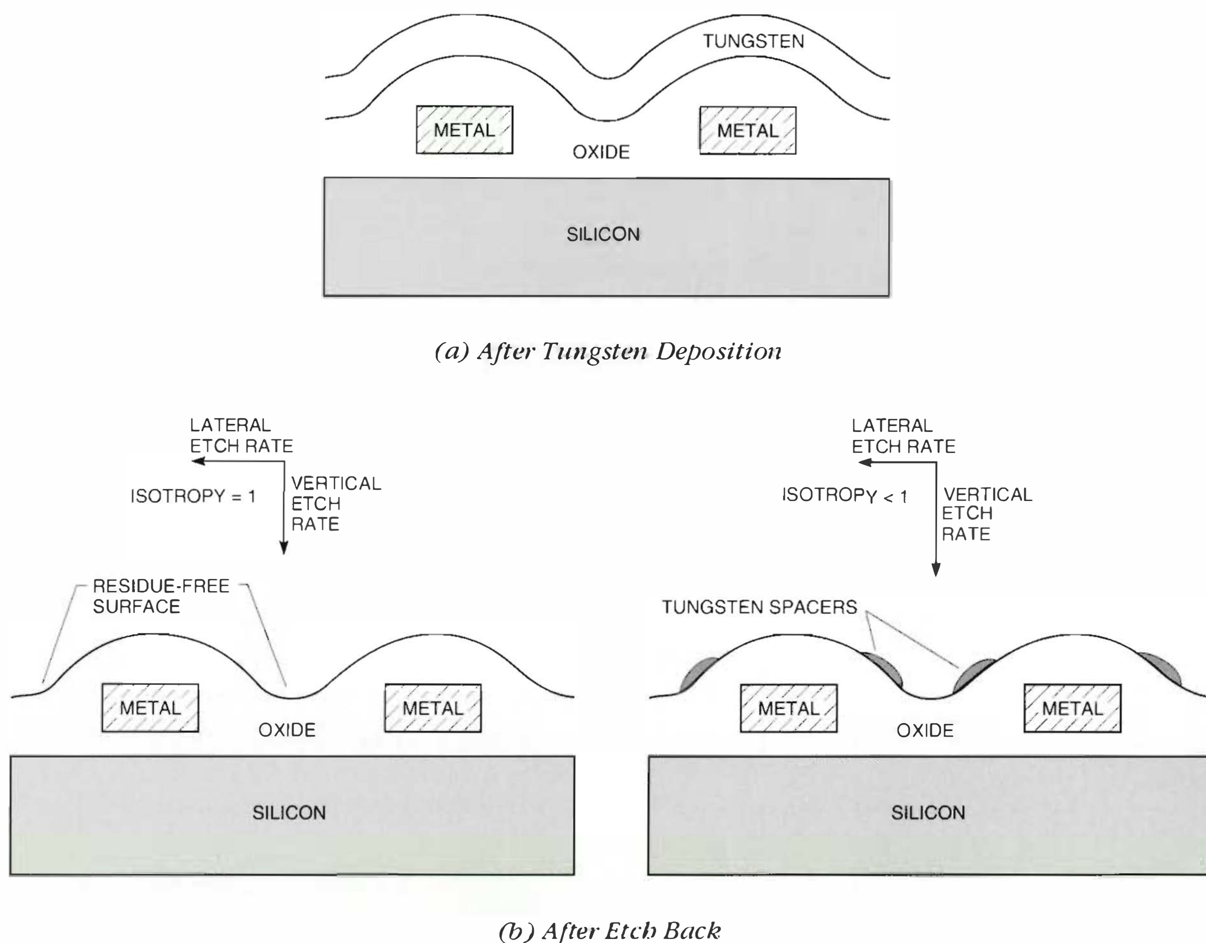
Acceptable TiN etch rates (approximately 145 nm per minute) and etch-rate uniformity (less than or equal to ±5 percent, 1 σ) were achieved. Since the adhesion layer is only approximately 120 nm thick, lower etch rates and higher etch-rate nonuniformities than those for the tungsten etch step can be tolerated. Wafer temperature control provides additional latitude against plug sidewall trenching.

Integration of Tungsten Plugs into CMOS-4 Technology

After the tungsten deposition and etch-back processes were developed, the overall plug formation process was integrated into the CMOS-4 technology. To ensure that electrical requirements were met and that adequate process latitude existed, the following factors were considered in the integration studies:

- Dielectric planarization
- $CoSi_2$ thickness
- Contact depth
- Contact overetch
- Adhesion layer
  - Sputter etch preclean
  - Deposition temperature
  - Substrate bias
  - Thickness
  - Material (TiN or TiN/Ti)

After the integration studies were completed, a relatively robust process was developed. At that time, it was determined that acceptable electrical



(a) After Tungsten Deposition

(b) After Etch Back

Figure 15 Comparison of Results of Tungsten Etch Isotropies

results can be obtained over relatively large process ranges.

Dielectric Process Development

The CMOS-4 process presented two specific technological challenges requiring dielectric development efforts. First, horizontal scaling without reducing metal thicknesses resulted in high aspect ratio spaces. The existing dielectric technology could not fill these spaces void-free. Second, the introduction of blanket tungsten plugs required the development of a dielectric planarization process.

In addition to meeting gap fill and planarization requirements, a dielectric film must also meet a number of electrical, mechanical, and deposition requirements. The required electrical characteristics for a dielectric film include low current leakage, high breakdown voltage, high electrical resistance, low dielectric constant, low mobile ion, and heavy metal concentration. Mechanical requirements include low moisture adsorption, low stress for cracking resistance, low particulate levels, and a low pinhole density. Important deposition requirements include high deposition rate, good uniformity, and low deposition temperature (not greater than 450 degrees Celsius) for prevention of metal hillock formation.

Gap Filling

The conventional silane-based oxides could not meet the CMOS-4 technology gap fill requirements due to poor conformality characteristics. A silane oxide profile is typically described as a "breadloaf," that is, the film is thicker on the top of a structure, but thinner on the sides and bottom, and forms cusps along the sidewall.[21] Silane oxides nucleate in the gas phase, which causes this reentrant type profile and limits the aspect ratio (height/width)

Table 5 Process Factors and Ranges for the Initial Tungsten Etch-back RSM Study and the Preferred Etch Conditions Found

Factor	Range	Preferred Settings
Tungsten Etch		
Total pressure	20–250 millitorr	85 millitorr
SF_6 flow rate	20–60 sccm†	60 sccm
Ar flow rate	15–60 sccm	60 sccm
RF power	200–500 watt	475 watt
Magnetic field	0–100 gauss	20 gauss
Electrode temperature	60°C	60°C
TiN Etch		
Total pressure	50–250 millitorr	85 millitorr
Cl_2 flow rate	5–35 sccm	10 sccm
Ar flow rate	50–120 sccm	115 sccm
RF power	150 watt	150 watt
Magnetic field	0–100 gauss	75 gauss
Electrode temperature	60°C	60°C

Notes:
*Table 5 is reprinted with permission from the IHS Publishing Group; see reference 19.
†sccm = standard cubic centimeters per minute

that can be filled without forming a void to approximately 0.5.[22]

To fill the aspect ratio of approximately 1.2 (0.9-μm height/0.75-μm width) for the CMOS-4 process, a tetraethylorthosilicate (TEOS)-based oxide was used. The conformality of TEOS oxides is much better than that of silane oxides. Because the organosilicon compounds produced during a TEOS-based CVD process have a significantly higher surface mobility, the reactive molecules diffuse on the surface before reacting, which results in better step coverage without cusps.[23] TEOS-based oxides can fill aspect ratios up to 1.0 void-free. When used in conjunction with profile-altering deposition/etch-back techniques, as in CMOS-4 technology, aspect ratios up to 1.8 can be filled.[22,24,25]

TEOS oxides are well suited for use as interlevel dielectrics. Oxides formed through the plasma dissociation of oxygen (O_2) in the presence of TEOS (PE-TEOS) are denser than silane oxides and therefore more resilient to moisture adsorption. PE-TEOS films are typically under low compressive stress, which results in higher cracking resistance than the tensile-stressed silane films. Due to lower deposition temperatures, TEOS oxides exhibit thermal stability and less hillock formation. Mobile ion and heavy metal concentrations are lower with TEOS oxides, and device testing indicates lower defect densities.[24,26]

In the CMOS-4 process, filling gaps between minimum-spaced metal lines was achieved with profile-altering techniques in conjunction with a PE-TEOS bulk dielectric. Wafers were moved back and forth between a series of deposition and etch-back steps in a load-locked, multichamber cluster tool. Deposition of a planarized dielectric was completed in one cassette-to-cassette operation.

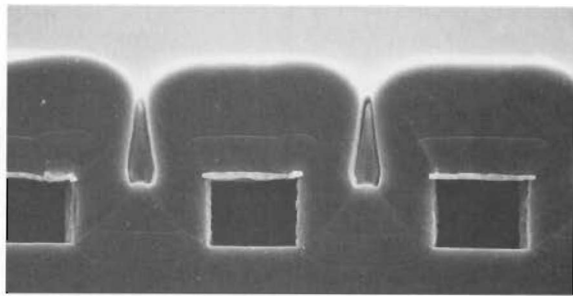
Gap Fill Process Flow

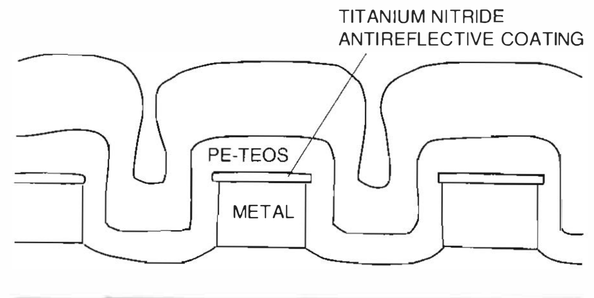
The gap fill process is shown in Figure 16. It begins with a PE-TEOS conformal deposition that is halted prior to reaching a thickness that would fill the smallest gaps. Next, an argon sputter etch is performed on the oxide film deposited in the previous step. The oxide removal rate is direction dependent, with the maximum removal occurring 45 degrees from vertical. The original 90-degree corners are beveled into positively tapered angles. Then an ozone-TEOS film is deposited to completely fill the remaining gaps. Ozone-TEOS is formed by a thermal reaction in which oxygen atoms are produced by the rapid decomposition of ozone. Ozone-TEOS film is not desirable as a bulk dielectric due to its characteristically low density and tensile stress. However, the superior step-coverage characteristics of ozone-TEOS allow it to fill very small gaps.

Following ozone-TEOS deposition, a second profile-altering etch back is performed. The ozone-TEOS is a sacrificial film that is removed in a CHF_3 chemistry until it remains only in the small gaps and as a spacer along the sidewalls of larger features. The combined effect of sputter etching and ozone-TEOS processing is a void-free surface with positively sloped sidewalls. Finally, a bulk PE-TEOS dielectric film is deposited over this smoothed surface to reach the desired dielectric film thickness.

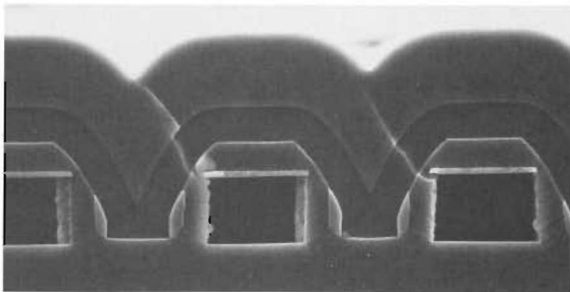
Planarization

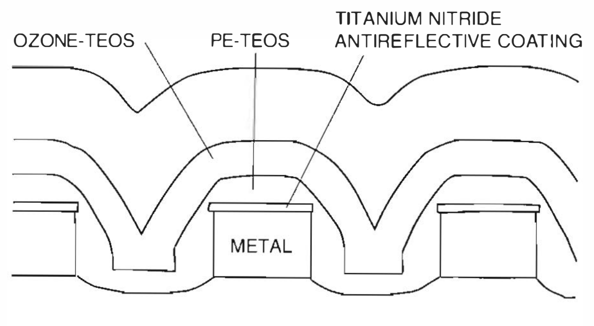
Planarization is a dielectric smoothing process that is performed to smooth or reduce the steps created by underlying interconnect features. A planarization process minimizes reflective notching, reduces the extent of metal overetch, increases the thickness of metal over underlying topography, and reduces interconnect defect densities. Enhanced planarization is required to successfully form tungsten plugs with a blanket etch-back process. The



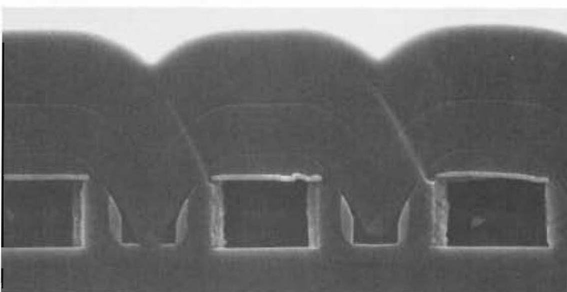


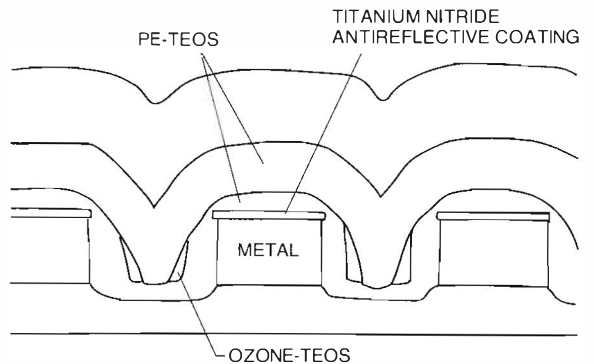
(a) Initial PE-TEOS Deposition





(b) Argon Sputter Etch and Ozone-TEOS Deposition





(c) Anisotropic Etch and PE-TEOS Cap

Figure 16 Schematic Drawing of and Photomicrograph of Gap-filling PE-TEOS Dielectric Deposition Process Stages

slope of the surface must be less than 30 degrees from the horizontal to ensure removal of tungsten stringers.

Two planarization techniques were considered for use in the CMOS-4 process. The first technique involved depositing a sacrificial planarizing boron oxide film and etching the planarized pattern into the oxide. The second technique was similar, except a spin-on-glass (SOG) process was used as the planarizing agent.

Initially, the boron oxide planarization was chosen for use in the CMOS-4 process. This technique was selected based on a perceived manufacturability benefit over the SOG process. The boron oxide planarization can be done in a single cluster tool that uses the same serial process as the gap fill process. One cassette-to-cassette operation can both fill the minimum gap and planarize the surface. On the other hand, SOG planarization requires two separate depositions, a spin coat, a cure, and an etch operation.

Boron Oxide Planarization

Boron oxide is a CVD film deposited by the plasma decomposition of trimethylborate (TMB) in the presence of O_2. The film has a low melting point and flows at deposition temperatures as low as 400 degrees Celsius. Boron oxide can be etched back in a CHF_3 plasma chemistry and a 1:1 boron oxide:oxide selectivity. Spaces up to 25 μm can be fully planarized with boron oxide.[27]

Boron Oxide Process Flow

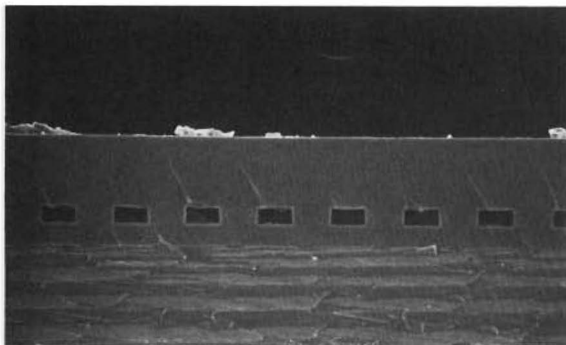
Upon completion of the gap fill process flow, the final bulk dielectric deposition step is targeted at approximately three times the desired final thickness. This overdeposition provides some initial smoothing of the underlying interconnect features and also fills spaces in the 2- to 5-μm range to eliminate formation of unwanted gaps. Next, two sequential boron oxide deposition and etch-back steps are performed. The boron oxide film flows as deposited, thereby smoothing and planarizing the topography. The isotropic etch back transfers the planarized surface into the underlying oxide. The photomicrographs in Figure 17 illustrate boron oxide planarization. The boron oxide film is a sacrificial film that is completely removed during this step. The deposition/etch sequence is repeated to further improve planarity. After all boron oxide is removed, the etch chemistry is switched to a higher-rate carbon tetrafluoride (CF_4) anisotropic etch process that removes the thick bulk deposition to the final desired thickness.

The boron oxide process is used in volume production for the dielectric between polysilicon and metal 1. For the metal 1 to metal 2 dielectric, the boron oxide process could not meet the wafer volume and uniformity requirements due to problems with equipment reliability, thickness variability, and low throughput. The SOG planarization scheme was selected as a more cost-effective process.

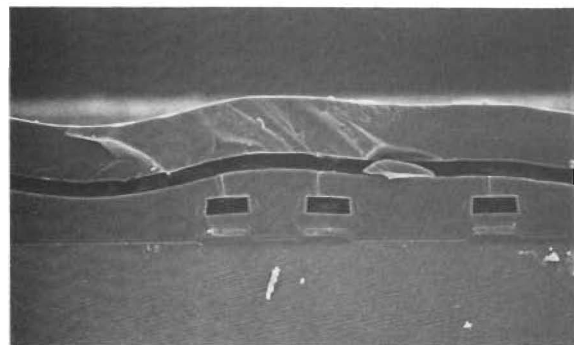
Spin-on-glass Planarization

The SOG process consists of a series of simple single-step operations. The overall integrated process characteristics exhibit good throughput and process control. The photomicrographs in Figure 18 illustrate SOG etch-back planarization.

SOG is a low-viscosity liquid polymer that is applied to the wafer using a spin-coating process similar to that used for photoresists. The SOG material used for CMOS-4 technology is a siloxane (methyl group containing polymer). Siloxane SOGs exhibit improved cracking resistance and lower dielectric

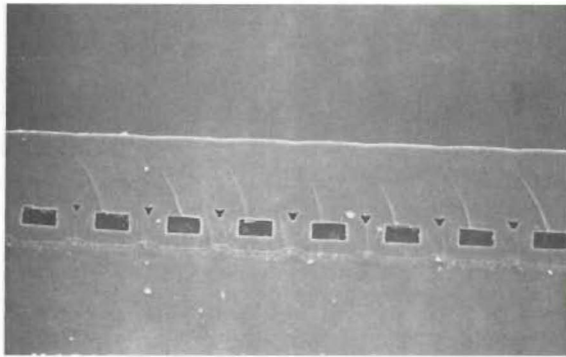


(a) Minimum-spaced Metal Lines

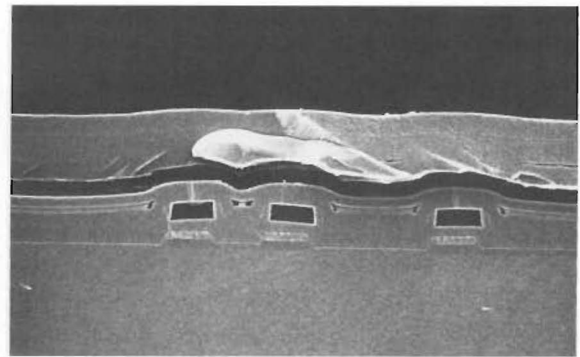


(b) Wide-spaced and Isolated Metal Lines

Figure 17 Photomicrographs Showing Boron Oxide Planarization



(a) Minimum-spaced Metal Lines



(b) Wide-spaced and Isolated Metal Lines

Figure 18 Photomicrographs Showing SOG Etch-back Planarization

constants as compared to other available SOG materials. Because it is a liquid, SOG can fill very small gaps and planarize topographical surfaces. The material is solidified into a glass by curing in a low-temperature furnace cycle. The cured SOG film has SiO_2-type mechanical and electrical properties, and can therefore be left behind as part of the bulk dielectric. Typically, a partial etch back of the material is performed that leaves SOG only in the gap areas.[28,29]

SOG Process Flow

The initial gap fill deposition is deposited thick enough to provide a buffer for the SOG etch-back overetch. SOG is then spun on to fill the larger spaces and planarize the surface. A low-temperature furnace cure is performed to remove the solvents from the SOG and transform the material from a liquid into a glass.

A partial etch back of the SOG is performed using a $CHF_3/CF_4/O_2$ chemistry. The selectivity of SOG to the underlying PE-TEOS is targeted at 1:1 and is controlled by the ratio of the CHF_3 and O_2 gas flows. Since the O_2 flow also affects the etch uniformity, trade-offs between selectivity and uniformity were necessary. SOG is etched completely from the tops of interconnect lines, but remains in the gaps of the larger-spaced lines. The etch back is targeted to remove SOG from locations at which contacts will be formed. Exposing SOG along the sidewalls of contacts can lead to problems with via "poisoning" and poor contact profiles. Finally, a PE-TEOS layer is deposited to achieve the desired dielectric thickness and encapsulate the SOG.

Summary

Digital's CMOS-4 on-chip interconnect technology is a three-level aluminum alloy metallization process, with planarized TEOS-based silicon dioxide dielectrics, tungsten-filled contacts and vias, and a minimum feature size of 0.75 μm. The process development goals required the maximum use of the existing manufacturing capability and the introduction of new process features. For photolithography, plasma etch, and PVD metallization, the 1.0-μm manufacturing equipment set and processes were modified and reoptimized for the submicron regime. In addition, two new process features, a blanket CVD tungsten process and a TEOS-based oxide planarization process, were developed and implemented in manufacturing to meet the CMOS-4 technology requirements.

References and Note

1. C. Kaanta et al., "Submicron Wiring Technology with Tungsten and Planarization," *Proceedings of the Fifth International IEEE VLSI Multilevel Interconnection Conference* (1988): 21-28.

2. A. Enver and J. Clement, "Finite Element Numerical Modeling of Currents in VLSI Interconnects," *Proceedings of the Seventh International IEEE VLSI Multilevel Interconnection Conference* (1990): 149-156.

3. P. Chien and M. Chen, "Proximity Effects in Submicron Optical Lithography," *SPIE Proceedings*, vol. 772 (March 1987): 35-40.

4. A. van Roosmalen, "Review: Dry Etching of Silicon Oxide," *Vacuum*, vol. 34, no. 3-4 (1984): 429-436.

5. J. Chang, "Selective Reactive Ion Etching of Silicon Dioxide," *Solid State Technology*, vol. 27, no. 4 (April 1984): 214-219.

6. C. Dobson, "Polymer Formation During Silicon Oxide Plasma Etching," *Microelectronic Manufacturing and Testing*, (October 1981): 18-20.

7. S. Roth, W. Ray, and G. Wissen, "In situ, Tapered Contact Etch," *Semiconductor International*, vol. 11, no. 6 (May 1988): 138-142.

8. Figure 5 is courtesy of D. Hardie and C. Jones from Digital's South Queensferry, Scotland facility.

9. A. Berti, "Characterizing a Titanium Nitride Reactive Deposition Process Using Design of Experiments," *Advanced Semiconductor Manufacturing Conference and Workshop* (1991): 179-184.

10. A. Aronson and I. Wagner, "Advanced Aluminum Metallization, Part II Planarization," *Transactions of the 47th and 48th Schools on Thin Film Technology: Advances in Magnetron Sputtering* (1990): DPS-VII-4.

11. G. Georgiou et al., "Influence of Selective Tungsten Deposition on Shallow Junction Leakage," in *Tungsten and Other Refractory Metals for VLSI Applications II*, edited by E. Broadbent (Pittsburgh, PA: Materials Research Society, 1987): 227-234.

12. C. Yang, J. Multani, D. Paine, and J. Bravman, "Junction Leakage of Selectively Deposited LPCVD Tungsten for Contact Fill Applications," *Proceedings of the Fourth International IEEE VLSI Multilevel Interconnection Conference* (1987): 200-207.

13. C. McConica, S. Chatterjee, and S. Sivaram, "Step Coverage Prediction during Blanket LPCVD Tungsten Deposition from Hydrogen, Silane and Tungsten Hexafluoride," *Proceedings of the Fifth International IEEE VLSI Multi-level Interconnection Conference* (1988): 268-276.

14. V. Rana, J. Taylor, L. Holschwandner, and N. Tai, "Thin Layers of TiN and Al as Glue Layers for Tungsten Deposition," in *Tungsten and Other Refractory Metals for VLSI Applications II*, edited by E. Broadbent (Pittsburgh, PA: Materials Research Society, 1987): 187-195.

15. P. Riley, T. Clark, E. Gleason, and M. Garver, "Implementation of Tungsten Metallization in Multilevel Interconnection Technologies," *IEEE Transactions on Semiconductor Manufacturing*, vol. 3 (November 1990): 150-157.

16. T. Clark, A. Constant, M. Chang, and C. Leung, *CVD Tungsten, Copper, and Other Refractory Metals for ULSI/VLSI Applications V*, edited by S. Wong (Materials Research Society, Pittsburgh, PA, 1990): 167-178.

17. T. Clark, M. Chang, and C. Leung, "Response Surface Modeling of High Pressure Chemical Vapor Deposited Blanket Tungsten," *Journal of Vacuum Science Technology B*, vol. 9 (May/June 1991): 1478-1486.

18. J. van Laarhoven, H. van Houtum, and L. de Bruin, "A Novel Blanket Tungsten Etchback Scheme," *Proceedings of the Sixth International IEEE VLSI Multilevel Interconnection Conference* (1989): 129-135.

19. T. Clark and P. Riley, "Multi-Chamber, Single-Wafer Integrated Tungsten CVD and Plasma Etchback," *Microelectronic Manufacturing and Testing*, vol. 13 (November 1990): 29-31.

20. K. Koller, H. Erb, and H. Korner, "Tungsten Plug Formation by an Optimized Tungsten Etch Back Process in Non Fully Planarized Topography," *Applied Surface Science*, vol. 53 (1991): 54-61.

21. C. Magnella, T. Ingwersen, and E. Fleck, "A Comparison of Planarization Properties of TEOS and SiH_4 PECVD Oxides," *Proceedings of the Fifth International IEEE VLSI Multilevel Interconnection Conference* (1988): 366-373.

22. D. Wang, K. Law, and J. Marks, "Planarized Dielectrics for 0.5 μm Applications," *ACET in Review* (Fall 1990): 11.

23. R. Levin and K. Evans-Lutterodt, "Step Coverage of Undoped and Phosphorus-Doped SiO_2,"

Journal of Vacuum Science and Technology, vol. B1, no. 1 (1983): 54-61.

24. J. Perchard et al., "Characterization of a Multiple-Step In-situ PECVD-TEOS Planarization Scheme for Submicron Manufacturing," Multichamber and In-Situ Processing of Electronic Materials, *SPIE Proceedings,* vol. 1188 (1989): 75-85.

25. G. Schwartz and P. Johns, "Gap-fill with PECVD SiO_2 Using Deposition/Sputter Etch Cycles," *Journal of the Electrochemical Society,* vol. 139, no. 3 (1992): 927.

26. F. Becker et al., "Low-pressure Deposition of High-quality SiO_2 Films by Pyrolysis of Tetraethylorthosilicate," *Journal of Vacuum Science Technology,* vol. B5 (1987): 1555-1563.

27. J. Marks, K. Law, and D. Wang, "In-Situ Planarization of Dielectric Surfaces Using Boron Oxide," *Proceedings of the Sixth International IEEE VLSI Multilevel Interconnection Conference* (1989): 89-95.

28. C. Chiang and D. Fraser, "Understanding of Spin-On-Glass (SOG) Properties from their Molecular Structure," *Proceedings of the Sixth International IEEE VLSI Multilevel Interconnection Conference* (1989): 397-403.

29. S. Gupta, "Spin-on Glass for Dielectric Planarization," *Microelectronic Manufacturing and Testing,* vol. 12, no. 5 (April 1989): 10-12.

Mary Beth Nasr
Ellen J. Mager

Implementation of Defect Reduction Strategies into VLSI Manufacturing

CMOS-4 technology combines a high-performance microprocessor with a fast, dense RAM. Consistently obtaining a specified die yield on CMOS-4 devices required the implementation of a series of defect reduction procedures. To achieve high yields, microcontamination and defect reduction plans needed to be in place well before initiation of product manufacturing. Levels of overall cleanliness had to be specified and controlled. Process equipment was monitored at the new particle level of 0.375 μm and greater to collect data. Defect density test reticles were designed and wafers were processed. Electrical results were then incorporated into a yield model and used to prioritize yield enhancement activities. Experiments were designed to reduce the defect levels of process areas, such as p-gate leakage and metal 2 short circuits.

Fourth-generation complementary metal-oxide semiconductor (CMOS-4) technology calls for a die area greater than 2 square centimeters (cm^2), geometries of 0.75 micron (μm), a gate oxide of 10.5 nanometers (nm), unique metallurgy, 1.7 million transistors, and 23 masking levels. These very large-scale integration (VLSI) process features required for chip performance dictate the need for increased defect reduction and microcontamination control in the semiconductor production environment.

Production of one fully functional CMOS-4 device is virtually impossible without substantial efforts in defect reduction and microcontamination control. Obviously, a single 0.75-μm particle in the active area of a die can create a short circuit and cause the entire device to fail. A 10.5-nm particle at gate oxide can have the same effect. A high level of sodium in rinse water can lead to premature gate oxide breakdown. In fact, there are approximately 250 processing steps that could contribute to the failure of a chip.

This paper describes the principles of microcontamination control and relates their application to VLSI manufacturing. It next discusses improvements that we implemented for wafer handling, cleaning, and monitoring. It then outlines the overall defect reduction techniques to increase product yield, focusing on efforts in the areas of p-gate leakage and metal 2 short circuits. The paper concludes with considerations for defect reduction in the next generation of CMOS technology.

Application of Microcontamination Control Principles

Implementation of a successful set of defect reduction procedures depends on understanding the principles of microcontamination control. This field of study encompasses a wide variety of areas. Microcontamination control seeks to minimize the presence of any substance, particle, monolayer, or ionic contaminant in the wafer production environment, that could cause a device to fail.

