Integrated Circuits and Systems

Yu Cao

Predictive Technology Model for Robust Nanoelectronic Design

Springer

Integrated Circuits and Systems

Series Editor
Anantha P. Chandrakasan
Massachusetts Institute of Technology
Cambridge, Massachusetts

For further volumes:
http://www.springer.com/series/7236

Yu Cao

Predictive Technology Model for Robust Nanoelectronic Design

Foreword by Chenming Calvin Hu

Springer

Yu Cao
School of ECEE
Arizona State University
Tempe, AZ, USA
ycao@asu.edu

ISSN 1558-9412
ISBN 978-1-4614-0444-6 e-ISBN 978-1-4614-0445-3
DOI 10.1007/978-1-4614-0445-3
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011931530

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To Xuejue,
your dance of the water
sings the pebbles of my life

Foreword

The minimum feature size of CMOS technology will approach 10 nm in 10 years. Such aggressive scaling will lead to wonderful benefits to consumers, businesses and the global society. Unfortunately, it will also lead to increased power dissipation, process variations and device drift, posing tremendous new challenges to designing robust circuits. Already, the design complexity and time are increasing at accelerating rates. The lure of early market entry pushes advanced design research to begin much earlier than the completion of device technology development. The need is even clearer where new devices, e.g. FinFET and post-silicon devices are involved. The concept of technology/circuit co-development is no longer just a good idea, it is a necessity.

This new paradigm requires predictive SPICE transistor models for future technology generations, including both nanoscale CMOS and post-silicon devices. SPICE models used in circuit design are traditionally extracted from measurements taken on working transistors generated by the technology development process. In stark contrast, predictive SPICE model is created before the physical transistor has been fabricated, thus allowing design research to get an important early start. A predictive model is critical to identifying emergent problems and enable early search for solutions. While integrated semiconductor companies already make significant efforts to generate predictive models, fabless companies and university researchers usually do not have access to them.

PTM, a canonical *Predictive Technology Model* of both transistors and interconnect, offers a generic, open-source tool for early stage design research. Based on the standard BSIM model, PTM projects technology scaling down to the 12 nm node. It has been adopted for a broad range of research on low-power design, design robustness, system integration, design tools, and for university teaching, worldwide. As PTM becomes the de facto device model for advanced design benchmarking, this book timely reveals the “hidden secrets” behind PTM. I am proud to have worked with Prof. Cao to develop the early PTM (then called BPTM) at UC Berkeley in the late 1990s. Prof. Cao has expanded PTM from a simple predictive model of conventional MOSFETs into a suite of predictive models ranging from models of very small MOSFETs and promising alternative devices to process variability and reliability models embedded into the device models. These predictive models are further incorporated into the design environment, through predictive PDKs.

This is the first book to help university researchers and industry practitioners to understand predictive modeling principles and to gain insights into future technology trends. As evidenced by the thousands of research publications based on the use of PTM, the understanding and insights provided by this book will have a far-reaching impact on future circuit design research and IC development.

Berkeley, California Chenming Calvin Hu

Preface

The story of PTM, standing for Predictive Technology Model, is dated back to the year of 1999, when IC designers were hectically migrating from 0.25 μm to 0.18 μm CMOS technology. At that moment, many new problems were emerging from the physical level, such as short-channel effects and crosstalk noise, posing significant challenges that slowed down the product development. PTM was proposed to help bridge the technology and design groups, such that these issues can be brought to the attention as early as possible in the design process. Enabled by PTM, the new concept of concurrent process-design development is then widely practiced by university and industry groups. PTM effectively enhances design productivity and catalyzes the silicon evolution into the nanoscale regime.

Ten years after the start, PTM has successfully developed state-of-the-art CMOS models toward the 10 nm node. They are well disseminated through the web interface, and adopted into university curriculums. The demand of predictive modeling becomes even stronger today, as we are facing much more complicated and more diverse technological choices, as well as much larger scale of integration. This book covers both the essence of modeling principles and the application of PTM in nanoelectronic design. The chapters are intended primarily for IC designers and EDA tool developers, who have the background in transistor physics and circuit performance analysis. The discussion will especially benefit those with research interests in the areas of technology scaling and compact modeling.

The book starts with the background and overview of PTM. Chapter 1 reviews the important issues as CMOS technology is scaling toward the 10 nm node. It motivates the shift of IC design paradigm, in which PTM is the essential component. Current PTM provides standard compact model of bulk CMOS devices, BSIM4, down to the 12 nm node. Chapter 2 presents the systematic approach to scale device model parameters for future bulk devices, based on the solid understanding of device physics and silicon data as a reality check. Furthermore, Chap. 3 deals with recent extensions of conventional CMOS devices, including strained Si, high-k/metal gate, and the double-gate structure. Below the 90 nm node, these non-traditional materials and structures are vitally important to enhance the device performance. Modeling solutions to them are compatible with standard CMOS model and circuit simulation tools.

With CMOS scaling approaches fundamental physics and manufacturing limits, process variability and reliability degradation becomes the key limiting factors for future integrated system design. Chapters 4 and 5 address these concerns by developing statistical modeling, extraction and simulation techniques. New compact models are proposed for emerging variability and reliability effects, such as NBTI, in order to support design exploration for reliability. Besides these parasitic effects of transistor scaling, interconnect parasitics play an increasingly significant role in contemporary IC design. Chapter 6 presents modeling results of wire capacitance, capturing the latest advancement in interconnect technology.

These device models provide the basis of design benchmarking and tool development. Using PTM, Chap. 7 quantitatively evaluates various technology factors in scaled CMOS design, helping shed light on the performance trend along the roadmap. Moreover, Chap. 8 describes a 45 nm predictive process design kits (PDK), which are the critical interface between circuit design and silicon fabrication. Under the increasing stress of the manufacturability, such a PDK facilitates designers assess layout dependent effects and manage their impact.

Beyond the 10 nm node, more radical solutions will be vital to meet the scaling criteria. While there have been significant accomplishments in scientific discovery, it is only the beginning of the engineering research that is required to transfer the science into device, circuit, and system integration. In Chap. 9, PTM outreaches the effort to the compact modeling of carbon nanotube devices, helping illustrate their enormous design potentials. Finally, Chap. 10 concludes the book with a brief outlook on future nanoelectronic modeling and design.

Tempe, AZ Yu Cao

Acknowledgements

I would like to thank many people who contribute to the PTM project. First and foremost, my grateful acknowledgement is made toward Professor Chenming Hu and the BSIM group at the University of California, Berkeley, for their insightful advices and guidance at the early stage. I also express my gratitude to Professor Dennis Sylvester, University of Michigan, Ann Arbor; Professor Michael Orshansky, University of Texas, Austin; and Professor Takashi Sato, Kyoto University, for their patient help to initialize this project. I am indebted to the visionary suggestions provided by Shekhar Borkar, Ram Krishnamurthy, and Jeff Parkhurst at Intel. Many students in my group have excellent contributions to PTM, including Wei Zhao, Wenping Wang, Chi-Chao Wang, Saurabh Sinha, Yun Ye, Asha Balijepalli, Min Chen, and Varsha Balakrishnan. The PTM project is made possible by the support from the Focus Center Research Program and Semiconductor Research Corporation.

Thanks are due to many university and industry researches who help improve PTM during its evolution. These people include: Judy An, Integrated Memory Logic; Dimitri Antoniadis, Massachusetts Institute of Technology; Robert Dutton and H.-S. Philip Wong, Stanford University; Nagib Hakim and Changhong Dai, Intel; Ali Icel, Global Foundries; Xia Li, Seung H. Kang, Matt Nowak, Qualcomm; Colin McAndrew, Freescale; Sani Nassif and Frank Liu, IBM; Jan Rabaey and Tsu-Jae King Liu, University of California, Berkeley; Vijay Reddy, Srikanth Krishnan, Keith Green, Texas Instruments; Kaushik Roy and Mark Lundstrom, Purdue University; Kishore Sinha, Synopsys; Yuan Xie and Vijay Narayanan, Pennsylvania State University; Yu Wang, Tsinghua University. I apologize in advance to anyone I overlooked. Their technical suggestions help lead PTM toward this stage.

Yu Cao

Contents

Chapter 1
Introduction

The scaling of CMOS technology has been the driving force of the semiconductor industry during past five decades, with the minimum feature size expected to reach 10 nm in 10 years [1]. Beyond that benchmark, the present scaling approach may have to take a different route, in order to overcome dramatic barriers in transistor performance degradation, power consumption, process and environmental variations, and reliability issues. For instance, Fig. 1.1 illustrates the scaling trends of the maximum on-state current (I_{on}) and the off-state leakage current (I_{off}), from a comprehensive set of published data [2–30]. From the 0.5 μm node to the 32 nm node, the increase in I_{on} is smaller than 3×; meanwhile, I_{off} increases by more than six orders! Such a dramatic reduction in the ratio of I_{on}/I_{off} significantly affects the drivability of the device, and further influences all aspects of circuit performance, such as data stability of on-chip memory.

To continue the success of integrated circuit (IC) design, the grand challenge to IC community is to identify unconventional materials and structures, such as carbon-based electronics, integrate them into the large-scale circuit architecture, and enable continuous growth of chip scale and performance [1, 31]. Different from previous design paradigm, today's competitive circuit design and research must begin before a future generation of CMOS technology is fully developed, in order to successfully manage the development cost and guarantee the time to market. Figure 1.2 highlights the paradigm shift toward concurrent technology and design research [32].

