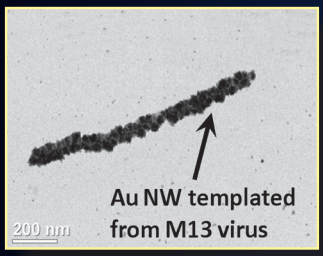
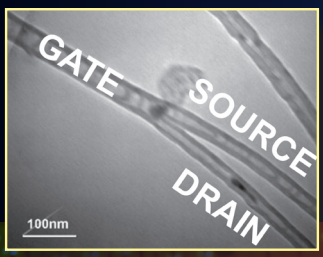
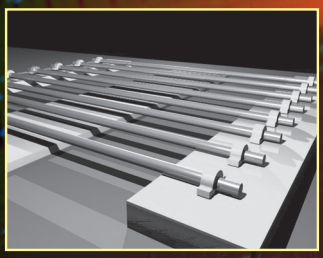
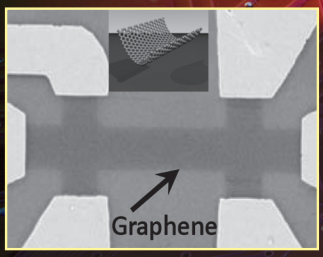
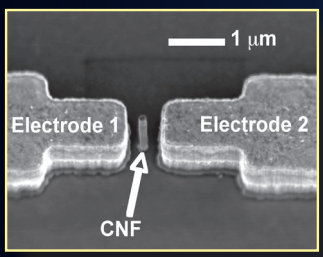




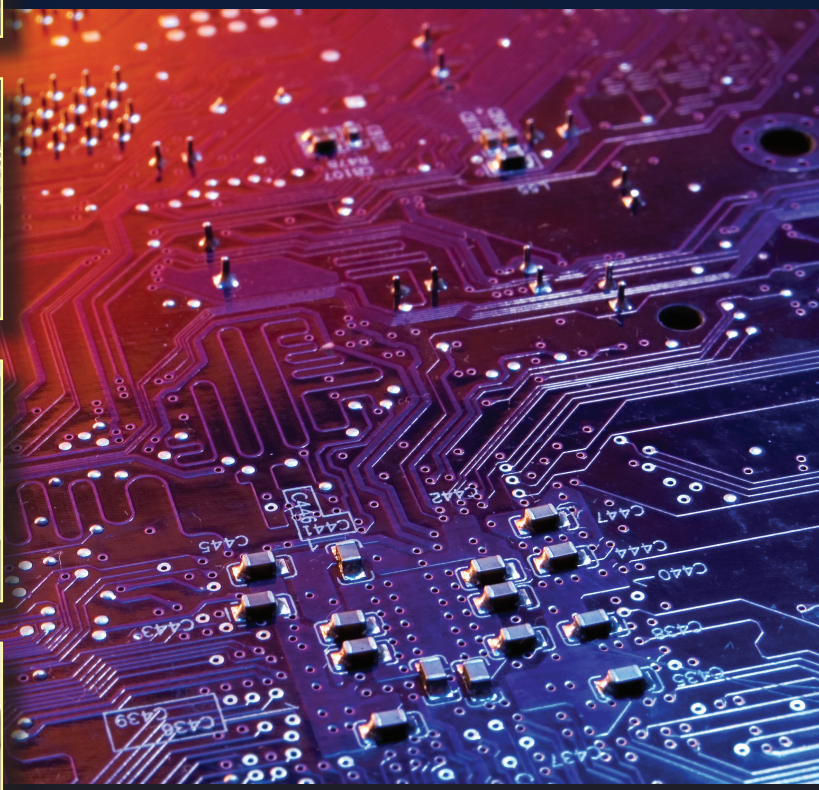






MICROELECTRONICS TO NANOELECTRONICS

Materials, Devices & Manufacturability



Edited by
ANUPAMA B. KAUL

CRC Press
Taylor & Francis Group

MICROELECTRONICS TO NANOELECTRONICS

Materials, Devices & Manufacturability

MICROELECTRONICS TO NANOELECTRONICS

Materials, Devices & Manufacturability

Edited by

ANUPAMA B. KAUL

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20120518

International Standard Book Number-13: 978-1-4665-0955-9 (eBook - PDF)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

In Loving Memory of

Mrs. S. P. Kaul

Justice J. N. Bhat

I dedicate this book to

Ashish, Ishani, and Arnav

Contents

About the Cover

Images from top to bottom: A transmission electron microscopy (TEM) image of a single viral-templated nanocrystalline gold nanowire. A genetically modified M13 virus with an affinity for gold was used to bind 5 nm gold nanoparticles. Electroless gold deposition was subsequently used to control the size and connectivity of the resulting nanowires. (Image courtesy of Professor Elaine Haberer, University of California–Riverside.)

TEM micrograph of a multi-walled carbon nanotube in a Y-junction form, which can be used as a prototypical nanoelectronic element constituting a switch or a transistor, depending on the transmission characteristics of the constituent branches. The Fe-Ti catalyst particles formed in the CVD growth, along with the presence of topological defects at the junction region, can influence the electrical transport characteristics. (Image courtesy of Professor Prabhakar R. Bandaru, University of California–San Diego, La Jolla.)

An artistic rendition illustrating an array of suspended nanowires assembled via a hybrid bottom-up/top-down approach: chemically synthesized nanowires integrated with lithographically patterned electrodes, which are promising for resonant sensing and high-speed nanomechanical computing. (Image courtesy of Professor Philip Feng, Case Western Reserve University, Cleveland, Ohio.)

A scanning electron microscopy (SEM) image of monolayer thick graphene sheet connected with electrical leads. Such devices show the ultimate in chemical sensitivity where adsorption of individual gas molecules can be detected. Top inset shows the hexagonal motif of carbon atoms in graphene which serves as the building block for other carbon-based nanomaterials such as CNTs and bucky balls. (SEM image courtesy of Dr. Konstantin Novoselov, University of Manchester, United Kingdom. Adapted from F. Schedin, et al. *Nature Materials* 6, 652, 2007. Copyright © Nature Publishing Group.)

A TEM micrograph of a singly wound coiled carbon nanotube (CNT) synthesized with thermal chemical vapor deposition (CVD) through the use of indium- and tin-based catalysts. (Image courtesy of Professor Prabhakar R. Bandaru, University of California–San Diego, La Jolla.)

SEM image of a single, vertically oriented carbon nanofiber (CNF) centered within high-aspect-ratio electrodes. Such structures, formed with low-cost, wafer-scale approaches using a hybrid combination of top-down and bottom-up nanofabrication techniques, are a stepping stone for the realization of controlled architectures for three-dimensional (3D) electronics. (Image courtesy of Dr. Anupama B. Kaul, Jet Propulsion Laboratory, California Institute of Technology, Pasadena.)

Background is an example of packaged and integrated components that have been created using micro- and nanofabrication techniques and form the backbone of the semiconductor industry for microprocessors and other system-level applications.

Foreword

It is a pleasure to write the Foreword for this book because of its timeliness and its excellent breadth and depth of coverage that capture most key facets of the field and also because its editor, Dr. Anupama Kaul, is an esteemed collaborator of mine in micro-electro-mechanical (MEMS) and nanotechnologies research. Indeed, her expertise and excellent reputation in the field derive from her extensive work on the development of carbon nanotube (CNT)- and carbon nanofiber (CNF)-based devices conducted at the Jet Propulsion Laboratory, California Institute of Technology. In particular, Dr. Kaul's contributions encompass nanoelectronic devices based on carbon-based nanomaterials for three application areas: (1) nano-electro-mechanical (NEM) switches and resonators; (2) physical sensors; and (3) optical absorbers aimed at applications in extreme environment electronics for planetary missions, miniaturized sensors as interfaced with vacuum micro-cavities for high-frequency vacuum electronics, and broadband optical absorbers operational from the UV to IR ranges. It is thus not surprising, given Dr. Kaul's excellent technical background and experience, that she is uniquely qualified to edit such an ambitious piece of work, gathering contributions from a stellar roster of top researchers in the field.

In this book, the reader will find that the science and engineering pertaining to virtually all areas of application where MEMS and nanotechnologies hold great potential to provide *enabling* advantages are addressed. This diversity of exposition intertwines the various relevant disciplines, e.g., from fabrication techniques to devices, circuits, and systems, and a plethora of applications from electronics to computing to science to sensing to energy. Thus, the book has something for everyone!

To place the book into perspective, we should recall that many attribute progress in the fields of MEMS and nanotechnologies to physicist Richard Feynman's 1959 contribution titled, *There is Plenty of Room at the Bottom.* Accordingly, the theme of miniaturization permeates these technologies. In this context, the book begins by addressing Moore's law and its impact on transistor scaling and reliability; then it proceeds to address MEMS technology and it spectacular successes typified by thermal ink-jet printing and timing resonators. Several authors deal with MEMS, but at a lower-size scale, and discuss nanoscale electromechanical devices enabled by nanowire structures before closing the MEMS topic with an in-depth exposition of silicon etching and etch techniques for NEMs and MEMs and their exploitation in materials and devices for ultra energy-efficient digital integrated circuits.