The facility in which CMOS-4 devices are manufactured was constructed for the production of the 1.0-μm CMOS-3 technology. However, because of the high capital costs associated with clean room construction, the facility was originally designed to meet the fabrication needs for three generations of CMOS technology: CMOS-3, CMOS-4, and CMOS-5. The original design specifications for the air quality of the clean room, the clean room suits, and the water, chemicals, and gases used in semiconductor manufacture are discussed in the following sections.

Clean Room Air

The ambient air quality is carefully controlled in a clean room. Typically, indoor ambient air contains more than one billion particles per cubic foot.[1] To

obtain a high yield on CMOS-4 devices, wafers must be processed in a well-characterized clean room.

The level of cleanliness within a clean room is defined by the clean room class, which is determined by the number of particles greater than 0.5 μm per cubic foot. The CMOS-4 devices are manufactured in a class 10 environment, i.e., in a clean room that has fewer than ten 0.5-μm particles per cubic foot. To obtain the class figure, air is measured with laser-scattering techniques when the clean room facility is at rest. The air filtration procedure employs high-efficiency particulate air (HEPA) filters with filter efficiencies greater than 99.9999 percent. Since HEPA technology is well advanced and understood, most of the particles in clean rooms are generated when process equipment and personnel are introduced. Airflow is maintained at vertical laminar to prevent particle deposition on the product wafers.

Clean Room Suits

The material used for the clothing worn by clean room workers was carefully selected. Humans can shed up to a million particles of a size greater than 0.5 μm every minute. To protect the wafers from microcontamination and maintain a class 10 environment, clean room workers must don special suits. This attire must cover the worker from head to toe.

When device line widths were greater than 1.0 μm, polyester fabrics were used to contain the particles emitted by clean room workers. The pore size of the best polyester fabric is 17 μm, and the filter efficiency at 0.5 μm is poor at less than 60 percent. With submicron line widths, new materials needed to be evaluated for their ability to contain particles. A new material, an expanded polytetrafluoroethylene, with a pore size of less than 1.0 μm and a filtering efficiency of greater than 99.99 percent, was selected as the garment material of choice.[2]

Ultrapure Deionized Water

Each wafer is exposed to hundreds of gallons of ultrapure deionized (DI) water during the 250 processing steps involved in producing CMOS-4 devices. The purity of the DI water, which is measured primarily by resistivity meters, is critical to obtaining a high yield on CMOS-4 devices. Not only must the number of particles be minimized, but the levels of cations, anions, total oxidizable carbon (TOC), silica, and bacteria must also be carefully regulated and monitored. For example, bacteria contain phosphorus and can be a source of uncontrolled dopant. Excess levels of TOC can double the rate of initial thermal oxidation. Silica is known to decrease the reliability of thermally grown oxides, and the presence of ionic contaminants can change semiconductor carrier lifetimes.[3] Table 1 lists the specifications for the DI water system used for CMOS-4 production.

Table 1 Deionized Water Specifications for CMOS-4 Technology

Resistivity	18.0 megohm per cm @ 25°C
Bacteria	0.05 colonies per milliliter maximum
Particles (>0.5 μm)	200 per liter maximum
Total organic carbon	50 ppb maximum
Silica	10 ppb maximum
All cations and anions	1.0 ppb maximum

Note:
ppb equals parts per billion

Chemicals and Gases

Wet chemistry is used to clean wafers and in the photolithographic process to develop and strip photoresist. The particle and impurity levels of the incoming chemicals used during wafer processing had to be specified and monitored. Studies have shown that bare silicon wafers placed in an ammonium hydroxide/hydrogen peroxide (NH_4OH/H_2O_2) solution exhibit a linear correlation between the metal content in the peroxide and the metal surface contamination. The same studies have also shown that iron and zinc are more important than other metals.[4]

Throughout the manufacturing process, gases are employed during gate oxide growth, metal depositions, and plasma etches. Both impurity and particle levels in these gases are critical. Impurity level changes can alter plasma etch-rate uniformity or could lead to corrosion. The process of reducing the impurity and particle levels of gases is better understood than that of wet chemicals. Implementing new filter technology and employing electropolished materials in gas distribution systems have reduced impurity levels.

Improvements to Wafer Cleanliness

Prior to implementation of a structured defect reduction plan to increase yield on CMOS-4 devices, three areas known to require better control were

improved at a minimal cost and effort. These were wafer handling, wafer cassette/box cleaning procedures, and particle per wafer pass monitoring.

Wafer Handling

Clean room workers use tweezers to handle wafers when they read wafer numbers and load wafers into equipment. A full wafer cassette contains 25 150-millimeter (mm) wafers, and each wafer is separated by only 5 mm of clearance. Wafer inspections of CMOS-3 devices revealed that scratches contributed 10 to 15 percent of the die loss.

Three corrective actions were implemented to reduce the number of scratches. Since the use of tweezers to handle wafers was the primary cause of scratches, their use on product wafers was banned from the production line. Vacuum wands were installed throughout the fabrication area. Vacuum wands restrict contact to the wafer backside only. With proper training in the use of wands, workers can achieve better vertical wafer control.

In addition to vacuum wands, automatic wafer transfer systems were installed in the production area. These mass transfer systems allow a worker to move an entire 25-wafer lot from one type of cassette to another, for example, to transfer wafers from polypropylene to quartz cassettes prior to a photoresist strip. This procedure eliminates manual roll transfers of wafers, which are known to generate particles.

Automatic wafer handlers were installed on engineering microscopes as were automated wafer sorters. Use of these devices eliminated a major source of scratches on experimental lots, which are inspected and sorted often. Sneeze guards installed at the microscopes were an added insurance against damage to any die from spittle.

Cassette/Box Cleaning Procedure

Several cleaning procedures for the wafer cassettes and boxes were implemented. A surfactant, which helps wet the surface, was added to the cleaning solution of the boat/box washing equipment. This improved the cleaning efficiency over the previously used method of DI water only. Wafer cassettes were cleaned more often. Cleaning cycles were added at several front-end operations, including well oxidation, initial oxidation, and first- and second-gate oxidation. In addition, weekly cleaning of cassettes dedicated to specific equipment was initiated.

Since wafer cassettes are known to become porous and in time to contaminate the wafer surface, a test was introduced to determine when boats start to degrade. Results indicated that if cassettes took longer than 30 seconds to rinse to resistivity in DI water, they should be replaced.

Particle per Wafer Pass Monitoring

Each piece of process equipment is monitored on a routine basis to detect particles added at each wafer pass; these checks are performed once per shift. The selected wafers are measured on a high-angle laser-scattering system designed for unpatterned wafer particle detection. The wafers are then processed through the production equipment. After final processing, the wafers are measured again for particles. These measurements are subtracted from the initial readings, and the difference is recorded on a trend chart.

For the 1.0-μm technology, bare silicon wafers were measured for particles greater than 0.5 μm, or half the minimum polysilicon line width. This size was chosen based on the premise that a conductive particle of this size could degrade device performance and reliability. Using the same reasoning, and prior to the implementation of CMOS-4 technology into manufacturing, the particle monitoring size was decreased to 0.375 μm.

Particle per wafer pass (PWP) monitoring must closely simulate the wafer processing environment to which product wafers will be exposed. If a process involves an oxide deposition, then the PWP process should also. However, it is important not to damage the wafer surface during the PWP run. For example, gases should be flowed, if possible, during a PWP process on an etcher, but a bias should not be applied. A bias might cause surface damage that could result in false particle counts.

A continuous production PWP program must be maintained on the process equipment. Small particles are held to a surface by strong van der Waals forces. These forces increase over time due to the particles conforming to the surface, thus increasing the contact area. Therefore, once particles are deposited on a wafer's surface, they are very difficult to remove.[5] Even if each process step contributed only five particles to each wafer, the cumulative effect of 250 process steps would be more than 1000 particles deposited on a fully processed wafer.

Defect Reduction Procedures to Increase Product Yield

All microcontamination control efforts would be fruitless without vehicles to assist in yield prediction and defect prioritization. The procedures developed for CMOS-4 technology are outlined in this section. Yield modeling and test chips are described elsewhere in this issue.[6] Therefore, the following discussion is brief.

General Yield Model

For the purposes of this paper, the Poisson yield model is used.[1] This model is very simple, but it can be used to illustrate some key points. The yield model is given by the following equation:

$$Y = e^{-AD}$$

where Y equals yield, A equals chip area in square centimeters, and D equals defect density per square centimeter.

It is easy to see that if the yield is equal to 50 percent, AD must be 0.69. Now if the chip area is increased by 50 percent, AD becomes 1.04, and the yield is 35 percent, if all else remains equal. However, each new CMOS technology reduces the line widths and decreases the film thicknesses. The size of a "killer" defect therefore decreases, which automatically increases the baseline defect density. For each successive generation of CMOS devices, substantial improvements are needed in the reduction of defect densities.

Test Chips

The test chips used during the manufacture of CMOS-4 devices were both full- and short-loop defect density test vehicles. Full-process test chips included snake structures to capture intralevel open circuits, comb patterns for intralevel short circuits, and capacitors for interlevel short circuits. The full-process test chips were run routinely to determine defect reduction priorities, as well as to assist in reducing defect levels.

Short-loop test chips, which are processed through 20 to 25 process steps, contain the same snakes, combs, and capacitors as the full-process test chips. Their purpose is to focus attention on certain layers, such as back-end levels, which are known to contribute many defects. These short-loop chips can be used in designed experiments to compare different processes. Short-loop chips also monitor shifts in defect density from week to week, since they can be processed with less than a two-week cycle time.

Defect Reduction Priorities

After electrical testing of failed structures, visual inspections were performed to identify various defect types. It is very common for large defect densities to be caused by more than one defect type. These inspections helped to design the experiments used to reduce the defect levels found in p-gate leakage and metal 2 short circuits.

In addition, laser- and holography-based automated inspection tools were initiated in-line at various process steps, including active area after strip inspect (ASI), polysilicon ASI, local interconnect ASI, tungsten plug 1 and plug 2 ASIs, and metal 1 and metal 2 ASIs. These inspections were performed routinely on all full- and short-loop test chips. The tungsten plug and local interconnect steps were chosen because they were not part of previous CMOS generations. The remaining steps were chosen because they were known areas of concern based on previous electrical results.

Defect reduction priorities are determined by incorporating electrical results from full-loop defect density test chips into a detailed yield model. Based on this information, yield enhancement activities are prioritized. Two of the areas selected for defect reduction were p-gate leakage and metal 2 short circuits.

P-gate Leakage Enhancements

Historical data obtained from CMOS-3 processing highlighted two potential contributors to high p-gate leakage values: surface damage and metallic contamination. The well etch process, which opens up the gate areas, was performed in a hexoid-configured, reactive ion etch (RIE) batch etcher. Designed experiments, therefore, examined ways to reduce potential lattice damage caused by the known physical etch process. The batch reactor processed 12 wafers at a time, and initial process development ensured complete oxide removal from all wafer surfaces. Based on uniformities both within the wafer and from wafer to wafer, a 30 percent overetch was chosen. The overetch procedure, however, exposed the silicon surface of the wafers to the plasma for an extended period of time, thus exacerbating any surface damage. Initial attempts to improve the gate leakage varied the overetch between 30 percent and 0 percent.

Results showed the p-gate leakage values consistently ran at a lower value with the decreased overetch, confirming a decrease in the amount of surface damage. Optimization work continued to lower the p-gate levels further. This work is outlined in the following sections.

Silicon Lattice Damage

With the improved overetch process, stacking faults and pits continued to be seen in the well region. This confirmed the presence of surface damage, which required further experimentation. Split lots were processed to examine the impact of changing power, bias, oxygen flow rate, and pressure on p-gate leakage and silicon lattice damage. Repeated attempts produced identical results; none of the changes affected p-gate defect density or visual surface damage. At this point, a split lot was designed to study various starting materials. The original starting material was compared to a polysilicon-backed starting material. Polysilicon is a known "getterer" of surface damage, that is, it attracts the damage to the backside of the wafer; it was expected that this type of starting material would show an improvement. The visual inspection of the polysilicon starting material found no stacking faults or pits in the well regions of the wafers. The p-gate defect density values also showed an improvement. Confirmation material verified the initial findings, and the new starting material was placed on the manufacturing line.

Metallic Contamination

In spite of significant improvements to the process, p-gate leakage values continued to show intermittent failures. The one area not previously investigated concerned the potential presence of metallic contamination at the wafer surface. Patterned and unpatterned wafers were sent for total reflectance X-ray fluorescence (TXRF) analysis. The surface analysis detected the presence of metallic contaminants: specifically cobalt, iron, and nickel. Iron and nickel are common elements in stainless-steel components, and it was discovered that the gas distribution tubes being used in the hexoid etcher were constructed of stainless steel. Replacement aluminum gas tubes were installed in the etcher, and additional surface analysis tests were taken. As expected, the iron and nickel elements were no longer detected; however, cobalt was present. Consequently, a complete wet clean of the etch system was performed, and one final set of etched wafers was analyzed by TXRF. The results confirmed the elimination of the cobalt contamination.

The results of TXRF analysis on all metallic contaminants are given in Table 2. Since wet cleans were performed routinely on the system, a trend chart of p-gate leakage was available to show the dates of all completed wet cleans. The trend chart showed that coincident with every wet clean was an improvement in p-gate defect density. The p-gate performance would begin to degrade whenever a metal 1 contact process was run in the same etcher. The contact etch process opened contacts to a cobalt silicide layer, which confirmed the contact etch process as the source of the variable cobalt levels. Cobalt cross-contamination of the gates was occurring whenever a well etch was processed immediately following a metal 1 contact etch, thus inducing p-gate variability. For this reason, it was decided to dedicate separate etch tools for the well etch process and for the metal 1 contact etch process.

Table 2 Results of TXRF Surface Analysis (Units of 10^{12} atoms per cm^2)

	Iron	Cobalt	Nickel
Initial wafer			
Wafer center	2	0.5	2
45 degrees from center	3	0.4	2
225 degrees from center	2	0	2
Aluminum gas tubes			
Wafer center	0	0.7	0
45 degrees from center	0	0.8	0
225 degrees from center	0	0.6	0
Post wet clean			
Wafer center	0	0	0
45 degrees from center	0	0	0
225 degrees from center	0	0	0

Single Wafer Etch System Evaluation

At the same time the metallic contaminant sources were isolated, a well etch process on a single wafer etcher was developed. Single wafer etchers have improved etch-rate uniformity control over batch etchers. Also, the single wafer etch process is more chemical (as opposed to physical) than the etch process used in batch systems. Optimization of process parameters (e.g., gas flows, pressure, power, and gap) was performed on patterned monitor wafers. As the final process step, a low-power surface cleanup was added to remove any remaining

surface contaminants from the wafer surface as well as from the top layer of damaged silicon. Cross-sectional micrographs were taken to verify the integrity of the slopes of the patterned lines.

Once the process was finalized on monitor wafers, full-process split lots were run through the line to compare the hexoid-configured etch process with the planar-configured etch process. The results of one of the splits are shown in Figure 1. A substantial improvement in p-gate defect density was obtained when using the single wafer etch process. Defect density levels on the batch reactor portion averaged 300 defects per cm^2, whereas the single wafer etcher portion averaged 15 defects per cm^2. Confirmation product lots were processed to ensure the probe yield was not adversely affected by the etch enhancements, and the well etch process was released again on the single wafer etch systems. The final requirement of this process release was the continued segregation of the well etch process and the metal 1 contact etch process. The metal 1 contact etch process remained on the batch etcher.

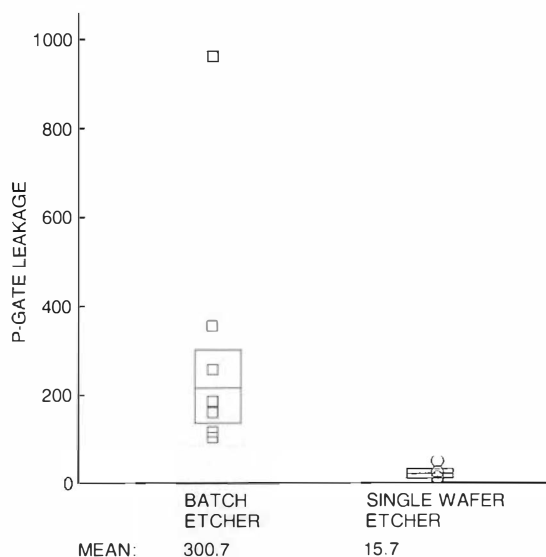


Figure 1 P-gate Area Defect Density Levels as a Function of Etcher Type

Reduction in Metal 2 Short Circuits

The implementation of the CMOS-4 metal process into manufacturing brought a new set of challenges. The estimated yield impact of the metal 2 short circuits on the first set of full-process lots was approximately 40 percent, based on a single-process-step defect density (D_{oi}) level of approximately 40 defects per 100-meter (m) length.

Equipment-induced Short-circuiting Mechanisms

PWP data on two sets of process equipment emphasized the need for concentrated particle reduction efforts. These two systems were metal deposition and dielectric deposition. Task forces, with members from defect reduction, process engineering, and equipment engineering, were organized and chartered with reducing these PWP numbers.

Metal Deposition System Upgrades A known yield limiter of the CMOS-3 process was the presence of a high number of titanium nitride (TiN) particles on the wafer surface. Information obtained from the equipment vendor and confirmed by Digital's semiconductor manufacturing fabrication plant in Scotland indicated that a new planar titanium target would provide a cleaner TiN film, thus decreasing metal short circuits and enhancing yield. The planar target, a rotating magnetic experimental (RMX) cathode, was able to decrease the yield loss attributed to TiN particles by 75 percent. Since several TiN layers were used in the CMOS-4 process, fewer particles on TiN film would definitely benefit the overall yield.

The PWP data in Figures 2 and 3 show how the particle levels dropped once the RMX cathode was installed. Data obtained on one of the initial RMX split lots showed a 50 percent reduction in metal short circuits. A substantial number of confirmation lots were processed to examine metal short circuits and electrical test data. Once the lots were analyzed, the decision was made to release the RMX cathode into production and an improvement in metal 2 short-circuit levels was realized.

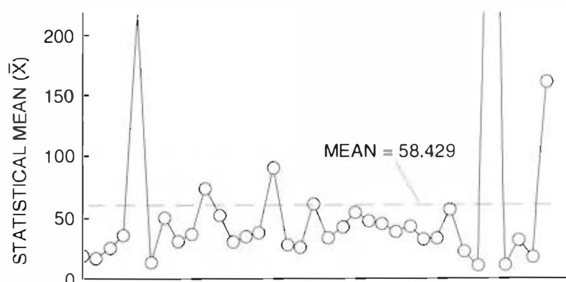


Figure 2 TiN Particles before Installation of RMX Cathode

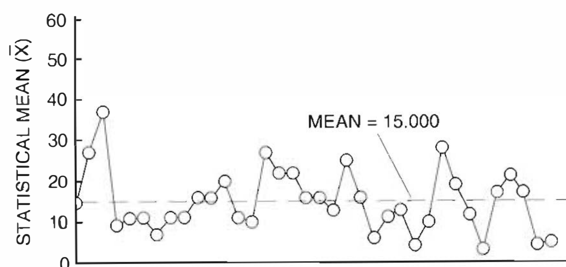


Figure 3 TiN Particles after Installation of RMX Cathode

Dielectric Deposition System Particle Reduction The dielectric deposition systems had the highest particle levels of any equipment in the fabrication area, as shown in Figure 4. Since dielectric particles can induce metal short circuits, and metal short circuits typically impact yield more than any other defect structure, a concerted effort was made to improve the particle stability. The four major areas of change were (1) the installation of a new type of O-ring, (2) the initiation of a continuous pump ballast, (3) pressure adjustments in the load lock, and (4) modifications to the "clean" recipes. These changes improved the average PWP count from 537.7 to 2.6 particles greater than 0.375 μm, as shown in Figure 5.

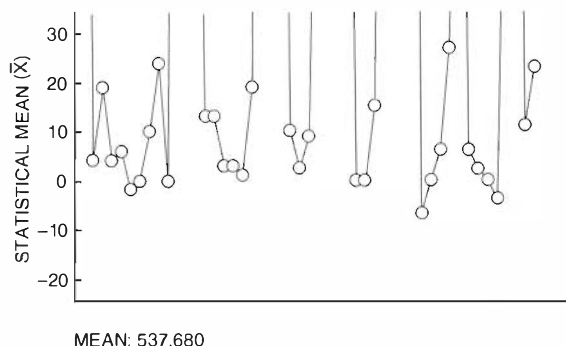


Figure 4 Initial Dielectric Particle Baseline

Process-induced Short-circuiting Mechanisms

Once the equipment PWP data improved to the levels shown in Figures 3 and 5, and the metal 2 short-circuit levels fell to approximately 10 defects per 100-m length, two systematic defects were uncovered: corrosion and grain-boundary stringers. Both defect types were noted during the inspection of failed sites on the test chip structure. The defects

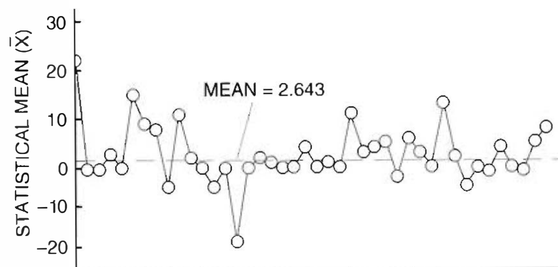


Figure 5 Improved Dielectric Particle Baseline

appeared intermittently (not all lots were affected to the same extent), and were clustered around the wafer edge. The metal 2 short circuits consistently were twice as high as the metal 1 short circuits; therefore, processes unique to the metal 2 process sequence were evaluated first. A lot history investigation performed on all lots exhibiting corrosion or grain-boundary stringers found no equipment commonalities, confirming possible process-induced mechanisms. Several processes unique to the metal 2 process flow include the intermetal dielectric, the metal 2 cutout sequence (used to remove metal from alignment targets allowing metal 2 alignment without manual intervention), and the tungsten process sequence.

Corrosion Figure 6 shows the corrosion on a metal line in the test circuit. A common cause of corrosion is the interaction of chlorine with moisture. The tungsten etch-back process was investigated because it uses chlorine as an etch gas. A short-loop monitor wafer experiment was designed to study the effects of various post-tungsten etch-back processing procedures on corrosion. The level of corrosion was determined by means of a patterned wafer defect detection tool. Practices commonly used in the semiconductor industry to eliminate chlorine-induced corrosion are a fluorine passivation process and/or an immediate water rinse. The data in Figure 7 shows that, without an immediate water rinse, the corrosion counts increase with time. After only 90 minutes, the rate of increase changed dramatically. Initially, the incorporation of a rinse provided more stability, but the corrosion levels were unacceptable after just 11 hours. The addition of a 5-second sulfur hexafluoride passivation process, however, completely prevented any corrosion.

To verify this data on test chips, several lots were split. The results from one of the splits are shown

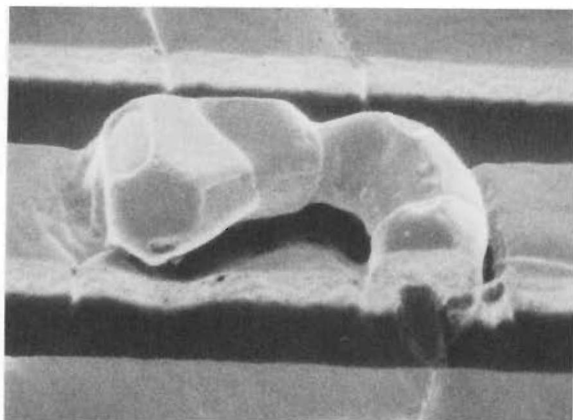


Figure 6 Corrosion on a Metal Line

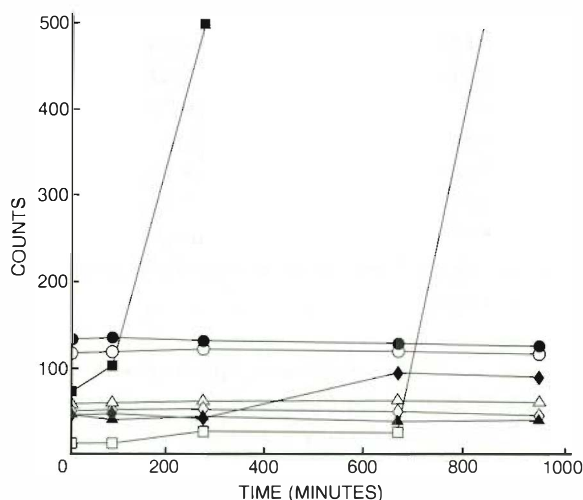


Figure 7 Particle Counts of Bare Aluminum Wafers as a Function of Time after Tungsten Etch Back

in Figure 8. The metal 2 short-circuit levels improved from 11.2 defects per 100 m to 4.1 defects per 100 m. Additional splits performed on product and test lots confirmed that the sulfur hexafluoride passivation process was as good as, if not better than, the original process. The electrical results

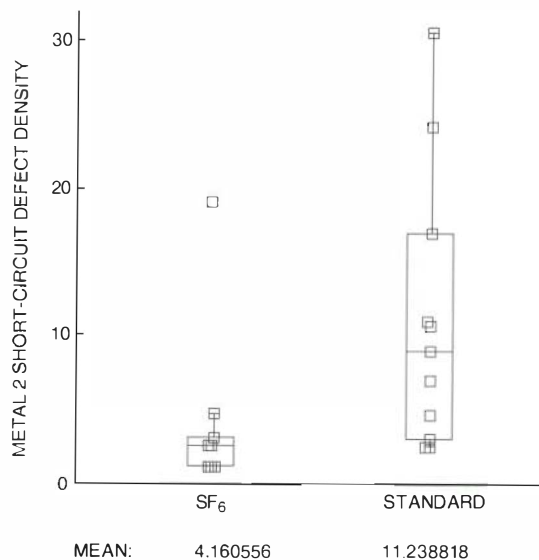


Figure 8 Results from Split Lot Test of Metal 2 Short Circuits

verified that the defects were a wet type of corrosion formed on metal 1 in the bond pads and guard rings and were caused by exposure to chlorine during the tungsten etch-back process at tungsten plug 2.

Grain-boundary Stringers A short-loop lot was designated to look for potential contributors to the grain-boundary stringers. It examined three factors: the dielectric film, the cutout process sequence, and the metal 2 deposition process. Results of the lot are shown in Figure 9. They indicate that the most significant factor affecting grain-boundary stringers was the cutout process sequence. Since the grain-boundary stringers were seen on an intermittent basis, the split was purposely designed to exaggerate the impact of the cutout process. This was achieved by reducing the metal overetch and processing the designated wafers through the photolithography and strip cycles twice. The metal 2 short circuits on these wafers were dramatically higher than those on the wafers that did not receive a cutout process; they improved from 27 defects per 100 m to 1.4 defects per 100 m. An example of a grain-boundary stringer is shown in Figure 10.

Since processing without a cutout process was not an option, a split lot was designed to study various versions of the cutout strip process. Although the electrical results of the split were not conclusive (all defect levels were excellent), a scanning

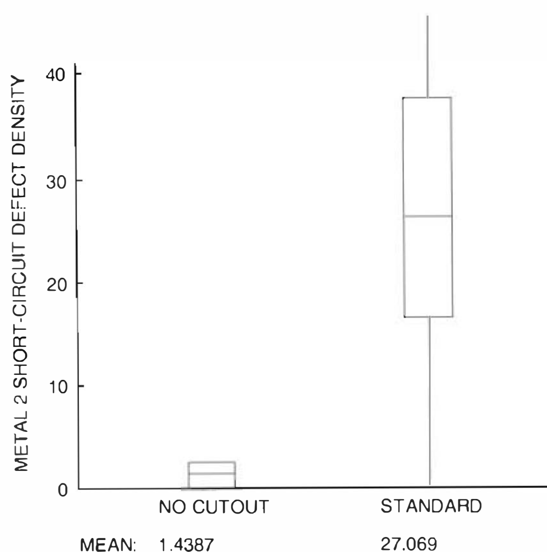


Figure 9 Results of Test on Grain-boundary Stringers

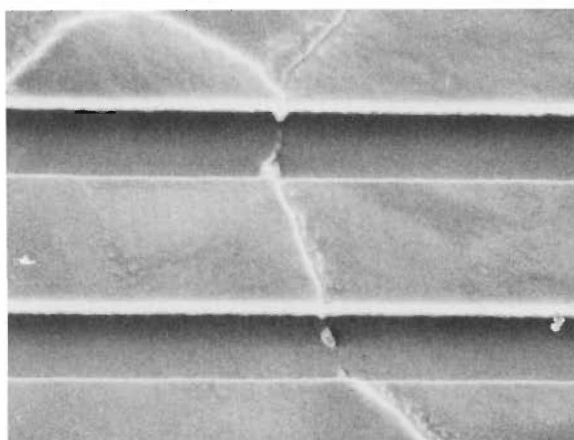


Figure 10 Photomicrograph of a Grain-boundary Stringer

electron microscope (SEM) analysis of the same die location on all splits showed a significant difference among splits. All splits incorporating a wet solvent had evidence of grain-boundary stringers, whereas the splits stripped only in a downstream plasma had none of the stringers. Therefore the cause of the grain-boundary stringers was the reaction of the grain boundaries with the solvent and subsequent water treatments.

The purpose of the solvent/water portion of the strip process was to assist in the removal of photoresist that had been plasma-hardened by the metal etcher. By transferring the metal etch process from a plasma etcher to a wet sink, the option of a downstream plasma strip became available. Several additional splits were processed to study the combination of a wet chemical cutout etch followed by a downstream plasma strip. They confirmed that the elimination of the interaction between the grain boundaries and the solvent/water combination resulted in a reduction in grain-boundary stringers. Metal 2 short circuits on one of the lots improved from 5.1 defects per 100 m to 3.2 defects per 100 m and showed a corresponding improvement in yield.