In this context, Predictive Technology Model (PTM), which bridges the process/material development and circuit simulation through device modeling, is essential to assessing the potential and limits of new technology and to supporting early design prototyping. PTM is the critical interface between technology innovation and IC design exploration, as shown in Fig. 1.3. Coupled with circuit simulation tools, they significantly improve design productivity, providing the insight into the relationship between technology/design choices and circuit performance. In order to guarantee the quality of the prediction, PTM should be scalable with latest technology advances, accurate across a wide range of process uncertainties and operation conditions, and efficient for large-scale computation. As semiconductor

Y. Cao, *Predictive Technology Model for Robust Nanoelectronic Design*,
Integrated Circuits and Systems, DOI 10.1007/978-1-4614-0445-3_1,

Fig. 1.1 The scaling trends of I_{on} and I_{off} [2–30]

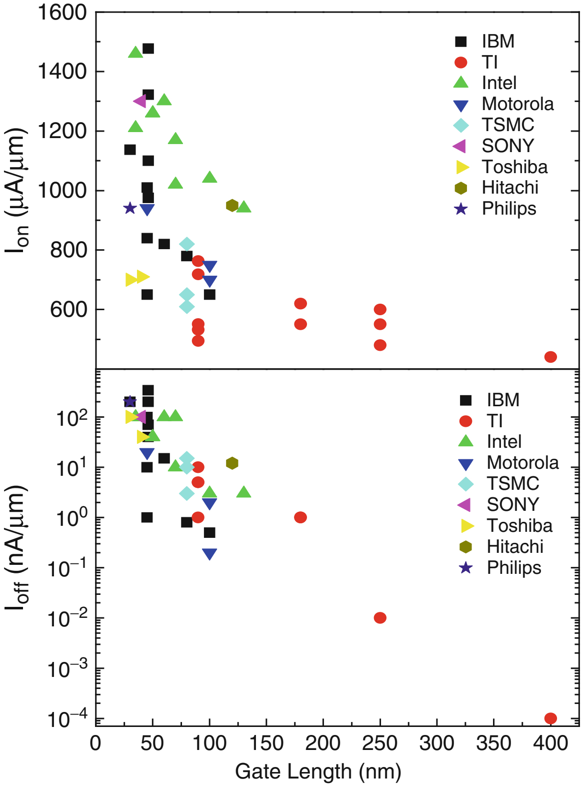
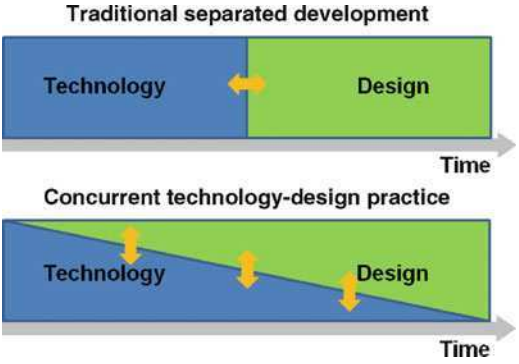


Fig. 1.2 The new paradigm of joint technology-design research


Traditional separated development
Technology
Design
Time
Concurrent technology-design practice
Technology
Design
Time

technology scales into the nanoscale regime, these modeling demands are tremendously challenged, especially by the introduction of alternative device materials and structures, as well as the ever-increasing amount of process variations.

This paper presents a comprehensive review on the development and latest results of Predictive Technology Model for nanoscale devices, covering end-of-the-roadmap and post-silicon technologies. Driven by the increasingly complex and diverse nature of the underlying technology, the overarching goal of PTM is to

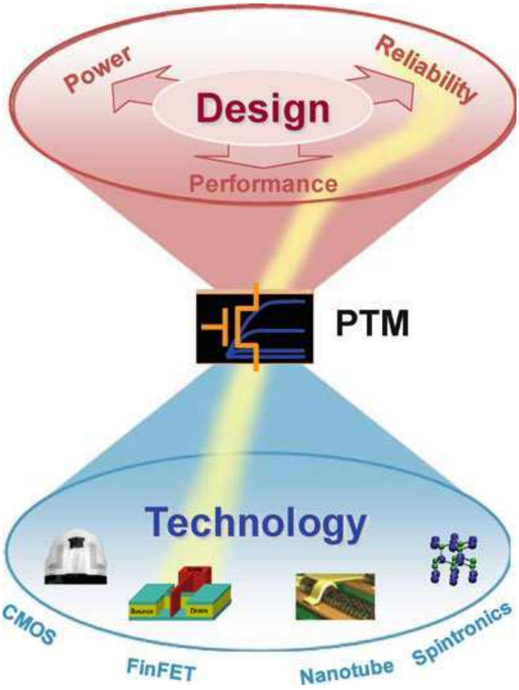


Fig. 1.3 PTM: a bridge between technological prediction and early stage design exploration

provide early comprehension of process choices and design opportunities, as well as to address key design needs, such as variability and reliability, for robust system integration. Specific topics include:

- *Predictive modeling of end-of-the-roadmap CMOS technology*: CMOS will arguably be the technology of choice for the next 15 years. To predict future technology characteristics, an intuitive approach would simply scale down the feature size and voltage parameters, such as supply voltage and threshold voltage (V_{th}), from an existing technology. However, this approach is overly simplified and underestimates the overall device performance toward the end of the roadmap [33]. During technology scaling, process developers will optimize many other aspects of the device beyond sole geometry scaling. For instance, the scaling of V_{th} not only requires the change of channel doping concentration, but also impacts other physical parameters, such as mobility, saturation velocity, and the body effect. These intrinsic correlations among physical parameters need to be carefully considered for an accurate prediction.
- *PTM for alternative materials and structures*: The scaling of traditional bulk CMOS structure is slowing down in recent years as fundamental limits are rapidly approached. For instance, short-channel effects, such as drain-induced-barrier-lowering (DIBL) and threshold voltage rolloff, severely increase leakage current and degrade the ratio of I_{on}/I_{off}. To overcome these difficulties and continue the path perceived by Moore’s law, new materials (e.g., strained silicon, metal gate, high-k dielectrics, low-resistance source/drain) and structures (e.g., double-gate device)

need to be adopted into conventional CMOS technology. Therefore predictive models for bulk CMOS technology should be updated to capture the distinct electrical behavior of these advances, guaranteeing start-of-the-art predictions and design benchmarking toward the 10 nm regime.

- *Modeling of CMOS variability and reliability*: While technology scaling can be extended with alternative materials and structures, CMOS technology will eventually reach the ultimate limits that are defined by both physics and the fabrication process. One of the most profound physical effects will result from the vastly increased parameter variations and reliability degradation due to manufacturing and environmental factors. These parameter fluctuations exacerbate design margins, degrade the yield, and invalidate current deterministic design methodologies. To maintain design predictability with those extremely scaled devices, predictive models should incorporate both static process variations and temporal shift of device parameters. They should be extended from the traditional corner-based approach to a suite of modeling efforts, including extraction methods, the decoupling of variation sources, and highly efficient strategies for the statistical design paradigm.
- *Process design kits (PDK) and design benchmark*: As technological and design issues become more complicated with scaled CMOS devices, design productivity continues to be a major challenge for the semiconductor industry. Improved design flow automation and reuse methodologies are well known approaches to deal with this problem. But the lack of standards for archiving design data has prevented these techniques from having a significant impact. Recent trends towards open frameworks and open PDK promise to provide the very standards needed to enable greater levels of automation and reuse. Based on PTM, the development of predictive PDK and open library makes widespread adoption of these standards possible, and allows designers to perform more realistic assessment of the trends and challenges in future IC design.
- *Predictive modeling of post-silicon devices*: Beyond the far end of the CMOS technology roadmap, several emerging technologies have been actively researched as alternatives, such as nano-tubes, nano-wires, and molecular devices. As demonstrated in the success of PTM for CMOS, the outreach of PTM to these revolutionary technologies will help shed light on design opportunities and challenges with post-silicon technologies beyond the 10 nm regime.

In nanoelectronic design, predictive device modeling plays an essential role in joint technology-design exploration. Solutions to those modeling challenges will ensure a timely and smooth transition from CMOS-based design to robust integration with post-silicon technologies.

References

1. International Technology Roadmap of Semiconductors, 2007. (available at http://www.itrs.net).
2. M. Khare, et al., “A high performance 90 nm SOI technology with 0.992μm^2 6T-SRAM cell,” *IEDM Tech. Dig*., pp. 407–410, 2002.