Subsequently, the interdisciplinarity and versatility of the field are revealed in chapters on viral-templated materials and devices—a technique exploited for uncovering new knowledge through explorations in biology.

At this point the book turns to CNTs, the quintessential nanotechnology materials. In particular, the principles and methods for integration of carbon nanotubes in miniaturized systems, heterogeneous integration of CNTs on complementary metal oxide semiconductor circuitry and sensing applications, and CNT Y-junctions pertaining to electronic systems implementation are addressed. As is well known, the miniaturization of electronic devices, circuits, and systems is accompanied by issues surrounding heat dissipation. This all-important aspect of exploiting nanotechnologies is addressed by exploring nanoscale effects in multiphase flows and heat transfer. The book concludes with discussions of nanoengineered materials for applications in electronics, biology, and energy harnessing.

As can be appreciated from the above summary, the book does indeed have something for everyone. While the theme of "making everything small," as envisioned by Feynman will undoubtedly continue to present a rich set of challenges, this book marks a milestone by capturing the current status of research and development. Therefore, it is of extreme value to both established researchers and newcomers who aim to discover and exploit the vast space available at the bottom!

Dr. Héctor J. De Los Santos
Karlsruhe Institute of Technology
Karlsruhe, Germany

Preface

Our desire for gadgets smaller, faster, and cheaper continues to fuel the miniaturization revolution that is now well into the nanoscale regime (defined as functional length scales smaller than 100 nm). In this regime, broadly defined as nanotechnology, novel properties often emerge that in many instances can be exploited to enhance the performance of devices, components, and systems for applications targeted toward a broad array of sectors spanning sustainable energy, climate change, homeland security, and healthcare. However, nowhere are the phenomenal advances in and benefits of miniaturization more apparent than in the electronics industry. Historically, the first digital computer, the Electronic Numerical Integrator and Computer (ENIAC) built in 1946, occupied an entire room, weighed 30 tons, and consumed 200 kilowatts of power. In contrast, a modern-day computer is handheld and operates on a 3- to 4-volt battery.

Needless to say, this 60-year journey in the miniaturization revolution has been filled with radical innovations in new materials, structures, devices, and integration schemes at every juncture that have helped shape the electronics industry into the trillion dollar empire it is today. One example of such innovation was replacing the vacuum tubes in the ENIAC with solid-state transistors, for which William Shockley, John Bardeen, and Walter Brattain received the Nobel Prize in Physics in 1956. While the smaller size of the transistor was a great benefit to circuit designers at the time, circuit complexity was limited by the unreliable and manually soldered connections, and the long wiring lengths also limited operational speeds. Yet another radical innovation came about in 1958, when Jack Kilby of Texas Instruments proposed making all the components from a single block of material instead of manufacturing discrete components; this made the circuits smaller and faster. Kilby's monolithic integration scheme and invention of the integrated circuit (IC) won him the Nobel Prize in Physics in 2000, over 40 years later.

Our voyage into the world of things smaller and faster does not seem to end here, however. While incremental advances have enabled a doubling of the density of ICs on a silicon (Si) chip every 2 years according to Moore's law, by 2018 however, we will find ourselves at a technological crossroads again, much like our predecessors. Critical dimensions in the Si transistor will approach the size of atoms and quantum mechanical effects will predominate, eventually prohibiting device operation altogether. Just as the forefathers of the modern-day computer did, scientists know that novel and drastically different solutions are necessary to overcome fundamental limitations and thus sustain electronics commerce and propel the larger global economy forward in the coming decades.

Such solutions and disruptive technologies will more than likely emerge from the realms of nanotechnology. A testament to nanotechnology's promise is perhaps best exemplified by the 2010 Nobel Prize in Physics awarded to Andre Geim and Konstantin Novoselov for their ground-breaking work on the nanomaterial graphene—a monolayer thick sheet of carbon atoms exhibiting remarkable electronic, mechanical, and optical properties; such properties are not only attractive for computation, but also exhibit far-reaching implications in other areas such as sensors, structural materials, and biology. The broad importance of nanotechnology to our overall economy is also mirrored by the global commitment to invest in nanotechnology research; for example, in the United States alone, federal support for the National Nanotechnology Initiative (NNI) formed in 1999 is expected to remain at $1.8 billion in 2013.

The purpose of this book is to provide a synopsis of the exciting work that is currently in progress in the area of micro- and nanotechnologies targeted primarily for electronics applications and pursued in academia, government laboratories, and industry. A broad spectrum of micro- and nanotechnologies is showcased here, from early laboratory investigations to more mature technologies that have already reached the commercial markets. Since nanotechnology is inherently multidisciplinary, this book is intended to serve as a reference for students in upper division undergraduate or graduate courses in electrical engineering, materials science, physics, chemistry, and bioengineering and for working professionals.

We start in Chapter 1 with a description of current state-of-the-art Si transistor technology in the context of Moore's law and the impact of continued scaling on transistor performance. With billions of transistors now in use in modern microprocessors, reliability issues in complementary metal–oxide semiconductor (CMOS) transistors are highlighted, and the near-term approaches used to mitigate these issues are also discussed. Highly scaled transistors are also more vulnerable to failure in harsh environments such as radiation, as discussed in Chapter 2. Nowhere is the need for radiation-hard electronics greater than in space-based electronic systems that must withstand large doses of integrated fluxes of ionizing and non-ionizing radiation over long mission durations; such issues are particularly apparent and exacerbated in scaled transistors, as the discussion in Chapter 2 highlights.

The Si microelectronics industry also gave birth to the field of micro-electro-mechanical systems (MEMS) when Kurt Petersen at International Business Machines (IBM) demonstrated the first micro-machined pressure sensor in 1970 using batch fabrication. Since then, many novel MEMS-based devices and components have been developed. They are no longer confined to the realms of the laboratory and have been commercialized and continue to generate revenue. Chapter 3 presents an overview of surface micro-machined ink-jet print heads as quintessential examples of one such successful MEMS technology. Ink-jet MEMS represents one of the most

profitable MEMS technologies and is now enabling applications in other areas such as microfluidics and biological and chemical sensing. Further, in Chapter 4 another example of MEMS components is provided, specifically mechanical resonators that surpass the performance of electronic resonators based on solid-state transistors. Such high-frequency mechanical resonators have applications serving as frequency standards and filters in communication systems. With dimensions of mechanical structures in MEMS now reaching the nanoscale regime, Chapter 5 reviews the design, fabrication, and characterization of nano-electro-mechanical (NEM) resonators that can serve as sensitive mass detectors that are particularly appealing for biological sensing applications.

Chapters 6 and 7 provide examples of the latest developments in nanofabrication derived from the top-down and bottom-up approaches, respectively. The top-down approach has traditionally sustained the microelectronics industry. Features are formed using combinations of deposition, lithography, and etching, but new techniques and tools are necessary to realize feature sizes in the nanoscale regime. Top-down pattern transfer etching techniques using state-of-the-art inductively coupled plasmas (ICPs) are discussed in Chapter 6. The chapter explains how highly anisotropic, high-aspect-ratio Si nanostructures are formed using wafer-scale approaches for MEMS and NEMS applications. Conversely, in the bottom-up approach, nanoscale building blocks can be hierarchically assembled to form complex structures, akin to the way nature uses proteins to build complex biological systems, for example. Chapter 7 presents an example of one such bottom-up approach. Select groups of viruses, each with unique structural traits, are employed to serve as templates for the bottom-up assembly of structures such as conductive nanowires for nanoelectronics, films for photovoltaics, and nanocomposites for sensors.

Chapters 8, 9, and 10 cover techniques and approaches used for integrating low-dimensionality, bottom-up synthesized materials such as CNTs (discovered in the early 1990s) with more mature technologies. Chapter 8 provides a comprehensive review of CNTs and the methods used to integrate them within the well-established semiconductor processing platform. Emphasis is placed on large-scale integration and manufacturability. Then, in Chapter 9, approaches used to integrate single-walled carbon nanotubes (SWCNTs) with CMOS circuitry are presented. The use of such hybrid SWCNT–CMOS components is discussed in the context of thermal and chemical sensing mechanisms; individual SWCNTs are decorated with deoxyribonucleic acid (DNA) to enhance sensitivity. In Chapter 10, the design, modeling, and analysis of NEMS-based ICs are discussed. These ICs appear to be more energy efficient than their CMOS-based counterparts and they rely on nanomechanical switches. Computation based on mechanical components dates as far back as the 1800s when Charles Babbage proposed the difference engine. It is interesting to note that we are now revisiting ideas formulated almost two

centuries ago in that mechanical structures, albeit in the nanoscale regime, are serving as building blocks for electronic computation.

While CNTs are linear and uniform, nonlinear CNT topologies such as SWCNT-based Y-junctions can be utilized to form novel architectures for electronic computation, which is the focus of Chapter 11. Such branched nanoelectronic architectures can constitute a switch or transistor, depending on the transmission characteristics of the constituent branches, and function vastly different than solid-state transistors.

The increased integration densities and higher clock speeds in Si ICs also created challenges for managing the large amounts of power or heat generated during microprocessor operation. In Chapter 12, thermal management schemes, specifically with phase change heat transfer techniques utilizing nanoscale materials and surfaces, are examined as viable cooling technologies. Finally, Chapter 13 surveys the broader applications of nanoengineered materials in the areas of electronics, biology, and energy harnessing.