From the initial stages of CMOS-4 manufacturing up to the present, an overall improvement in metal 2 defect levels has been seen. The test chip metal 2 short-circuit D_{oi} levels have diminished from approximately 40 defects per 100 m to approximately 4 defects per 100 m. Corresponding metal 2 visual defect inspection data has decreased from approximately 1.5 defects per cm^2 to approximately 0.2 defects per cm^2.

Future Considerations

For the successful production of future generations of CMOS technology, improvements need to be made in the areas of general microcontamination, wafer handling, defect review, and data management. The needed improvements are outlined below.

Microcontamination

Initial results from particle studies on PWP measurements from current equipment indicate that 70 percent of the total particles are between 0.25 μm and 0.375 μm. As discussed previously, once deposited on wafers, these small particles are very difficult to remove. Efforts must be made to isolate and eliminate the source of these particles before the equipment is introduced into manufacturing. To provide early identification, in-situ particle measurement for process tools needs to be incorporated as a part of the tool specification.

In addition, as geometries shrink and gate oxides become thinner, the importance of understanding the effect of nonparticulate contamination such as zinc, sodium, and hydrocarbons on device yield is critical. The cost of reducing such contaminants from the wafer environment is astronomical. We must learn which contaminants are harmful and at what level. Surface analysis techniques, such as

TXRF and Auger electron spectroscopy, need to be incorporated into the manufacturing environment to provide the necessary data.

Wafer Handling

As die size continues to increase, wafer size also increases. As technology moves toward 200- and 300-mm wafers, manual handling of wafers is all but impossible. One broken wafer will result in the loss of thousands of dollars. Automated handling of wafers must be incorporated across each process step.

Defect Review

As the killer defect size continues to decrease, visual inspections with optical microscopes will lose their value. As many as 50 percent of the electrical failed combs on our 0.5-μm process are not observed during optical microscopic review. This has highlighted the need for two things: test chips designed for failure analysis and easy-to-use in-line inspection SEMs.

Test chips can no longer be designed without the active involvement and input of defect reduction personnel. A test chip must not only be used to prioritize defect reduction efforts, but it must also help to determine and isolate defect sources. A test chip that is difficult to inspect provides only half the information needed to reach the ultimate goal of high yield.

In-line inspection SEMs are required to review defects found during process inspections and to analyze process problems. They should be capable of energy dispersive X-ray spectroscopy (EDXS) analysis to provide information on the elemental components of the defects. In addition, the transfer of electrical test data to an SEM is required so that failing locations can be easily reviewed.

Data Management

The future of defect reduction efforts depends on the ability to manage and analyze large quantities of data. Defect inspection tools perform full wafer inspections in less than five minutes. A thorough understanding of statistical methods, such as control charts and sampling procedures, is required to determine the extent of defect review efforts. Automatic storage of images on optical disks for subsequent review is another key area. Efforts are ongoing with Digital's Campus-based Engineering Center in Vienna to determine the benefits of using fractal analysis on airborne particle data.

Summary

The success or failure of a semiconductor production facility lies in the ability to control contamination and reduce defects. Efforts are under way to improve Digital's ability to obtain high yields on current and future microprocessors. Semiconductor companies must join with inspection and analytical equipment manufacturers, as well as with research institutes, to develop the required tools to support future technologies.

Acknowledgments

Many people contributed to the successful implementation of the defect reduction work highlighted in this paper. We would like to thank all members of the manufacturing process and equipment groups who worked long hours to qualify the CMOS-4 process. Special thanks to Tom Estelle and Jan Haanpaa for their contributions in coordinating the work outlined in this paper.

References

1. C. Osburn et al., "The Effects of Contamination on Semiconductor Manufacturing Yield," *The Journal of Environmental Sciences* (March/April 1988): 45-57.

2. W. Whyte and P. Bailey, "Particle Dispersion in Relation to Clothing," *The Journal of Environmental Sciences* (March/April 1989): 44.

3. P. Burggraaf, "Ultrapure Water Focus," *Semiconductor International* (May 1987): 129-132.

4. M. Meuris et al., "Correlation of Metallic Impurity Content of ULSI Chemicals and Defect-Related Breakdown of Gate Oxides," *Proceedings of the 178th Meeting of the Electrochemical Society* (Washington, D.C., 1991): 488.

5. R. Bowling, "An Analysis of Particle Adhesion on Semiconductor Surfaces," *Journal of the Electrochemical Society: Solid-State Science and Technology,* vol. 132, no. 9 (September 1985): 2209.

6. R. Collica, X. Dietrich, R. Lambracht, and D. Lau, "A Yield Enhancement Methodology for Custom VLSI Manufacturing," *Digital Technical Journal,* vol. 4, no. 2 (Spring 1992, this issue): 83-99.

Randall S. Collica
X. Joseph Dietrich
Rudolf Lambracht Jr.
David G. Lau

A Yield Enhancement Methodology for Custom VLSI Manufacturing

Integrated circuit yield enhancement is a complex issue due to the many steps involved in the manufacturing process and the number of variables governing the overall yield. The task is further compounded by industry technology goals for continually improving performance achieved by decreasing minimum feature size, increasing chip area, and incorporating more on-chip functionality from generation to generation. In the final analysis, the cost of producing chips is directly related to the yield, hence the necessity for a comprehensive yield improvement strategy.

The industry technology goal for continually improving complementary metal-oxide semiconductor (CMOS) very large-scale integration (VLSI) chip performance has been achieved by decreasing minimum feature size and incorporating more on-chip functionality in a larger chip area. At Digital, four generations of CMOS technology have been developed. Each generation possesses successively more complex manufacturing processes as well as more individual process steps to fabricate the chips. These complex processes and additional steps have increased the number of variables that have the potential for affecting yield.

Digital's Alpha 21064 microprocessor chip[1] has a peak operating frequency of 200 megahertz (MHz), contains 1.7 million transistors, and has chip dimensions 1.68×1.39 square centimeters (cm^2). The Alpha microprocessor is the highest performance, highest density, and largest chip currently manufactured in volume by Digital. With these chip complexities, achieving chip yield goals is imperative for successful cost-effective manufacturing.

In the manufacture of integrated circuits, yield is defined as the fraction of the total number of die sites introduced into processing that are completed as fully functional chips. The cost of producing a fully functional chip is inversely proportional to the yield. In the CMOS-4 manufacturing line, 22 wafers are processed through all the steps as a single lot. On a lot basis,

$$\text{Cost of chip} = \frac{\text{Cost of lot}}{\text{Yield of lot} \cdot \text{Total number of die sites in lot}}$$

Hence, improving the yield directly affects reducing the cost of production per chip. This motivation for yield enhancement applies to all product chips for the life of the respective chip.

Once the product chip is introduced, demand for the chip increases up to some level and then declines to the end of its life. During a product's development, circuit and system design teams require quantities of chip prototypes for design verification and debug purposes. The rapid increase in demand after product introduction requires a steep yield learning curve and hence rapid yield enhancement to supply production quantities.

This paper discusses the yield enhancement methodology used to evaluate processing, process equipment, manufacturing methods, design, and testing in relation to yield. It describes the yield test vehicles, designed experiments, yield models, and special-purpose analytical tools to identify and prioritize defects and to focus resources on appropriate defect inspection and failure analyses tasks. Before our discussion of the specific techniques, we present a brief overview of the methodology to enhance yield.

Overview of Yield Methodology

The yield enhancement methodology applied to chip manufacturing at Digital involves the use of test chip data, product chip data, static random-access memory (SRAM) data, yield models, special-purpose analytical tools, and defect inspection to perform defect identification and prioritization. The information gained is relayed to process engineering for rapid yield learning and the design of experiments for yield improvement. In addition, yield, development, and design teams use the

information to estimate the manufacturability of future products.

Figure 1 presents the overall methodology and the relationships of the various stages of yield analysis. After a chip has been designed, its design is converted into a chip layout to produce the reticle set for the chip. Test chip wafers from process manufacturing are electrically tested. Parametric data, defect structure data, and SRAM test vehicle functionality data are measured on test chips. The layout information for the chip, the measurements from defect structures on the test chip, and SRAM functional yield are used to model yield.

The yield model generates a priority list (defect pareto) of the yield-limiting defect mechanisms for the layout of that particular chip. If the layout of the chip is determined to be especially sensitive to a particular defect type, the chip layout can be altered to optimize for higher yield without changing the design functionality. In addition, in the early stages of design, yield information from a previous generation technology can be used to forecast the yield of chips to be manufactured in a succeeding generation of technology. The effects of choosing different redundancy schemes can also be forecast.

Product wafers from process engineering are also electrically tested, and parametric data and product functionality data are measured on these wafers. The results of measurements at electrical test, the lot process records, and defect inspection data undergo parametric and product yield analyses to determine the factors affecting yield. All results of analyses are used by process engineers for experimental design and yield learning. The use of test chips is one of the starting points in early process defect learning and is described next.

Test Chip

Test vehicles and structures on the test chip provide the process and electrical information needed to estimate chip yield. The CMOS-4 yield test chip is shown in Figure 2. The chip is divided into three functional areas:

1. Defect density test structures
2. 128 kilobit (Kb) SRAM
3. Scribe lane structures

The defect density structures are used to determine the component defect densities for the integrated

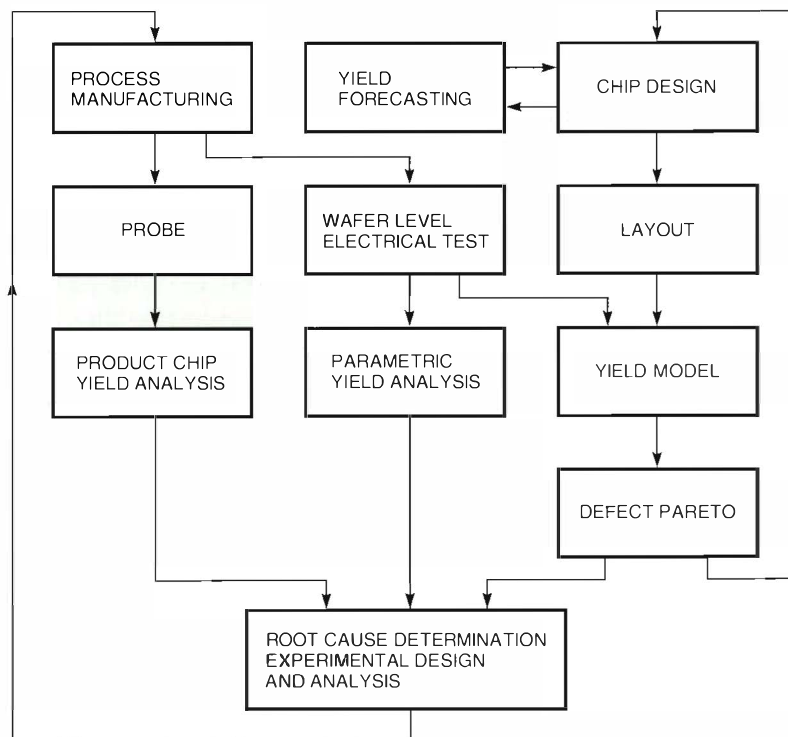


Figure 1 Yield Methodology Flow Diagram

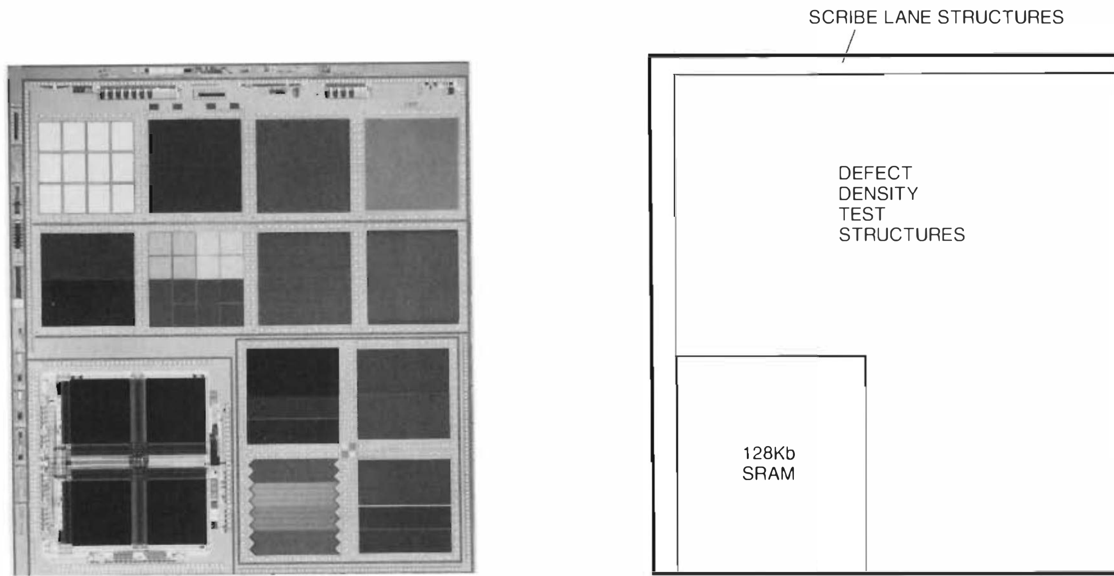


Figure 2 Yield Test Chip for the CMOS-4 Process

process, and the scribe lane structures are used to determine the parametric yield. Since the SRAM and the defect density structures are both on the chip, the SRAM failure modes can be correlated with the component defect levels measured on the defect density structures. In addition, with the component defect densities factored into the yield model, a defect pareto can be generated that prioritizes the defect mechanisms according to their yield-limiting effect on the SRAM.

The defect density structures have the following qualities:

1. Simple design, with large area structures for determining specific component defect densities
2. Testable in-line
3. Compatible with the yield model
4. Inspectable in-line

The CMOS-4 technology integrates approximately 250 process modules on a single complex chip.[1] If a process aberration or defect renders a chip inoperable, it is very difficult to diagnose at which process step the defect occurred.

The defect structures partition the integrated process into critical physical process features that are more easily characterized than an entire complex chip. A number of defect structures are designed; each one is designed to detect a specific defect type. The complete set of defect structures combines to represent the full process.

The defect structures test for the following general electrical faults:

1. Intralayer short circuits and open circuits
2. Interlayer short circuits
3. Contact/via chain open circuits

Intralayer short circuits and open circuits are measured using snake/comb structures. For example, Figure 3 shows a schematic diagram of a metal 1 (M1) snake/comb structure that tests for M1 short circuits and open circuits. Testing the continuity of the snake detects open circuits in the M1 line. Testing for continuity between the snake and combs detects M1 short circuits.

Figure 4 shows an example of an interlayer short-circuit test using an M1 to metal 2 (M2) capacitor. A test for continuity between M1 and M2 conductor plates detects short circuits in the dielectric layer between M1 and M2.

Contact/via integrity is tested using chains of contacts or vias. Figure 5 shows how an M2 to M1 via chain is tested for continuity to check for any chain open circuits.

Table 1 lists the yield test chip defect structures and respective electrical faults or defect mechanisms that they detect.

The test structures must be testable in-line, that is, they must allow electrical tests to be performed

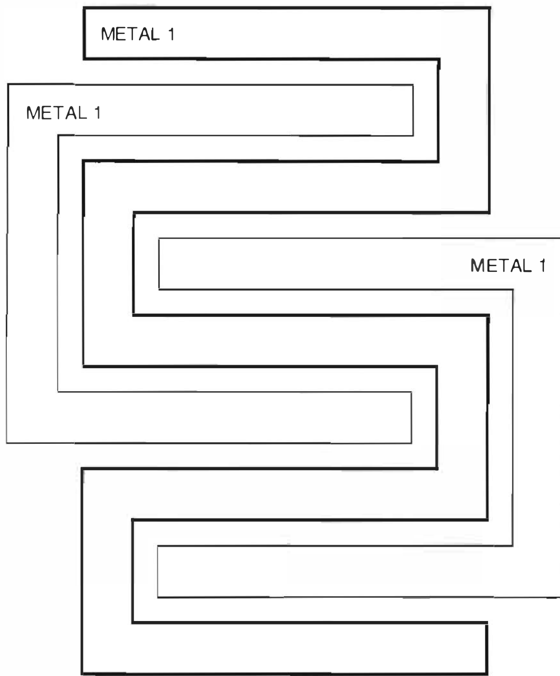
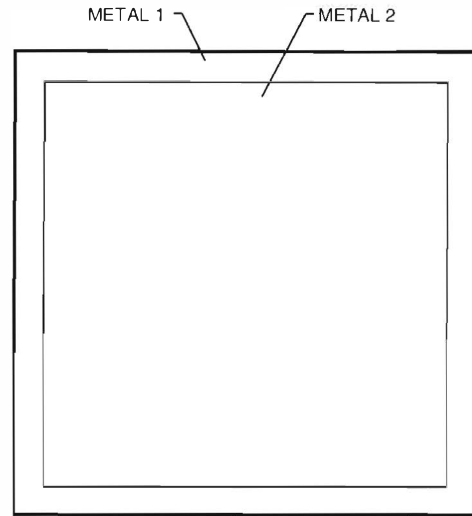
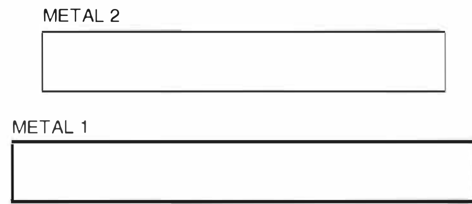


Figure 3 Metal 1 Snake/Comb Structure

after M1, M2, or metal 3 (M3) patterning. Consequently, structures that do not require process steps after M1 must be designed only in those layers up to and including M1. For example, M1 to polysilicon contact chains are connected to pads by M1 and not by M1 to M2 vias in conjunction with M2. The former chain is testable in-line after M1 pattern-

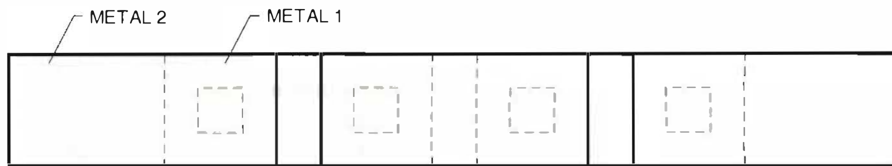


(a) Top View

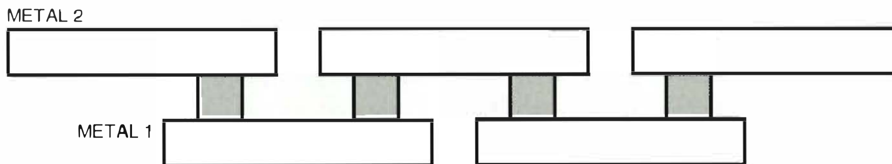


(b) Side View

Figure 4 Metal 2 over Metal 1 Capacitor



(a) Top View



(b) Side View

Figure 5 Metal 2 to Metal 1 Via Chain

Table 1 Defect Structures and Defect Types

Structure	Electrical Fault to Detect
Dielectrics	
M3/M2/M1/polysilicon/substrate capacitor	Gate oxide, dielectrics 1,2,3 short circuits
M3/M2/M1/polysilicon/well capacitor stack	Gate oxide, dielectrics 1,2,3 short circuits
Bird's beak structure — polysilicon plate minimum pitch active area stripes in well	Field/active area over periphery short circuits
Bird's beak structure — polysilicon plate over minimum pitch active area stripes in substrate	Field/active area periphery short circuits
Contacts	
M1 to n+ contact chain	M1 to n+ open circuits
M1 to p+ contact chain	M1 to p+ open circuits
M1 to n+ polysilicon contact chain	M1 to n+ polysilicon open circuits
M1 to p+ polysilicon contact chain	M1 to p+ polysilicon open circuits
M2 to M1 contact chain	M2 to M1 open circuits
M3 to M2 contact chain	M3 to M2 open circuits
Polysilicon to local interconnect to n+ chain	Polysilicon to local interconnect to n+ open circuits
Interconnect	
Polysilicon snakes and combs	Polysilicon open circuits and short circuits
M1 snakes and combs	M1 open circuits and short circuits
M2 snakes and combs	M2 open circuits and short circuits
M3 snakes and combs	M3 open circuits and short circuits
N+ snakes and combs	N+ open circuits and short circuits
P+ snakes and combs	P+ open circuits and short circuits
Local interconnect to n+ combs	Local interconnect to n+ short circuits
N-channel gate meander	N-gate meander short circuits
P-channel gate meander	P-gate meander short circuits
Local interconnect snakes and combs	Local interconnect open circuits and short circuits
Local interconnect to n+ polysilicon combs	Local interconnect to n+ polysilicon short circuits

ing; the latter has to continue processing to be tested after M2 patterning.

The structures were designed with minimum pitch metal lines and minimum surrounds for contacts to give them the greatest sensitivity to defects. In addition, the structures covered an area large enough to detect the minimum defect density desired. The yield model required that the yield test chip contain four or more instances of each type of defect structure to determine the clustering parameter in the yield model.

The structures are integrated onto the chip to facilitate manual or automated visual inspection of failing test sites. For example, in the process flow, M1 is patterned several process steps after polysilicon is patterned. To keep the polysilicon layer in view, some polysilicon snake/comb structures must not be covered with any M1 structures. Should a failing polysilicon comb be found at M1 electrical test, that failing comb can then be visually inspected to identify the type of defect causing the comb to fail.

The scribe lane contains the minimum set of electrical test and process monitor structures required to characterize and monitor the process in a manufacturing mode. The scribe lane exists on all chips. Table 2 lists parameters measured from the scribe structures. If a critical parameter does not comply with its specification on more than a certain number of die sites on a wafer, that wafer is rejected. Hence, scribe lane structures determine parametric yield loss.

Electrical Failure Specifications

To estimate the defect density from an electrical test structure monitor, specification limits must be established that determine a fault. Usually, the DC parametric testing applies either a voltage or a

Table 2 Scribe Lane Electrical Test Parameters

Transistor threshold voltages
Transistor saturation region currents
Transistor leakage currents
Transistor effective channel lengths
Diode breakdown voltages
Field transistor threshold voltages
Interconnect sheet resistances
Gate oxide thickness
Contact resistances
Dielectric leakage currents

current value. Either the leakage current or voltage is then recorded or the resistance is computed on metal interconnect vias, on contact chains, or on serpentine lines. A fault at a gate capacitor is recorded if the level of current passed after a voltage ramp is 1 microampere (μA) or more. The fault-level specifications are periodically reviewed and changed as necessary.

SRAM Analysis for Process Fault Signatures

The SRAM is a useful circuit vehicle for yield analysis. On the yield test chip, the probe yield of the SRAM measures the capability of the full, integrated process to yield product. The regular array of the SRAM and its bit mapping capability allow some correlation of failure modes to defect mechanisms. The use of SRAMs processed with the defect test structures allows the determination of process defects that affect circuit failures. A relatively large SRAM array (i.e., 128Kb or larger) typically captures most of the faults within the memory cells, which comprise up to 90 percent of the memory chip layout. These cells are tested after the basic continuity and leakage requirements have been met during the SRAM wafer level testing. Since the memory cell layout is a regular array, certain defect mechanisms have specific functional failure patterns within the circuit. Figure 6 is a typical bit fail map showing the types of patterns analyzed with pattern recognition techniques.[2] These patterns were analyzed for their signature of possible defect mechanism with a probability of failure (POF) matrix.[3]

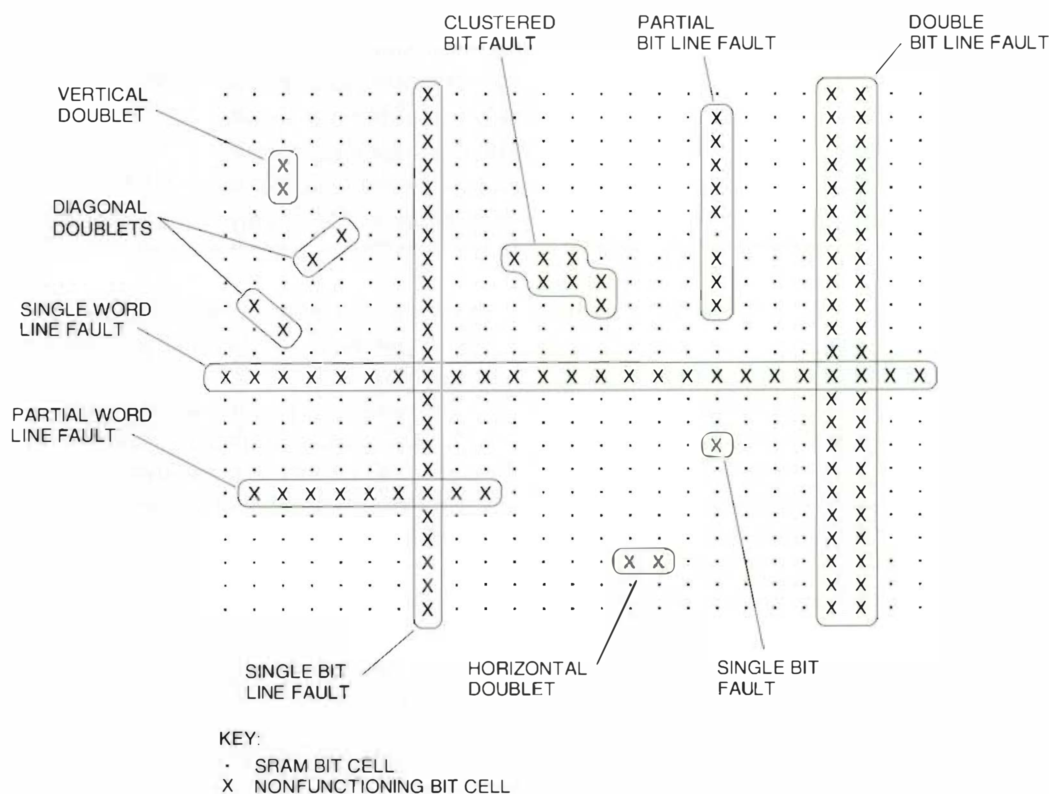


Figure 6 Bit Map of SRAM Failure Patterns

Table 3 gives the POF matrix for a 128Kb SRAM. The rows of the matrix are the individual defect types monitored in the process using the special test structures described above. The columns of the matrix are the types of SRAM failure patterns gathered from pattern recognition programs. Each cell in the matrix is the average probability that a defect of that type will cause a circuit fault type. This analysis was originally done using Monte Carlo simulation techniques on the memory cell using the VLASIC yield simulator.[4,5] This matrix is currently being used to compile defect statistics from the defect-sensitive test structures together with the SRAM pattern fail data.

An example of SRAM analysis shows the yield and failure diagnosis of column failures. Since the SRAM columns are designed in the M2 interconnect layer, the level of M2 to M2 short circuits measured from the defect test monitors correlates to the level of column failures. Figure 7 shows the level of single and double column failures and the level of lateral M2 short circuits on a per lot basis. SRAM analysis is used because the defects accumulated during the manufacturing process do not have a strong electrical fault signature on a large, complex microprocessor circuit.

The yields of the SRAM chips are typically compared to those forecast with the yield model. The test structures described earlier are used collectively in the yield model, which is discussed in the following section.

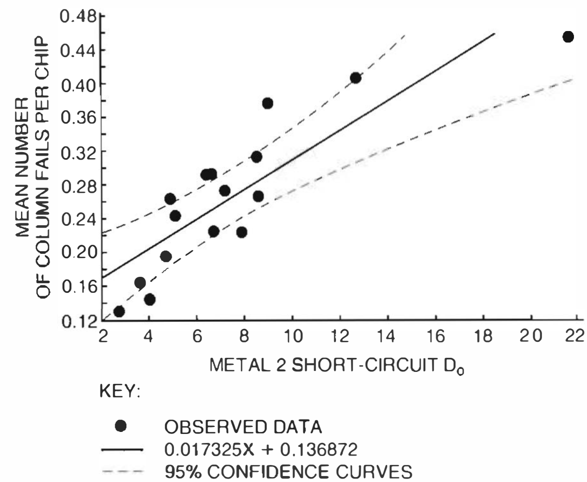


Figure 7 Comparison of SRAM Single and Double Column Failures to Metal 2 Short Circuits

Single Layer Yield Model

The single layer yield model identifies the defect types that have the greatest impact on yield. Process engineering uses this information, known as the defect pareto, to improve the yield by designing experiments to reduce the defect density of the highest priority defects. Design engineering uses this information to lay out product chips so that they are less sensitive to these defect types.

Table 3 128Kb SRAM Probability of Failure Matrix

	Fault Type				
Defect Type	Row Fails	Partial Row Fails	Double Column Fails	Single Column Fails	Partial Column Fails
Dielectric 1 (flat)	0.20000	0.0000	0.00000	0.00000	0.00000
Dielectric 2 (worst-case step)	0.00000	0.0000	0.00000	0.95000	0.00000
Dielectric 2 (flat)	0.00000	0.0000	0.00000	0.05000	0.00000
Polysilicon short circuit	0.15640	0.0242	0.02160	0.18315	0.00000
M1 short circuit	0.59000	0.0000	0.17410	0.20453	0.00000
M2 short circuit	0.00000	0.0000	0.73221	0.22870	0.00000
Polysilicon open circuit	0.19602	0.0681	0.00000	0.00000	0.00000
M1 open circuit	0.36310	0.6020	0.00000	0.00000	0.00000
M2 open circuit	0.00000	0.0000	0.12100	0.21950	0.63900
N-gate area	0.00000	0.0000	0.00000	0.00000	0.00000
P-gate area	0.00000	0.0000	0.00000	0.00000	0.00000

The single layer yield model uses the negative binomial yield equation (1) to predict the yield impact on a given product for each defect type.[6,7,8]

$$\text{yield}_i = (1 + A_{c_i} D_{o_i} / \alpha_i)^{-\alpha_i} \quad (1)$$

where A_{c_i} is the critical area of the chip, D_{o_i} is defect density, α_i is a measure of spatial distribution of defects,[9] and i is an index used to denote a specific defect type.