3. R. A. Chapman, et al., “High performance sub-half micron CMOS using rapid thermal processing,” *IEDM Tech. Dig.*, pp. 101–104, 1991.
4. Y. Taur, et al., “High performance 0.1 μm CMOS devices with 1.5 V power supply,” *IEDM Tech. Dig.*, pp. 127–130, 1993.
5. M. Rodder, Q. Z. Hong, M. Nandakumar, S. Aur, J. C. Hu, and I. C. Chen, “A sub-0.18 μm gate length CMOS technology for high performance (1.5 V) and low power (1.0 V),” *IEDM Tech. Dig.*, pp. 563–566, 1996.
6. L. Su, et al., “A high-performance sub-0.25 μm CMOS technology with multiple thresholds and copper interconnects,” *VLSI Symp. Tech. Dig.*, pp. 18–19, 1998.
7. M. Hargrove, et al., “High-performance sub-0.08 μm CMOS with dual gate oxide and 9.7 ps inverter delay,” *IEDM Tech. Dig.*, pp. 627–630, 1998.
8. S. Yang, et al., “A high performance 180 nm generation logic technology,” *IEDM Tech. Dig.*, pp. 197–200, 1998.
9. P. Gilbert, et al., “A high performance 1.5 V, 0.10 μm gate length CMOS technology with scaled copper metalization,” *IEDM Tech. Dig.*, pp. 1013–1016, 1998.
10. T. Ghani, et al., “100 nm gate length high performance/low power CMOS transistor structure,” *IEDM Tech. Dig.*, pp. 415–418, 1999.
11. K. K. Young, et al., “A 0.13 μm CMOS technology with 193 nm lithography and Cu/low-*k* for high performance applications,” *IEDM Tech. Dig.*, pp. 563–566, 2000.
12. S. Tyagi, et al., “A 130 nm generation logic technology featuring 70 nm transistors, dual V_t transistors and 6 layers of Cu interconnects,” *IEDM Tech. Dig.*, pp. 567–570, 2000.
13. K. Ichinose, et al., “A high performance 0.12 μm CMOS with manufacturable 0.18 μm technology,” *VLSI Symp. Tech. Dig.*, pp. 103–104, 2001.
14. S. Thompson, et al., “An enhanced 130 nm generation logic technology featuring 60 nm transistors optimized for high performance and low power at 0.7–1.4 V,” *IEDM Tech. Dig.*, pp. 257–260, 2001.
15. M. Celik, et al., “A 45 nm gate length high performance SOI transistor for 100 nm CMOS technology applications,” *VLSI Symp. Tech. Dig.*, pp. 166–167, 2002.
16. V. Chan, et al., “High speed 45 nm gate length CMOSFETs integrated into a 90 nm bulk technology incorporating strain engineering,” *IEDM Tech. Dig.*, pp. 77–80, 2003.
17. K. Mistry, et al., “Delaying forever: Uniaxial strained silicon transistors in a 90 nm CMOS technology,” *VLSI Symp. Tech. Dig.*, pp. 50–51, 2004.
18. S. Mayuzumi, et al., “Extreme high-performance n- and p-MOSFETs boosted by dual-metal/high-*k* gate damascene process using top-cut dual stress liners on (100) substrates,” *IEDM Tech. Dig.*, pp. 293–296, 2007.
19. A. Pouydebasque, et al., “High density and high speed SRAM bit-cells and ring oscillators due to laser annealing for 45 nm bulk CMOS,” *IEDM Tech. Dig.*, pp. 663–666, 2005.
20. W.-H. Lee, et al., “High performance 65 nm SOI technology with enhanced transistor strain and advanced-low-*k* BEOL,” *IEDM Tech. Dig.*, pp. 56–59, 2005.
21. S. Tyagi, et al., “An advanced low power, high performance, strained channel 65 nm technology,” *IEDM Tech. Dig.*, pp. 1070–1072, 2005.
22. M. Rodder, et al., “Oxide thickness dependence of inverter delay and device reliability for 0.25 μm CMOS technology,” *IEDM Tech. Dig.*, pp. 879–882, 1993.
23. M. Rodder, A. Amerasekera, S. Aur, and I. C. Chen, “A study of design/process dependence of 0.25 μm gate length CMOS for improved performance and reliability,” *IEDM Tech. Dig.*, pp. 71–74, 1994.
24. M. Rodder, S. Aur, and I.-C. Chen, “A scaled 1.8 V, 0.18 μm gate length CMOS technology: Device design and reliability considerations,” *IEDM Tech. Dig.*, pp. 415–418, 1995.
25. M. Rodder, et al., “A 1.2 V, 0.1 μm gate length CMOS technology: Design and process issues,” *IEDM Tech. Dig.*, pp. 623–626, 1998.
26. M. Mehrotra, et al., “A 1.2 V, sub-0.09 μm gate length CMOS technology,” *IEDM Tech. Dig.*, pp. 419–422, 1999.

27. A. H. Perera, et al., “A versatile 0.13 μm CMOS platform technology supporting high performance and low power applications,” IEDM Tech. Dig., pp. 571–574, 2000.
28. N. Yanagiya, et al., “65 nm CMOS technology (CMOS5) with high density embedded memories for broadband microprocessor applications,” *IEDM Tech. Dig.*, pp. 57–60, 2002.
29. S. Thompson, et al., “A 90 nm logic technology featuring 50 nm strained silicon channel transistors, 7 layers of Cu interconnects, low-*k* ILD, and 1μm^2 SRAM cell,” *IEDM Tech. Dig.*, pp. 61–64, 2002.
30. P. Bai, et al., “A 65 nm logic technology featuring 35 nm gate lengths, enhanced channel strain, 8 Cu interconnect layers, low-*k* ILD and 0.57 pmZ SRAM Cell,” *IEDM Tech. Dig.*, pp. 657–660, 2004.
31. B. H. Calhoun, Y. Cao, X. Li, K. Mai, L. T. Pileggi, R. A. Rutenbar, and K. L. Shepard, “Digital circuit design challenges and opportunities in the era of nanoscale CMOS,” *Proceedings of the IEEE*, vol. 96, no. 2, pp. 343–365, February 2008.
32. S. Jha, “Challenges on design complexities for advanced wireless silicon systems,” *Design Automation Conference*, 2008.
33. W. Zhao, Y. Cao, “New generation of predictive technology model for sub-45 nm early design exploration,” *IEEE Transactions on Electron Devices*, vol. 53, no. 11, pp. 2816–2823, November 2006.

Chapter 2
Predictive Technology Model of Conventional CMOS Devices

Bulk CMOS has been the dominant device structure for integrated circuit design during the past decades, because of its excellent scalability. It is expected that such a device type will continue toward the 10 nm regime. To efficiently predict the characteristics of future bulk CMOS, the scaling trends of primary model parameters, such as the threshold voltage and gate dielectric thickness, need to be identified; their association in determining major device characteristics should be well included for accurate model projection. In this chapter, a new generation of Predictive Technology Model (PTM) for conventional CMOS technology is presented to accomplish these goals. Based on a set of essential device models and early stage silicon data, PTM of bulk CMOS is successfully generated down to the 12 nm node. The accuracy of PTM predictions is comprehensively verified with published silicon data: the error of I_{on} is below 10% for both NMOS and PMOS devices. By tuning only ten primary model parameters, PTM can be easily customized to cover a wide range of process uncertainties. Furthermore, PTM correctly captures the sensitivity to process variations.

2.1 PTM in Light of CMOS Scaling

The relentless scaling of CMOS technology has accelerated in recent years and will arguably continue toward the 10 nm regime [1]. In the nanometer era, physical factors that previously had little or no impact on circuit performance are now becoming increasingly significant. Particular examples include process variations, transistor mobility degradation, and power consumption. These new effects pose dramatic challenges to robust circuit design and system integration. To continue the design success and make an impact on leading products, advanced circuit design exploration must start in parallel with, or even earlier than silicon development. This new design paradigm demands predictive MOSFET models that are reasonably accurate, scalable with main process and design knobs, and correctly capture those emerging physical effects.

Y. Cao, *Predictive Technology Model for Robust Nanoelectronic Design*,
Integrated Circuits and Systems, DOI 10.1007/978-1-4614-0445-3_2,

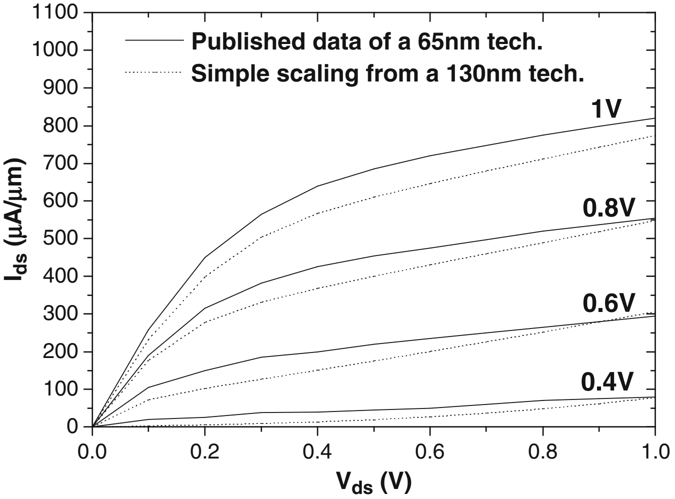


Fig. 2.1 A simple method fails the I-V prediction (Adapted from [8])

To predict future technology characteristics, an intuitive approach would simply scale down the geometry and voltages from an existing technology. For instance, based on the standard MOSFET model, BSIM4 [2], we can shrink the parameters of effective gate length (L_{eff}), equivalent electrical oxide thickness (T_{oxe}), threshold voltage (V_{th0}), drain and source paratactic resistance (R_{dsw}), and supply voltage (V_{dd}) to the target values, while keeping all the other parameters unchanged. However, as shown in Fig. 2.1, this approach is too simple to capture the basic MOSFET behavior. In Fig. 2.1, the I-V characteristics of a preliminary 65 nm technology are predicted based on a well-characterized 130 nm technology by scaling L_{eff}, T_{ox}, V_{th0}, R_{dsw} and V_{dd}. Compared to published measurement data, this simple prediction underestimates the overall performance. This observation matches the fact that during technology scaling, process developers will optimize many other aspects of the device beyond simple geometry scaling, in order to meet all performance criteria.