Acknowledgments

I would like to express my sincere gratitude to all the distinguished contributors in this book for their comprehensive coverage of topics of current research in the exciting area of micro- and nanotechnologies that fill a wide variety of applications in electronics. I would also like to particularly thank Dr. Konstantin Novoselov (2010 Nobel Laureate, University of Manchester, UK), Professor Philip Feng (Caltech and Case Western Reserve), Professor Prabhakar Bandaru (University of California–San Diego), Professor Elaine Haberer (University of California–Riverside), and Professor Mehmet Dokmeci (Northeastern University) for providing the images that appear in the collage on the front cover.

In this regard, I would also like to acknowledge the support of my colleagues at the Jet Propulsion Laboratory (JPL), particularly Krikor G. Megerian and Robert Kowalczyk, for their technical assistance that enabled the images shown in the front cover to be made.

In addition, I would also like to thank Ashley Gasque, associate editor at CRC Press for her guidance and assistance in the publication of this book and Joselyn Banks-Kyle, project coordinator and Iris Fahrer, project editor at CRC Press for their editorial assistance in arranging the materials for this book.

I also gratefully acknowledge my family for their support through the years. I thank them from the bottom of my heart for their love, patience, and joyful spirit that sustain me and for which I am truly grateful.

Editor

Dr. Anupama B. Kaul obtained her BS degrees with honors in physics, as well as engineering physics, from Oregon State University. Her MS and PhD degrees were in materials science and engineering, with minors in electrical engineering and physics, from the University of California–Berkeley. Presently, she is a program director at the National Science Foundation (NSF) in the ECCS Division within the Engineering Directorate, where she is serving under the Intergovernmental Personnel Act (IPA) from the Jet Propulsion Laboratory (JPL), California Institute of Technology (Caltech). Prior to joining JPL, Dr. Kaul held industrial research positions at Motorola Labs and the R&D Division of Hewlett-Packard Company, where she worked on RF micro-electro-mechanical-systems (MEMS) components for wireless applications, and surface micro-machined ink-jet print heads, respectively. Her research interests are in characterizing the properties of functional micro- and nanoscale materials, developing bottom-up assembly and top-down nanofabrication techniques, and integrating such materials into novel devices and components for applications in electronics, sensing and energy harnessing.

While at JPL, Dr. Kaul has supported programmatic-review activities for technology development in support of various NASA proposal calls in planetary science. She has written more than 70 journal, conference and NASA technology brief publications, and received 6 NASA Patent Awards, a NASA Team Accomplishment Award, and holds 9 issued and pending patents. Dr. Kaul is a senior member of the Institute of Electrical and Electronics Engineers (IEEE) and has also contributed to several invited book chapters and has served as a guest editor for various journals.

Presently, Dr. Kaul serves as the American editor for *Nanoscience and Nanotechnology Letters*, associate editor of *Reviews in Advanced Sciences and Engineering* and is on the editorial boards of the *Journal of Nanoengineering and Nanomanufacturing* and the *Open Process Chemistry Journal*. Dr. Kaul has organized symposia for the Materials Research Society (MRS) on nanomaterials and devices, and serves as one of the track chairs in nanoelectronics for the IEEE NANO 2012. In addition, Dr. Kaul has also organized and chaired sessions for other conferences sponsored by the Nano Science and Technology Institute (NSTI) and SPIE. She is listed in the *Who's Who in America* and the *Who's Who in Science and Engineering*. Dr. Kaul has served as a reviewer on proposals for the NSF, NASA, and JPL, in addition to being a technical reviewer for leading journals within the Nature Publishing Group, the American Chemical Society, IEEE, and the MRS. Dr. Kaul was selected to participate in the 2012 US Frontiers of Engineering Symposium by the National Academy of Engineering. She is the Nanoelectronics Track

Chair for the IEEE NANO 2012 sponsored by the IEEE and has given invited and keynote talks at various international conferences sponsored by SPIE, MRS and NSTI.

Contributors

Prabhakar R. Bandaru
Department of Mechanical and Aerospace Engineering
University of California, San Diego
La Jolla, California

Debjyoti Banerjee
Department of Mechanical Engineering
Texas A&M University
College Station, Texas

Kaustav Banerjee
Department of Electrical and Computer Engineering
University of California, Santa Barbara
Santa Barbara, California

Paul Benning
Hewlett-Packard Company
Corvallis, Oregon

Rob N. Candler
Department of Electrical Engineering
University of California, Los Angeles
Los Angeles, California

Chia-Ling Chen
Department of Mechanical and Aerospace Engineering
University of California, Los Angeles
Los Angeles, California

Michelle Chen
Department of Physics and Engineering
Point Loma Nazarene University
San Diego, California

Daniel S. Choi
Department of Chemical and Materials Engineering
University of Idaho
Moscow, Idaho

Hamed F. Dadgour
Department of Electrical and Computer Engineering
University of California, Santa Barbara
Santa Barbara, California

Michael F.L. de Volder
Imec and KULeuven
Heverlee, Belgium

Mehmet R. Dokmeci
Harvard–MIT Health Sciences Technology
Harvard University
Cambridge, Massachusetts

Philip X.-L. Feng
Department of Electrical Engineering and Computer Science
Case Western Reserve University
Cleveland, Ohio

Elaine D. Haberer
Department of Electrical Engineering and Materials Science and Engineering Program
University of California, Riverside
Riverside, California

A. John Hart
Department of Mechanical Engineering
University of Michigan
Ann Arbor, Michigan

M. David Henry
Department of Applied Physics
California Institute of Technology
Pasadena, California

Matthew A. Hopcroft
Hewlett-Packard Laboratories
Palo Alto, California

Allan H. Johnston
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Bongsang Kim
Sandia National Laboratories
Albuquerque, New Mexico

Amit Marathe
Microsoft Corporation
Silicon Valley Campus
Mountainview, California

Eric R. Meshot
Department of Mechanical Engineering
University of Michigan
Ann Arbor, Michigan

Tanya Nigam
Global Foundries
Sunnyvale, California

Naresh Pachauri
Department of Chemical and Materials Engineering
University of Idaho
Moscow, Idaho

Sei Jin Park
Department of Mechanical Engineering
University of Michigan
Ann Arbor, Michigan

Susan Richards
Hewlett-Packard Company
Corvallis, Oregon

Leif Z. Scheick
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Axel Scherer
Department of Applied Physics
California Institute of Technology
Pasadena, California

Donghyun Shin
University of Texas at Arlington
Arlington, Texas

Navdeep Singh
Department of Mechanical Engineering
Texas A&M University
College Station, Texas

Sameer Sonkusale
Department of Electrical and Computer Engineering
Tufts University
Medford, Massachusetts

James Stasiak
Hewlett-Packard Company
Corvallis, Oregon

Sameh H. Tawfick
Department of Mechanical Engineering
University of Michigan
Ann Arbor, Michigan

Kok-Yong Yiang
Global Foundries
Sunnyvale, California

Zhikan Zhang
Department of Chemical and Materials Engineering
University of Idaho
Moscow, Idaho

1

Moore's Law: Technology Scaling and Reliability Challenges

Tanya Nigam, Kok-Yong Yiang, and Amit Marathe

CONTENTS

ABSTRACT Semiconductor technology involves continued scaling of semiconductor processes to the deep sub-micron and nanometer levels according to Moore's law, as well as the addition of new and complex materials and process modules. Aggressive scaling entails numerous challenges involving power dissipation, variability, reliability, yield, and manufacturing. Often,

the scaling of a technology is performed without proportionately reducing the power supply voltage to enable higher performance. Such an approach presents great challenges to device engineers, reliability engineers, and process integration engineers. As a result, trade-offs are usually required among reliability, design, and process development. The chapter highlights these numerous challenges in technology scaling for transistors and interconnects with special emphasis on the key reliability issues and the different wear-out mechanisms. The reliability of each new process module and how it interacts with other modules will be critical to the final reliability of the entire process. As the technology becomes more complex and aggressively scaled, reliability becomes critical as the technology is pushed to the limits to squeeze out every ounce of performance. Since generic technology reliability specifications can often be limiting and overly punitive, appropriate product reliability models are required to fully enable designs without compromising reliability.

1.1 Introduction: Technology Scaling Challenges

Silicon-based semiconductor device dimensions have been scaled continuously over the last 35 years. Current CMOS-based technologies have device dimension in the sub-100-nm range with gate dielectric thicknesses in the 1- to 2-nm range. Such advances largely followed the industry's governing tenet, i.e., Moore's law, which states that the number of transistors on a chip doubles every 2 years, as shown in Figure 1.1. To meet these needs, device dimension have shrunk 0.7 times per generation to improve performance by doubling frequency and reducing gate delay (Borkar, 1999; Paul et al., 2006, Taur et al., 1997).

Table 1.1 shows the scaling of typical device dimensions for different CMOS generations (Ghai et al., 2003). For the 70-nm technology node, the typical gate length is only 35 nm while the electrical gate oxide thickness (t_{OX}) is 1.6 nm and the source drain extension (SDE) depth is 17 nm. Modern microprocessors are manufactured with billions of transistors. Keeping power dissipation, variability, and reliability under control is therefore critical.