The critical area of the chip is the area that is most susceptible to a particular type of defect. Critical area can be measured in various units, including square centimeters, meters, and numbers of contacts. The defect density is the number of defects per unit area where the units of area are the same as the critical area. Defect density is independent of product type, and depends only on the cleanliness of the fabrication. The spatial distribution of defects is independent of product type (depending on chip size[10]) and depends to some degree on the cleanliness of the fabrication.[11]

Critical Area$_i$ Extractor

Integrated circuits are designed using computer-aided design (CAD) technology. The geometries that eventually become the wires, transistors, resistors, and other circuit elements are stored in a layout artwork file. Design Rule Check software calculates critical areas for different defect types. The algorithms for this usually involve Boolean algebraic operations and sizing of layout geometries.

One example of a critical area extraction is to define a temporary layer to be the intersection of the M1, polysilicon, and M1 contact layers. The number of geometries on this temporary layer is the total number of M1 to polysilicon contacts on the product chip. The critical area of the M1 to polysilicon open-circuit defect type is related to this number (there are additional steps to the algorithm that eliminate counting redundant contacts). Other algorithms are used to calculate the critical areas for the other defect types.[12,13]

Table 4 gives the units of critical area for the four basic defect types. These area extractions are made on both the test pattern and the product chip and are used in the negative binomial yield equation as the critical areas.

Table 4 Critical Areas by Defect Type

Defect Type	Units of Critical Area
Interlayer open circuits (contact chains)	Number of nonredundant contacts
Interlayer short circuits (caps)	Area of overlap between sequential interconnecting layers
Intralayer open circuits (snakes)	Hundreds of meters of minimum width interconnect
Intralayer short circuits (combs)	Hundreds of meters of minimum spacing interconnect

Computation of Alpha$_i$

The negative binomial yield model requires the calculation of a parameter that is usually referred to as α. As stated previously, α_i is a measure of the spatial distribution of defects on a wafer.[9] Small values of α indicate that defects are more likely to be clustered in isolated areas on the wafer. High values of α indicate that defects occur across the wafer in a more uniform fashion. A single test pattern has multiple copies of the same test structure on it. Each of these test structures is independently testable so that the number of failing test structures on a test pattern can be counted. From this data, a distribution can be created of number of occurrences as a function of number of failing test structures. The mean and variance of this distribution can be computed and then α can be calculated from the following equation (2).[14]

$$\alpha_i = m_i^2 / (v_i - m_i) \quad (2)$$

where m_i is the average number of failing test structures per test pattern, v_i is the variance of failing test structure per test pattern, and i is an index indicating the particular defect type.

Defect Density$_i$ Calculation

The binomial yield equation can be solved for defect density$_i$ if the yield$_i$, α_i, and critical area$_i$ are known. Therefore, the defect density can be calculated for the test pattern because the yield$_i$ and α_i are measured directly at electrical test, and the critical area$_i$ is calculated from the area extraction software.

Product Yield$_i$ Calculation

Defect density$_i$ and α_i are assumed to be the same for all products and test patterns. This assumption states that the amount and the spatial distribution of defects are not product dependent. Their values are established from the electrical testing of the test pattern. The critical area$_i$ is product specific and

determines how defect density$_i$ and α_i affect the yield for that particular product. The critical area$_i$ for the product is extracted from the layout artwork file. These three values are then inserted into the negative binomial yield equation, and the yield$_i$ of the product is solved.

Defect Pareto

After all the values for yield$_i$ have been calculated, they are listed in order of increasing yield to create the defect pareto. An example of a defect pareto from which the values for yield$_i$ have been removed is given in Table 5. Information from the defect pareto is relayed to process and design engineering to complete the process enhancement and design for manufacturability.

Composite Layer Yield Model

The modeled yields of each defect type in the single layer yield model are assumed to be independent of each other. The product of the modeled yields represents the overall modeled yield.

Table 5 Typical Yield and Defect Pareto

Defect	D_o	Units	α	Estimated Yield
				0.xxx
M2 short-circuited lines	11.350	100 m	0.108	0.xxx
M1/polysilicon contact open circuits	3.397	ppm	.490	0.xxx
M1 short-circuited lines	4.949	100 m	0.150	0.xxx
Local interconnect-polysilicon short-circuited lines	5.376	100 m	0.059	0.xxx
M2 open lines	0.941	100 m	1.000	0.xxx
Active area p+ short circuits	4.804	100 m	0.098	0.xxx
Active area n+ short circuits	4.700	100 m	0.080	0.xxx
M2/M1 contact open circuits	0.266	ppm	0.061	0.xxx
M1/n+ contact open circuits	0.113	ppm	0.056	0.xxx
Polysilicon short-circuited lines	1.739	100 m	0.035	0.xxx
Polysilicon open lines	2.610	100 m	0.027	0.xxx
M1/p+ contact open circuits	0.117	ppm	0.034	0.xxx
Active area n+ open circuits	0.831	100 m	0.017	0.xxx
Local interconnect-n+ active area short-circuited lines	1.674	100 m	0.101	0.xxx
Active area p+ open circuits	0.458	100 m	1.000	0.xxx
Dielectric 2 (capacitor)	0.316	cm^2	0.045	0.xxx
Local interconnect short-circuited lines	2.110	100 m	0.109	0.xxx
Dielectric 3 (capacitor)	0.260	cm^2	0.030	0.xxx
Dielectric 1 (flat)	0.047	cm^2	1.000	0.xxx
Local interconnect open lines	0.397	100 m	1.000	0.xxx
M3/M2 contact open circuits	0.201	ppm	0.027	0.xxx
Local interconnect-polysilicon-active area contact open circuits	0.007	ppm	0.004	0.xxx
M3 short-circuited lines	3.596	100 m	0.146	0.xxx
M1 open lines	0.138	100 m	0.003	0.xxx
Dielectric 2 (worst-case step)	0.661	cm^2	0.055	0.xxx
M3 open lines	2.338	100 m	0.013	0.xxx

Notes:
100 m = defects per 100 meters of length
ppm = parts per million defective
cm^2 = defects per square centimeter

yield = product of $yield_i$ for all values of i

where i is an index describing each defect type.

Product Chip Yield Analysis

The continuous production of high-yield product wafers requires understanding the impact of design, manufacturing, processing, process equipment, and testing on yield. Root cause analysis of a low-yield wafer with 54 chips, each containing 1.7 million transistors, capable of operating at 200 MHz, is a difficult task. The following section describes the testing methodology applied by the product yield enhancement engineer.

After the wafers have completed the fabrication process, all die are electrically tested. The electrical test code starts with simple continuity checks (open circuits and short circuits) on small areas of the die. Tests that require limited functionality are performed early in the test sequence. Electrical tests incrementally cover a larger area and more functionality. First-pin fail identity and parameter value (voltage, current, or test vector) are retained for each die.

The code written in the test sequence allows the yield engineer to determine the cause of the failure. For example, high current failures (short circuits) can often be correlated to metal short circuits caused by inadequate metal etch, poor planarization, or particle deposition. Electrical testing that can identify a specific area within a die as the probable failure site aids in analysis of the failure. If pin 1 is short circuited to ground (V_{SS}), it is probable that the cause of the failure is in close proximity to pin 1. Visual inspection and the scanning electron microscope (SEM) are often employed for this type of analysis. If no cause is found, more intensive failure analysis is pursued.

Functional testing in a production environment limits the amount of data that can be stored. The testing often stops at the first test failed. It is therefore important that initial functional tests require only minimal functionality. Stored data that identifies the failing pin and test vector provides the ability to perform commonality studies on manufacturing data. If a failure mode can be isolated, analysis is simplified.

Tests that are similar or specific to an area are grouped in bins. For example, a die that fails because of a short-circuited pad is collected in a bin labeled "CS"; a die that fails a vector for floating point is stored in a bin labeled "FBOX" (functional failure in the module where the floating point is processed); fully functional die are stored in bins labeled "$$." These bins can be analyzed for trends and commonalities.

The probe failure bins can be analyzed by several methods. Composite wafer mapping is useful for describing the failure pattern. A composite map graphically displays a probe bin for the entire lot as a percent of die failing a certain type of test. Composite maps can quickly show if lots have common causes of failure. For example, after extensive analysis, the cause for a certain failure pattern can be correlated to a certain process operation. Analysis of composite maps from other lots quickly reveals if they were affected by a similar process operation. An example of a composite map for silicide over growth is shown in Figure 8.

BINCODES: CS

	1	2	3	4	5	6	7	8	9
1			0	13	0	0	0		
2		0	13	0	13	0	13	0	
3	38	13	0	0	100	100	0	0	13
4	0	0	13	88	88	88	88	0	0
5	13	0	13	100	100	88	88	0	0
6		0	0	75	75	75	0	0	
7			0	13	50	50	13		

Figure 8 Composite Wafer Map for Silicide over Growth on CMOS-4 Process

Another method to correlate probe results to a process change is to plot probe data on a cumulative sum control chart. Typically, probe data is plotted as a response to the sequence of a lot (wafers are processed in lots of 22 each) through fabrication processes. The effectiveness of employing these control charts relies on two factors: the sequencing of lots must be randomized from one fabrication process to the next, and process changes must be meticulously recorded. When a slope inflection point correlates to the date of a process change, a highly probable cause for change in probe data can be identified, as shown in Figure 9. If two or more systems are used for a particular process interchangeably, a control chart for each system can be generated. For example, an aluminum etcher with low etch rate may consistently cause more CS probe failures than another aluminum etcher with high etch rate.

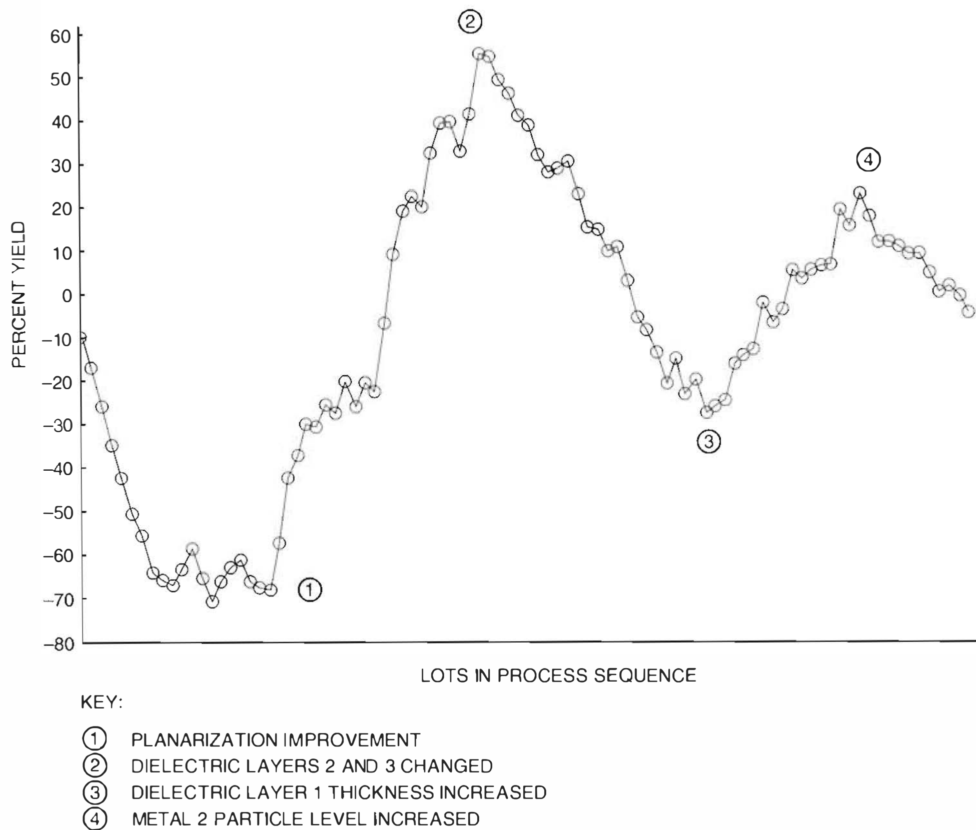


Figure 9 CMOS-4 Yield Cumulative Sum Control Chart

Recording lot process history is crucial in performing probe yield analysis. As a CMOS-4 wafer is processed, hundreds of parameters are recorded. All of these process parameters must be easily accessible to software tools for analysis. Equally important, all the process parameters must be under statistical process control. Poor probe yields with occasional high-yielding lots indicate a poorly controlled process. Probe analysis of a process with wide variability often reveals several causes and pertains only to a single lot. Probe analysis of a process under statistical control normally identifies limited causes and represents the majority of the lots.

Commonality studies to investigate process-related causes for differences between high- and low-yielding lots are often performed. Software tools extract all process-related parameters concerning the two groups of lots. Each group is analyzed for a common parameter that is different from the other group. Commonality studies that identify processing differences between high- and low-yield lots are often confirmed by experimental analysis.

The CMOS-4 process was debugged and qualified using a 128Kb SRAM. As discussed earlier, the SRAM is designed to offer increased analysis capability over custom-designed CPUs. As the process converts to product, actual product yields may differ from projected product yields. The yield engineer needs to understand the similarities and differences among the chips. The SRAM yield determined by the lower level processes (under M1) is very similar to the Alpha 21064 chip yield because large areas of the 21064 chip are designed the same as the SRAM. The upper layers of the 21064 chip can deviate in layout from the SRAM and can respond differently to variations in the process.

Due to the circuitry of the CMOS-4 die, foreign particle control and monitoring are critical. Nearly every process step deposits some particles. Particle size and type are important parameters to correlate to yield. These parameters are correlated with the aid of the defect density structures. The final and most significant correlation must be done to product yield. Visual inspection and characterization of particles on failing die compose the first-order

analysis. The second order of analysis is to correlate particle size and distribution (measured with automated laser inspection tools) to yield. This analysis defines the type of particles and the particle size that will impact yield most significantly. Furthermore, it uncovers the sources of the defects that most contribute to yield loss. Yield engineers can then prioritize defect reduction.

Yield Forecasting

The yield model is often used during feasibility studies to forecast the yield of a planned chip prior to its design. Typically, the process of predicting the yield of a new planned integrated circuit chip starts by examining the basic layout of a chip. As shown in Figure 10, the structure of a chip is partitioned into functional subblocks. By understanding how much a subblock will change from its previous use, yield engineers can estimate changes to the new subblock. Frequently, a subblock will not change enough to cause its yield to be significantly different from the layout used in a previous chip. These subblocks are available in a library of artwork layouts kept in the form of their extracted critical or susceptible areas.[15]

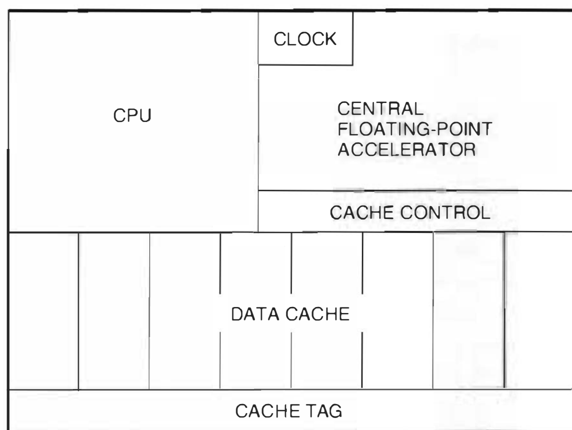


Figure 10 Simplified Chip Layout

In the next step, the critical areas estimated from each subblock are used to obtain an estimate for the entire chip. Cache memories are usually added to the random logic areas by taking critical areas of memory support circuitry and adding the multiplication of the cell areas to the number of total bits required in a cache array. Once the total critical area estimates are complete, the defect density goals and targets for each layer are used to project a yield estimate for the foreseeable life of the product.

If the subblock is new, an estimate can often be made by understanding the type of logic or circuitry being considered. If the circuitry is pure random logic, buses, or memory-like (e.g., in a cache controller), then the critical area estimate of that subblock will assume the artwork properties of these circuits. This scenario may appear overly complex, however, when the chip being estimated is 2 to 3 cm^2 in footprint area and is tightly packed with minimum ground rule artwork, this complex procedure is necessary for a reasonable yield estimate. A reasonable estimate is considered to be within ± 20 percent of the actual yield; if one used more simplified approaches, errors up to 300 percent could easily occur.

Figure 11 compares the forecasted yield and the actual yield for the 128Kb SRAM and the Alpha 21064 microprocessor. The vertical axis is the forecasted chip yield normalized to a relative scale. The open circle is the 128Kb SRAM chip forecast, and the closed square is the actual SRAM yield. The Alpha model yield is plotted with a closed circle, and the Alpha actual yield is plotted with an open square. The estimates were made approximately 4 to 6 months prior to the product chip being prototyped in the fabrication facility. The projected estimates

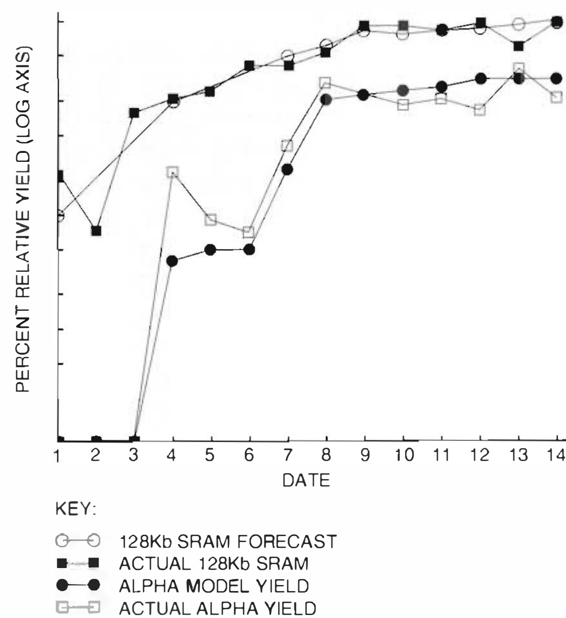


Figure 11 CMOS-4 Yield Forecast Projections

are relatively close to the actual chip yields during the same manufacturing time frame.

Redundancy Yield Models of Caches and Memory Chips

Relatively large cache RAMs are being used today on microprocessor chips to increase the processor's bandwidth performance. The bit capacity typically ranges from 100Kb to 1.5 megabits. This cache memory size can use up to 50 percent of the entire footprint area of a microprocessor chip. Because of these dimensions and bit capacities, on-board spares are sometimes used to provide fault tolerance for the processor's cache memory. If the number of repairable faults is less than or equal to the number of available spares, the chip can be repaired to a fully functioning device.

Redundancy Yield Model The redundancy yield model consists of two sets of parameters. One set characterizes the process and the other set describes the product. The set of defect densities modeled in an older generation process is shown in Table 6; this set of parameters characterizes the random defects in the process. These defect densities can be expanded in mathematical form to

$$\hat{\lambda} = \mathbf{A}_c \cdot \hat{D} \quad (3)$$

where $\hat{\lambda}$ is a vector representing the four fault types of memory pattern bit failures as given below.

1. Single bits
2. Single word lines
3. Single bit lines
4. Chip kill

These fault patterns are the parameters describing the random faults on the memory portion of the product. $\mathbf{A}_c$ is the critical area matrix of the product, and $\hat{D}$ is the vector of defect densities at all modeled layers. The critical area matrix relates the defect densities modeled in the process to the cache memory circuit fault patterns of the product as given in Table 7. These arrays of numbers represent the sensitivity of the cache memory circuit to random defects. The sensitivity to defects is obtained by calculating the critical area as described in the yield model section of this paper.

The probability of failure for a given size defect is the fraction of defects of that size which has been determined to have caused a fault. These probabilities of failure have been analyzed using Monte Carlo simulation techniques using the VLASIC yield simulator as described by Walker.[3] By establishing libraries of critical areas for different circuits, a cache memory can be simulated to the number of total bits required by the design. The random logic critical areas are thus lumped into the chip-kill category as seen in Table 7. In this fashion, the mean number of fails per chip for each circuit fault type can be computed. In many circumstances, the mean number of fails per chip can be obtained from existing memories that are similar in design to the future cache memory. Fault statistics can then be collected, and the failure distribution can be modeled by using the negative binomial

Table 6 Defect Density Types

D_o Parameter	Units
Dielectric 1 (flat)	cm^2
Dielectric 2 (worst-case step)	cm^2
Dielectric 2 (flat)	cm^2
Polysilicon short circuit	100 m
M1 short circuit	100 m
M2 short circuit	100 m
Polysilicon open circuit	100 m
M1 open circuit	100 m
M2 open circuit	100 m
N-gate area	cm^2
P-gate area	cm^2
Gate meander	100 m
Gate bird's beak	100 m
M2/M1 contact	ppm
M1-polysilicon contact	ppm
M1–n+ contact	ppm
M1–p+ contact	ppm
Active area short circuits n+	100 m
Active area open circuits n+	100 m
Active area short circuits p+	100 m
Active area open circuits p+	100 m
Dielectric 3 (cap)	cm^2
Dielectric 3 (worst-case step)	cm^2
M3/M2 contact	ppm
M3 open circuit	100 m
M3 short circuit	100 m

Notes:
100 m = defects per 100 meters of length
ppm = parts per million defective
cm^2 = defects per square centimeter

Table 7 Critical Area Matrix for Processor Chip with Cache Memory

	Fault Type			
Defect Type	Single Cells	Single Word Lines	Single Bit Lines	Chip Kill
Dielectric 1	0.016560			0.020250
Dielectric 2 (worst-case step)	0.088800	0.000050	0.002430	0.141500
Dielectric 2 (flat)				0.018167
Polysilicon short circuit	0.016060			0.033670
M1 short circuit	0.052460	0.000304		0.170150
M2 short circuit	0.060980		0.000149	0.157040
Polysilicon open circuit	0.048640			0.114590
M1 open circuit	0.065060	0.000374		0.890415
M2 open circuit	0.065600		0.000175	0.116670
N-gate area				
P-gate area	0.014060			0.023650
Gate meander	0.037340			0.026550
Gate bird's beak	0.009220			0.002770
M2/M1 contact	0.230700			0.108990
M1-polysilicon contact	0.296600			0.082260
M1-n+ contact	0.426600			0.227200
M1-p+ contact	0.218400			0.142300
Active area short circuits n+	0.002733			0.012060
Active area open circuits n+	0.002730			0.012060
Active area short circuits p+	0.012420	0.000081		0.077647
Active area open circuits p+	0.012420	0.000081		0.077674
Dielectric 3 (cap)				
Dielectric 3 (worst-case step)				0.003353
M3/M2 contact				0.021534
M3 open circuit				0.000250
M3 short circuit				0.000335

distribution with $\hat{\lambda}$ and α as parameters.[10,16,17] This is accomplished through use of the probability model given as

$$P(X = x) = \frac{\Gamma(x + \alpha)}{x!\Gamma(\alpha)} \cdot \frac{(\lambda/\alpha)^x}{(1 + \lambda/\alpha)^{x+\alpha}} \qquad (4)$$

A nonlinear least-squares technique is used to fit the parameters to the observed distribution; these parameters can be used instead of defect densities if desired.[17] This alternative can give the model more flexibility, depending on which data is most appropriate to use for an estimate. Examples of the fitted distributions of single cell failures, column failures, and double column failures are given in Table 8.

The estimates of chip yield using different combinations of redundant spares and turning off banks of the cache memory have also been used in the past. Figure 12 shows the yield as a function of redundant spares and sets of banks where the total set is eight and the desired number of good banks is at least six out of eight. These computations are again performed using a model described by Stapper.[10,17,18] The yields were estimated approximately one year before actual manufacturing data existed for that product. These types of analyses are part of the feasibility studies that help the design engineers determine the optimum product yields. This model has been expanded and modified to include clustering of faults within a chip when a chip becomes rather large.[18]

Design for Manufacturability

The yield equation (1) shows that yield is a function of both critical area and defect density. Since criti-

Table 8 Fitted Distributions of SRAM Functional Failures

Faults per Chip	Observed	Predicted	Residuals
0	0.626866	0.629694	−0.002829
1	0.199627	0.200209	−0.000582
2	0.082090	0.087091	−0.005001
3	0.033582	0.041282	−0.007700
4	0.016791	0.020374	−0.003583
5	0.013060	0.010293	0.002766
6	0.007463	0.005281	0.002182
7	0.003731	0.002739	0.000993
8	0.007463	0.001432	0.006031
9	0.005597	0.000753	0.004844
10	0.001866	0.000398	0.001468
11	0.000000	0.000211	−0.000211
12	0.000000	0.000113	−0.000113
13	0.001866	0.000060	0.001806
14	0.000000	0.000032	−0.000032
15	0.000000	0.000017	−0.000017

Parameter	Final Value	Standard	Lower 95%	Upper 95%
λ	0.709781	0.019695	0.667540	0.752021
α	0.575939	0.027410	0.517150	0.634728

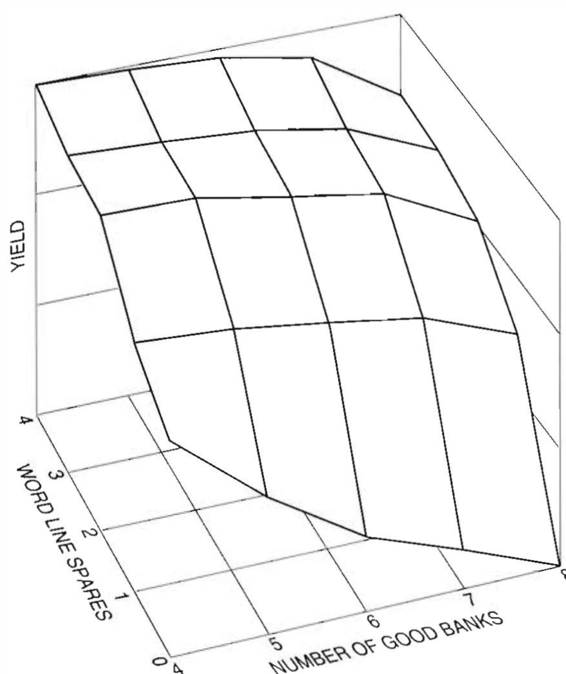


Figure 12 Chip Yield Estimate with Spare and Bank Redundancy

cal area is extracted from the circuit layout, it follows that changing the layout can affect the yield. However, the total number of die manufactured on a wafer is also a function of area, and changing the layout can also affect the number of die that can fit on a wafer. The design for manufacturability (DFM) method quantitatively analyzes these two relationships and maximizes the number of good die per wafer.

Figure 13 is a flowchart that shows the role the single layer yield model plays in process enhancement and design for manufacturability. Two loops have been outlined in the diagram: the process enhancement loop and the design for manufacturability loop.

DFM currently uses three software tools: the yield model, the cell counter, and the die counter. The yield model has been made available to design engineers for their use to model yield based on layout artwork. This software tool can be used in the design phase to evaluate the manufacturability of chip subblocks. The subblocks that produced the greatest number of good die per wafer can then be used in the chip design. Since M2 short circuits frequently cause defects, the work to date has focused on reducing the critical area of M2 short circuits. The idea is extendible to other circuit features such as line widths and contacts.

The cell counter routine counts the number of times that a specified portion of a circuit is repeated in a product layout. Critical area extractions can require large amounts of CPU time, but the cell counter allows the extraction to be performed on a small portion of the chip quickly. The result can then be multiplied by the result of the cell counter to model the yield. The cell counter can also be used during layout to determine the effect on yield of increasing or decreasing the size of the product. This technique is especially effective with very repetitive layout, for example, SRAMs.

The third software tool included in DFM is the die counter. This routine counts the number of die or chips that can fit on a wafer given the die size. The die counter is used during layout along with the yield model to optimize the number of good chips per wafer.

Conclusions

Enhancement of integrated circuit yield at times requires many and diverse analytical techniques

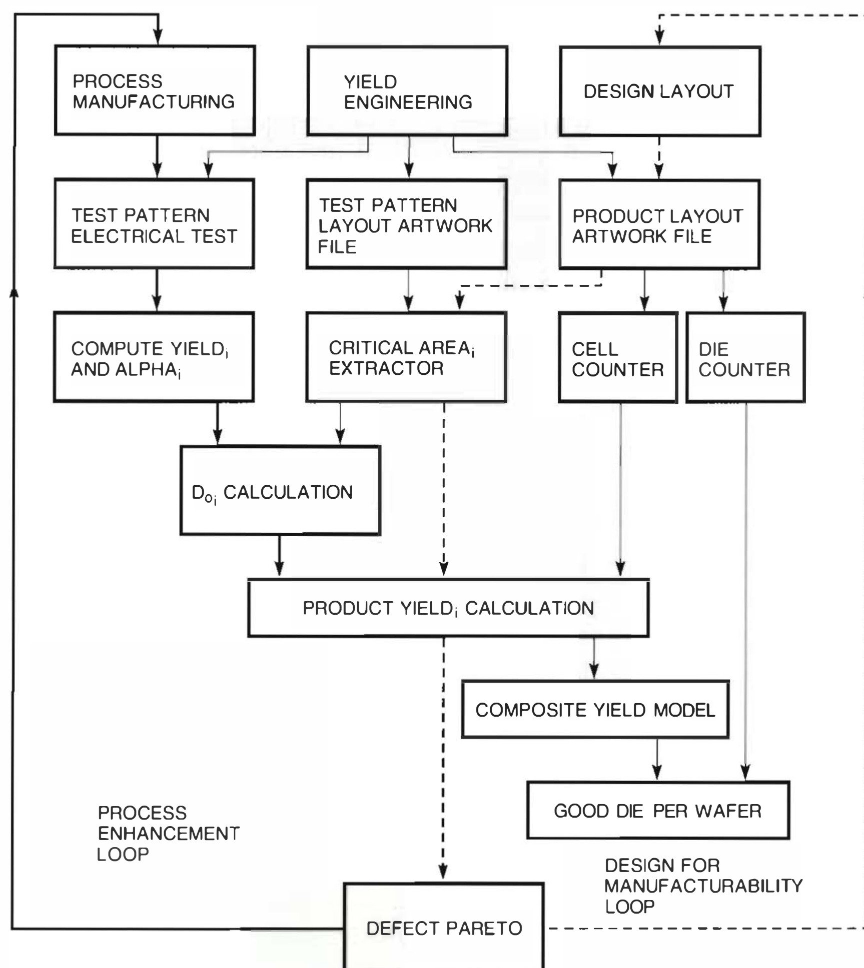


Figure 13 Yield Model/Design for Manufacturability Flow Diagram

for the detection and elimination of yield-limiting mechanisms. The techniques described in this paper discuss some of the analytical tools necessary for the successful manufacture of very large, complex CMOS digital circuits. These tools have been in use at Digital for approximately ten years, and the tool set continues to evolve with each new generation of technology. As these techniques are refined and new ones are developed, the overall understanding of integrated circuit yield and yield loss is increased.