An improved predictive method was presented by Berkeley Predictive Technology Model (BPTM) [3]. Based on BSIM3 model, BPTM includes more physical parameters into the prediction. Their values are empirically extracted from published data during early stage technology development. Although BPTM provides reasonable models for technology nodes from 180 to 45 nm, its empirical nature constrains the physicality and scalability of the predictions. As the model file for each technology node is independently fitted, the overall scaling trend is not smooth from BPTM, as shown in Fig. 2.2. Furthermore, intrinsic correlations among physical parameters are not sufficiently considered. For instance, the scaling of V_{th0} not only requires the change of channel doping (N_{ch}), but further affects other physical parameters, such as mobility (μ_0), saturation velocity (V_{sat}), the body effect, etc. Insufficient modeling of these correlations limits the prediction accuracy

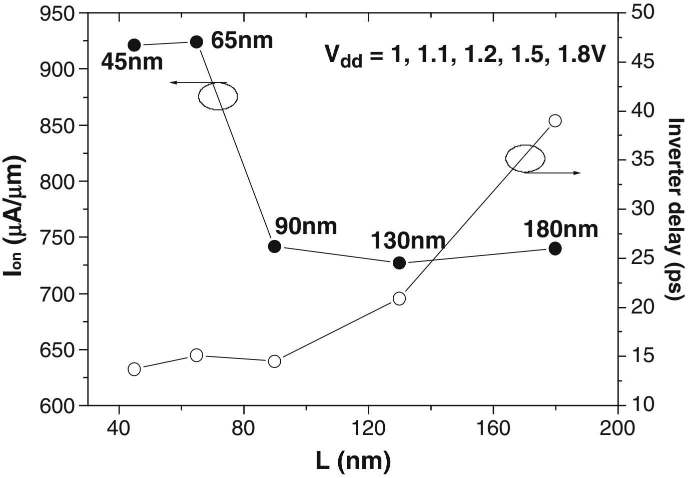


Fig. 2.2 The prediction in BPTM is not smooth with scaling (Adapted from [8])

of process sensitivities. As process variations become increasingly significant in scaled CMOS technology, it is critical to include these parameter correlations into future predictive models, such that robust circuit design can be correctly guided [4].

In this context, a new generation of PTM is developed to overcome these shortcomings. Two cornerstones ensure the accurate and smooth prediction:

1. Essential device physics that governs key device characteristics and parameter correlations. PTM identifies a set of simplified equations for critical electrostatic behavior and carrier transport, rather than the full set of BSIM models. Such simplification allows more transparent correlation between model parameters and device performance; it further facilitates physical prediction of the scaling trends. Given the expectations of device geometry and voltage conditions, these models help project the underlying physical parameters to be tuned.
2. Silicon data from previous technology generations and early stage technology development. A comprehensive collection of published data from various sources provides a practical ground to predict the evolution of CMOS technology. It reflects the limits of CMOS manufacturability and fabrication cost during technology scaling, especially in the definition of device geometries. By recognizing these engineering limits, prediction of PTM is realistic and reasonable.

Based on these principles, first, new physical models are integrated into the predictive methodology to correctly capture the correlations among model parameters. These models include V_{th0} dependence on N_{ch}, mobility degradation, and velocity overshoot. Second, based on comprehensive studies of published data over various technology generations, i.e., from 250 nm node to 45 nm node, the scaling trends of key physical parameters are extracted. By integrating these results into PTM, both nominal and variational transistor characteristics are predicted, following the traditional trend of scaling. Smooth and accurate predictions are obtained from 250 to 12 nm nodes, with L_{eff} down below 10 nm. Compared to various published data, the error in the prediction of I-V characteristics is less than 10%. PTM can be conveniently customized by adjusting only ten primary parameters, in order to cover a wide range of process uncertainties. Using PTM, the impact of process variations is further investigated for nanoscale CMOS design. Overall, this chapter develops a solid predictive base for exploratory circuit design with extremely scaled bulk CMOS. The following chapter (Chap. 3) will further describe how PTM incorporates physical models for new technology advances, such as strained silicon, high-k dielectrics and metal gate, in order to make a far-reaching impact on future design.

2.2 Predictive Methodology

2.2.1 Parameter Taxonomy

Based on our previous work on BPTM, it is recognized that the appropriate categorization of transistor model parameters is crucial for an efficient and physical prediction [3, 5, 6, 7]. Although there are typically more than 100 parameters in a compact transistor model to calculate the I-V and C-V characteristics, only about ten of them are critical to determine the essential behavior of a nanoscale transistor. The performance of a transistor is less sensitive to the rest of secondary parameters. Based on their physical meanings, these first order parameters are listed in Table 2.1 [5–7], including technology specifications as well as process and physical parameters. Such taxonomy keeps the physics of scaling while reducing the complexity of prediction. Furthermore, this categorization is relatively independent on model formats as those key parameters are mostly shared among different transistor models to represent the underlying silicon technology. Accurate modeling and prediction of their values is the key to the development of PTM. In this work, BSIM4 is used as the model basis while the predictive methodology is general enough to be applied to other model formats [8].

In addition to predicting nominal values, it becomes increasingly important to capture process sensitivities as well. As process variations are vastly exacerbated at future technology nodes, current deterministic design paradigm needs to be shifted towards a statistical design flow in order to reduce design uncertainties [1, 4]. Thus, physical correlations among main model parameters, such as the transport behavior [9–11], should be explicitly expressed in compact models for both accurate

Table 2.1 Primary parameters in the development of new PTM

Technology specifications	V_{dd}, V_{th0}, T_{oxe}, L_{eff}, R_{dsw}
Process parameters	N_{ch}, E_{ta0}
Physical parameters	K_1, μ_0, V_{sat}

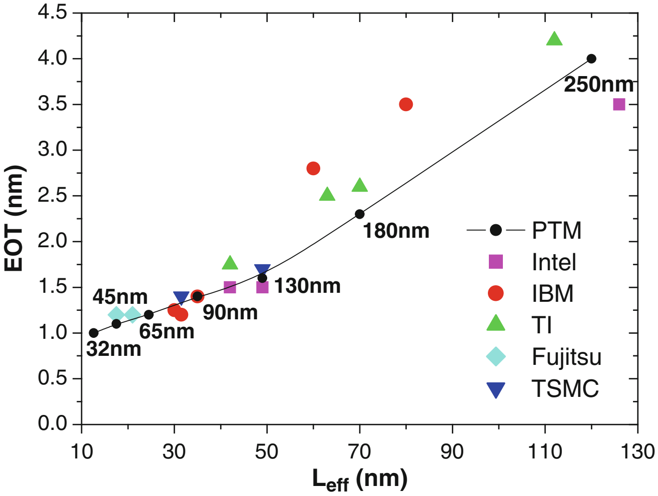


Fig. 2.3 The trend of EOT scaling (Adapted from [8])

technology extrapolation and robust design exploration. While such a consideration is absent in BPTM [3], the new generation of PTM identifies those critical correlations, particularly the interactions among $L_{eff,}$ V_{th}, mobility, and saturation velocity.

2.2.2 Prediction of Model Parameters

As presented in Table 2.1, the first group of parameters is related to the process specifications in technology scaling, including V_{dd}, T_{oxe}, L_{eff}, V_{th0} and R_{dsw}. Their nominal values are determined by literature survey from published industry data, including the ITRS [1]. Based on the collected data, Fig. 2.3 presents the trend of equivalent oxide thickness (EOT). EOT is steadily scaling down, although the pace may slow down in recent years. The trend of V_{dd} and V_{th} scaling is plotted in Fig. 2.4, where the value of V_{th} is extracted from the sub-threshold I-V curves, using the constant current definition. Due to the concern of sub-threshold leakage, V_{th} stays almost the same in the nanoscale. The fifth technology parameter, R_{dsw}, is extracted by fitting the I-V curves in the linear region, after the low-field mobility, μ_0, is predicted (i.e., Eqs. 2.1 and 2.2). The trend of R_{dsw} is shown in Fig. 2.5. The reduction of R_{dsw} becomes more difficult in short-channel devices and results in a constant scaling as the data shows. These trends, which are supported by experimental data, are then integrated into PTM to predict the nominal values during CMOS technology scaling.

Values of technology specifications not only define the basic characteristics of a process; they further determine other important electrical details of a transistor.