1.1.1 Power Dissipation

The active power dissipated in a CMOS chip is given by

$$Power = C \times V_{DD}^2 \times f \tag{1.1}$$

where, C is the capacitance, V_{DD} is the supply voltage, and f is the frequency of the circuit.

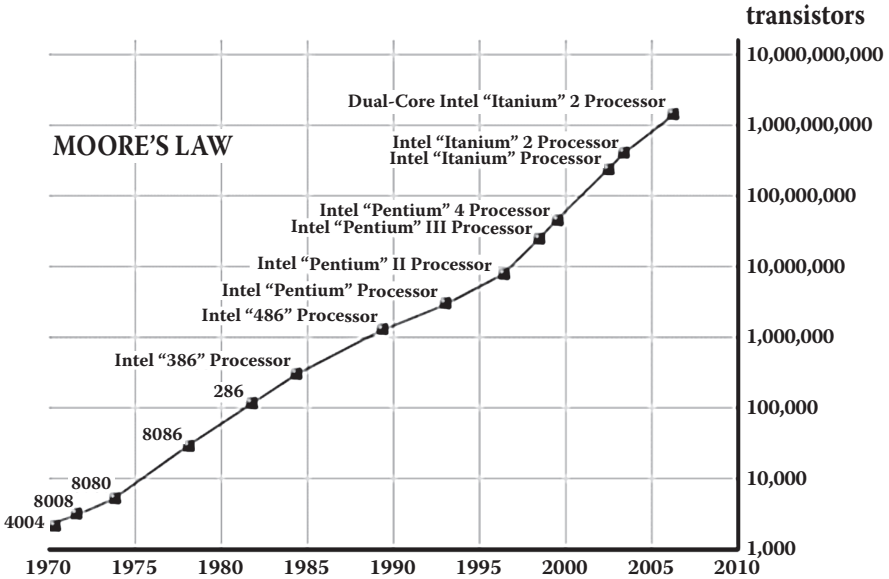


FIGURE 1.1
Plot of CPU transistor counts against dates of introduction. The line corresponds to exponential growth, with transistor count doubling every 2 years. (From http://download.intel.com/technology/silicon/Neikei_Presentation_2009_Tahir_Ghani.pdf)

There are two approaches to technology scaling: (1) constant power scaling where V_{DD} is not scaled and a reduction in capacitance is negated by an increase in f, and (2) reducing power where dissipation by scaling V_{DD} by 0.7 times, leading to a 50% reduction in active power for the scaled technology. The semiconductor industry has followed constant power scaling

TABLE 1.1

Scaling Projection of Transistor Parameters for Different Technology Generation Levels

	Generation Level (nm)				
Parameter	**180**	**130**	**100**	**70**	**Scaling Factor**
L_{GATE} (nm)	100	70	50	35	0.7×
V_{DD} (V)	1.5	1.2	1.0	0.8	0.8×
t_{ox} (e) (nm), t_{ox} (phys) (nm)	3.1, 2.1	2.5, 1.5	2.0, 1.0	1.6, 0.6	0.8×
SDE depth (nm)	50	35	24	17	0.7×
SDE under diff (nm)	23	16	11	8	0.7×
L_{MET} (nm)	55	40	27	20	0.7×
Channel doping ($\times 10^{18}$ cm^{-3})	1	1.6	2.6	4	$1/(0.8)^2 = 1.6\times$
I_{DSAT} (relative)	1	1	1	1	1×
I_{OFF} (nA/μm), 25°C	20	40	80	160	2×

Source: Ghai, T. et al. 2000. *Proceedings of Symposium on VLSI Circuits.* IEEE, Honolulu, pp. 174–175. With permission.

as V_{DD} scaling poses a challenge. V_{DD} scaling directly impacts the gate delay and increases sub-threshold leakage current, which in turn increases (static) power dissipation due to leakage. On the other hand, constant power scaling leads to gate leakage increase as the physical gate oxide thickness is scaled continuously to meet the performance requirements. Challenges associated with sub-threshold leakage and gate oxide leakage are discussed further.

1.1.2 Sub-Threshold Leakage

The sub-threshold current flows from the source to the drain of a transistor due to the diffusion of the minority carriers for gate-to-source voltages (V_{GS}) below the threshold voltage (V_{TH}). It depends exponentially on both V_{GS} and V_{TH} and is a strong function of temperature. Ideally, the ratio of V_{TH}/V_{DD} is kept below 0.25 so that the gate overdrive capability of the scaled device can be maintained and CMOS circuit performance is not compromised (Ghai et al., 2003; Taur et al., 1997).

In a long-channel device, V_{TH} does not depend on the drain bias or on the channel length. However, in short-channel devices, source and drain depletion regions penetrate significantly into the channel and control the potential and the field inside the channel. This is known as the short channel effect (SCE). As a result of SCEs, V_{TH} reduces via (1) a reduction in channel length (V_{TH} roll-off), and (2) an increase in drain bias (drain induced barrier lowering or DIBL; Taur and Ning, 1998). This results in increased sub-threshold currents in short-channel devices. In order to keep SCEs under control, both the gate oxide thickness and the depletion width of the transistor must be reduced. The latter requires tailoring of the channel doping profile by implanting retrograde wells while the former directly leads to reliability challenges associated with gate oxide thickness scaling.

1.1.3 Gate Leakage

Gate leakage increases exponentially with a decrease in the gate oxide thickness and an increase in the potential drop across oxide. It exhibits a weak temperature dependence. Gate current is primarily due to the tunneling of electrons (or holes) from the silicon bulk through the gate oxide potential barrier into the gate (or vice versa). Figure 1.2 shows how reducing the gate oxide thickness leads to an increase in tunneling current (Massoud et al., 1996). Gate leakage is critical during the off-state of the devices and results in the standby power dissipation of the chip.

The maximum tolerable gate leakage current for a 10 mm^2 chip is 1- to 10-A/cm^2 at V_{DD} = 1 V. As shown in Figure 1.2, SiON-based sub-2-nm oxides exceed this threshold. Based on Table 1.1, a gate oxide thickness of 1 nm and below is required for sub-50-nm technologies. Therefore, the use of conventional SiON-based gate stacks becomes a major challenge.

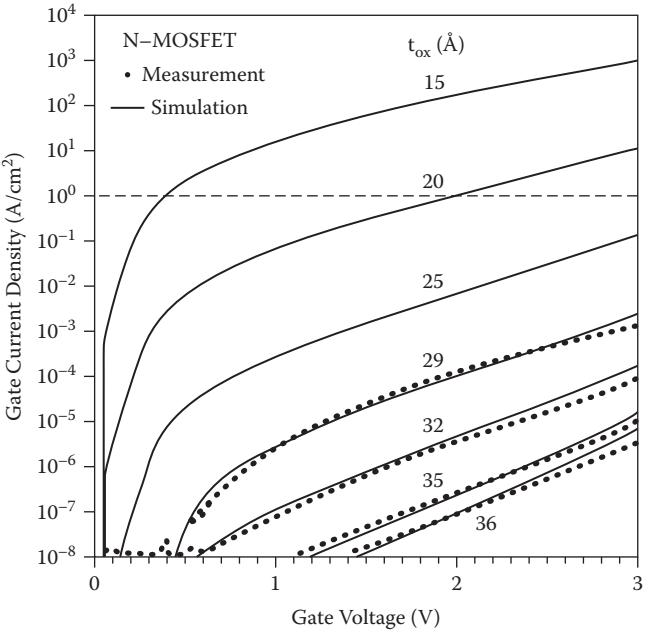


FIGURE 1.2
Measured and simulated I_G–V_G characteristics under inversion conditions for different oxide thicknesses. Dotted line indicates 1-A/cm² limit for leakage current discussed in text. (From Lo et al., 1997. With permission.)

One possible solution is to use dielectric films with higher dielectric constants (known as high-K materials, e.g., HfO_2 or Al_2O_3 etc.) such that the physical gate stack thickness can be increased, leading to a lower gate leakage current despite a lower equivalent gate oxide thickness as measured in inversion. The SiO_2 equivalent oxide thickness is:

$$EOT = t_{HK} \times \varepsilon_{SiO_2} / \varepsilon_{HK} \tag{1.2}$$

where ε_{SiO2} and ε_{HK} are the dielectric constants of SiO_2 and the high-K material, respectively.

Integrating high-K dielectric into CMOS technology is a major challenge and extensive work has been done to screen and select high-K dielectrics (Wilk et al., 2001; Huff and Gilmer, 2004; Gusev, 2006). Over the last 10 years, hafnium oxide with a dielectric constant of 25 (HfO_2, HfSiON, HfON, and HfSiO) emerged as a viable candidate to replace SiON. In addition to replacing the gate dielectric, it is also essential to replace the poly-Si gate with a metal gate to reduce the effective decrease in inversion capacitance due to depletion of the poly-Si at the high fields typically present in modern devices.

1.1.4 Variability

Power dissipation for scaled technologies is further complicated by process and device variability. For a review on this subject, please refer to Bernstein et al. (2006). Controlling physical dimensions in the manufacturing process now commonly involves atomistic-level constraints. Delay and power variability in CMOS devices are influenced by many factors.