References

1. D. Dobberpuhl et al., "A 200MHz 64b Dual-Issue CMOS Microprocessor," *IEEE International Solid-State Circuits Conference Digest of Technical Papers* (February 1992): 106-107.

2. P. Gangatirkar, R. Presson, and L. Rosner, "Test/Characterization Procedures for High Density Silicon RAMs," *Proceedings of the International Solid-State Circuits Conference* (February 1982).

3. D. Walker, "Yield Analysis for Fault-Tolerant Arrays," CMU Research Report No. CMUCAD-88-46, October 1988.

4. D. Walker, *Yield Simulation for Integrated Circuits* (Kluwer Academic Publishers, 1987).

5. D. Walker, "Experience with the VLASIC System in Defect Probability Prediction," CMU Research Report No. CMUCAD-90-41, September 1990.

6. J. Cunningham, "The Use and Evaluation of Yield Models in Integrated Circuit Manufac-

turing," *IEEE Transactions on Semiconductor Manufacturing,* vol. 3, no. 2 (May 1990).

7. T. Okabe, M. Nagata, and S. Shimada, "Analysis of Yield of Integrated Circuits and a New Expression for the Yield," *Electrical Engineering in Japan,* vol. 92 (December 1972).

8. C. Stapper, "LSI Yield Modeling and Process Monitoring," *IBM Journal of Research and Development,* vol. 20 (May 1976).

9. A. Rogers, *Statistical Analysis of Spatial Dispersion* (London: Pion Ltd., 1974).

10. C. Stapper, "Large Area Fault Clusters and Fault Tolerance in VLSI Circuits: A Review," *IBM Journal of Research and Development,* vol. 33, no. 2 (March 1989): 162-173.

11. R. Collica, "The Effect of the Number of Defect Mechanisms on Fault Clustering and its Detection Using Yield Model Parameters," *IEEE Transactions on Semiconductor Manufacturing,* vol. 5, no. 3 (August 1992).

12. J. Pineda de Gyvez and J. Jess, "Systematic Extraction of Critical Areas from IC Layouts," *International Workshop on Defect and Fault Tolerance in VLSI Systems* (October 1989).

13. A. Ferris-Prabhu, "Modeling the Critical Area in Yield Forecasts," *IEEE Journal of Solid-State Circuits,* vol. SC-20, no. 4 (August 1985).

14. C. Stapper, "The Effects of Wafer to Wafer Defect Density Variations on Integrated Circuit Defect and Fault Distributions," *IBM Journal of Research and Development,* vol. 29, no. 1 (January 1985).

15. A. Ferris-Prabhu, "Role of Defect Size Distribution in Yield Modeling," *IEEE Transactions on Electron Devices,* vol. ED-32, no. 9 (September 1985).

16. S. Kikuda et al., "Optimized Redundancy Selection Based on Failure-Related Yield Model for 64-Mb DRAM and Beyond," *IEEE Journal of Solid-State Circuits,* vol. 26, no. 11 (November 1991).

17. C. Stapper, "On Yield, Fault Distributions, and Clustering of Particles," *IBM Journal of Research and Development,* vol. 30, no. 3 (May 1986).

18. C. Stapper, A. McLaren, and M. Dreckmann, "Yield Model for Productivity Optimization of VLSI Memory Chips with Redundancy and Partially Good Product," *IBM Journal of Research and Development,* vol. 24, no. 3 (May 1980).

19. C. Stapper, "Small Area Fault Clusters and Fault Tolerance in VLSI Circuits," *IBM Journal of Research and Development,* vol. 33, no. 2 (March 1989): 174-177.

Daniel B. Jackson
David A. Bell
Brian S. Doyle
Bruce J. Fishbein
David B. Krakauer

Transistor Hot Carrier Reliability Assurance in CMOS Technologies

Hot carrier-induced degradation of MOS transistors is an essential consideration in the development of CMOS processes. Most empirical approaches that characterize transistor hot carrier lifetime only provide indications of relative degradation; they do not make a connection between circuit operation and hot carrier degradation under experimental stress conditions. Digital's Advanced Semiconductor Development Group has devised a physically based method for ensuring that the hot carrier lifetime of transistors produced by a new process technology is acceptable. The models used incorporate degradation under three voltage bias conditions and allow for the effect of dominant manufacturing variations on transistor hot carrier lifetime. The method also takes into account the sensitivity of the circuit design to transistor hot carrier degradation. This hot carrier reliability assurance gives developers the ability to predict circuit hot carrier lifetime and thus allows them to maximize transistor performance.

Hot carrier-induced degradation of metal-oxide semiconductor (MOS) transistors is an essential consideration in complementary metal-oxide semiconductor (CMOS) technology development. The reduction of metal-oxide semiconductor field-effect transistor (MOSFET) channel dimensions to micron and submicron sizes has placed increasing importance on the reliability of the gate oxide and its interface with the underlying silicon. Achieving optimum transistor performance while maintaining the necessary circuit reliability is fundamental to manufacturing high-performance CMOS microprocessors.

The hot carrier-induced transistor degradation arises from the high energy acquired by channel carriers, either electrons or holes, as they move from the MOSFET source to drain.[1-8] The electric field through which the carriers move has increased with succeeding generations of process technology because transistor dimensions have scaled faster than operating voltages. The high energy of the channel carriers leads to a gradual degradation of the transistor characteristics through charge trapping in the gate oxide and generation of interface states. This hot carrier-induced degradation can become large enough to cause circuit failure.

Consequently, much work has been devoted to understanding hot carrier degradation, with a view toward developing accurate prediction techniques for both transistor and circuit lifetime, such as those discussed in this paper.[5-8] Substantial progress has also been made in increasing the hot carrier robustness of transistors through drain junction design and gate oxide optimization; a discussion of this is beyond the scope of this paper however. Most of the effort to date has been focused on n-channel transistors, where, for current semiconductor process technologies, the degradation is more severe than in p-channel transistors. However, for technologies with an effective channel length of less than 0.5 micron (μm), p-channel hot carrier effects will become increasingly important. This paper thus focuses on n-channel transistors.

A number of empirical approaches have evolved in the semiconductor industry to characterize transistor hot carrier lifetime. Most of these approaches provide indications of relative degradation, allowing the comparison of different transistor designs. However, these schemes do not make a connection between circuit operation and hot carrier degradation under experimental stress conditions. The lack of a method for predicting circuit lifetime can

lead to a conservative approach of compromising transistor performance in order to improve transistor hot carrier lifetime, whether this improvement is warranted or not.

This paper presents a physically based method for determining whether the hot carrier-induced degradation in transistor characteristics is acceptable. The procedure is divided into two parts, an experimental measurement of transistor degradation, and a determination of the maximum permissible transistor degradation for continued circuit operation. For convenience, we will refer to both transistor lifetime and circuit lifetime. The circuit lifetime is the length of time a circuit will operate in conformance to its stated specifications. The transistor lifetime is the time it takes a transistor under stress to reach a chosen degree of degradation. The Circuit Considerations section shows that these two lifetimes are not necessarily the same. In practice, of course, it is the circuit lifetime that is important. A significant contribution of the work described in this paper is the ability to predict circuit lifetime from transistor lifetime.

The method begins with experimental measurements of transistor degradation under static stress conditions, which results in three fundamentally different types of damage to the transistor. Using these measurements, a model is developed that allows the determination of transistor lifetimes under dynamic bias conditions for an assumed maximum acceptable degradation. Also, for a given circuit, the method specifies the worst-case transistor produced as a result of manufacturing process variations in such parameters as gate oxide thickness and channel length. The circuit-dependent part of the method relies on two important quantities: the set of worst-case, time-varying biases seen by transistors in the circuit, and the maximum amount of transistor degradation that the circuit can tolerate and still remain functional.

Physical Mechanisms—Measurement of Transistor Degradation

Hot carrier degradation in a MOS transistor is usually localized to the region where most of the voltage drop between the drain and the source occurs, i.e., between the pinch-off point and the drain junction, as shown in Figure 1. The size of the high field region near the drain junction is typically on the order of 0.1 μm. In this region, the charge carriers in the inversion layer are accelerated by the high field and become energetic or "hot." Since the mean free path of an electron in silicon is small, approximately 60 angstroms, most electrons lose the excess energy they acquire by moving through the high field region via collisions with lattice phonons. However, some fraction of the electrons will traverse the high field region without suffering enough collisions to lose all of the energy gained from the electric field. These electrons are the hot carriers that cause degradation in transistor characteristics.

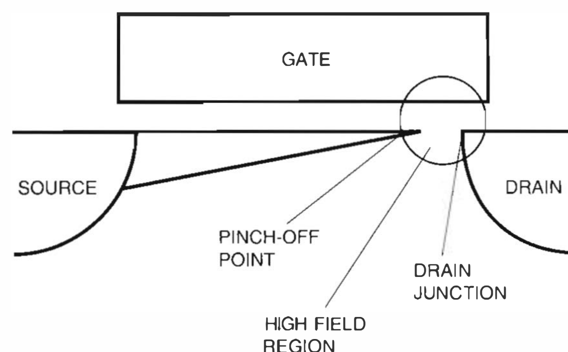


Figure 1 Schematic of a Transistor in Saturation Showing the Device Biased into Pinch-off

Under conditions that generate hot carriers, a relatively large number of "lucky" electrons (those suffering few collisions) can gain the 1.7 electron volts in energy necessary for electron hole avalanche generation. The generated holes move out through the substrate and can be measured as a substrate current I_B. Fewer of the lucky electrons will gain the 3.1 electron volts in energy necessary to overcome the silicon/silicon dioxide (Si/SiO_2) barrier and pass into the gate dielectric. At sufficiently high voltages, the holes created in the avalanche can also have enough energy to be injected into the gate dielectric of the transistor. Because the avalanche-generated holes can be measured as a substrate current, substrate current is frequently used as a measure of the driving force for hot carrier degradation. It gives a convenient measure of the number and energy of high energy carriers in the pinch-off region.

Location of Hot Carrier Damage

The injection of charge into the oxide is, by itself, not a cause for concern. Measurements using the

floating gate technique, for example, show that this current is on the order of picoamperes per μm of gate width for electron injection at high drain voltages V_{DS}.[9] For hole injection, the current is several orders of magnitude smaller.

However, a problem does arise from the small fraction of the injected charge trapped in the oxide, N_{ox}, and the interface states generated by the hot carriers, N_{ss}. The oxide and interface damage causes changes in the linear region transconductance g_m and/or the threshold voltage V_T, as well as in the saturated drain current I_{DSAT} of the transistor. The I_{DSAT} of the transistor is the drain current with the gate and drain voltages equal to the positive power supply voltage V_{DD} and the source voltage equal to the negative power supply voltage V_{SS}.

Oxide traps and interface states that are uniformly distributed along the channel can be identified from the poststress current versus voltage characteristics.[10] A uniform N_{ss} causes a change in the subthreshold current characteristics, and a uniform N_{ox} causes a shift in the V_T. However, hot carrier stress damage is not uniform. It is localized to the region around the drain junction shown in Figure 1. The degradation resulting from this nonuniform damage is limited primarily to gate voltages V_{GS} above V_T, i.e., the drain current I_D is reduced when V_{GS} is greater than V_T, irrespective of whether N_{ox} or N_{ss} is created.[11]

This interpretation is supported by two-dimensional simulations.[12] The similarity in effects of interface states and oxide traps has resulted in a poor understanding of the types of damage that occur in hot carrier stressing and, in turn, has led to difficulties in predicting transistor lifetimes under dynamic stress conditions.

Kinetics of Damage Evolution

Figure 2 shows typical stress results at five different drain voltages. The gate voltage in each case corresponds to the peak substrate current. Although many criteria are used to monitor stress damage, e.g., V_T, g_m, and linear I_D, the criterion used to obtain the results presented in this section is the change in the saturated drain current when both V_{DS} and V_{GS} equal V_{DD}, I_{DSAT}. From a circuit standpoint, I_{DSAT} has been identified as one of the most meaningful hot carrier monitors. Figure 2 illustrates that, at all stress voltages, the degradation as a function of time t obeys the equation

$$\Delta I_{DSAT} = K \cdot t^n, \tag{1}$$

where the constants K and n are empirically determined from experimental data; the value of n is usually between 0.3 and 0.7. From curves such as those shown in Figure 2, it is possible to interpolate or extrapolate to a certain level of degradation and obtain the transistor lifetime at a given voltage. The choice of the transistor lifetime criteria, in this case 5 percent change in I_{DSAT}, is discussed more fully in the Circuit Considerations section.

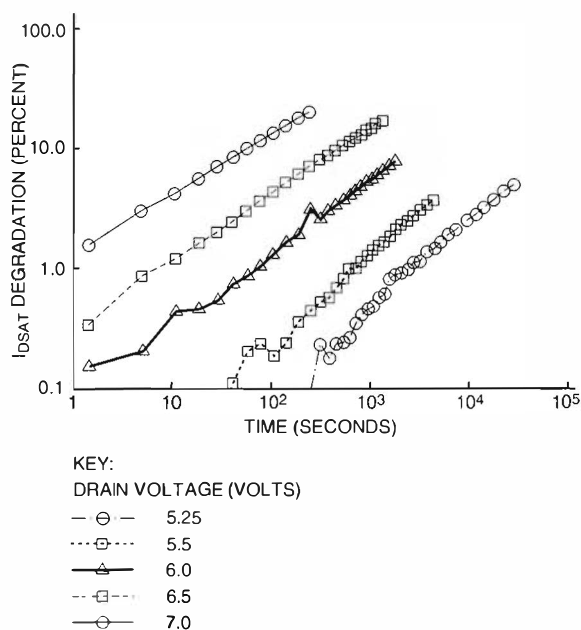


Figure 2 Degradation as a Function of Time for Devices Stressed at Different Drain Voltages

Three Types of Stress Damage

In the stressing of n-channel metal-oxide semiconductor transistors, there are three major regions of damage in the gate voltage range.[4-8] The three regions are distinguished by the charge injected into the oxide during hot carrier stress. Figure 3 shows the gate current I_G as a function of gate voltage for a fixed drain voltage. The figure is divided into three regions: the low gate voltage region, region I, in which holes are the predominant component of the gate current; the medium gate voltage region, region II, where both electrons and holes are injected in approximately equal numbers; and the high gate voltage region, region III, where electrons are the main current species. In region I, the predominant damage mechanism is the genera-

tion of electron traps in the bulk oxide by the injected holes, $N_{ox,h}$. In region II, the dominant damage mechanism is the generation of interface states N_{ss}. In region III, the dominant damage mechanism is the generation of electron traps in the bulk oxide by the injected electrons, $N_{ox,e}$. In circuit operation, all three types of damage occur. The relative importance of each depends not only on the voltage of operation, but also on the relative rate of degradation in each region as shown in the section Dynamic Hot Carrier Effects.

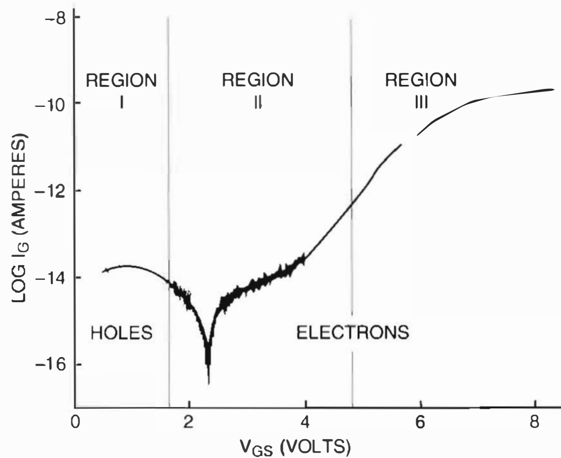


Figure 3 Gate Current as a Function of Gate Voltage for a Device Biased at High Drain Voltages

Medium Gate Voltages The most widely recognized damage mechanism occurs in the medium gate voltage region and is caused by interface state generation. Curve (b) in Figure 4 shows the amount of degradation suffered as a function of gate voltage for a series of devices stressed at a fixed drain voltage. The amount of degradation peaks at the same gate voltage as the peak in substrate current, depicted in curve (a). A direct relationship exists between the degradation and the peak substrate current during stress.[3] The transistor lifetime under stress conditions that generate interface states is given by

$$\frac{1}{\tau_{N_{ss}}} = A \cdot I_B^m, \qquad (2)$$

where $\tau_{N_{ss}}$ is the transistor lifetime for stress in region II of Figure 3, and I_B is the substrate current. The constants A and m are established by fitting the experimental data; the value of m is usually about 2.9. Figure 5 shows the dependence of $\tau_{N_{ss}}$ (obtained from data similar to that in Figure 2) on I_B. Under these gate voltage conditions, transistor lifetime can be predicted simply by establishing the values of A and m in equation (2) and extrapolating to the known substrate current at the operating drain voltage.

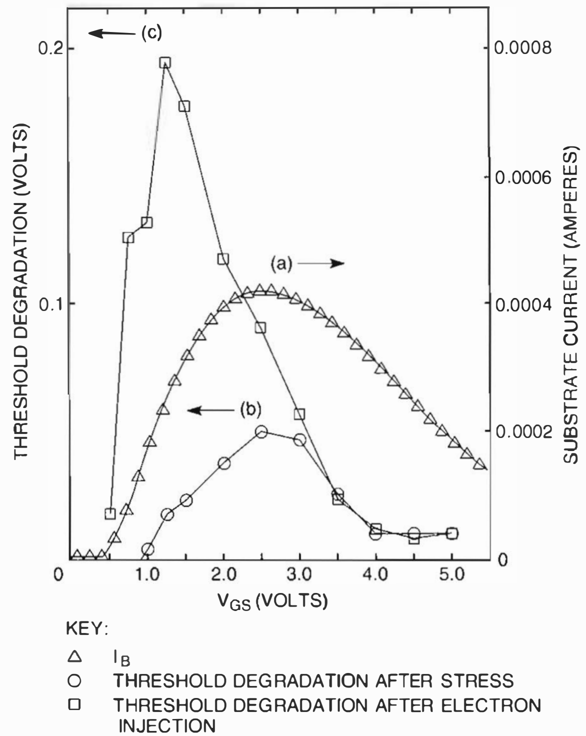


Figure 4 Degradation as a Function of the Gate Voltage

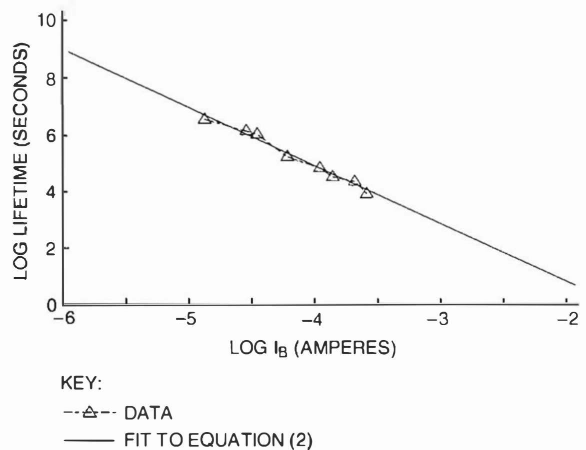


Figure 5 Transistor Lifetime as a Function of the Substrate Current

Low Gate Voltages The second type of damage results from the injection of avalanche-generated hot holes into the oxide. These injected holes can create neutral electron traps and hole traps.[5,7,13] The hole traps do not significantly affect transistor lifetime estimates and will not be discussed extensively in this paper. It was not evident in previous work that electron traps were being created. Since the traps were neutral, their effects on the current versus voltage characteristics were not observed. However, if the stressed transistors of Figure 4 are injected with electrons (brief stress with high V_{GS} equal to V_{DS}), the neutral electron traps become charged and contribute to the degradation, as shown in curve (c). The lifetime due to electron traps created at low gate voltage obeys

$$\frac{1}{(\tau_{Nox,h} \cdot I_D)} = B \cdot (I_B/I_D)^n, \quad (3)$$

where $\tau_{Nox,h}$ is the transistor lifetime for stress in region I of Figure 3.[7] I_B and I_D are the substrate and drain currents during stress. The constants B and n are determined by fits to the experimental data; the value of n is usually in the range of 7 to 12. Figure 6 shows the transistor lifetime for oxide trap damage created at low gate voltages as a function of the ratio of the substrate current to the drain current. Using equation (3), the transistor lifetime for electron trap damage at low gate voltage can be found at any given drain voltage.

High Gate Voltages The third type of stress damage in NMOS transistors occurs at high gate voltages, under conditions that inject electrons into the oxide (region III of Figure 3). This damage, caused by electron trapping, was first identified through the different gradient obtained when plotting degradation against time, as shown in Figure 6. Similar to equation (1), equation (4) for the time behavior is

$$\Delta I_{DSAT} = D \cdot t^k, \quad (4)$$

where the constants D and k are determined by fitting the data; the value of k is usually between 0.15 and 0.35. The same procedure used to obtain the results presented in Figure 2 was used to predict a transistor lifetime for this type of damage. That is, apply stress to a number of devices at different drain voltages (with V_{GS} equal to V_{DS}), and carry out the extrapolation/interpolation to obtain the transistor lifetimes at the operating drain voltage. The transistor lifetime can be plotted as

$$\frac{1}{(\tau_{Nox,e} \cdot I_D)} = C \cdot (I_G/I_D)^l, \quad (5)$$

where $\tau_{Nox,e}$ is the transistor lifetime under stress in region II of Figure 2. I_D and I_G are the drain and gate currents during stress. The constants C and l are determined by fits to the experimental data; the value of l is usually about 2.9. Figure 7 indicates that the data for these stresses is linear, where V_{GS} is equal to V_{DS}, allowing for the extrapolation and prediction of transistor lifetimes at different drain voltages for this type of stress damage.

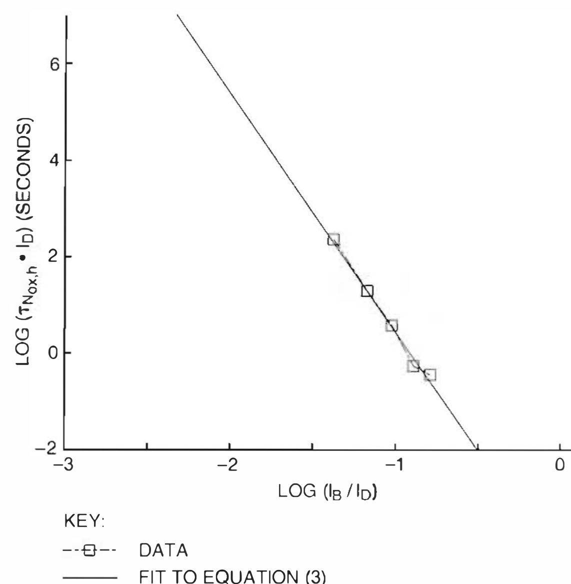


Figure 6 Transistor Lifetime for Oxide Trap Damage Created at Low Gate Voltage

Process Variations

In addition to understanding the effects of different applied voltages on the transistor, it is important to consider the effect of manufacturing process variability on hot carrier reliability. For a given source/drain design, the major cause of that variability is variation in the transistor effective channel length L_{eff}. Of secondary importance is variation in the gate oxide thickness t_{ox}. From a performance standpoint, it is desirable to make L_{eff} and t_{ox} as small as possible in order to get the highest saturated drain current I_{DSAT}, and thereby the highest speed. However, as L_{eff} and t_{ox} decrease, the transistor lifetime rapidly decreases.

Because the damage region remains roughly constant in size, as the channel length decreases a

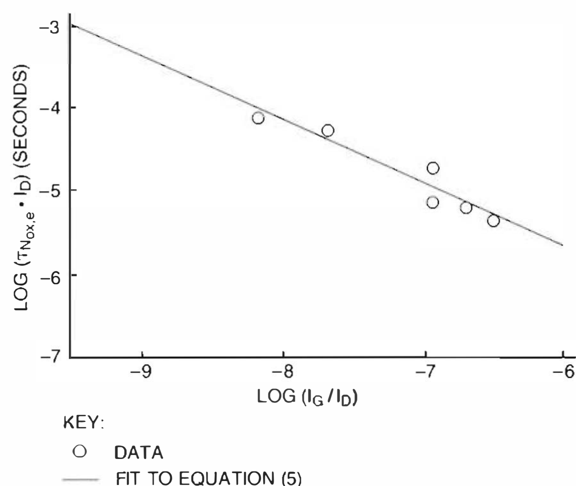


Figure 7 Transistor Lifetime for Oxide Trap Damage Created for Equal Gate and Drain Voltage Stress

larger fraction of the channel is damaged and the transistor lifetime is shorter. The direct effect of L_{eff} on transistor lifetime is approximately

$$\tau \propto 1 / L_{eff}^{n_l}, \qquad (6)$$

where n_l is usually about 3.[14] There is also an implied transistor lifetime dependence on L_{eff} through the increase in I_B that results from a smaller L_{eff}. Taking this into account, the net dependence of transistor lifetime on effective channel length is approximately

$$\tau \propto 1 / L_{eff}^{7}. \qquad (7)$$

The effect of variations in t_{ox} is accounted for in the changes in I_B, I_G, and I_D that result.

One of the keys to designing the source/drain process is to optimize the moderately doped drain (MDD) region for minimum peak substrate current under operating conditions and thus maximize the hot carrier reliability. However, once the MDD implant and diffusion process is fixed, the variation in transistor hot carrier lifetime for a given L_{eff} is relatively small. Typically, we find that for a stable process, transistor lifetime variation is less than a factor of two from one integrated circuit manufacturing lot to another.

Dynamic Hot Carrier Effects

An accurate transistor lifetime prediction model for hot carrier reliability must consider the actual stress conditions to which MOSFETs are subjected under normal circuit operation. Thus, although accurate transistor hot carrier lifetime models exist for static bias conditions, dynamic bias conditions must also be taken into account.[15,16] Initial attempts to predict transistor dynamic stress lifetimes were based on quasi-static sums of transistor static stress lifetimes.[17,18] However, these initial predictions were much longer than transistor lifetime measured under dynamic stress, because they only considered the contribution from interface state generation and omitted the effects of bulk oxide trapping.[19,20] While the enhanced dynamic stress degradation has generated much debate and numerous explanations, the effect can be explained by considering a quasi-static sum of the three damage mechanisms detailed in the previous section.[19,21]

An Accurate Dynamic Hot Carrier Lifetime Model

Our model, based on transistor static stress lifetimes, differs from previous static-based models in that it takes into account all three types of damage. During dynamic stress, with the drain voltage constant at some high voltage, a dynamic gate voltage subjects the MOSFET to the three types of damage shown in Figure 8. (The stress waveform shown in Figure 8 is for illustrative purposes only and does not necessarily reflect an actual circuit waveform.)

If the instantaneous values of I_D, I_B, and I_G are known for the dynamic stress, we can calculate the contributions of the three damage mechanisms by integrating equations (2), (3), and (5) over the time period T of the dynamic stress waveform. The

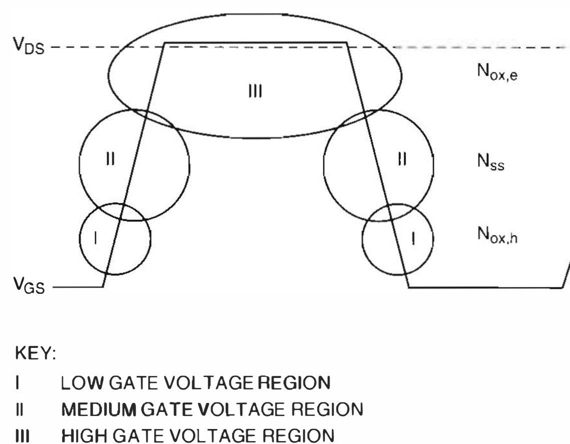


Figure 8 Schematic Representation of Dynamic Stress Waveform and Damage Regions

resulting expressions for the lifetimes due to the three physical mechanisms are given by

$$\frac{1}{\tau_{N_{ss}}} = \frac{A}{T}\int_0^T I_B^m \cdot dt, \qquad (8)$$

$$\frac{1}{\tau_{N_{ox,h}}} = \frac{B}{T}\int_0^T (I_B/I_D)^n \cdot I_D \cdot dt, \text{ and} \qquad (9)$$

$$\frac{1}{\tau_{N_{ox,e}}} = \frac{C}{T}\int_0^T (I_G/I_D)^l \cdot I_D \cdot dt. \qquad (10)$$

Treating $1/\tau$ as a damage function, the total dynamic stress damage can be modeled as

$$\frac{1}{\tau_{\text{dynamic}}} = \frac{1}{\tau_{N_{ss}}} + \frac{1}{\tau_{N_{ox,h}}} + \frac{1}{\tau_{N_{ox,e}}}. \qquad (11)$$

Example of the Dynamic Hot Carrier Lifetime Model To demonstrate the application of equation (11), consider the dynamic stress waveform in Figure 8 applied to n-channel MOSFETs. V_{GS} was pulsed between zero volts and V_{DS} with 2-microsecond rise/fall times, a 20-microsecond period, and 10 and 50 percent duty cycles. Stresses were performed for different values of V_{DS}, which was held constant during each stress.

Figure 9 shows the dynamic stress results, as a function of V_{DS}, compared to the transistor dynamic stress lifetimes predicted by a quasi-static interpretation of equation (2) and data similar to that shown in Figure 5. The measured transistor dynamic lifetime is noticeably shorter than that predicted by the quasi-static sum. The transistor dynamic stress lifetimes can be as much as two orders of magnitude lower than $\tau_{N_{ss}}$. These results demonstrate that consideration of only one damage mechanism, in this case the interface states created by medium gate voltage stress, is insufficient for predicting transistor dynamic stress lifetimes. The contributions of $N_{ox,h}$ and $N_{ox,e}$ must also be included, as indicated by equation (11).