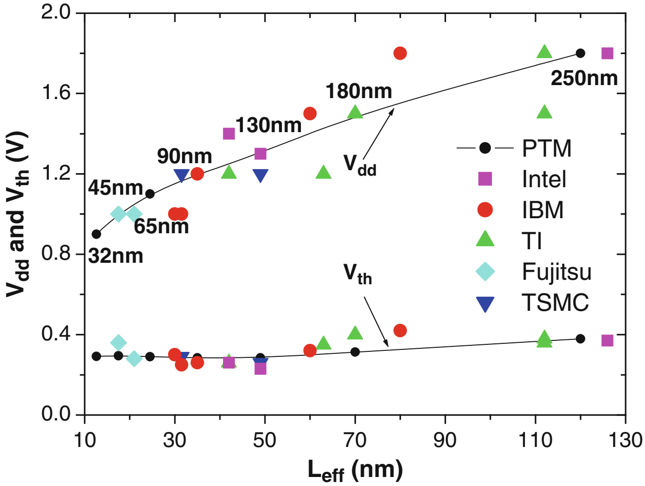


Fig. 2.4 The trends of V_{dd} and V_{th} scaling (Adapted from [8])

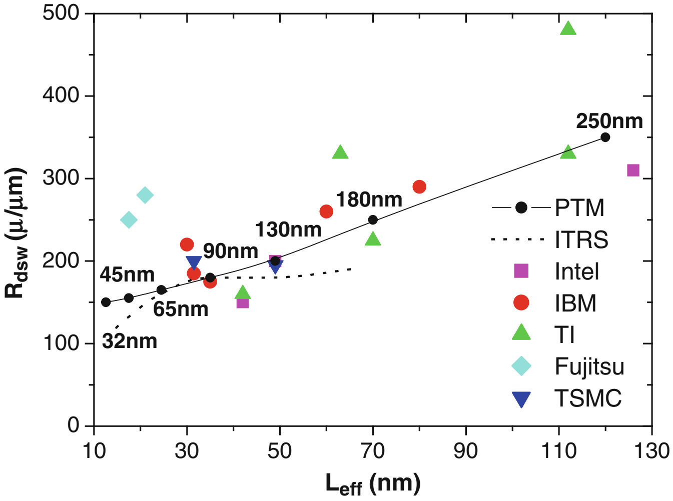


Fig. 2.5 The trend of R_{dsw} scaling (Adapted from [8])

In particular, channel doping concentration, N_{ch}, is mainly defined by the threshold voltage. Exact value of N_{ch} is extracted from published data of V_{th0} in [12–27], using the V_{th} model in BSIM [2]. Figure 2.6 illustrates the trend of N_{ch} scaling. Based on N_{ch}, the main coefficient for the body effect of V_{th}, K_1, is also estimated with analytical models [2]. Furthermore, to model the V_{th} behavior of short-channel transistors, drain-induced-barrier-lowering (DIBL) must be accounted for.

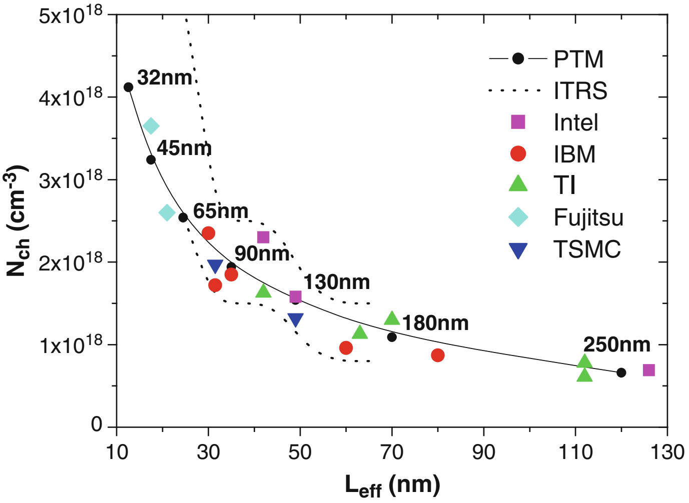


Fig. 2.6 The trend of N_{ch} scaling (Adapted from [8])

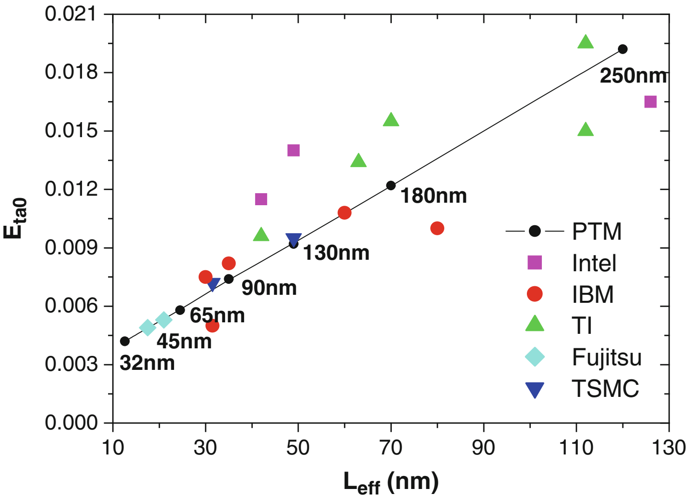


Fig. 2.7 The trend of DIBL coefficient E_{ta0} (Adapted from [8])

To the first-order, this effect is captured by E_{ta0}, which is a model parameter for the DIBL effect. Its value is extracted from published data of V_{th} roll-off [12–27]. A clear trend of E_{ta0} is illustrated in Fig. 2.7.

The amount of channel doping, N_{ch}, is actually important for both threshold voltage and the transport property in a conductive channel, i.e., effective carrier mobility (μ_{eff}) and the saturation velocity (V_{sat}). For example, low field carrier mobility degrades as N_{ch} increases, so does also the effective carrier mobility;

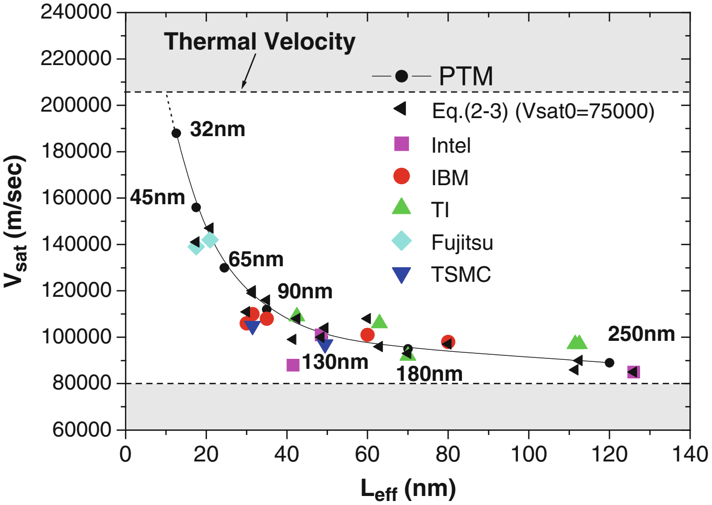


Fig. 2.8 The trend of V_{sat} scaling from traditional velocity saturation to the overshoot region

V_{sat} also depends on N_{ch} and L_{eff} due to the phenomenon of velocity overshoot [9]. To account for these effects, the following formulas are adopted in the new PTM to estimate V_{sat} and μ_0 respectively [8, 9]:

$$\text{NMOS}: \quad \mu_0 = 1150 \cdot \exp\left(-5.34 \cdot 10^{-10}\sqrt{N_{ch}}\right) \tag{2.1}$$

$$\text{PMOS}: \quad \mu_0 = 317 \cdot \exp\left(-1.25 \cdot 10^{-9}\sqrt{N_{ch}}\right) \tag{2.2}$$

$$V_{sat} = V_{sat0} + 0.13\mu_{eff}\sqrt{\tau\mu_{eff}kT/q} \cdot \left(V_d \middle/ L_{eff}^2\right) \tag{2.3}$$

Equations 2.1 and 2.2 are based on the physical model of mobility [9–11]; the coefficient values are extracted from advanced silicon data. Equation 2.3 of velocity overshoot is a simplified solution of the energy-balance equation in [9]. These equations describe the important dependence on N_{ch} and are compatible with the current BSIM framework. The value of V_{sat} is extracted from published I-V data, particularly the saturation current I_{on}; its trend during scaling is plotted in Fig. 2.8. The effect of velocity overshoot is pronounced as technology scales down to sub-100 nm regime. Figure 2.8 also demonstrates excellent model prediction by Eq. 2.3 with the extracted V_{sat}.

Combining these steps together, the ten primary parameters, e.g., V_{dd}, T_{oxe}, L_{eff}, V_{th0}, R_{dsw}, N_{ch}, E_{ta0}, K_1, μ_0 and V_{sat} can be extrapolated towards future technology nodes. Furthermore, their values can be adjusted to cover a range of process uncertainties, e.g., from one company's to another one's, or from intrinsic process variations. In general, the error by only considering these primary parameters can be reduced to 5%, as demonstrated in [8]. This is further verified by comparing the model predictions with published data, as shown in Sect. 2.3.

The rest of model parameters are secondary ones, without explicit methods to predict their values. To improve the accuracy of predictions, they are further classified into two groups, depending on their importance in the determination of transistor performance. The first group is not as critical as the primary parameters, but still has an observable impact on I-V characteristics. They are related to the determination of short channel effects (e.g., D_{vt0} and D_{vt1} are short channel effects coefficients and their values are extracted from published data of V_{th} roll-off [12–27]), subthreshold behavior (D_{sub}, N_{factor}, V_{off}, C_{dsc}, C_{dscd}), mobility (μ_a, μ_b), and Early voltage. During the scaling of CMOS technology, their values may change from one generation to the next, but are relatively stable within one generation. In this context, their values are fit from experimental data for each technology node and then fixed over a range of process conditions. The remaining secondary parameters have little impact on transistor performance. Thus, for the purpose of early prediction, it is reasonable to leave these parameters unchanged. Finally, the parameters for parasitic C-V characteristics are extrapolated based on BSIM models.