Variability can be temporal or spatial in nature. Temporally, the variability can occur from nanoseconds (such as in the SOI history effect; Asenov et al., 2003) to years (such as in process centering). Aging-induced variation arising from wear-out mechanisms has a negative impact on performance. Bias temperature instability (BTI) and hot electron effects both elevate device thresholds and degrade device and circuit performance (Nigam et al., 2009; Frank et al., 2006). Electromigration (EM) slowly erodes interconnect admittance, becoming more severe below 65 nm because of higher interconnect current densities. Time-dependent variability is a strong function of capacitive loading and the ratio of p-FET to n-FET device widths (beta ratio), how often and how long the device is on (activity factor), and the chip environmental (voltage and temperature) operating conditions of a given circuit over the lifetime of a product.

Spatial variation refers to lateral and vertical differences from intended device dimensions and film thicknesses (Stine et al., 1997). Spatial variation modes exist between devices, between circuits, between chips, and across wafers, lots, and lifetimes of various fabrication systems. Such a variation is often referred to as extrinsic variability. Additionally, intrinsic device variations also occur due to atomic-level differences between devices that exist even though the devices may have identical layout geometries and environments. These stochastic differences appear in the dopant profiles, film thickness variations, and line-edge roughness. A typical example of such variations is V_{TH} variation due to the atomistic nature of the dopants in MOSFETs. The implant and annealing processes result in the placement of a random number of dopants in the channel (described by a Poisson distribution).

Variations are classified as (1) those that involve the chip mean, (2) those that vary within the chip but have local or chip-to-chip correlation, and (3) those that vary randomly from device to device. Chip mean variations can be caused by spatial variation (L_G, V_{TH}, and t_{OX}) and temporal variation (operating temperature, activity factor, and device degradations such as BTI and HCI). Within-chip variation comes from pattern-density or layout-induced transconductance on die hot spots and hot spot-induced BTI. Finally, device-to-device variation comes from atomistic variations in dopant distributions, line edge roughness, and BTI- or (hot carrier injection) HCI-induced V_{TH} distributions.

1.2 Transistor Reliability

Reliability is the probability that a product will perform a required function under stated conditions for a stated period of time. A typical example of reliability is gate oxide integrity. An oxide is defined as reliable if it maintains its insulating properties for 10 or 25 years at a specified bias, temperature, chip area, and failure fraction.

Reliability studies typically require accelerated testing conditions such that the physical mechanism responsible for breakdown can be studied in a time frame much shorter than the targeted lifetime. Based on stress at accelerated conditions, projections are made toward the use conditions. Intrinsic reliability studies revolve around generation of material defects that lead to product failure. Since defect generation is random in nature, the statistical nature of defect generation and its impact on reliability must be understood. Strong et al. (2009) performed an extensive review of device reliability.

1.2.1 Time-Dependent Dielectric Breakdown

Time-dependent dielectric breakdown (TDDB) occurs during the off-state when the voltage across the gate dielectric is high. TDDB failure was traditionally catastrophic and caused the gate dielectric to lose its insulating properties after the breakdown event, leading to a functional failure of the chip. As technology is scaled downward, TDDB is no longer automatically considered catastrophic since the dielectric does not fully lose its insulating properties for sufficiently thin gate oxides.

Typically, a dielectric breakdown leads to the formation of a leakage path through the oxide. Figure 1.3 illustrates the defect generation and eventual breakdown of a dielectric. As a high voltage stress is applied, defects are randomly generated in the bulk of the oxide (Stage 1). During TDDB, bulk defects are created due to the breaking of Si-O bonds and the defects are permanent under typical operating conditions.

Once a critical density of defects is reached, a localized leakage path is formed (Stage 2, often called soft breakdown or SBD). According to Weir et al. (1997), the formation of such a leakage path does not lead to a complete loss of insulating properties. If the stress continues, gate leakage further increases and finally the dielectric breaks down, resulting in ohmic conduction through the gate stack (Stage 3, hard breakdown or HBD).

Since the defect generation and breakdown path creation are random in nature, TDDB requires statistical description by means of a distribution. Dielectric breakdown is described very well by Weibull statistics. Gate oxide breakdown represents the weak link phenomena as the different areas of the gate oxide are competing with each other for the formation of breakdown path. The area with highest density of defects leads to breakdown. Weibull statistics is described by

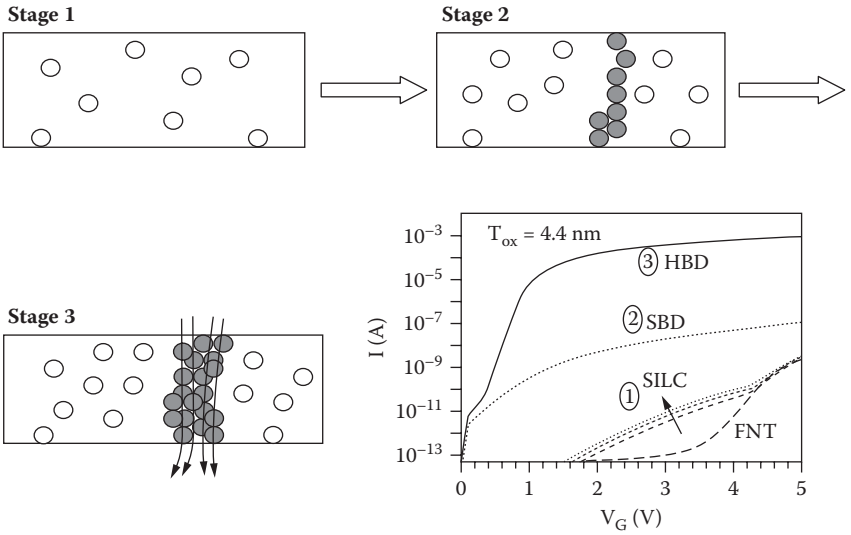


FIGURE 1.3
Defect generation in a vertical cross-section of gate oxide. Stage 1: defects are generated in the oxide. Stage 2: a local conducting path is formed (soft breakdown or SBD). As the stress continues, a final hard breakdown with complete loss of insulating properties occurs. The current voltages for the three stages of oxide wear-out are also shown.

$$F(t_{BD}) = 1 - \exp(-(\frac{t_{BD}}{\tau})^{\beta}) \tag{1.3}$$

where $F(t_{BD})$ is the cumulative failure probability and t_{BD} is the time to breakdown. The characteristic time to breakdown τ is the time for 63rd percentile and β is the Weibull shape factor. An important property of Weibull slope is that it describes three different failure rates that occur during product life. The extrinsic failure is described by Weibull shape $\beta < 1$ which corresponds to decreasing failure rate, random fails for $\beta = 1$ with constant failure rate, and intrinsic fails for wearout with $\beta > 1$. Typical extrinsic and intrinsic Weibull distributions are shown in Figure 1.4.

The observed intrinsic distribution of oxide breakdown can be modeled using percolation theory (Degraeve et al., 1998; Stathis et al., 1999). Under the percolation approach, the defect generation continues until a critical defect density is reached. To model the probability for reaching a critical defect density, defects are randomly generated using Monte Carlo (MC) simulation until a local connecting path is formed as shown in Figure 1.5a. Using this approach, the reduction in Weibull slope for thinner gate oxides was predicted for the first time in agreement with the experimental findings (see Figure 1.5b).

A three-dimensional (3D) analytical model has also been proposed to model the defect generation in gate oxide (Sune 2001; Krishnan and Nicollian, 2007.

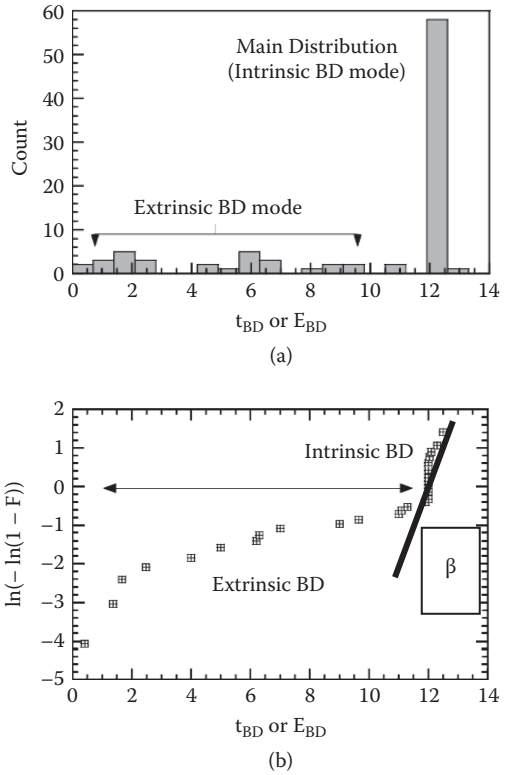


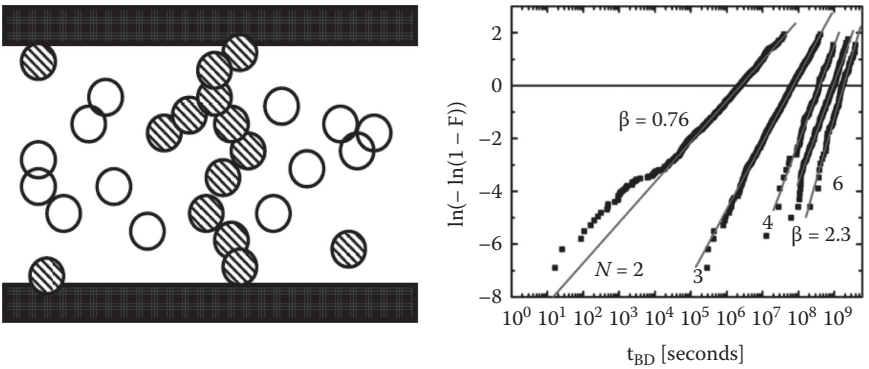
FIGURE 1.4
Probability distribution (a) and Weibull distribution (b) of gate oxide breakdown data. The sharp part of the distribution corresponds to the intrinsic breakdown while shallow distribution is extrinsic. The Weibull shape factor is given by β and τ corresponds to l n (–l n (1 – F)) = 0.