The quasi-static contributions of the three types of damage, N_{ss}, $N_{ox,h}$, and $N_{ox,e}$, are shown in Figures 10 and 11, with 10 and 50 percent duty cycles, respectively. The curves shown were calculated from the static stress results of the previous section and equations (8)–(10), except for $N_{ox,e}$. The calculation of the quasi-static contribution of $N_{ox,e}$ requires knowledge of the instantaneous gate current which, in this case, requires time-consuming measurement techniques. Fortunately, the dynamic stress waveform chosen allows us to make simplifying approximations in the calculations of $\tau_{N_{ox,e}}$.

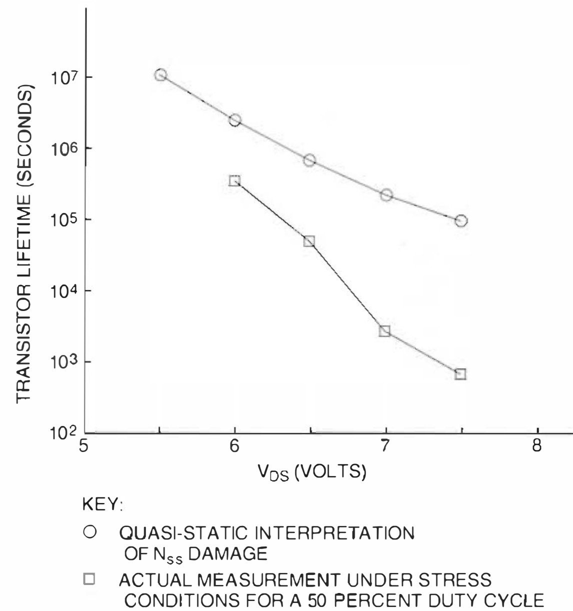


Figure 9 Transistor Dynamic Stress Lifetime as a Function of Drain Voltage

Since V_{DS} is constant during the dynamic stress,

$$\tau_{N_{ox,e}} = \frac{\tau_{N_{ox,e}(\text{static})}}{d}, \qquad (12)$$

where $\tau_{N_{ox,e}(\text{static})}$ is the transistor static stress lifetime when V_{GS} equals V_{DS}, and d is the effective duty cycle in percent. In our example, we assume that the time during which V_{GS} equals V_{DS} is large compared to the time spent during V_{GS} transition in the high V_{GS} region. Additionally, I_G drops off quickly with reduced V_{GS}. Thus, d may be approximated by the fraction of the time period that V_{GS} equals V_{DS}. As measured on an oscilloscope, the effective duty cycles were 7 and 45 percent for settings of 10 and 50 percent, respectively.

The dynamic stress model of equation (11) was applied with the data from Figures 10 and 11 to yield the transistor dynamic stress lifetime curves in Figures 12 and 13. Figures 12 and 13 also include the measured transistor dynamic stress lifetime curves for both 10 and 50 percent duty cycles and reveal an excellent match between the predicted and measured transistor lifetimes. These figures show that the inclusion of N_{ss}, $N_{ox,e}$, and $N_{ox,h}$ fully accounts for the enhanced dynamic stress degradation.

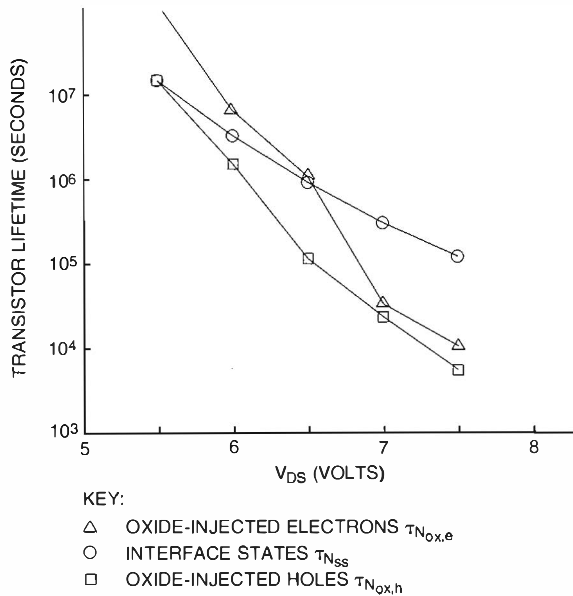


Figure 10 Calculated Contributions from the Three Types of Dynamic Stress Damage (10 Percent Duty Cycle)

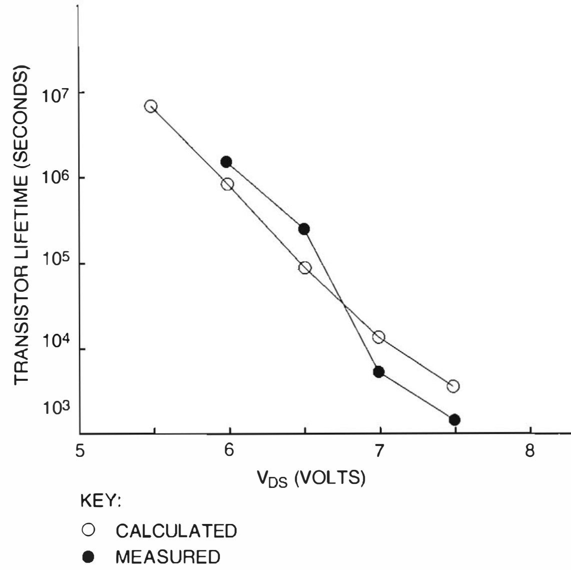


Figure 12 Measured and Calculated Transistor Dynamic Stress Lifetimes (10 Percent Duty Cycle)

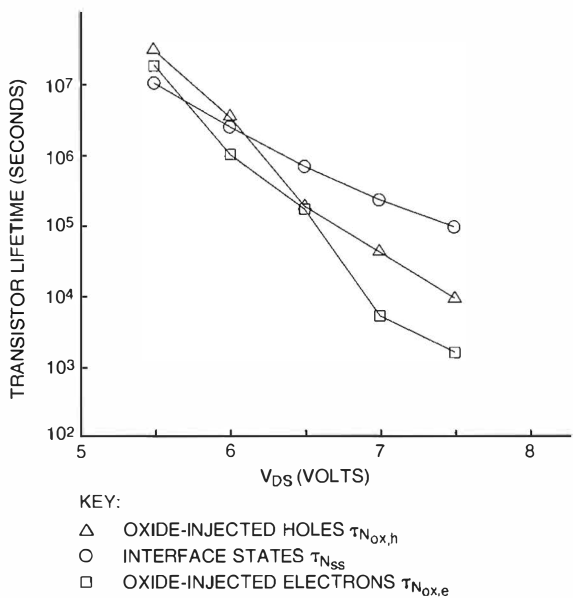


Figure 11 Calculated Contributions from the Three Types of Dynamic Stress Damage (50 Percent Duty Cycle)

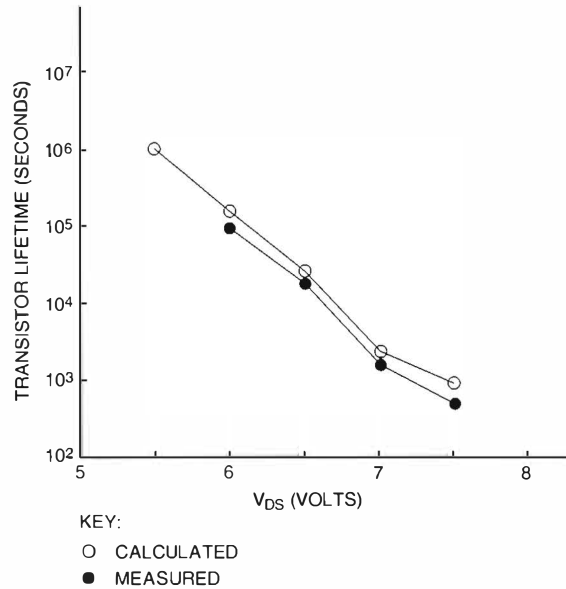


Figure 13 Measured and Calculated Transistor Dynamic Stress Lifetimes (50 Percent Duty Cycle)

Discussion of the Dynamic Lifetime Model Further examination of the contributions of each of the three types of damage illustrates their effect on transistor dynamic lifetimes relative to stress waveforms. Figures 10 and 11 indicate that $N_{ox,e}$ has the lowest transistor lifetime at high V_{DS}. Thus, the transistor dynamic lifetime is sensitive to the duration of stress in the region where V_{GS} equals V_{DS}; the contribution of $N_{ox,e}$ decreases with the duty cycle.

In the case of a 10 percent duty cycle, the contribution of $N_{ox,e}$ is reduced to the contribution level of $N_{ox,h}$. As shown in Figures 12 and 13, the measured and calculated transistor dynamic stress lifetimes are 3 to 10 times greater for the 10 percent duty cycle than for the 50 percent duty cycle.

The results presented in this section are for one CMOS technology. The relative importance of the three types of damage can be different in other CMOS technologies. Thus, different technologies can have different dependencies on the gate and drain voltage waveforms. These differences can be analyzed through the use of a SPICE postprocessor, which calculates the damage integrals of equations (8)-(10) from the gate and drain voltage waveforms.

Circuit Considerations

The transistor lifetime integrals of equations (8)-(10), along with equation (11) for the combined transistor lifetime, contain the essential transistor physics. However, a number of circuit-related considerations are important in assessing the hot carrier reliability of a process technology. These include the actual waveforms experienced by the transistor, the implications of speed binning, and the amount of degradation that will cause a circuit to fail. These topics are discussed in this section.

In-circuit Waveforms

The technique described in the previous section for determining hot carrier transistor lifetime under dynamic bias conditions can be applied to transistors in integrated circuits, if the gate and drain voltage waveforms are accurately known. The waveform shapes depend on the circuit type of which the device is a part, and also on the magnitude of switching transients and power supply noise, which can elevate the drain voltage above V_{DD}, the on-chip positive power supply voltage. The following discussion of the factors contributing to integrated circuit waveform amplitude and timing uses CMOS microprocessor design as an example.

Maximum Node Voltage Switching transients and power supply ringing on an integrated circuit can raise internal node voltages well above the nominal positive power supply voltage V_{DD}. Table 1 categorizes these effects for a microprocessor chip with a nominal power supply voltage of 3.3 volts, and shows that voltages as high as 4.3 volts are expected. In addition to the 5 percent power supply tolerance based on ripple and drift considerations, the on-chip positive and negative power supplies, V_{DD} and V_{SS}, experience ringing due to inductance in the package. The ringing worsens with increases in the package inductance and the rate at which current is drawn into the chip (dI/dt). With advances in technology, increases in clock frequency will result in more severe ringing unless accompanied by a reduction in the package inductance or an increase in the amount of on-chip decoupling capacitance. The combined effects of power supply ripple and inductive ringing can be modeled by a sine wave superimposed on the static supply with the appropriate amplitude and a frequency several times greater than the clock rate (depending on the number of clock phases).

Table 1 Contributions to the Maximum Node Voltage

Contribution	Increment (volts)	Subtotal (volts)
Nominal Power Supply	3.30	3.30
5 Percent Power Supply Tolerance	0.165	3.465
On-chip Ringing	0.175	3.64
20 Percent Capacitive Coupling	0.66	4.30

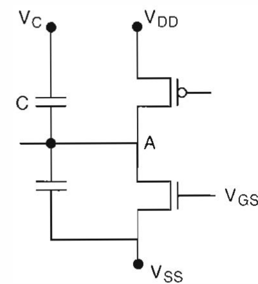
Although power supply tolerance and power supply ringing cause increases in the on-chip power supply voltage to which all nodes are exposed, some internal nodes can be booted above V_{DD} by capacitive coupling of voltage transitions on nearby nodes. The magnitude of the voltage rise above V_{DD} depends on the ratio of coupling capacitance to total node capacitance, and on the magnitude of the nearby transition. Two factors limit the maximum voltage: clamping by p^+/n diodes in p-channel pull-up transistors, when the drain voltage exceeds V_{DD} by a forward diode drop; and turning on of p-channel pull-up transistors, when the node voltage exceeds V_{DD} by a threshold voltage drop. Simulations of capacitive coupling events in CMOS microprocessors have shown that n-well resistance severely limits the diode clamping ability, and that p-channel conduction is typically weak. Furthermore, some circuit types, e.g., virtual ground circuits, contain nodes without p-channel pull-up transistors.

The voltage rise above V_{DD} on internal nodes is often limited primarily by the amount of capacitive

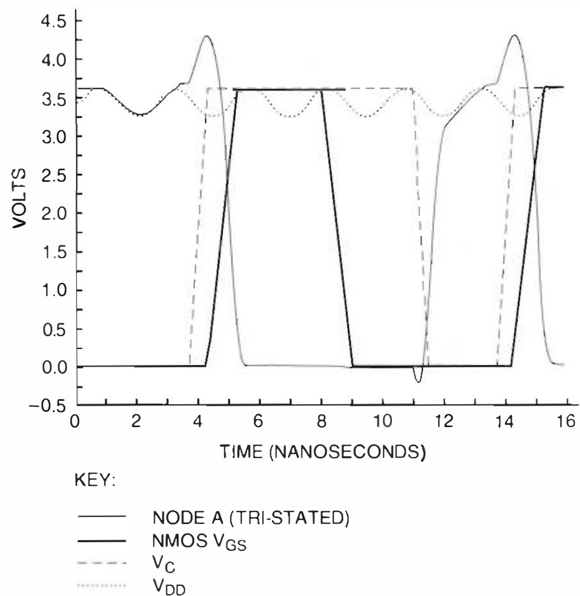
coupling. Therefore, it is important to know the maximum possible coupling. Examination of the circuits in the CMOS microprocessor example shows that not all nodes experience capacitive coupling, and only a small fraction of the nodes are expected to reach 4.3 volts during the part of their operating cycle during which hot carrier damage occurs. In some cases, capacitive coupling is restricted by noise margin considerations. For the microprocessor example, the ratio of coupling capacitance to total nodal capacitance is limited to a maximum of about 20 percent. A noise margin limit to coupling capacitance provides an upper limit for the node voltage. In general, though, it is necessary to independently determine the extent of coupling for each node in the circuit that may be affected by hot carrier degradation.

Waveform Timing The relative timing of gate and drain voltage waveforms and the extent of capacitive coupling are both circuit dependent. Therefore, it is necessary to find the worst-case combination of these factors in order to determine the circuit hot carrier lifetime for a particular chip. Analysis of four circuit types used in CMOS microprocessors (complementary drivers, pass transistors connected to storage nodes, precharge circuits, and virtual ground circuits) shows that the worst-case hot carrier conditions typically occur in precharge circuits and in tri-state driver circuits. Figure 14(a) is an example of a precharge circuit with capacitive coupling between the output (node A) and a nearby node at voltage V_C that switches from V_{SS} to V_{DD}. The worst-case waveform for this circuit is shown in Figure 14(b). This waveform was constructed using the circuit simulator SPICE, assuming worst-case capacitive coupling for that circuit type (20 percent) and phasing between the power supply noise and the coupling event to produce the maximum drain voltage. The factors listed in Table 1 increase the n-channel drain voltage during the low and medium gate voltage part of the waveform. This is the region in which hole-generated electron traps $N_{ox,h}$ and interface states N_{ss} are created in the gate oxide.

Figure 15 shows a waveform for a pass transistor connected to a storage node (capacitive coupling and ringing effects not included) that differs from the precharge circuit described in the previous paragraph in the timing between the gate and drain signals. The longer duration of simultaneously high gate and drain voltages in this circuit make electron-generated trap damage $N_{ox,e}$ potentially worse than for the precharge circuit.



(a) Circuit Diagram



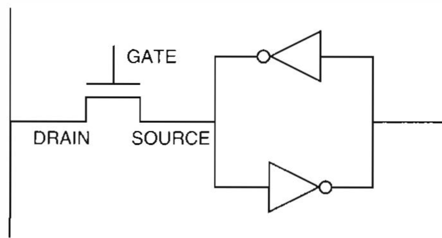
(b) Waveform

Figure 14 Circuit and Waveform of a Precharge Circuit with Capacitive Coupling

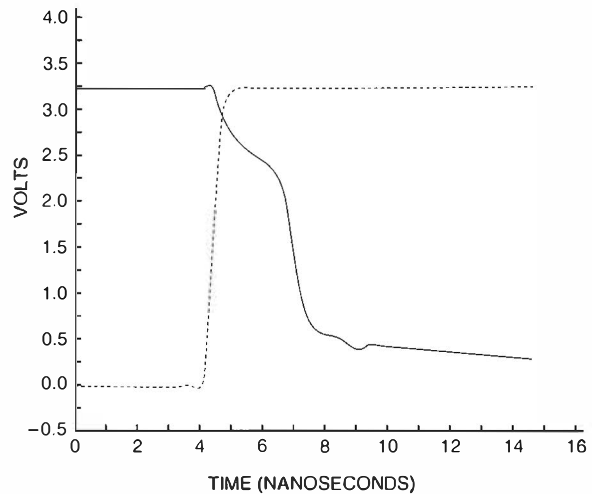
It is possible to establish broad circuit categories that are particularly susceptible to hot carrier damage. However, the complex dependence of waveform shape on circuit layout and timing requires that many nodes be examined individually to determine if they present hot carrier problems. An automated software tool such as a SPICE postprocessor can be very useful for this purpose, given that all aspects of the waveforms are modeled accurately.

Lifetime Criteria

The previous discussion applies to any transistor lifetime criterion used. To determine if a process technology produces transistors with acceptable hot carrier degradation, it is necessary to establish reasonable criteria from a circuit perspective.



(a) Circuit Diagram



(b) Waveform

Figure 15 Circuit and Waveform of a Pass Gate to a Storage Node Showing the Timing of the Gate and Drain Voltage Signals during a Read Operation

These criteria depend on the circuit design and the stated performance. For typical microprocessor circuits, the performance is stated in terms of speed. Thus for microprocessors, the most important consideration is usually how much transistor degradation will cause the speed of a part to fall below the minimum of the speed bin into which it would initially be placed. A secondary consideration is a degradation of parameters that results in loss of functionality, independent of speed.

Circuit Speed From the standpoint of hot carrier degradation, the single most important parameter affecting circuit speed is transistor saturated drain current. Therefore, the key hot carrier limit is that the degradation in transistor saturated drain current not exceed some value, usually stated as a percent of the unstressed saturated drain current.

As shown in this section, the limiting case is the slowest part in the fastest speed bin. If the fastest speed bin has a tested speed of 0.5 nanosecond faster than the stated performance, then the fractional degradation of the slowest transistor falling into the bin must be small enough that the overall circuit speed drops by less then 0.5 nanosecond. For a typical 12-nanosecond complex instruction set computer (CISC) microprocessor, this would be a few percent change in saturated drain current.

Hot carrier degradation of transistors results in a change in reverse saturated drain current that is different from the change in forward saturated drain current. Reverse saturated drain current is the transistor saturated drain current measured with the source and drain interchanged relative to the stress configuration. Thus, an additional criterion is necessary for reverse saturated drain current degradation. The sensitivity to reverse saturated drain current degradation is seen only in cases where the transistor operates in both forward and reverse directions, e.g., in pass transistors. Although generally not used in speed-sensitive situations, these transistors are sized to provide extra margin, if they are so used. Thus, the allowable percent degradation in reverse saturated drain current is typically several times that permitted for forward saturated drain current.

Circuit Functionality In addition to parameters that affect circuit speed, hot carrier stress also causes degradation in parameters that can affect circuit functionality, independent of speed. Most important among these are threshold voltage and transistor off-state current.

Transistor threshold voltage shifts can alter inverter noise margins and indirectly affect off-state current. Transistor threshold voltages are also important in cases in which matched performance of a pair of devices is required, e.g., in a memory sense amplifier. Because the threshold voltage of n-channel transistors usually increases with time (perhaps after an initial decrease of a few millivolts), off-state current does not increase. For a 3.3-volt nominal power supply and transistors with ±0.5-volt threshold voltages, a threshold shift of a few tens of millivolts is acceptable. (An increase in threshold voltage also affects transistor saturated drain current, but this is taken into

account by a separate limit on transistor saturated drain current.)

The degradation in p-channel devices is usually not important for technologies with channel lengths larger than 0.5 μm. However, very short channel length p-channel devices usually suffer a decrease in threshold voltage, which leads to increased off-state current and reduced noise margins. These effects will be of increasing importance in sub-0.5-μm technologies.

Table 2 summarizes the important transistor degradation parameters for circuit lifetime of a high-performance microprocessor.

Table 2 Typical Permitted Degradation in Device Parameters for High-performance Microprocessor Circuits

Parameter	Transistor Lifetime Drifting
Forward I_{DSAT}	3–10% Shift
Reverse I_{DSAT}	10–25% Shift
Threshold	10–100 mV Shift
Off-state Current	Absolute Limit

Note that the off-state current absolute limit may be difficult to measure, particularly at operating temperature. This limit can be incorporated into a limit in the decrease in magnitude of p-channel threshold voltage. Off-current degradation is usually not important for n-channel devices.

Speed Binning Implications

It is customary to bin the chips into one or more discrete speed categories or bins. The circuit lifetime of a given speed bin is equal to the time it would take for a worst-case device in that bin to fail. Because slower bins have a lower I_{DSAT}, i.e., a longer effective channel length, and therefore have relatively long transistor lifetimes, the lifetime issue is more of a constraint for fast bins. For the fast bins, the choice of where in the I_{DSAT} distribution the hot carrier reliability should be assessed is thus a major concern. This section discusses how to determine the worst-case transistor for a given speed bin, the relationship between this transistor and the circuit hot carrier lifetime of the bin, and the selection of I_{DSAT}.

Although it might appear that the circuit lifetime is determined by the fastest chip, which degrades more rapidly, in fact, the slowest chip in a bin will be the first to fail. Figure 16 shows the expected I_{DSAT} as a function of time for several starting I_{DSAT} values. A transistor with a higher I_{DSAT} degrades more rapidly, relative to the initial I_{DSAT}, than a transistor that began with a lower I_{DSAT}. However, even after degradation, the curves never cross, i.e., a transistor that initially had a higher I_{DSAT} continues to have a higher I_{DSAT}. If these two parts are put into the same speed bin, the faster part will take longer to fail, because it will take longer for I_{DSAT} to drop below the critical value for that speed bin. Thus, very fast chips will be reliable as long as they are not put into a very fast bin. The circuit hot carrier lifetime of a particular speed bin is limited by the circuit lifetime of the most marginal, i.e., the slowest, chip in that speed bin.

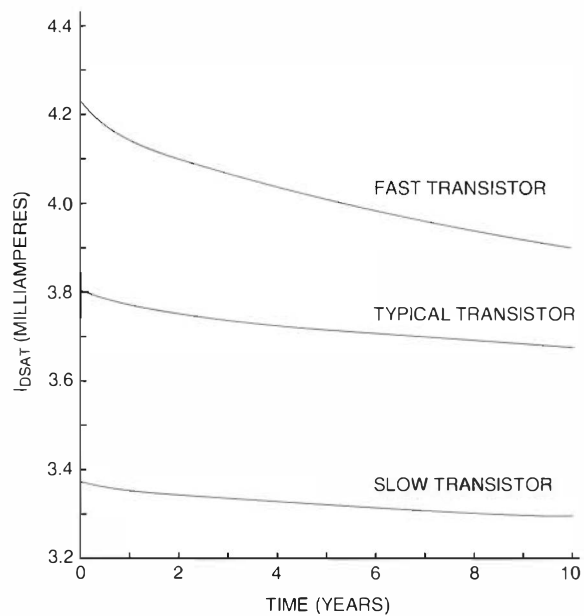


Figure 16 Predicted Saturated Drain Current as a Function of Time

One approach to verifying the reliability of a fast speed bin is to compare the distribution of the I_{DSAT} values for chips in that bin with that of chips in the next slower bin. This distribution should be compared with I_{DMAX}, which is the highest I_{DSAT} value that just meets the reliability requirements for a discrete transistor. If chips with an I_{DSAT} equal to I_{DMAX} normally fall into the fastest speed bin, then that speed bin will be reliable even if it also contains parts with an I_{DSAT} greater than I_{DMAX}. Although the faster parts degrade more quickly, because they have more margin to degrade, they do not limit the reliability of the speed bin. Thus, chip and transistor lifetimes may not be the same; if the

chip has extra margin in the speed bin, the chip hot carrier lifetime will be longer than the transistor lifetime.

The above argument applies provided there are no systematic changes in the parasitic capacitances. However, such changes, if they occur, do not present a serious problem because the key thicknesses and widths that determine the capacitances are normally monitored in production. As a result, circuit reliability is limited by the slowest device with the lowest I_{DSAT} in the fastest speed bin. Therefore, transistors used for hot carrier evaluation should have an I_{DSAT} corresponding to the slowest chip in the fastest bin, and the lifetime criteria are determined by that chip.

Conclusions

This paper presents a physical model for hot carrier degradation incorporating three damage modes:

- Medium Gate Voltages, N_{ss}
- Low Gate Voltages, $N_{ox,h}$
- High Gate Voltages, $N_{ox,e}$

A quasi-static sum of the contributions of each of these damage mechanisms accurately predicts the transistor hot carrier lifetime for any specified waveform and technology. In addition, models were developed that account for the effect on transistor hot carrier lifetimes of the dominant manufacturing variations. These models show that circuit waveforms, including power supply ringing, can have a substantial effect on circuit hot carrier lifetime. The limiting case from the standpoint of speed binning is the slowest transistor falling into the fastest speed bin. The paper gives an example of circuit lifetime criteria for a high-performance microprocessor. From these conclusions, it is possible to outline the following procedure for transistor hot carrier reliability assurance for a CMOS technology used to fabricate a particular chip:

1. For the chip to be fabricated, determine the transistor lifetime criteria. The microprocessor is only one example; other circuits may have quite different lifetime criteria. In practice, it may be better to specify the transistor degradation prior to circuit design and incorporate the lifetime criteria into the design process; i.e., as a part of the design process, make sure the design works, given the specified degradation.
2. Characterize transistor degradation under the three gate voltage ranges indicated in Figure 3 to determine the coefficients for the damage integrals, equations (8)-(10), for each of the previously determined transistor lifetime criteria. The results must then be scaled for transistors with I_{DSAT}, or L_{eff} and t_{ox} determined in step 3. Alternatively, if step 3 has already been performed, use transistors that match either the desired I_{DSAT}, or the L_{eff} and t_{ox}.
3. Determine either the I_{DSAT}, or the L_{eff} and t_{ox}, corresponding to the slowest transistor in the fastest speed bin. This can be done by simulation, or by processing a characterization lot with transistor parameters deliberately varied in order to determine the speed impact.
4. Choose several representative parts of the chip circuitry to determine worst-case, time-dependent transistor biases. For each part, using the coefficients for the damage integrals determined in step 2, calculate the transistor lifetimes. There will be several, one for each criterion. The shortest transistor lifetime is the limiting case.

The two key elements in this procedure are the three damage integrals with the combined transistor dynamic stress lifetime and the close connection with the circuit design process. In practice, several iterations of the procedure will probably take place as the transistor design and production process are optimized for maximum circuit performance and yield, consistent with the necessary reliability goals. For the best results, this needs to be done concurrently with circuit design to be sure the appropriate criteria are optimized.

Acknowledgments

The contributions of Kaizad Mistry to the understanding of the physics of degradation under dynamic stress, and Frank Fox to determining the circuit implications of hot carrier degradation have been indispensable. In addition, the contributions of Chris Conran, Hamid Partovi, Shu-Fung Hsia, Lynn Mohan, and Ann Chen are gratefully acknowledged.

References

1. P. Cotrell, R. Troutman, and T. Ning "Hot-Electron Emission in N-Channel IGFET's," *IEEE*

Transactions on Electron Devices, vol. ED-26 (1979): 520.

2. T. Ninget et al., "1 μm MOSFET VLSI Technology: Part IV—Hot-Electron Design Constraints," *IEEE Transactions on Electron Devices,* vol. ED-26 (1979): 346-353.

3. E. Takeda and N. Suzuki, "An Empirical Model for Device Degradation Due to Hot-Carrier Injection," *IEEE Electron Device Letters,* vol. EDL-4 (1983): 111-113.

4. C. Hu et al., "Hot-Electron-Induced MOSFET Degradation—Model, Monitor, and Improvement," *IEEE Transactions on Electron Devices,* vol. ED-32 (1985): 375-385.

5. P. Bellens, P. Heremans, G. Groeseneken, and H. E. Maes, "Hot Carrier Effects in n-Channel MOS Transistors Under Alternate Stress Conditions," *IEEE Electron Device Letters,* vol. EDL-9 (1988): 232-234.

6. B. Doyle, M. Bourcerie, J. Marchetaux, and A. Boudou, "Interface State Creation and Charge Trapping in the Medium-to-High Gate Voltage Range ($V_d/2 > V_g > V_d$) During Hot Carrier Stressing of n-MOS Transistors," *IEEE Transactions on Electron Devices,* vol. ED-37 (1990): 744-755.

7. B. Doyle and K. Mistry, "The Generation and Characterization of Electron and Hole Traps Created by Hole Injection During Low Gate Voltage Hot Carrier Stressing in n-MOS Transistors," *IEEE Transactions on Electron Devices,* vol. ED-37 (1990): 1869-1879.

8. B. Doyle and K. Mistry, "A Lifetime Prediction Method for Oxide Electron Trap Damage Created During Hot Electron Stressing of n-MOS Transistors," *IEEE Electron Device Letters,* vol. EDL-12 (1991): 178-180.

9. T. Hori and H. Iwasaki, "Improved Hot-Carrier Immunity in Submicrometer MOSFET's with Reoxidized Nitrided Oxides Prepared by Rapid Thermal Processing," *IEEE Electron Device Letters,* vol. EDL-120 (1989): 64-67.