The predictive methodology was first implemented using Verilog-A, since the physical models (i.e., Eqs. 2.1–2.3) are currently not available in the standard model format. After generating the PTM for each technology node, the Verilog-A models can be mapped to standard BSIM4 models for nominal performance prediction, so that designers can directly use them with available circuit simulators. In addition, the Verilog-A format is also compatible with SPICE simulation tools, such that circuit designers can use them directly. Presently, PTM model files for 130 to 12 nm technology generations are available. For easy access, a webpage was established to release the latest models (http://ptm.asu.edu) [8].

2.3 Evaluation of PTM

2.3.1 Verification and Prediction of I-V Characteristics

About twenty sets of published I-V data from the 250 nm node to the 45 nm node at room temperature are collected to verify the prediction by PTM. Using the methodology presented above, we are able to generate corresponding PTM model files. By tuning ten primary parameters, the predicted I-V characteristics are then compared to published data for verification. The parameter tuning steps are explained below. First, V_{dd}, T_{oxe}, L_{eff} and V_{th0} are directly adjusted to the published values. Then N_{ch} is reversely calculated from V_{th0}, using analytical models [2]. Based on N_{ch}, μ_0 and V_{sat} can be calculated with Eqs. 2.1–2.3. Finally, R_{dsw} is extracted from the linear region of I-V curves. Figures 2.9 and 2.10 illustrate two examples at

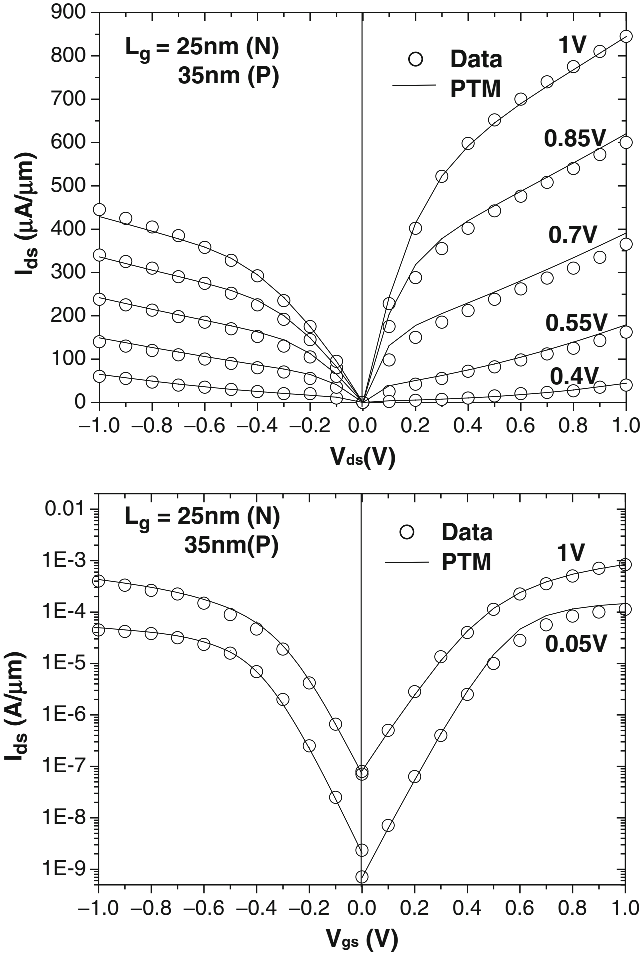


Fig. 2.9 The verification of 45 nm PTM with [13] (Adapted from [8])

45 and 65 nm nodes, respectively. Predicted I-V curves are compared to the measured silicon data from [13] and [14]. Excellent agreement between prediction and published data is achieved in both sub- and super- threshold regions. More comprehensive verifications are listed in Table 2.2 [12–27]. Without any further model optimization, the error of I_{on} predictions is smaller than 10%, for both NMOS and PMOS transistors. Such an excellent matching proves the physicality and scalability of PTM.

Based on the successful verifications, PTM for 130 to 12 nm technology nodes have been generated and released at http://ptm.asu.edu. Figure 2.11 illustrates the trend of nominal I_{on} and I_{off}. Figure 2.12 illustrates the trend of nominal CV/I and switch power (CV_{dd}^2). Table 2.3 further highlights the major characteristics of

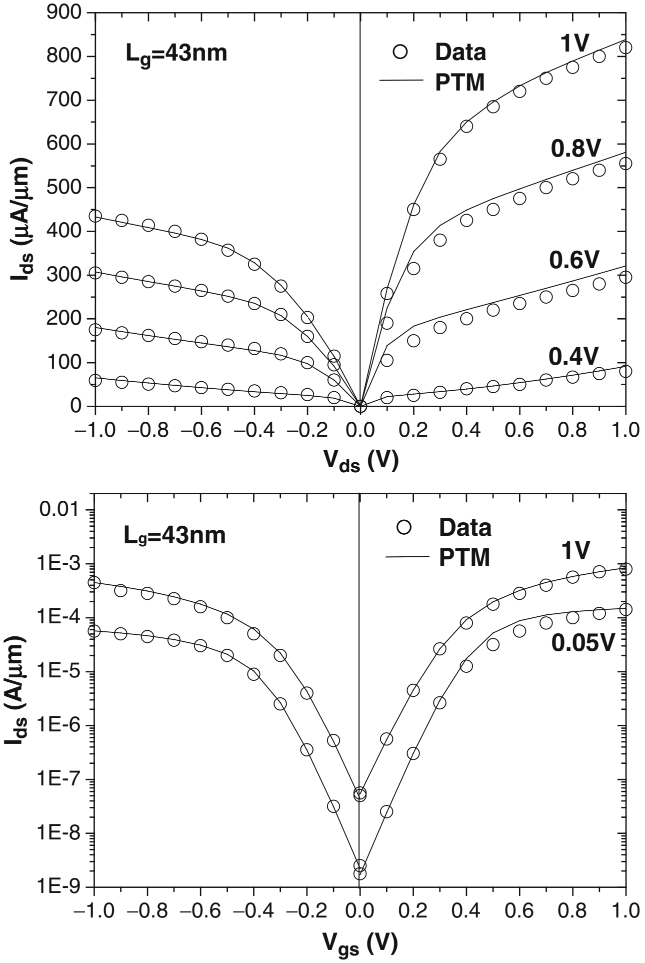


Fig. 2.10 The verification of 65 nm PTM with [14] (Adapted from [8])

PTM predictions for technology scaling. Note that the threshold voltage remains almost unchanged due to the leakage concern (Fig. 2.4). With continuous efforts, PTM will be extended toward the 12 nm technology node and below.

2.3.2 Impact of Process Variations

According to the ITRS, similar or larger amount of process variations are expected at future technology nodes. What matters is not only the amount of variations, but also the sensitivity to variations. In the nanometer regime, the sensitivity of transistor performance on process variations becomes more significant and is

Table 2.2 Evaluation of PTM predictions with published data (adapted from [8])

Data source	V_{dd} (V)	T_{oxe} (nm)	L_{eff} (nm)	V_{th} (V)	R_{dsw} (Ω/μm)	I_{on} (μA/μm)	I_{on} (Pred.)	I_{off} (nA/μm)	I_{off} (Pred.)	Error of I_{on} (%)
[12]	1	1.85	21	0.28	280	940	950	150	120	1
[13]	1	1.85	17	0.36	250	845	855	80	20	1
[14]	1	1.9	30	0.30	220	820	845	50	40	3
[15]	1.2	2.05	32	0.29	200	1090	1187	80	50	9
[16]	1	1.85	32	0.25	185	1005	1045	160	130	4
[17]	1.2	2.05	35	0.26	175	1160	1210	130	100	4
[18]	1.2	2.4	42	0.26	160	1000	995	70	30	−1
[19]	1.4	2.15	42	0.26	150	1120	1205	10	10	8
[20]	1.3	2.15	49	0.23	200	1155	1145	130	140	−1
[21]	1.2	2.35	49	0.26	195	930	970	100	60	4
[22]	1.5	3.6	60	0.32	260	820	855	230	130	4
[23]	1.8	4.3	80	0.42	290	780	775	0.6	0.6	−1
[24]	1.2	3.3	63	0.35	330	586	555	5	4	−5
[25]	1.5	3.4	70	0.40	225	750	755	1	1	1
[26]	1.5	4	112	0.36	330	615	570	1	1	−7
[27]	1.8	4.3	126	0.37	310	690	690	1	1	0
[26]	1.8	5.0	112	0.38	480	605	580	0.6	1	−4

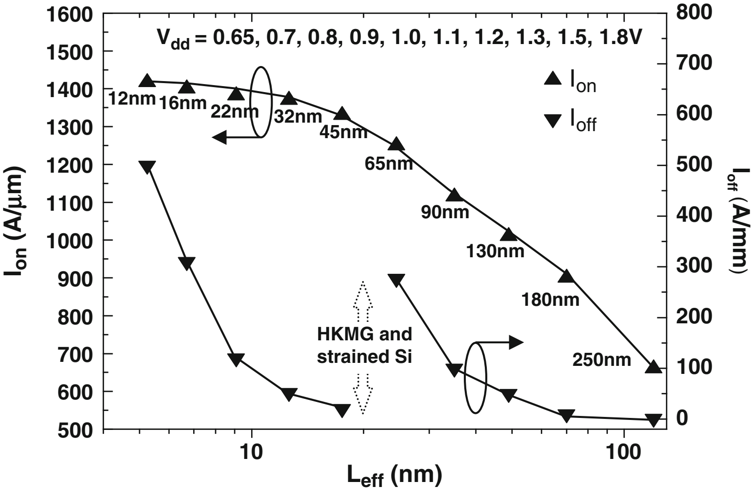


Fig. 2.11 Predictions of the scaling of nominal I_{on} and I_{off}. The jump in I_{off} is due to the adoption of high-k/metal gate and stained Si technology, as described in Chap. 3

critical for robust CMOS design. One particular phenomenon is velocity overshoot (Eq. 2.3). Figure 2.8 illustrates the trend of V_{sat} for successive technology nodes. When L_{eff} is larger than 100 nm, V_{sat} can be treated as a constant value, e.g., about 80,000 m/s. However, as L_{eff} scales below 100 nm, V_{sat} can no longer be