The analytical model is similar to the MC approach and also predicts the decrease in Weibull slope as a function of thickness. The analytical expression for Weibull shape factor that decreases as a function of oxide thickness is given by:

$$\beta = \alpha \frac{t_{OX}}{a_0} \tag{1.4}$$

where α is the defect generation rate, t_{OX} is the gate oxide thickness, and a_0 is the defect size.

The percolation approach using MC simulations that was applied successfully to single SiO_2 layers can also be extended to any dual or multilayer system as in high-K gate stacks (Nigam et al., 2009). For this work, the kinetic MC technique was used so that the defect generation rate in the two layers could be varied independently. Results of the simulation are shown in Figure 1.6.



(a) (b)

FIGURE 1.5
(a) Defect generation and percolation path formation. (b) Weibull distribution as a function of gate oxide thickness. A reduction in Weibull slope is observed as the number of layers in the gate oxide is reduced. N^*a_0 indicates physical oxide thickness; a_0 represents defect size.

To gain insight into the nature of dielectric stack breakdown with a non-uniform defect generation rate, we recorded the breakdown path during kinetic Monte Carlo (kMC) simulations for a stack with a three-layer high K (HK) and two-layer interfacial layer (IL), shown in Figure 1.6a. Depending on the failure fraction, the density of defects in the HK varies significantly. For very small areas or, equivalently, high failure fractions, the defect density in the HK layer is very high and we are limited by the IL (and the Weibull slope is that of the IL).

For low failure fractions or large areas, where lucky events dominate, the breakdown path resembles the case of a uniform oxide with few defects in the HK layer, in agreement with Krishnan and Nicollian (2007). In this case, the Weibull slope is determined by the complete stack thickness. Figures 1.6b through e show the defect distribution around the breakdown path. Figure 1.6f shows the 300× higher defect density at breakdown in the HK layer as compared to the IL for the kMC simulation of Figure 1.6a.

The above approach of using a percolation model for multiple layers may explain the low Weibull slopes observed for HK dielectrics (Bersuker et al., 2008; Degraeve et al., 1999) and the assessment that the IL determines TDDB for these stacks. This is true only for small area devices.

The possibility of a large number of defect paths in the HK layer prior to stack breakdown was proposed by Okada et al. (2007), but its impact on the statistics of the HK stack breakdown was not addressed. The dual layer model of Nigam et al. (2009) showed that if defect generation rate in the HK is higher than that in the IL, the TDDB distribution becomes bimodal. For large areas or small failure fractions, the Weibull slope increases to a value consistent with the complete stack thickness and measured data. This significantly enhances TDDB predictions for typical products that use high-K gate

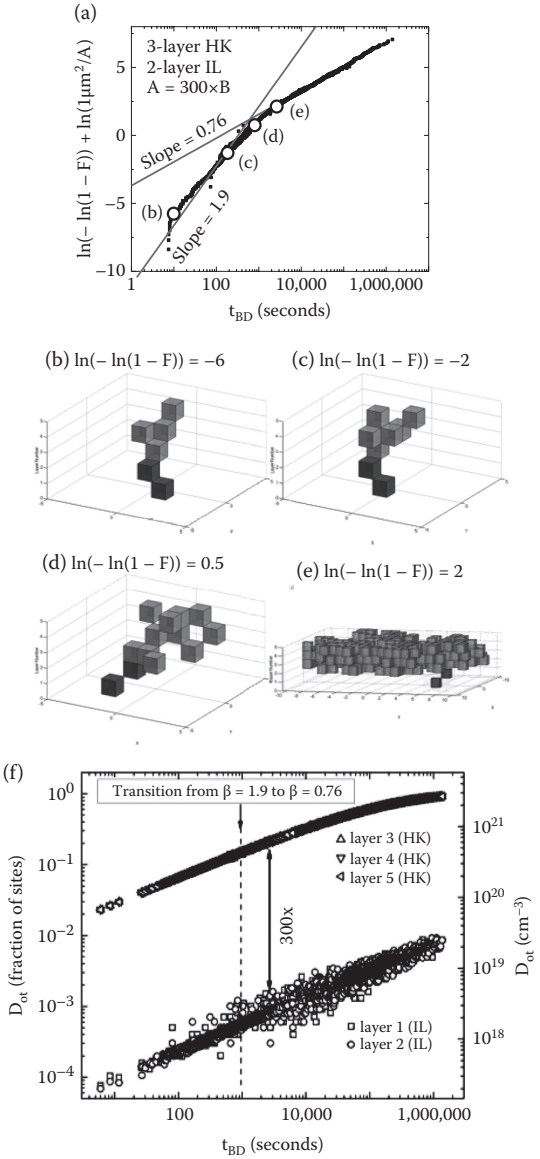


FIGURE 1.6
(a) Simulated TDDB distribution for a dielectric stack with a two-layer IL and three-layer HK. The Weibull slope for large failure fractions is that of a two-layer dielectric; the slope for small failure fractions is that of a five-layer dielectric. (b)–(e) Breakdown path shapes typical for various failure fractions (f). (f) D_{ot} in the HK layer and IL as a function of t_{bd}.

stacks. Application of the percolation concept to gate oxide breakdown has been very successful in extending gate oxide scaling without compromising product reliability.

1.2.2 Bias Temperature Instability

Bias temperature instability (BTI) occurs during an off-state condition with a uniform field across the oxide. It causes a shift in FET parameters such as threshold voltage (V_{TH}), saturation regime drain current (I_{DSAT}), and linear regime drain current (I_{DLIN}) according to Schroder and Babcock (2003). BTI is a major challenge as it occurs at low fields and is enhanced at higher temperatures. It is observed for both PMOS (negative bias temperature instability) and NMOS (positive bias temperature instability) devices as technology scales have evolved. BTI is a strong function of gate stack processing conditions (Schroder and Babcock, 2003; Kerber and Cartier, 2009).

1.2.2.1 Negative Bias Temperature Instability

For PMOS devices, BTI has become a reliability concern because of the switch to surface channel devices instead of buried channel devices. Reliability margins due to negative bias temperature instability (NBTI) decreased further as more nitrogen was incorporated in the oxides to prevent boron penetration and to increase the dielectric constants. NBTI is caused by the generation of interface states (N_{IT}) due to the presence of cold holes that weaken the Si-H bond and cause its dissociation and charge trapping in the oxide close to the Si–SiO_2 interface. Significant work has involved modeling and explaining NBTI (Schroder and Babcock, 2003; Mitani, 2004; Mahapatra et al., 2003).

N_{IT} generation has been modeled using a reaction–diffusion process in which the presence of cold holes triggers an electrochemical reaction coupled to the diffusion of hydrogenated species in the gate oxide (Alam, 2003; Jeppson and Svensson, 1977; Chakravarthi et al., 2004). The electrochemical reaction leads to the breaking of Si-H bonds at the Si–SiO_2 interface. The interface generation rate is initially controlled by the electrochemical reaction process and subsequently by the diffusion of hydrogen species. In additional to N_{IT} generation, significant charge trapping is also observed during NBTI stress and may be attributed to two different mechanisms: (1) positive fixed-charge formation as a by-product of reaction–diffusion processes and (2) charge trapping into pre-existing defects in the oxide. While process (1) occurs sequential to N_{IT} generation, process (2) occurs parallel to N_{IT} generation.