10. E. Nicollian and J. Brews, *MOS Physics and Technology* (New York: John Wiley and Sons, 1982): 534-537.

11. T. Poorter and P. Zoestbergen, "Hot Carrier Effects in MOS Transistors," *Proceedings of the IEEE International Electron Devices Meeting (IEDM)* (1984): 100-104.

12. H. Haddara and S. Cristoloveanu, "Two-Dimensional Modeling of Locally Damaged Short-Channel MOSFET's Operating in the Linear Region," *IEEE Transactions on Electron Devices,* vol. ED-34 (1987): 378-385.

13. Reply to "Comments on 'The Generation and Characterization of Electron and Hole Traps Created by Hole Injection During Low Gate Voltage Hot Carrier Stressing in n-MOS Transistors'" by B. Doyle et al., *IEEE Transactions on Electron Devices,* vol. ED-39 (1992): 460-464.

14. K. Mistry and B. Doyle, "An Empirical Model for the L_{eff} Dependence of Hot Carrier Lifetimes of n-Channel MOSFETs," *IEEE Electron Device Letters,* vol. EDL-10 (1989): 500-502.

15. K. Mistry and B. Doyle, "A Model for AC Hot-Carrier Degradation in n-Channel MOSFET's," *IEEE Electron Device Letters,* vol. 12 (1991): 492-494.

16. K. Mistry, B. Doyle, A. Philipossian, and D. Jackson, "AC Hot Carrier Lifetimes in Oxide and ROXNOX n-Channel MOSFET's," *IEEE International Electron Devices Meeting (IEDM) Technical Digest* (1991): 727-730.

17. C. Hu et al., "Hot-Electron-Induced MOSFET Degradation—Model, Monitor, and Improvement," *IEEE Transactions on Electron Devices,* vol. 32 (1985): 375-385.

18. E. Takeda and N. Suzuki, "An Empirical Model for Device Degradation Due to Hot-Carrier Injection," *IEEE Electron Device Letters,* vol. 4 (1983): 111-113.

19. W. Weber, C. Werner, and G. Dorda, "Degradation of N-MOS-Transistors after Pulsed Stress," *IEEE Electron Device Letters,* vol. 5 (1984): 518.

20. W. Weber, C. Werner, and A. Schwerin, "Lifetimes and Substrate Currents in Static and Dynamic Hot-Carrier Degradation," *IEEE International Electron Devices Meeting (IEDM) Technical Digest* (1986): 390.

21. W. Weber, "Dynamic Stress Experiments for Understanding Hot Carrier Degradation Phenomena," *IEEE Transactions on Electron Devices,* vol. 35 (1988): 1476.

J. Joseph Clement
Eugenia M. Atakov
James R. Lloyd

Electromigration Reliability of VLSI Interconnect

Increased speed, reduced line widths, larger chip size, and additional levels of interconnect are all factors that contribute significantly to the improved performance and functionality of VLSI circuits. At the same time, these factors place growing demands on interconnect reliability. Therefore, careful characterization of the interconnect reliability is important in achieving VLSI performance and reliability goals. A scaling model was developed and used to examine factors essential to assuring electromigration reliability in Digital's CMOS-4 technology and in the Alpha 21064 microprocessor, which uses this technology.

Background

For complex, very large-scale integration (VLSI) circuits, the individual components that make up the circuit must be extremely reliable to assure acceptable overall reliability. Since a chip is expected to operate for many years, testing to characterize the reliability of circuit components such as interconnects must be performed under accelerated test conditions. To evaluate circuit reliability by extrapolating from data on components tested at accelerated rates, it is essential to have dependable models.

Electromigration is one of the primary failure mechanisms in the polycrystalline aluminum-alloy thin films that are widely used for interconnects on VLSI circuits.[1] Although much has been written on the effect of stress conditions on interconnect failure due to electromigration, little information is available on the use of lifetime measurements on simple test structures to evaluate the complex combination of interconnects on a VLSI chip.

In this paper, we describe work performed to characterize the reliability of the interconnect in chips manufactured in Digital's CMOS-4 technology. We place particular emphasis on a model for scaling test structure data to chip level. Before presenting the model, we discuss the physics of electromigration and electromigration testing, as background information for the reader.

Electromigration

Electromigration is the mass transport of metal atoms from collisions with the current conduction electrons. The momentum exchange resulting from these collisions creates a net flux of metal atoms in the direction of the electron flow and thus biases the normal random atomic diffusion. At sites of atomic flux divergence, where the number of metal atoms coming in does not equal the number going out, either material depletion or material accumulation occurs. Corresponding voids or hillocks form in the metal line and cause open or short circuits and, ultimately, circuit failure.

For typical circuit operating temperatures, the diffusivity of metal atoms is much higher in the grain boundaries than through the lattice of the grain itself. (A grain is an individual crystal in a polycrystalline film.) Hence, the atomic mass transport occurs primarily along grain boundaries. Therefore, microstructural inhomogeneities, such as variations in grain size, can cause a flux divergence and thus be sites of potential failure.

Enhanced mass transport along grain boundaries is one reason that electromigration is a more important failure mechanism for thin films than for bulk conductors. In thin films, the film thickness dictates that the grain size be much smaller than the grain size for bulk material. Thus, a greater proportion of the thin-film cross section can be composed of high-diffusivity grain boundaries.

A second reason electromigration is a concern for thin films is related to current density. The thin-film interconnect on integrated circuits is in intimate thermal contact with the underlying silicon substrate, which acts as a big heat sink. Therefore, thin films can withstand higher current densities than bulk wires, without incurring thermal

damage. Maximum operating current densities in VLSI circuits are as much as 100 to 1,000 times higher than normal for bulk wires. With reduced line widths and increased speed of operation, the trend is toward increased current densities in interconnects. This combination of high current densities and high-diffusivity paths along grain boundaries promotes electromigration in polycrystalline thin-film interconnects.

The addition of small amounts of alloying elements can substantially improve the electromigration performance of thin-film aluminum (Al) conductors. In particular, alloying with copper (Cu) has proven to be a popular, cost-effective way to improve lifetimes.

Electromigration Testing

Testing to evaluate electromigration reliability is accelerated using stress conditions for temperature and current that are higher than the expected operating conditions. Thus, an accurate physical model is essential for extrapolating test results to actual operating conditions.

Due to random microstructural variations, nominally identical interconnect test structures will not all fail at the same time. Typically, several identical samples are stressed together, resulting in a distribution of failure times. Knowledge of the correct failure distribution is critical for predicting electromigration reliability.

Electromigration reliability testing is usually performed by stressing packaged test lines in an oven with a constant current, while monitoring the voltage in situ. Depending on the metallization, failure may result from open circuits or an increase in resistance caused by voiding, or from the formation of extrusions that short-circuit to adjacent lines. General guidelines for electromigration test structure design and lifetime measurements are available.[2,3]

Care must be taken in the selection of the stress current and temperature, so that the failure mechanism that operates under stress conditions is the same one that works at expected operating conditions. The temperature must be kept low enough to remain in the range that grain boundary diffusion dominates. At very high temperatures, lattice diffusion becomes significant.

In addition, the current must be limited to prevent excessive Joule heating. Joule heating creates temperature gradients along the metal line. Since the diffusivity of the metal atoms has an Arrhenius temperature dependence, a temperature gradient itself can cause a flux divergence and lead to failure.

Stress Acceleration Model

A model proposed by Shatzkes and Lloyd relates the time-to-failure t_f to the temperature T and the current density j by

$$t_f = A(T/j)^2 \exp(H/kT) \qquad (1)$$

where A is a constant that depends on material properties, k is Boltzmann's constant, and H is the activation energy.[4] This formulation differs slightly from an earlier, largely empirical model proposed by Black, which did not include the pre-exponential T^2 term.[5]

The activation energy measured for aluminum and aluminum-alloy thin films is typically in the range of 0.5 to 0.8 electron volt (eV). This is significantly less than the activation energy for lattice diffusion measured in single-crystal aluminum, which is approximately 1.4 eV. These measurements indicate that the mass transport takes place along grain boundaries.

Length/Width Scaling Model

To characterize the electromigration reliability of the interconnect on VLSI chips, we developed a scaling model that considers line length, line width, and tungsten-filled vias.

The Lognormal Failure Distribution

Electromigration failure times are generally represented as a lognormal distribution, whereby a plot of the logarithms of the failure times relative to the cumulative percentage order on a normal probability scale approximates a straight line, as illustrated in Figure 1. The lognormal distribution is characterized by two parameters: the median time-to-failure, or logarithmic mean, θ, and the slope, or logarithmic standard deviation, γ. The lognormal cumulative density function $F(t)$ is

$$F(t) = \Phi((\ln t - \theta)/\gamma) \qquad (2)$$

where $\Phi(x)$ is the standard normal cumulative distribution function.

If p represents the cumulative fraction failed, and t_p is the corresponding time, then

$$\ln t_p = \theta + \gamma \Phi^{-1}(p) \qquad (3)$$

where $\Phi^{-1}(x)$ is the inverse of the standard normal cumulative distribution function. For p approximately equal to 0.16 or 0.84, $\Phi^{-1}(p)$ equals -1 and

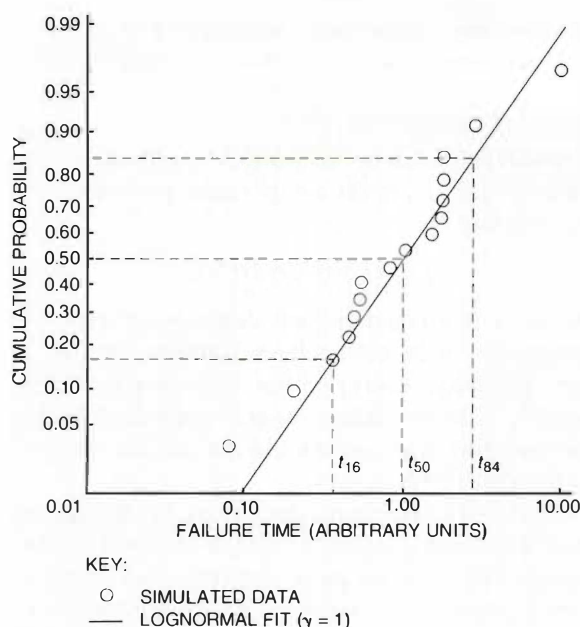


Figure 1 Lognormal Failure Distribution

+1, respectively; for p equal to 0.5, $\Phi^{-1}(p)$ equals 0. Therefore,

$$\ln t_{16} = \theta - \gamma \qquad \ln t_{84} = \theta + \gamma \qquad \ln t_{50} = \theta \quad (4)$$

These three points in time are indicated in Figure 1.

Length Scaling Effect

A major conceptual problem with the lognormal distribution, however, is that it does not scale with line length. That is, if the failure times for a given length line are lognormally distributed, the failure times of shorter or longer lines cannot also be lognormally distributed. This problem has significant repercussions in extrapolating data on test devices to predict the reliability of VLSI circuits, particularly for early failures at and below the 1 percent level.

Consider an ensemble of nominally identical lines of length l that fail over time with a distribution described by $F(t)$. We could, in principle, combine these lines to form new lines of length $L = N_l$, where N is the number of original units that make up the new line. The failure of each new line would be determined by the earliest failure time of the component elements, assuming that the unit elements behave independently. The failure distribution of the new ensemble is given by

$$G(t) = 1 - (1 - F(t))^N \quad (5)$$

If $F(t)$ is lognormal, then $G(t)$ cannot be, and vice versa.

For the particular case where $F(t)$ is lognormal, $G(t)$ has become known as the multilognormal or multiple lognormal (MLN) distribution, and is given by

$$G(t) = 1 - (1 - \Phi((\ln t - \theta)/\gamma))^N \quad (6)$$

Figure 2 contains a plot of the the MLN distribution for $N = 1$ and $N = 100$. Also shown in this figure is a lognormal curve fitted to the t_{16} and t_{84} points of the MLN curve at $N = 100$. The logarithmic standard deviation of this fitted curve, σ, is a function of γ and N.

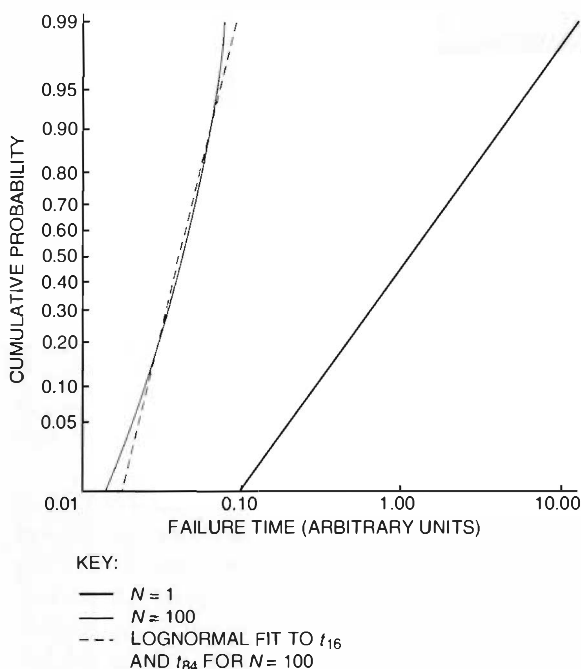


Figure 2 Multilognormal Failure Distribution

There are some important characteristics to glean from this figure. For small sample sizes, that is, over the range of cumulative percent failures between 0.05 and 0.95, the MLN distribution is nearly indistinguishable from the lognormal. However, the difference between the two distributions is more significant for early failures, at and below the 1 percent level. These early failures are of concern for reliability. Also note in Figure 2 that as N increases, the median time-to-failure t_{50} decreases, and the σ of the lognormal fitted curve also decreases. This behavior of the model agrees

with experimental observations on the effect of line length on t_{50} and σ.[6]

Moreover, recent work modeling electromigration failure in fine lines yields a failure distribution that is well approximated by the MLN distribution.[7] The agreement of the modeling results with experimental findings strongly supports the use of the MLN distribution for length scaling.

The scaling model as described by the MLN distribution in equation (6) has three adjustable parameters: θ and γ of the failure distribution for the elemental failure unit, and N, which is the total line length L divided by the characteristic elemental length l. Since a line-width dependence of the length effect has been demonstrated, some or all of these parameters must be a function of line width, as well.[6]

Width Scaling Effect

For line widths somewhat greater than the average grain size, we can expect a continuous "network" of grain boundaries along the length of the line. Since the microstructural defects leading to electromigration failure are associated with the grain boundaries, when the line width is much larger than the grain size, we might expect multiple defects to align at a failure site. The lifetime of the line, therefore, is determined by the least severe defect existing along the width. Thus, as the width increases, the probability of aligning with less severe defects rises, and an increase in lifetime is expected.

However, the expectation that the lifetime will decrease with line width does not hold for very narrow lines when a "bamboo" microstructure develops, in which a single grain boundary traverses the width of the line. As the line width decreases, the likelihood of having a single grain span the entire width of the line increases. When the line width is comparable to the average grain size, the line consists of bamboo and nonbamboo segments, as shown in Figure 3. As the line width decreases further, the fraction of the line that is bamboo increases, that is, the length and number of nonbamboo segments decreases. Since electromigration proceeds primarily along grain boundaries, failure correlates with the length and number of nonbamboo segments. Hence, as the line width decreases below the average grain size, the electromigration lifetime improves.

The dependence of electromigration lifetime on line width has been well established.[8,9,10] In the bamboo region, the lifetime increases very rapidly as the line width decreases. For line widths somewhat greater than the average grain size, the lifetime gradually increases with line width. There is, therefore, a minimum in lifetime in relation to the line width when the line width is comparable to the average grain size.

Grain size in polycrystalline thin films is influenced by a number of factors including the substrate, the method by which the film is deposited, the deposition conditions, and the grain growth due to postdeposition annealing. In addition, grain growth and grain boundary movement during postpatterning annealing can be even more significant in fine lines.[11]

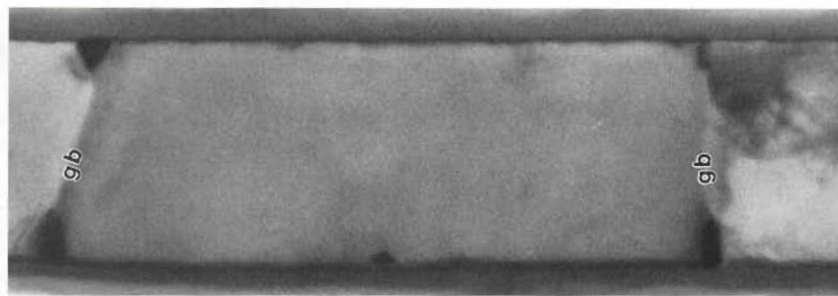
In terms of the scaling model, the expected improvement in lifetime with reduced line width is a result of two effects on the model parameters. Research suggests that the susceptibility of a grain cluster to failure decreases as the length of the cluster decreases.[9,12,13] If the time-to-failure of the unit element is related to the length of the nonbamboo grain clusters, θ increases as the line width decreases.

More importantly, as the line width decreases, bamboo segments constitute a larger percentage of the line. Thus, the number of microstructural defects per unit length decreases; that is, the characteristic unit length l increases. In our model, for a line of a given length L, as l increases, the number of elemental failure units N decreases. Consequently, t_{50} and σ increase, provided γ does not change significantly. The increase in t_{50} and σ with decreasing line width given by the model is in complete agreement with the results obtained from electromigration life tests. Because the slope of the failure distribution is increasing together with t_{50} as the line width decreases, it is generally recognized that improvement in the early failure times at and below the 1 percent level is much less substantial than the increase in t_{50}.[10]

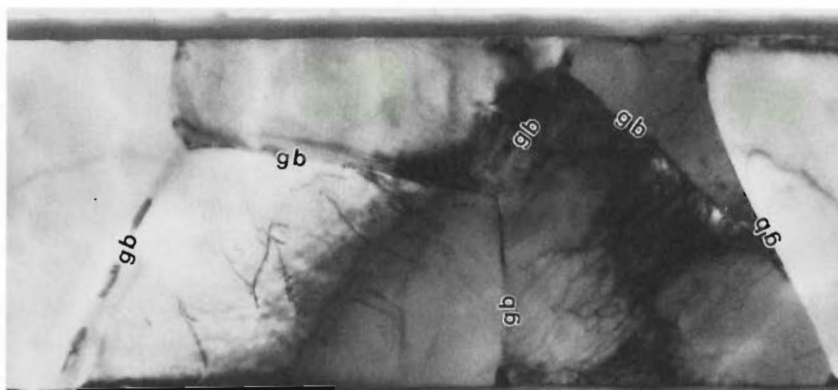
The chip interconnect reliability is determined by the time-to-first-failure of its component lines of various lengths and widths. Therefore, the probability of chip failure can be written as

$$H(t) = 1 - \prod_n (1 - F_n(t))^{L_n/l_n} \qquad (7)$$

where L_n is the total length of the line of a particular width, and l_n and $F_n(t)$ are the characteristic unit length and unit failure distribution for that line width. If the total interconnect length on the chip is distributed fairly evenly between narrow and wide lines, the chip reliability will be dominated not by



(a)



(b)

Figure 3 Comparison of Bamboo and Nonbamboo Segments:
(a) A Bamboo Microstructure in a 1.25-micron-wide Line and
(b) A Nonbamboo Grain Cluster Network Found in a 4.25-micron-wide Line
(Note that gb indicates the location of a grain boundary.)

narrow (bamboo) lines, but by the lifetime of the wide lines (one to two times the average grain size) for equivalent current densities. Thus, a qualification plan based on testing minimum-width lines exclusively is inadequate for advanced VLSI circuit technologies and may lead to overly optimistic reliability estimates.

Length/Width Scaling Model Parameters

In principle, it should be possible to uniquely determine the three model parameters θ, γ, l for a given width, knowing t_{50} and σ from lifetime measurements on two different length lines. However, in practice, this is not the case because of the statistical uncertainty in estimating t_{50} and σ. Estimates of t_{50} and σ extrapolated from test data range over intervals bounded by the choice of confidence limit. This statistical uncertainty is unavoidable but can be reduced by using larger sample sizes. Nevertheless, we can determine the model parameters within the constraints of these limits.

For example, consider testing two groups of lines with the same line width but with different lengths, such that $L_2 = 10L_1$ (i.e., $N_2 = 10N_1$). For each sample group, a range for σ and t_{50} can be extracted from the resulting failure-time distributions. In Figure 4, the calculated γ for the elemental failure distribution is shown as a function of N_1 for the upper and lower confidence limits of σ. The band of parameter values lying between these two curves represents the allowable combinations of γ and N that will fit the estimates of σ. Similarly, two curves were calculated for the upper and lower confi-

dence limits on the ratio of the t_{50} values for the two lines. The parameter space defined by the intersection of these two bands, as shown by the cross-hatched area in Figure 4, indicates the combination of model parameters that will fit the test data.

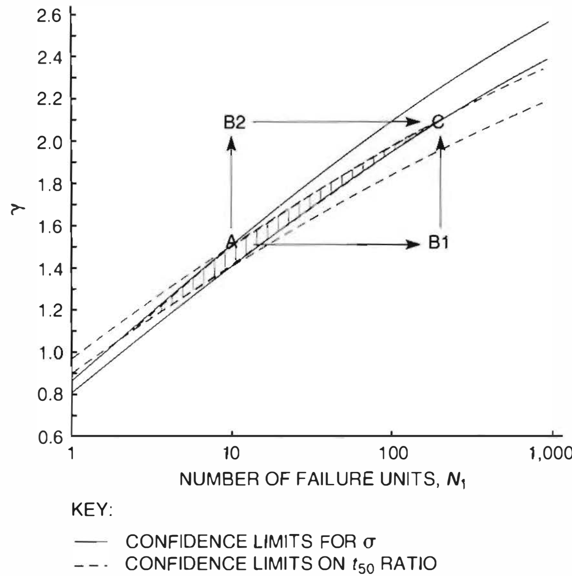


Figure 4 Allowed Values of Model Parameters γ and N Resulting from Statistical Uncertainty in t_{50} and σ from Lifetime Measurements

To determine appropriate model parameters for our metallization, we systematically studied the effect of interconnect length and width on lifetime.[14] The test samples were processed through two levels of metallization; however, only lines in the first-level metal were tested. Each level of metal was an aluminum alloy containing 1 weight-percent copper (Al:1%Cu), 7500 angstrom units (Å) thick, capped with 400 Å of titanium nitride (TiN).

The interlevel dielectric was a planarized plasma-enhanced tetraethylorthosilicate (PE-TEOS) oxide process 7500 Å thick, and the wafers were passivated with a layer of undoped oxide 7100 Å thick. The average metal grain size was determined from transmission electron microscopy (TEM) analysis to be approximately 3.5 microns (μm).

The effect of line length on lifetime is demonstrated in Figure 5. t_{50} decreases by roughly a factor of 20 as the line length increases from 1.05 to 18.9 millimeters (mm) for 1.25-μm-wide lines. The error bars on the data represent the 90 percent confidence limits. Also shown in Figure 5 is the increase in σ observed as the line length increases. The model parameters were determined from this test data; the ranges of these parameter values are given in Table 1. The solid lines in Figure 5 were calculated using the model.

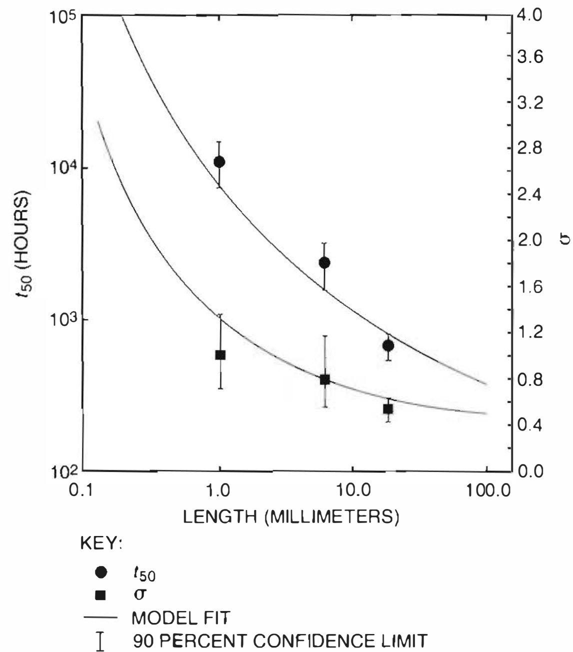


Figure 5 t_{50} and σ as Functions of Length for 1.25-micron-wide Lines

Table 1 Model Parameters as a Function of Line Width

Parameter	Line Width 1.25μm	Line Width 4.25 μm
l μm	2100–2700	26–124
θ ln (hours)	7.92–8.18	6.00–7.15
γ	0.975–1.05	0.975–1.05

The test results shown in Figure 6 demonstrate the increase in lifetime with decreasing line width expected in the bamboo region. This data shows that t_{50} decreases by roughly a factor of 100 as the line width decreases from 4.25 to 1.25 μm for 1.05-mm-long lines. The increase in σ observed as the line width decreases is also shown in Figure 6.

Since we do not yet have data on different lengths at line widths other than 1.25 μm, all three

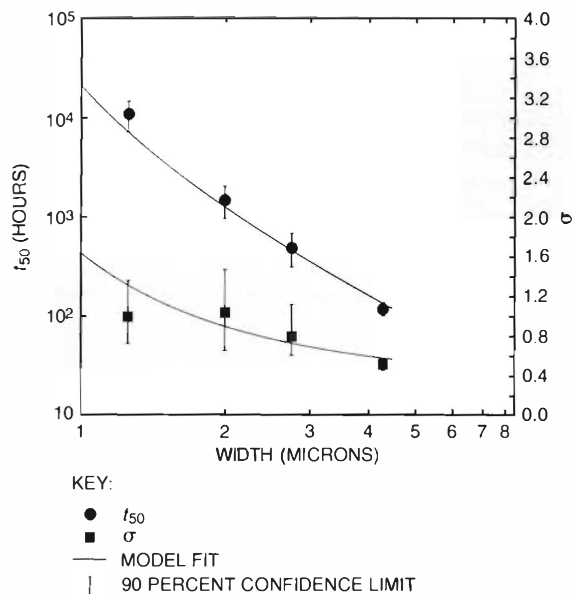


Figure 6 t_{50} and σ as Functions of Width for 1.05-millimeter-long Lines

model parameters could not be independently determined at other widths. Since θ is expected to decrease and l to increase as the length and number of nonbamboo segments decreases with line width, we chose to fix γ. Parameter value ranges were extrapolated from the 4.25-μm-wide line data and are included in Table 1. The lines in Figure 6 are calculated from the model using the parameters extracted from the 1.25- and 4.25-μm line widths and interpolating for intermediate line widths. The test results for intermediate line widths of 2.0 and 2.75 μm are shown in Figure 6 to lie on the calculated curve, as well.

For line widths larger than 4.25 μm, the lifetime does not continue to decrease. Lifetime measurements on 5.75-μm-wide lines are comparable to the 4.25-μm-wide lines. The lifetime gradually increases as the line width increases, for line widths that are greater than the grain size.

Tungsten-filled Vias

In CMOS technology development, the lateral dimensions of the vias have shrunk more quickly than the vertical thickness of the interlevel dielectric. It is difficult to reliably sputter-deposit aluminum metallization to cover the steep, vertical sidewalls of small-diameter, high-aspect-ratio vias. As a result, techniques for filling high-aspect-ratio vias with vertical sidewalls have been developed in recent years using tungsten from low-pressure chemical vapor deposition (LPCVD).[15]

The electromigration lifetime of via chains for nonfilled vias decreases with reduced hole diameter.[16] This is a direct consequence of poor metal sidewall coverage for small-diameter, high-aspect-ratio vias.[16,17]

Although filling the via with LPCVD tungsten improves sidewall coverage, the presence of the tungsten "plug" between the two aluminum lines introduces a discontinuity in the electromigration-induced mass transport of aluminum that may lead to failure. Therefore, it is very important to characterize the electromigration reliability of tungsten-filled vias.

Testing was performed on metal-1 to metal-2 tungsten-filled via chains (M1 and M2) fabricated using standard CMOS-4 processing through all three metal levels. Both levels of metal were 7500 Å thick Al:1%Cu with a 400 Å TiN cap. The interlevel dielectric was a planarized PE-TEOS oxide 7500 Å thick. The total dielectric thickness over the M2 was approximately 2.5 μm. The via diameter was 0.75 μm, and there were eight vias in each chain. The widths of the M1 and M2 interconnections were identical, nominally 1.88 μm, as were the lengths of the metal links, which measured approximately 170 μm. The links were long enough to eliminate the chance of lifetime enhancement due to the effect of back pressure in short lines.[13,18]

Electromigration lifetime measurements were carried out at temperatures of 200 and 220 degrees Celsius (°C) and currents of 6 and 8 milliamperes (mA). A 10 percent change in resistance was the failure criterion. The lifetime is found to be approximately proportional to the inverse of the square of the current density. This dependence is in agreement with other test results on tungsten-filled vias and with observations for single-level lines.[18]

The extrapolated activation energy is nearly 1.2 eV. This value is considerably higher than that expected for grain boundary transport, and approaches the value for lattice diffusion in single-crystal aluminum (about 1.4 eV). However, since the line width is much smaller than the average grain size for these thin films, the M1 and M2 interconnects will have bamboo microstructures. Therefore, we would expect very little mass transport via grain boundaries. Instead, lattice diffusion, or possibly diffusion along the aluminum-oxide interface, would be the primary transport mechanism.

We also considered the possibility of local heating at the via because of the higher resistivity of the via-fill material and the reduced cross-sectional area for current flow. Three-dimensional finite element models of the via structure were used in simulations to calculate the current density distribution and temperature profiles resulting from Joule heating, as shown in Figure 7. These simulations indicate that in the entire structure, the temperature increase above ambient is less than 3 °C under the worst-case current stress of 8 mA. Due to the good heat conductivity of aluminum, the temperature gradients are small, and the effect on electromigration lifetime is negligible.