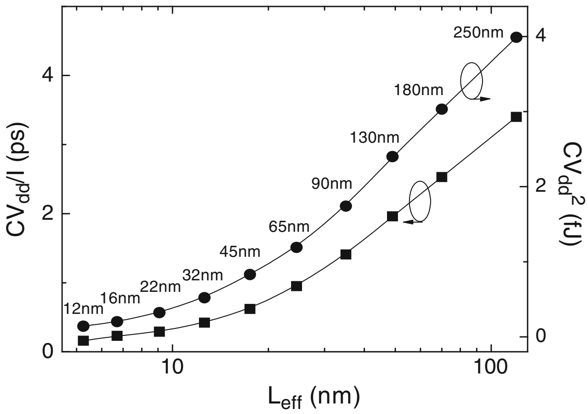


Fig. 2.12 The prediction of nominal CV/I and CV_{dd}^2

Table 2.3 The summary of PTM predictions for NMOS devices.

Tech.node (nm)	V_{dd} (V)	T_{oxe} (nm)	L_{eff} (nm)	V_{th} (V)	R_{dsw} (Ω/μm)	I_{on} (μA/μm)	I_{off} (nA/μm)	CV/I (ps)
12	0.65	0.6	5.25	0.265	135	1417	500	0.16
16	0.7	0.7	6.7	0.285	140	1400	310	0.23
22	0.8	0.8	9.1	0.31	145	1382	120	0.29
32	0.9	0.9	12.6	0.292	150	1370	52	0.42
45	1.0	1.0	17.5	0.295	155	1330	20	0.62
65	1.1	1.2	24.5	0.290	165	1250	277	0.95
90	1.2	1.4	35	0.284	180	1105	100	1.31
130	1.3	1.6	49	0.284	200	1000	50	1.96
180	1.5	2.3	70	0.309	280	890	10	2.53
250	1.8	4.0	120	0 379	350	610	1	3.34

approximated as a constant. Even though mobility (μ_{eff}) decreases with technology scaling due to higher N_{ch}, V_{sat} increases because of the inversely quadratic dependence on L_{eff} (Eq. 2.1) due to velocity overshoot. As a consequence, I_{on}, which is somewhat proportional to V_{sat}, is more sensitive to variations of L_{eff}, mobility, and V_{dd} in the nanoscale (Eq. 2.3). When the channel length is further reduced, the importance of velocity overshoot may degrade due to the ballistic transportation and the source-injection limit [2].

The importance of velocity overshoot in the study of process variations is further illustrated in Fig. 2.13. Figure 2.13 decomposes the variation of I_{on} into various physical mechanisms at the 45 nm node, for the variation of L_{eff}. Without considering DIBL and velocity overshoot, I_{on} is relative insensitive to L_{eff} variations as a result of pronounced velocity saturation in a nanoscale transistor. However, V_{th} of a nanoscale transistor changes when there exists the variation of L_{eff}, i.e., DIBL. For example, −20% L_{eff} variation will result in approximate 18% higher I_{on} due to DIBL. An additional amount of 27% I_{on} variation can be observed if velocity

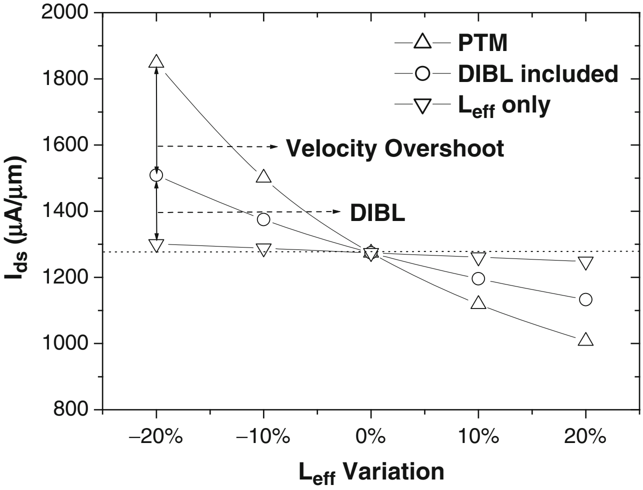


Fig. 2.13 The impact of L_{eff} variation at 45 nm (Adapted from [8])

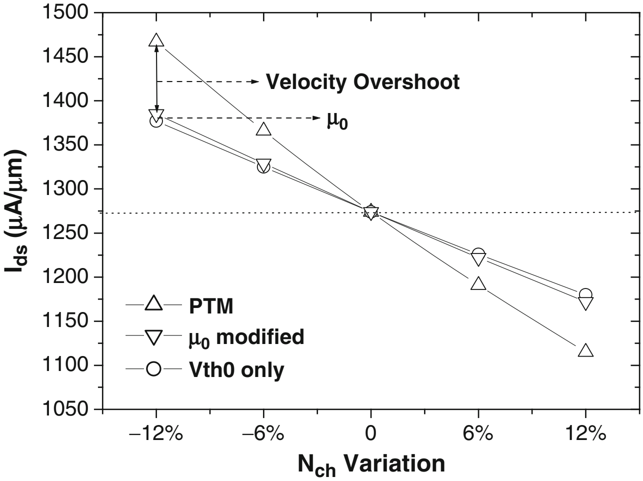


Fig. 2.14 The impact of N_{ch} variation at 45 nm (Adapted from [8])

overshoot is included (Fig. 2.13). Therefore, it is critical to include these physical models in prediction, in order to provide correct guidance to robust design explorations.

Besides L_{eff} variation, the random fluctuation of channel doping concentration is another leading source of process variations. When N_{ch} deviates from the target value, not only V_{th0}, but also K_1 (the body effect), μ_0 (mobility) and V_{sat} will change accordingly. Figure 2.14 shows the impact of N_{ch} variation on I_{on}. Similar to

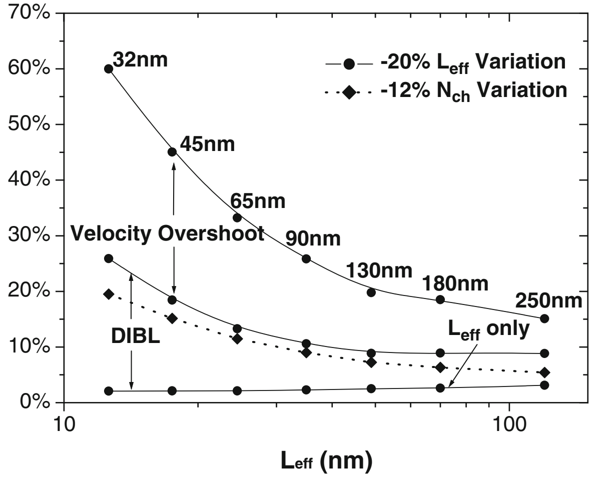


Fig. 2.15 The impact of L_{eff} variation on I_{on} during CMOS technology scaling (Adapted from [8])

Fig. 2.13, the sensitivity of I_{on} on N_{ch} variation increases when additional physical mechanisms are included. Considering the dependence of μ_0 and V_{sat} on N_{ch}, −12% N_{ch} variation leads to 15% increase in I_{on} at 45 nm node. These physical correlations were not considered in previous BPTM, which could cause significant underestimation of performance variability.

The overall map of process sensitivities is shown in Fig. 2.15 across technology generations from 130 to 32 nm. Due to increasing process sensitivities, the variation of I_{on} becomes larger during technology scaling, even if the normalized process variation remains constant, e.g., −20% and −12% for L_{eff} and N_{ch} variation, respectively (Fig. 2.15). For future technology generations, L_{eff} will continue to be the dominant factor affecting performance variation, because of its role in velocity and the DIBL effect. Second to L_{eff} variation, the impact of N_{ch} variation also keeps increasing as technology scales. Figure 2.15 shows the decomposition of the impact of L_{eff} variations during technology scaling. It reveals that velocity overshoot plays a more important role than DIBL for nanoscale MOSFET. Therefore, physical modeling of velocity overshoot is necessary in variation-aware design. Since PTM can be easily customized by tuning L_{eff}, T_{oxe}, R_{dsw}, V_{th0}, E_{ta0}, V_{dd}, and the other primary parameters, robust circuit design research under different conditions are fully supported.

In summary, a new generation of PTM was developed for 130 to 12 nm bulk CMOS technology [8]. As compared to previous BPTM, the new predictive methodology has better physicality and scalability over a wide range of process and design conditions. Both nominal values and process sensitivity are captured in the new PTM for robust design research. Excellent predictions have been verified with published transistor data. The importance of physical correlations among parameters and the impact of process variations have been evaluated. Model files for bulk CMOS down to the 12 nm node are available at http://ptm.asu.edu. These predictive model files enable early stage circuit design for end-of-the-roadmap technologies. Feedbacks from both industrial and academic researchers will be very helpful to improve the accuracy and flexibility of PTM.