NBTI relaxes after the stress bias is removed (Chen et al., 2002). The cause of recovery is still controversial. It can be attributed to both re-passivation of interface states by the available hydrogenated species in the oxide and/or de-trapping of the trapped charge in the oxide. NBTI relaxation makes measurement of the true NBTI component very challenging because delays

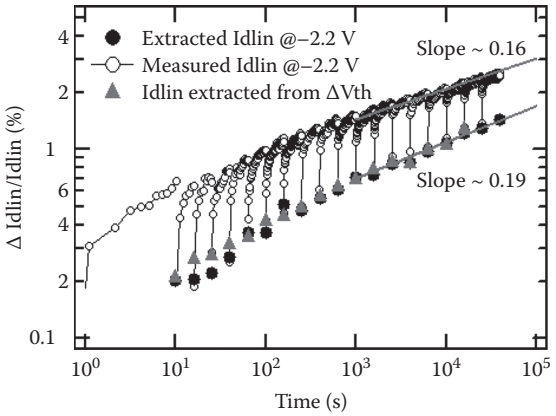


FIGURE 1.7
Continuous measurement of $\%I_{dlin}$ during stress with a $V_{DD} = -0.1$ V (open circle). Closed circle shows $\%I_{dlin}$ obtained from an IV measurement. Also shown is the extracted $\%I_{dlin}$ from a V_{TH} measurement (triangles). Based on the reaction–diffusion model, the slope of 0.16 corresponds to the diffusion of molecular hydrogen and not atomic hydrogen, consistent with the obtained activation energy.

in the range of a few milliseconds already cause significant relaxation. This directly impacts the ability to measure the true NBTI component.

Typically the impact of accelerated voltages on I_{DLIN} is measured after a certain delay. In order to minimize the amount of recovery, an "on-the-fly" technique in which I_{DLIN} is measured during stress (open circles) was used, as shown in Figure 1.7. The figure also shows the amount of I_{DLIN} degradation measured after a delay of 1 second (closed circles). Delay reduces the magnitude of damage and increases the time dependence of the NBTI-induced degradation. The higher slope obtained due to measurement delay can be explained by the stress time dependence of relaxation.

Note the delay between two stress cycles is fixed while the stress time increases exponentially; hence the ratio of t_{Relax}/t_{Stress} decreases as the stress progresses. Since the fractional recovery shows a universal curve as a function of t_{Relax}/t_{Stress} (Denis et al., 2006; Grasser et al., 2007), a lower t_{Relax}/t_{Stress} corresponds to less fractional recovery at longer stress times and a smaller impact of relaxation. This has significant implications for predicting product level reliability as discussed in Section 1.2.2.3.

1.2.2.2 Positive Bias Temperature Instability

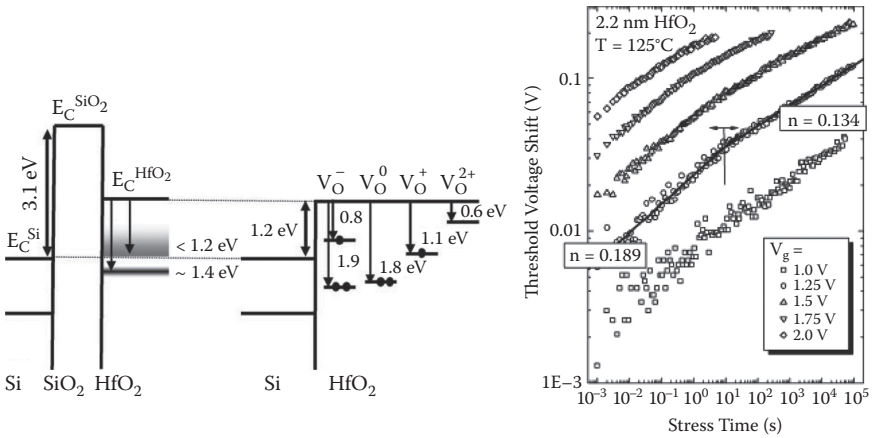
For NMOS devices, the switch to higher dielectric constant materials leads to charge trapping in high-K materials. A high-K gate stack is a multilayer structure consisting of an interfacial SiO_2 layer (IL), a deposited high-K layer, a metal gate, and a poly-Si layer. The presence of IL is critical for the success

of a high-K gate dielectric. Increasing the thickness of the IL improves reliability of high-K gate stacks.

Electron trapping in SiO_2–HfO_2 dual-layer gate stacks has been studied intensely because it is strongly enhanced with HK–MG stacks. Electron trapping has been attributed to defects in the IL (oxygen deficiency-related defect precursors; Heh et al., 2006), defects at the interface between the two dielectrics (Casse et al., 2006), defects in the HfO_2 layer (Kerber et al., 2003), and defects at the HfO_2–TiN interface (Torii et al., 2003).

Extensive work based on frequency-dependent charge pumping indicates that the defects are related to oxygen vacancy in HfO_2. Charge pumping measurements suggest the presence of trap levels at energies $Ev - Ec$ HfO_2 ~1.4 eV and additional levels at $Ev - Ec$ HfO_2 <1.2 eV, where Ec HfO_2 is the energy of the bottom of the HfO_2 conduction band, and Ec [Ev] is the energy of the bottom [top] of the Si conduction [valance] band as shown in Figure 1.8a.

These numbers fall well within the range of calculated electron energy levels for the various charge states of the oxygen vacancy in HfO_2 (Robertson, 2006). Since the electron trapping is due to tunneling of the carrier through the IL, increasing the thickness of IL reduces the magnitude of positive bias temperature instability (PBTI) while increasing the thickness of high-K material increases PBTI. The shift in NMOS device characteristics after PBTI-like stress is shown in Figure 1.8b.



(a) (b)

FIGURE 1.8
(a) Left: band diagram containing the energy level position of the electron traps as derived from spectroscopic ACP experiment. Right: calculated energy levels for oxygen vacancies in various charge states. (From Cartier, E. et al. 2006. *Proceedings of International Electronics Development Meeting,* IEEE, San Francisco, 321–324; Robertson, J. 2006. *Rep. Progr. Phys.,* 69, 327–396. With permission.) (b) Threshold voltage shift during PBTI stress at 125°C for 2.2-nm thick HfO_2 layer. (From Kerber, A. et al. 2008. *IEEE Trans. Electron. Dev.,* 55, 3175–3183. With permission.)

1.2.2.3 Relaxation in Bias Temperature Instability

It has been shown that BTI shows significant relaxation after stress is removed (Chen et al., 2002; Kerber and Cartier, 2009). Depending on the temperature and stress time, a V_{TH} recovery up to 80% is possible. Logarithmic time dependence is observed during relaxation for both NBTI and PBTI (see Figure 1.9a) unlike the stress phase in which a power law is observed.

If the fractional recovery is plotted as a function of relaxation time over stress time (t_R/t_S) a universal curve is obtained (Graser et al., 2007). For NBTI, the recovery is attributed to the re-passivation of the generated interface states (Alam, 2003; Chakravarthi et al., 2004) during the zero bias state and/or hole de-trapping (Huard et al., 2006) upon removal of stress. For PBTI, recovery is attributed to the de-trapping of trapped electrons. This implies that for a circuit level stress (such as a ring oscillator) in which a device is continuously switching, a lower BTI is expected and measured as shown in Figure 1.9b (Nigam, 2008).

1.2.3 Hot Carrier Injection

Hot carrier injection (HCI) occurs during the on-state condition with a high voltage on the drain that leads to a non-uniform field across the oxide. As with BTI, HCI also causes a shift in FET parameters such as threshold voltage (V_{TH}), saturation regime drain current (I_{DSAT}), and linear regime drain current (I_{DLIN}). HCI modeling and data have changed as the technology has scaled.

Hot carriers are generated by a high lateral electric field in the channel. When the mean kinetic energy of the carrier is higher than the lattice

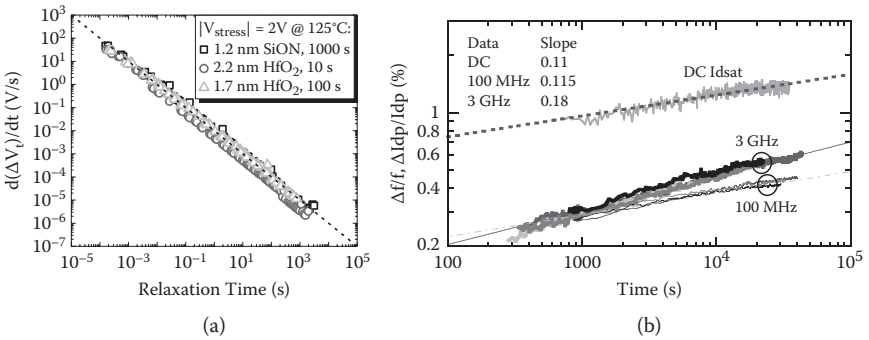


FIGURE 1.9
(See color insert.) (a) Threshold voltage recovery for NMOS and PMOS with 2.2-nm and 1.7-nm HfO2 gate dielectric and 1.2 nm SiON. The recovery follows a log(t) dependence for both NBTI and PBTI. (From: Kerber, A. et al. 2008. *IEEE Trans. Electron. Dev.*, 55, 3175–3183. With permission.) (b) Impact of AC stress using RO. Ring oscillator frequency degradation at 25OC along with I_{dsat} degradation for DC stress on PMOS transistor. Lower degradation is measured for RO as compared to DC stress. (From Nigam, T. 2008. *IEEE Trans. Dev. Mater. Reliability*, 8, 72–78. With permission.)