Figure 8 shows a micrograph of a scanning electron microscope (SEM) image of a stressed via that has been cross-sectioned using a focused ion beam (FIB). The transport of aluminum away from the aluminum-tungsten interface in the direction of electron current flow is apparent. The FIB/SEM analysis of stressed via chains shows voiding at both interfaces, with no indication of preferential failure at either the top or the bottom interface.

Clearly, via failure is caused by void formation resulting from the migration of aluminum from the tungsten plug. In the future, the reliability of submicron via interconnects may be improved by developing processes to replace the tungsten with aluminum to fill the via.

Assuming that the failure distribution for a single via is lognormal, the failure distribution for the via chain is also MLN. In this case, since there is no apparent lifetime dependence on the direction of current flow, i.e., from M1 to M2 or vice versa, the value of N is the number of vias in the chain, i.e., $N=8$. For a single via, θ and γ can then be determined from via chain test results using the MLN distribution as given in equation (6).

The chip scaling model can be easily modified to include vias as separate failure elements. The resulting probability for chip failure is given by

$$H(t)=1-\prod_{n}(1-F_n(t))^{L_n/l_n}\prod_{m}(1-F_m(t))^{N_m} \quad (8)$$

where N_m is the number of vias of a particular type, and $F_m(t)$ is the failure distribution for a single via.

Electromigration Reliability Qualification

The Alpha 21064 microprocessor is the largest and fastest chip built using the CMOS-4 technology. Careful analysis of this chip determined the number of vias and the total length of lines (as a function of line width) that carry currents near the electromigration design rule limit. Interconnects with currents much less than the design rule limit have negligible impact on the electromigration reliability. With this information, we can then calculate the appropriate performance requirements for the test structures to meet our chip reliability goal.

Our reliability goal is to assure a less than 1 percent probability of chip failure in 10 years under worst-case operating conditions. The stress acceleration model described in equation (1) is used to extrapolate from accelerated stress conditions to worst-case operating conditions. Scaling of test data to the chip level is accomplished by use of the multilognormal approximation to the failure distribution.

The entire circuit can be considered to consist of several component groups, where each group includes a particular type of interconnect element, e.g., a certain via type or line width. The lognormal unit failure distribution for the ith group $F_i(t)$ is characterized by γ_i and θ_i. Rewriting equation (8) in a slightly more compact form, the cumulative probability of chip failure

$$H(t)=1-\prod_{i}S_i(t) \quad (9)$$

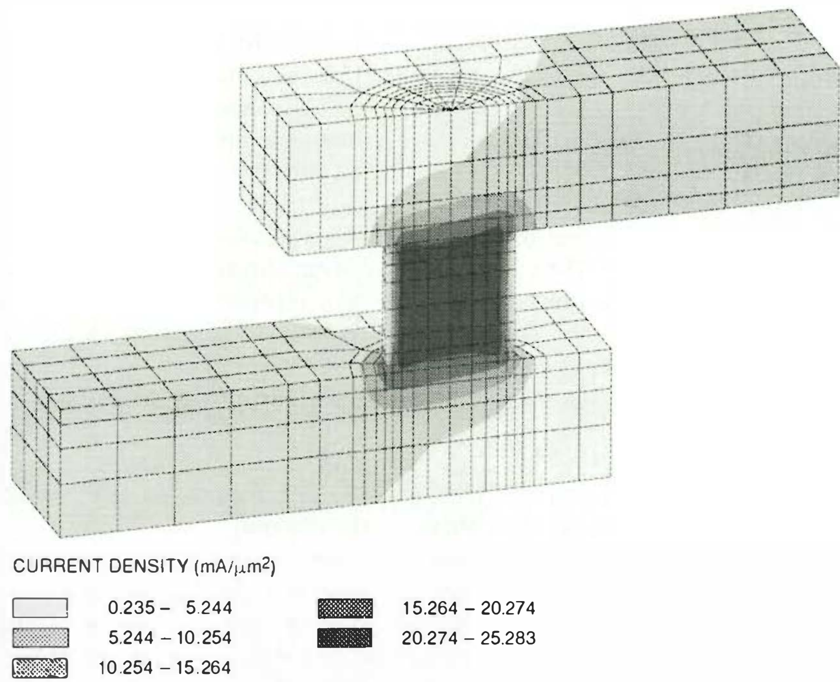
where

$$S_i(t)=(1-F_i(t))^{N_i} \quad (10)$$

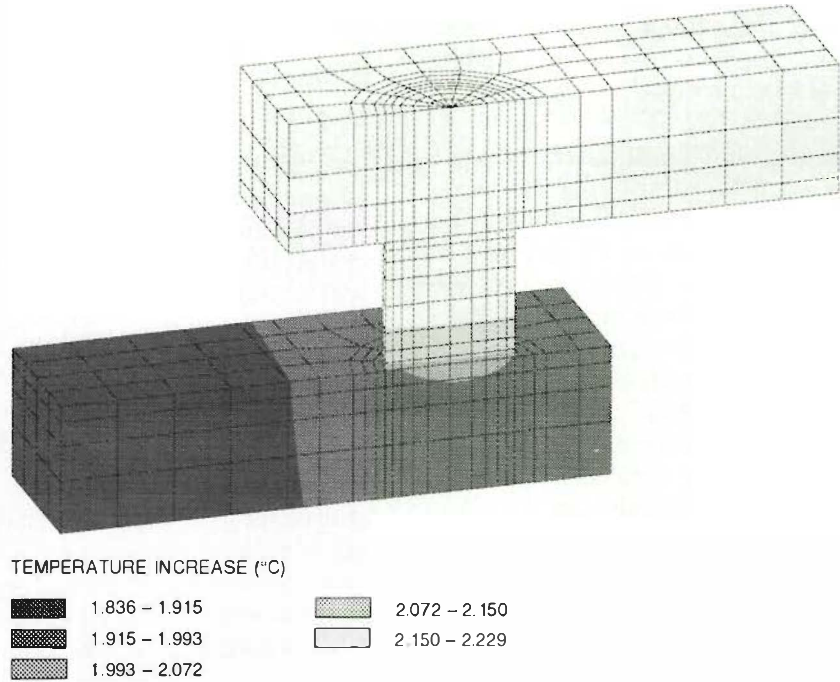
and N_i is the number of components from the ith group on the chip. For vias, N_i is the number of vias; for lines, N_i is the total length divided by the characteristic unit length, which depends on line width. Since the reliability goal is for $H(t)$ to be less than 0.01 for 10 years under worst-case operating conditions, then for each group, $F_i(t)$ is much less than 0.01.

As mentioned previously, uncertainty in the values for γ, θ, and l stems from the uncertainty inherent in estimating t_{50} and σ to a given level of confidence from experimental data. It is illuminating to examine the impact of this inherent uncertainty has on the performance requirements for the test structures to meet the chip reliability goal.

Consider the simple case of only one component group, namely a single line width. The required test structure $t_{50,test}$ relative to the time until 1 percent cumulative failure for the chip occurs $t_{01,chip}$ is plotted in Figure 9. This ratio is graphed as a function of the ratio of the number of units in the test structure to that of the chip, N_{test}/N_{chip}, for a given combination of N_{test} and γ. For this illustration, we use



(a)



(b)

Figure 7 Finite Element Simulations of Current Density and Temperature Distributions in Tungsten-filled Vias

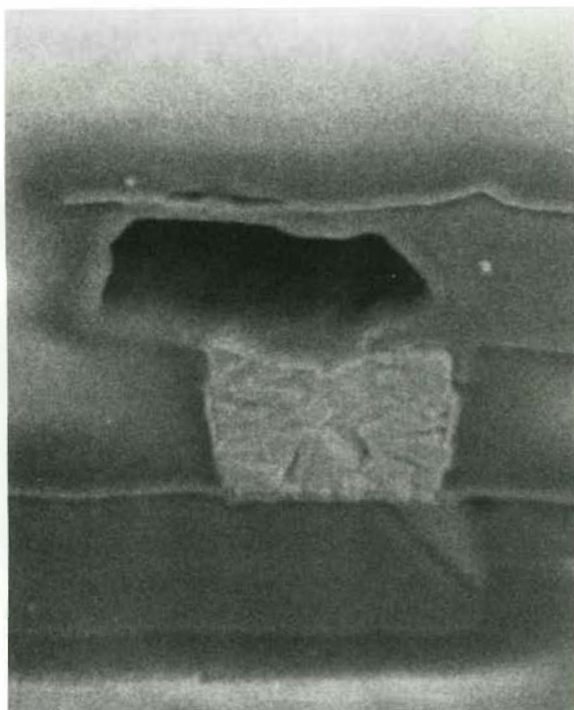


Figure 8 Secondary Electron Micrograph of Electromigration Damage at a Tungsten-filled Via

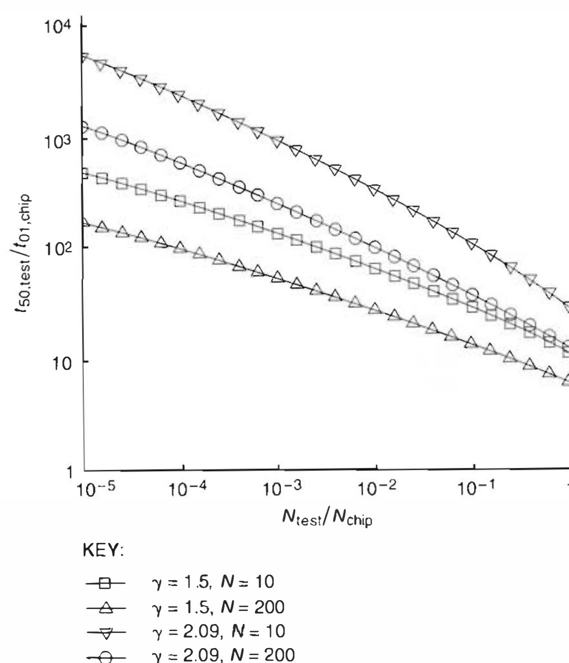


Figure 9 Requirements on t_{50} for the Test Structure Relative to t_{01} for the Chip as a Function of the Ratio of the Number of Elements in the Test Structure to That of the Chip

points from the example shown in Figure 4 for comparison.

Points A and C both lie on the edge of the allowed parameter space in Figure 4, but γ and N are both larger for C. Figure 9 shows that the requirements on $t_{50,test}$ are greater for C than for A. However, the relative effect of increasing N or γ separately is not clear. Increasing N alone, that is, comparing point B1 to A, actually results in a decrease in the test requirements, as shown in Figure 9. This effect weakens as the N_{test}/N_{chip}, ratio approaches unity. Thus, the test line should be as long as practical to mitigate the impact of uncertainty in l. The increase in $t_{50,test}/t_{01,chip}$, therefore, is a result of increasing γ. The sensitivity to increasing γ alone can be seen in Figure 9, by comparing point B2 to A.

To assure that the chip reliability goals are met, the most conservative combination of parameter estimates should be used. These values are the most stringent in setting and meeting the test performance requirements. The most rigorous test performance requirements are set by using the largest possible value for γ. In the example presented in Figure 9, this value of γ implies the largest value of N, i.e., the smallest value for l, within the ranges set by the confidence limits on σ and t_{50}. The most stringent criterion for meeting the test performance requirements is to use the lower limit of the confidence interval for t_{50} from the test data, which corresponds to the minimum θ value consistent with the values of γ and N.

Electromigration lifetimes, and thus the scaling model parameters, are a sensitive function of the microstructure of the conductor film.[19] Therefore, since the effect of normal microstructural variations cannot be unambiguously determined a priori, a number of lots must be tested to assess the effects of lot-to-lot variations.

Each lot undergoes a statistical test of the hypothesis

$$H(10 \text{ years}) \leq 0.01 \qquad (11)$$

where $H(10 \text{ years})$ denotes the cumulative probability of chip failure in the first 10 years of continuous operation under worst-case conditions. To pass the test, the probability of accepting the hypothesis when $H(10 \text{ years})$ is greater than 0.01 must be less than 0.1. This criterion gives at least 90 percent confidence that a test group passing the statistical test, i.e., for which the hypothesis is accepted,

- Does indeed come from a lot that meets the requirement of not failing more than 1 percent of the time in 10 years
- Is not a statistical fluke

The statistical analysis procedures used to implement this model in the electromigration qualification testing are coded in a software tool. The software extracts the most conservative MLN model parameters from the failure-time distribution measured for every type of structure tested for each lot. This tool can be used to perform the statistical test just described to verify that the reliability goal has been met.

Summary

Interconnect electromigration reliability becomes increasingly important with each step in the evolution of CMOS technology. Therefore, it is necessary to rigorously characterize the various components of the circuit metallization and to develop dependable models to relate test device data to long-term chip reliability.

We have presented a scaling model for relating the results of accelerated electromigration life tests on test structures to the overall chip reliability. This model was used as the basis for formulating qualification requirements for electromigration reliability assurance of the CMOS-4 process technology and the Alpha 21064 microprocessor.

Acknowledgments

We would like to acknowledge the contributions of the following individuals: Barbara Miner, Steven Bill, and Jamie Rose for TEM work, Aldo Pelillo for FIB/SEM work, Ahsan Enver for finite element simulations, John Kitchin, Bill Martin, Dave Dunnell, Dave Foggo, Terry Spooner, Lesley Elliott, and Professor Carl Thompson of the Massachusetts Institute of Technology.

References

1. T. Kwok and P. Ho, "Electromigration in Thin Films," *Diffusion Phenomena in Thin Films and Microelectronic Materials,* edited by D. Gupta and P. Ho (Park Ridge, New Jersey: Noyes Publications, 1988): 369-431, and the references therein.

2. P. Ghate, "Electromigration-induced Failures in VLSI Interconnects," *Proceedings of the Twentieth Annual International Reliability Physics Symposium* (1982): 292-299.

3. H. Schafft, T. Staton, J. Mandel, and J. Shott, "Reproducibility of Electromigration Measurements," *IEEE Transactions on Electron Devices,* vol. ED-34, no. 3 (March 1987): 673-680.

4. M. Shatzkes and J. Lloyd, "A Model for Conductor Failure Considering Diffusion Concurrently with Electromigration Resulting in a Current Exponent of 2," *Journal of Applied Physics,* vol. 59, no. 11 (June 1986): 3890-3893.

5. J. Black, "Mass Transport of Aluminum by Momentum Exchange with Conducting Electrons," *Proceedings of the Sixth Annual International Reliability Physics Symposium* (1967): 148-159.

6. B. Argarwala, M. Attardo, and A. Ingraham, "Dependence of Electromigration-induced Failure Time on Length and Width of Aluminum Thin-film Conductors," *Journal of Applied Physics,* vol. 41, no. 10 (September 1970): 3954-3960.

7. J. Lloyd and J. Kitchin, "The Electromigration Failure Distribution: The Fine Line Case," *Journal of Applied Physics,* vol. 69, no. 4 (February 1991): 2117-2127.

8. S. Vaidya, D. Fraser, and A. Sinha, "Electromigration Resistance of Fine-line Al for VLSI Applications," *Proceedings of the Eighteenth Annual International Reliability Physics Symposium* (1980): 165-170.

9. E. Kinsbron, "A Model for the Width Dependence of Electromigration Lifetimes in Aluminum Thin-film Stripes," *Applied Physics Letters,* vol. 36, no. 12 (June 1980): 968-970.

10. J. Cho and C. Thompson, "Grain Size Dependence of Electromigration-induced Failures in Narrow Interconnects," *Applied Physics Letters,* vol. 54, no. 25 (June 1989): 2577-2579.

11. D. Walton, H. Frost, and C. Thompson, "Computer Simulation of Grain Growth in Thin-film Interconnect Lines," *Materials Research Society Symposia Proceedings,* vol. 225 (1991): 219-224.

12. E. Arzt and W. Nix, "A Model for the Effect of Line Width and Mechanical Strength on Electromigration Failure of Interconnects with 'Near-bamboo' Grain Structures," *Journal of Materials Research,* vol. 6, no. 4 (April 1991): 731-736.

13. I. Blech, "Electromigration in Thin Aluminum Films on Titanium Nitride," *Journal of Applied Physics,* vol. 47, no. 4 (April 1976): 1203-1208.

14. E. Atakov and J. Clement, "Electromigration Failure Time Distribution: Scaling from a Test Structure to the VLSI Chip," *Proceedings of the Ninth VLSI Multilevel Interconnection Conference* (1992): 366-371.

15. M. Garver et al., "CMOS-4 Back-end Process Development for a VLSI 0.75-μm Triple-level Interconnection Technology," *Digital Technical Journal,* vol. 4, no. 2 (Spring 1992, this issue): 51-72.

16. F. Matsuoka et al., "Electromigration Reliability for a Tungsten-filled Via Hole Structure," *IEEE Transactions on Electron Devices,* vol. ED-37, no. 3 (March 1990): 562-567.

17. H. Rathore, "Via Resistance as a Technique to Control the Electromigration of Non-overlap Via Holes," *Proceedings of the Twentieth Annual International Reliability Physics Symposium* (1982): 77-80.

18. J. Estabil, H. Rathore, and F. Dorleans, "The Effect of Metal Thickness on Electromigration-induced Shorts in Submicron Technology," *Proceedings of the Twenty-ninth Annual International Reliability Physics Symposium* (1991): 57-63.

19. P. Smith, J. Lloyd, and G. Prokop, "Lot-to-lot Variations in Electromigration Performance for Thin Film Microcircuits," *Journal of Vacuum Science and Technology,* vol. A2, no. 2, (April-June 1984): 220-223.

Further Readings

The Digital Technical Journal *publishes papers that explore the technological foundations of Digital's major products. Each* Journal *focuses on at least one product area and presents a compilation of papers written by the engineers who developed the product. The content for the* Journal *is selected by the Journal Advisory Board. Digital engineers who would like to contribute a paper to the* Journal *should contact the editor at RDVAX::BLAKE.*

Topics covered in previous issues of the *Digital Technical Journal* are as follows:

PATHWORKS: PC Integration Software
Vol. 4, No. 1, Winter 1992

Image Processing, Video Terminals, and Printer Technologies
Vol. 3, No. 4, Fall 1991

Availability in VAXcluster Systems/ Network Performance and Adapters
Vol. 3, No. 3, Summer 1991

Fiber Distributed Data Interface
Vol. 3, No. 2, Spring 1991

Transaction Processing, Databases, and Fault-tolerant Systems
Vol. 3, No. 1, Winter 1991

VAX 9000 Series
Vol. 2, No. 4, Fall 1990

DECwindows Program
Vol. 2, No. 3, Summer 1990

VAX 6000 Model 400 System
Vol. 2, No. 2, Spring 1990

Compound Document Architecture
Vol. 2, No. 1, Winter 1990

Distributed Systems
Vol. 1, No. 9, June 1989

Storage Technology
Vol. 1, No. 8, February 1989

CVAX-based Systems
Vol. 1, No. 7, August 1988

Software Productivity Tools
Vol. 1, No. 6, February 1988

VAXcluster Systems
Vol. 1, No. 5, September 1987

VAX 8800 Family
Vol. 1, No. 4, February 1987

Networking Products
Vol. 1, No. 3, September 1986

MicroVAX II System
Vol. 1, No. 2, March 1986

VAX 8600 Processor
Vol. 1, No. 1, August 1985

Subscriptions to the *Digital Technical Journal* are available on a yearly, prepaid basis. The subscription rate is $40.00 per year (four issues). Requests should be sent to Cathy Phillips, Digital Equipment Corporation, MLO1-3/B68, 146 Main Street, Maynard, MA 01754, U.S.A. Subscriptions must be paid in U.S. dollars, and checks should be made payable to Digital Equipment Corporation.

Single copies and past issues of the *Digital Technical Journal* can be ordered from Digital Press at a cost of $16.00 per copy.

Technical Papers by Digital Authors

C. Brench, "Optimizing EMI Shield Design with Numerical Techniques," *Conference Proceedings of the Applied Computational Electromagnetics Society* (March 1992).

J. Clement and J. Lloyd, "Numerical Investigations of the Electromigration Boundary Value Problem," *Journal of Applied Physics* (February 1992).

S. Das, "The MR Head: An Emerging Trend in Hard Disk Drive Head Development," *Electrochemical Society International Symposium on Magnetic Materials, Processes and Devices* (1991).

S. Das, "Selective Laser Microetching," *Electrochemical Society International Symposium on Electrochemical Microfabrication* (October 1991).

C. Derry, "Transition from Quality Circles to Total Involvement in Semiconductor Manufacturing," *Quality and Reliability Engineering International*, vol. 7, no. 6 (November-December 1991).

B. Doyle, "Recovery of Hot-Carrier Damage in Reoxidized Nitride Oxide MOSFETs," *IEEE Electron Device Letters* (January 1992).

B. Doyle, B. Fishbein, and K. Mistry, "NBTI-Enhanced Hot-Carrier Damage in p-Channel MOSFETs," *IEEE International Electron Devices Meeting* (December 1991).

A. Gardel and P. Deosthali, "Using Simulation to Test the Robustness of Various Existing Production Control Policies," *1991 Winter Simulation Conference* (December 1991).

S. Heng, "Temperature Mapping of Localized Hot Spots on Microelectronic Chip Surfaces," *ASME Transactions: Journal of Electronic Packaging* (September 1991).

S. Heng and H. Pei, "Air Impingement Cooled Pin-Fin Heat Sink for Multi-Chip Unit," *National Electronic Packaging and Production Conference* (February 1991).

A. Mason, M. Zurko, C. Kahn, P. Karger, and D. Bonin, "A VMM Security Kernel for the VAX Architecture," *IEEE Symposium on Research in Security and Privacy* (May 1990).

K. Mistry, B. Doyle, A. Philipossian, and D. Jackson, "AC Hot Carrier Lifetimes in Oxide and ROXNOX N-Channel MOSFETs," *IEEE International Electron Devices Meeting* (December 1991).

K. Ramakrishnan, "A Model of Naming for Fine Grained Service Specification in Distributed Systems," *Proceedings of the Eleventh International Conference on Distributed Computing Systems* (May 1991).

K. Ramakrishnan, "A Naming System for Feature-based Service Specification in Distributed Operating Systems," *Proceedings of the 1991 ACM SIGSMALL/PC Symposium on Small Systems* (June 1991).

J. Rose, B. Minor, and A. Pelillo, "TEM: A Review and a Comparison with High Resolution SEM," *International Symposium for Testing and Failure Analysis* (November 1991).

K. Springer, "A Forward Error Correcting Code for Gigabit Fiber Optic Links," *Proceedings of SPIE—The International Society of Optical Engineering* (September 1991).

R. Ulichney, "On the Manipulation of Power Spectra of Halftone Patterns," *IST (SPIE) Seventh International Congress on Advances in Non-Impact Printing Technology* (October 1991).

D. Uttley and S. Coverdale, "ReTAB for Test and Failure Analysis," *Proceedings of the Seventeenth International Symposium for Testing and Failure Analysis* (November 1991).

N. Veisfeld, "Microbeam Surface Analysis Methods: Applications and Limitations in an Industrial Environment," *Proceedings of the Microbeam Analysis Society of America* (Fall 1991).

A. Vitale, "An Algorithm for High Accuracy Name Pronunciation by Parametric Speech Synthesizer," *Journal of Computational Linguistics* (September 1991).

X. Wang, "Test Generation Based on Dynamic Search Space Reductions," *International Phoenix Conference on Computers and Communications* (March 1990).

W. Zahavi, "Capacity Planning/Needs Analysis: The Economy Version," *Computer Measurement Group Conference* (December 1991).

Digital Press

Digital Press is the book publishing group of Digital Equipment Corporation. The Press is an international publisher of computer books and journals on new technologies and products for users, system and network managers, programmers, and other professionals. Proposals and ideas for books in these and related areas are welcomed.

The following book descriptions represent a sample of recent publications available from Digital Press.

MOTIF PROGRAMMING: The Essentials...and More
Marshall Brain, 1992, softbound, 632 pages, Order No. EY-J816E-DP-EEB ($29.95).

A straightforward and easy-to-understand introduction to Motif application development, this book will ease you into Motif programming as smoothly and quickly as possible. It starts with an introduction to event-driven programming and proceeds to discuss three concepts essential to Motif programming: resources, callbacks, and containers. Advanced topics will expose the reader to all of the Motif widgets, the capabilities of the X and Xt layers, the X drawing model, and the process of application design in Motif.

VMS FOR ALPHA PLATFORMS INTERNALS AND DATA STRUCTURES: Preliminary Edition, Volume 1
Ruth E. Goldenberg and Saro Saravanan, 1992, softbound, 416 pages, Order No. EY-L466E-P1-EEB ($30.00).

This volume is the first publication to explain VMS operating system support for Alpha, Digital Equipment Corporation's recently announced 64-bit RISC architecture. Volume 1 contains twelve chapters and an appendix. Subsequent volumes will be published as more chapters become available.

THE X WINDOW SYSTEM SERVER: X Version 11, Release 5
Elias Israel and Erik Fortune, 1992, softbound, 534 pages, Order No. EY-L518E-DP-EEB ($45.95).

This technical reference covers every aspect of the sample server developed by the MIT X Consortium. It provides essential information to knowledgeable X users who want to learn about the basic interactions between client and server—including developers who want to port, extend, tune, or test a server. Examples, guidelines, and tutorials reinforce material on theory and practice.

SOFTWARE IMPLEMENTATION TECHNIQUES: VMS, UNIX, OS/2, and MS-DOS
Donald E. Merusi, 1992, softbound, 743 pages, Order No. EY-J822E-DP-EEB ($39.95).

Written for programmers and software developers, this book eases the transition from one operating system to another. Topics help you understand the capabilities needed to design applications that run on one or more of these systems or to port existing applications to a new system. This book will reduce the time you need to gather information to peform a specific task, and diminish your frustration level by showing you where to turn for additional help.

ALL-IN-1: A Technical Odyssey
Tony Redmond, 1992, softbound, 550 pages, Order No. EY-H952E-DP ($44.95).

This extensive treatment of Digital Equipment Corporation's office automation tool addresses the needs of system managers, application programmers, and technically oriented users who work with ALL-IN-1. Based on the author's ten years of experience in developing ALL-IN-1 subsystems and in customizing its application to specific customer sites, the presentation extends beyond the product documentation to explore the deep and distant corners of the product. The wealth of examples of actual installation and customization experiences help communicate how to best use ALL-IN-1 on VAX, DOS PC, and Apple Macintosh computers.

RELIABLE COMPUTER SYSTEMS: Design and Evaluation, Second Edition
Daniel P. Siewiorek and Robert S. Swarz, 1992, hardbound, 908 pages, Order No. EY-H880E-DP-EEB ($64.95).

This major revision of the 1982 publication, *The Theory and Practice of Reliable System Design,* provides an up-to-date and comprehensive guide to the design, evaluation, and use of reliable computer systems. It includes case studies of systems from manufacturers such as Tandem, Stratus, IBM, and Digital as well as the special Galileo fault protection and AT&T telephone switching processor systems.

ALPHA ARCHITECTURE REFERENCE MANUAL
Edited by Richard L. Sites, 1992, softbound, 600 pages, Order No. EY-L520E-DP-EEB ($34.95).

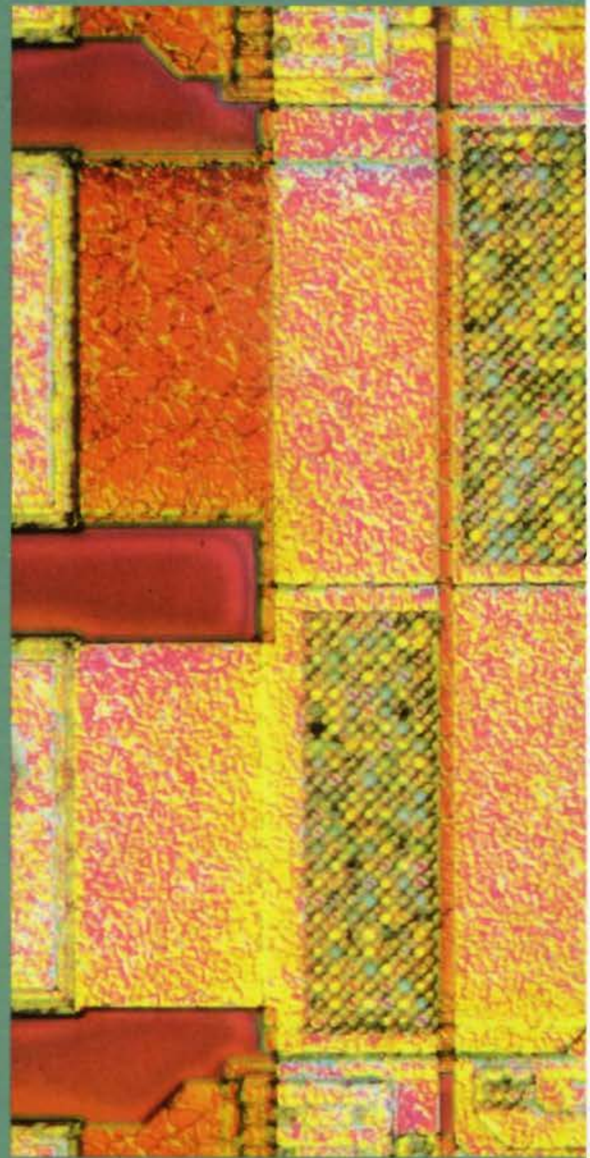
Written by the engineers who developed the Alpha specifications, this book contains complete descriptions of the common architecture required for all implementations and the interfaces required to support the OSF/1 and OpenVMS operating systems. Of great interest to software developers, system managers, system analysts, and system programmers, this is the authoritative reference on Alpha, Digital Equipment Corporation's new 64-bit architecture.

To receive a copy of the latest catalog or further information on these or other publications from Digital Press, please write:

Digital Press
Department EEB
1 Burlington Woods Drive
Burlington, MA 01803-4597

Or, you can order a Digital Press book by calling DECdirect at 800-DIGITAL (800-344-4825). When ordering be sure to refer to Catalog Code EEB.

digital™



ISSN 0898-901X

Printed in U.S.A. EY-L521E-DP/92 08 02 18.0 Copyright © Digital Equipment Corporation. All Rights Reserved.