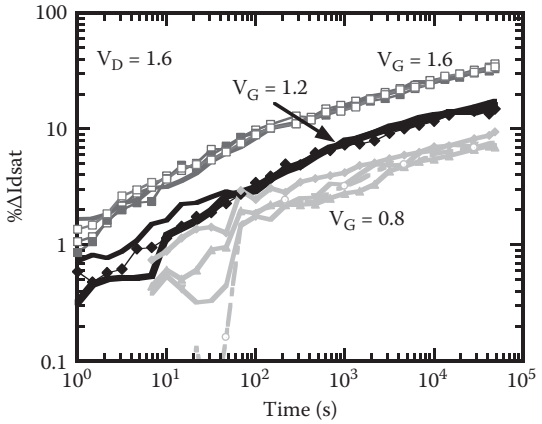


FIGURE 1.10
HCI degradation increases as a function of V_{GS} for a fixed V_{DS}. Channel length for these NMOS devices is 40 nm.

temperature, a carrier is "hot." The generated hot carriers can be injected into the oxide causing bulk defect generation or charge trapping. They can also lead to N_{IT} generation near the drain. Typically the damage due to HCI is highly localized. HCI increases as the channel length (L_G) is reduced. To reduce the HCI-induced device parameter shift, the lightly doped drain (LDD) was introduced. The main goal was to reduce the peak lateral electric field as the technology was scaled. Junction optimization significantly impacts HCI.

Typically HCI is described by the electron distribution function (EDF). As carriers travel through the channel, they gain energy due to the high fields and lose energy due to scattering events such as electron–phonon scattering and electron–electron (e-e) scattering. The combined effect of these two mechanisms is a wide distribution of electron energy. As the channel lengths are scaled, the pinch-off region approaches the dimension of the mean free path of the carrier and for sub-100-nm technologies, the carriers travel quasi-ballistically. The presence of e-e scattering in sub-100-nm device dimension leads to a shift in worst case HCI from $V_G = V_{DS}/2$ to $V_G = V_{DS}$ (see Figure 1.10).

At low V_{DS}, e-e scattering broadens the EDF tail as compared to the thermally distributed tail and exhibits a strong dependence on I_{DSAT}. In recent years, HCI-related degradation has received less attention due to the reductions of operating voltages, and it is expected that HCI damage will reduce as well. During circuit operation, HCI occurs only during switching. Therefore significant relief is expected as compared to the DC degradation (Nigam et al., 2009).

1.3 Back End of Line: Interconnect Technology

Advanced integrated circuits (ICs) require elaborate wiring (or interconnect) systems to distribute power, grounding, and various clock and input and output (I/O) signals to and from transistor devices. To maintain the cost and performance benefits associated with reduced transistor feature size and higher on-chip device density, the interconnect architecture must correspondingly increase in complexity and density; this is achieved by reducing the geometrical dimensions of the wirings and increasing the number of interconnect layers. The downsides to the "interconnect scaling," however, are an increase in wiring resistance (due to the smaller cross-sectional area) and the parasitic capacitance of wires that exert serious impacts on dynamic power consumption,* self-heating, and signal propagation speeds in the form of increased resistance–capacitance (RC) delay.

It was generally recognized that interconnect RC delay would effectively limit further gains in transistor speeds at the 0.25-μm technology node if no new interconnect materials with lower resistance and permittivity (K value) were used to replace the traditional Al and SiO_2 (Ho et al., 2002). To meet future performance requirements, the National Technology Roadmap for Semiconductors (NTRS) as early as 1994 began to predict the accelerated adoption of low-K interconnect materials. However, the commercialization of low-K materials to be more challenging than anticipated—only at the 0.13-μm technology node (in 2002) were higher conductivity Cu and low-K materials introduced. Even then, the implementation of even lower K materials (with nanometer porosities) has been slow and problematic, in part due to the reliability and yield issues associated with the integration of these materials. This delayed the production predictions of the International Technology Roadmap for Semiconductors (ITRS) by many years.

1.4 Evolution of Interconnect Materials and Patterning Schemes

Three metal candidates (Cu, Ag, and Au) exhibit lower bulk resistivities than Al, and the only practical option in terms of cost and manufacturability is Cu. Due to the high diffusivity of Cu into the dielectrics, a thin conformal barrier (liner) on the sidewalls and trench bottom of the interconnect wire is required. Since Cu does not form a volatile by-product, it is very difficult to

* The equation for dynamic power consumption P is generally $P = \alpha CfV^2$, where α is the wire activity, f is the transmission frequency, V is the power supply voltage, and C is the cumulative capacitances of the devices and interconnect wirings (Maex et al., 2003).

TABLE 1.2

Spun-On and Chemical Vapor-Deposited Low-K Dielectric Materials

Spun-On	Chemical Vapor-Deposited
B-staged polymers (CYCLOTENE™, SiLK™)	Fluorine-doped oxide (FSG)
Hydrogen silsesquioxane (HSQ)	Carbon-doped oxide (SiOCH)
Polybenzoxazole-based oxazole (OxD)	Parylene-N, parylene-F
Methyl silsesquioxane (MSQ) (LKD™)	
Divinylsiloxane-benzocyclobutene (BCB)	
Poly-tetrafluoroethylene (PTFE) (Teflon™)	
Fluorinated poly(arylene ether) (FLARE™)	

etch. This led to the development and adoption of the damascene approach whereby the wiring pattern is created in the dielectric by dry etching before filling with Cu and planarized to remove excess metal on top of the dielectric. This is in contrast to the traditional subtractive approach by which dielectric (SiO_2) is deposited on patterned Al metal and planarized.

The choice of low-K dielectric was less clear. Throughout the 2000s, research focused on different materials and deposition techniques for low-K materials, and a proliferation of spun-on and chemical vapor-deposited (CVD) dielectrics with bulk permittivities (K values) of 2.5 to 3.7 surfaced (Table 1.2). For these materials, the low K values (compared to ~4.2 for SiO_2) are achieved by incorporating atoms and bonds with lower polarizability, reducing the atomic and bond density, or both. Since 2005, the industry gradually gravitated toward CVD SiOCH because of its close resemblance to SiO_2 in terms of chemical, thermal, and mechanical properties.

While the further lowering of K values below 2.5 is possible by introducing nanometer porosity or macroscopic air gaps into the low-K dielectrics, these schemes create additional integration issues related to moisture uptake, damages arising from etch and chemical-mechanical polishing (CMP), and insufficient mechanical strength to withstand the forces during dicing, packaging, and assembly.

1.5 Challenges in Interconnect Reliability

More than three decades of continual CMOS scaling have now pushed existing interconnect materials to their reliability limits. The damascene fabrication process and materials fundamentally changed the stress states and interfaces of Cu wires compared to the subtractive patterning process for Al. In addition, the minimum thickness of low-K dielectrics between wires have scaled to deep submicron regimes (<50nm). Although this is nearly 50 times

the gate oxide thickness, low-K dielectrics are far from the quality of gate oxides in terms of electrical, thermal and mechanical characteristics.

One major interconnect challenge associated with the adoption of new Cu and low-K materials is electromigration (EM). In Al interconnects, momentum exchange between the current-carrying electrons and host metal lattice can cause the Al ions to drift in the direction of the electron current, causing the formation of a void at flux divergent sites (i.e., regions of tensile stress). At locations of high compressive stress, extrusions and hillocks can also form, causing electrical shorts to neighboring wires. EM reliability has been shown to depend on the grain size and crystallographic texture of the Al (Vaidya and Sinha, 1981; Campbell et al., 1993; Toyoda et al., 1994; Knorr et al., 1996; Rodbell et al., 1996).

Cu has a higher melting point than Al, and at operating conditions is supposed to offer a 40× increase in reliability (Besser et al., 2000). However, while Al readily forms a stable native oxide that acts as a slow diffusion site, Cu requires a barrier (at the side walls and trench bottom) and a dielectric capping layer (at the top) to prevent diffusion into the low-K dielectric; these interfaces of Cu to barrier and capping layer provide a fast diffusion pathway for Cu to electromigrate (Hu et al., 2004).

Although the fundamental failure mechanism and processing of Al and Cu metallizations are different, their effective atomic diffusivity D_{eff} follows a similar equation (below) and is the sum of the individual diffusion coefficients along grain boundaries (D_{gb}), barrier interface (D_b), and capping layer interface (D_c). The diffusivities are functions of grain size (d) and the width (w) and height (h) of the wire; δ_{gb}, δ_b, and δ_c are the dimensions of the grain boundary and interface layers.

$$D_{eff} = D_{gb} \frac{\delta_{gb}}{d} + D_b \delta_b \left(\frac{2}{w} + \frac{1}{h} \right) + D_c \frac{\delta_c}{h} \tag{1.5}$$

With geometrical scaling, EM performance is expected to degrade due to the relative increase in Cu interface area-to-volume ratio. The lower modulus low-K dielectrics also provide less backflow stress (i.e., reduced Blech effect) that may increase the risk of electromigration failure. Optimizing Cu interconnect reliability therefore includes, but is not limited to, optimizing the interfaces and process temperatures, using alternative capping materials such as CoWP, CoSnP and Pd (Hu et al., 2002), and doping the Cu with alloy elements (such as Al). A very thin (~10 nm) CoWP layer is reported to be sufficient in reducing most interfacial diffusion (Hu et al., 2003).

Another failure mechanism associated with interconnect metals is stress migration (SM) defined as the movements of metal atoms under the influence of mechanical stress gradients. Cu wires exist in a state of tensile stress due to the different materials and process temperatures of chip fabrication.