References

1. International Technology Roadmap of Semiconductors, 2007. (available at http://www.itrs.net)
2. "BSIM4 Manual," University of California, Berkeley, 2005.
3. Y. Cao, T. Sato, M. Orshansky, D. Sylvester, and C. Hu, "New paradigm of predictive MOSFET and interconnect modeling for early circuit simulation," *CICC*, pp. 201–204, 2000.
4. D. Boning and S. Nassif, "Models of process variations in device and interconnect," *Design of High-Peformance Microprocessor Circuits*, Chapter 6, pp. 98–115, IEEE Press, 2000.
5. M. Miyama, S. Kamohara, M. Hiraki, K. Onozawa, and H. Kunitomo, "Pre-silicon parameter generation methodology using BSIM3 for circuit performance-oriented device optimization," *IEEE Trans. Semiconductor Manufacturing*, vol. 14, no. 2, pp. 134–142, May 2001.
6. M. Orshansky, J. An, C. Jiang, B. Liu, C. Riccobene and C. Hu, "Efficient generation of pre-silicon MOS model parameters for early circuit design," *IEEE J. Solid-State Circuits*, vol. 36, no. 1, pp. 156–159, Jan. 2001.
7. K. Vasanth, et al., "Predictive BSIM3v3 modeling for the 0.15-0.18 μm CMOS technology node: A process DOE based approach," *IEDM Tech. Dig.*, pp. 353–356, 1999.
8. W. Zhao, Y. Cao, "New generation of predictive technology model for sub-45 nm early design exploration," *IEEE Transactions on Electron Devices*, vol. 53, no. 11, pp. 2816–2823, Nov. 2006. (Available at http://ptm.asu.edu)
9. D. Sinitsky, "Physics of future very large-sclae integration (VLSI) MOSFETs," Ph. D. dissertation, Univ. of California, Berkeley, 1997.
10. G. M. Yeric, A. F. Tasch, and S. K. Banerjee, "A universal MOSFET mobility degradation model for circuit simulation," *IEEE Trans. Computer-Aided Design*, vol. 9, no. 10, pp. 1123–1126, Oct. 1990.
11. Y. M. Agostinelli, G. M. Yeric, and A. F. Tacsh, "Universal MOSFET hold mobility degradation models for circuit simulation," *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems*, vol. 12, no.3, pp. 439–445, Mar. 1993.
12. H. Ohta, et al., "High performance 30 nm gate bulk CMOS for 45 nm node with Σ-shaped SiGe-SD," *IEDM Tech. Dig.*, pp. 6–10, 2005.
13. K. Goto, et al., "High performance 25 nm gate CMOSFETs for 65 nm node high speed MPUs," *IEDM Tech. Dig.*, pp. 623–626, 2003.
14. Z. Luo, et al., "High performance and low power transistors integrated in 65 nm bulk CMOS technology," in *IEDM Tech. Dig.*, 2004, pp. 661–664.
15. C. C. Wu, et al., "A 90-nm CMOS device technology with high-speed, general-purpose, and low-leakage transistors for system on chip applications," *IEDM Tech. Dig.*, pp. 65–68, 2002.
16. V. Chan, et al., "High speed 45 nm gate length CMOSFETs integrated into a 90 nm bulk technology incorporating strain engineering," *IEDM Tech.* Dig., pp. 77–80, 2003.
17. S.-F. Huang, et al., "High performance 50 nm CMOS devices for microprocessor and embedded processor core applications," *IEDM Tech. Dig.*, pp. 237–240, 2001.
18. M. Mehrotra, et al., "60 nm gate length dual-Vt CMOS for high performance applications," *VLSI Tech. Symp.*, pp. 124–125, 2002.
19. S. Thompson, et al., "An enhanced 130 nm generation logic technology featuring 60 nm transistors optimized for high performance and low power at 0.7-1.4 V," *IEDM Tech. Dig.*, pp. 257–260, 2001.

20. S. Tyagi, et al., “A 130 nm generation logic technology featuring 70 nm transistors dual Vt transistors and 6 layers of Cu interconnects,” *IEDM Tech. Dig.*, pp. 567–570, 2000.
21. K. K. Young, et al., “A 0.13 μm CMOS technology with 193 nm lithograghy and Cu/low-k for high performance applications,” *IEDM Tech. Dig.*, pp. 563–566, 2000.
22. M. Hargrove, et al., “High-performance sub-0.08 μm CMOS with dual gate oxide and 9.7 ps inverter delay,” *IEDM Tech. Dig.*, pp. 627–630, 1998.
23. L. Su, et al., “A high-performance sub-0.25 μm CMOS technology with multiple threshold and copper interconnects,” *VLSI Tech. Symp.*, pp. 18–19, 1998.
24. M. Rodder, et al., “A 1.2V, 0.1 μm gate length CMOS technology: design and process issues,” *IEDM Tech. Dig.*, pp. 623–626, 1998.
25. M. Rodder, et al., “A 0.10 μm gate length CMOS technology with 30Å gate dielectric for 1.0V-1.5V applications,” *IEDM Tech. Dig.*, pp. 223–226, 1997.
26. M. Rodder, et al., “A sub-0.18 μm gate length CMOS technology for high performance (1.5V) and low power (1.0V),” *IEDM Tech. Dig.*, pp. 563–566, 1996.
27. M. Bohr, et al., “A high performance 0.25 μm logic technology optimized for 1.8V operation,” *IEDM Tech. Dig.*, pp. 847–850, 1996.

Chapter 3
Predictive Technology Model of Enhanced CMOS Devices

The scaling of traditional bulk CMOS structure has slowed down in recent years as fundamental physical and process limits are rapidly approached. For instance, short-channel effects, such as drain-induced-barrier-lowering (DIBL) and threshold voltage (V_{th}) rolloff, severely increase leakage current and degrade the I_{on}/I_{off} ratio (Fig. 2.11) [1]. To overcome these difficulties and continue the path projected by Moore's law, new materials need to be incorporated into the bulk CMOS structure, including high-permittivity (high-k) gate dielectrics, metal gate electrodes, low-resistance source/drain, and strained Si channel for high mobility [2, 3]. Furthermore, more flexible process choices, such as multiple-V_{th}, are required in today's integrated circuit design, in order to satisfy various design needs (e.g., low power vs. high performance). These technology evolutions should be incorporated into PTM to facilitate contemporary design exploration.

Beyond the 32 nm technology generation, more radical solutions will be vital to meet the scaling criteria of off-state leakage. The FinFET, or the double-gate device (DG), is considered as the most promising alternative technology to bulk CMOS structure [2, 4]. Predictive models for bulk CMOS technology are updated in this chapter to capture the distinct electrical behavior of these alternative materials, structures and device choices.

3.1 Strain Engineering in Scaled CMOS

During the past decades, the miniaturization of device feature sizes has driven the improvement in transistor performance [5, 6]. Meanwhile the channel doping has to keep increasing in order to meet the scaling criteria of threshold voltage. However, increased doping levels degrade carrier mobility and reduce the driving current. In addition, the reduction in channel length does not help improve carrier velocity anymore as the limit of ballistic transportation is gradually approached. In this context, strain technology, which alters the band structure and reduces the effective

Y. Cao, *Predictive Technology Model for Robust Nanoelectronic Design*,
Integrated Circuits and Systems, DOI 10.1007/978-1-4614-0445-3_3,

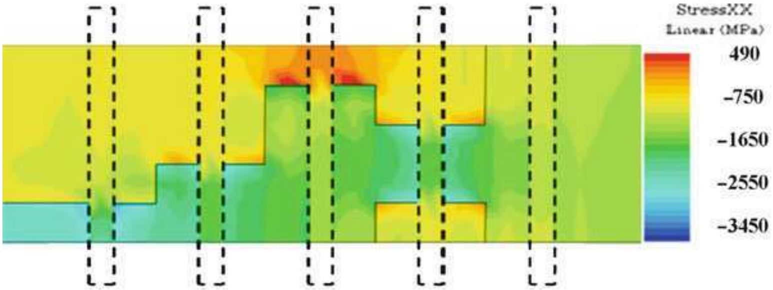


Fig. 3.1 The non-uniform stress distribution in a 45 nm layout under restrictive design rules (Adapted from [5])

mass and scattering rate, is essential to elevate carrier mobility for continual scaling. There are two types of stress: biaxial stress and uniaxial stress, both of which result in significant mobility enhancement [3, 7]. Due to lower integration complexity and smaller threshold voltage shift, uniaxial stress has been adopted since the 90 nm node [8]. The major fabrication steps involve eSiGe technology or Dual stress liner (DSL). The eSiGe technology embeds SiGe in the source and drain to introduce compressive stress, while DSL introduces the stress by depositing a highly stressed silicon nitride liner over the entire wafer [9].

In strained silicon technology, the exact amount of mobility enhancement depends on both the applied stress level in the fabrication (e.g., determined by the Ge composition for eSiGe technology) and circuit layout parameters, such as transistor length and source/drain size [10–12], because of the non-uniform stress distribution in the channel region. Figure 3.1 illustrates the simulation results of stress distribution in a 45 nm standard cell under restrictive design rules, using Taurus-Medici [13] and Taurus-Tsuprem4 [14]. The stress level is widely different across the cell, depending on transistor size, layout pitch, etc. Such non-uniformity results in pronounced variations in transistor and circuit performance and increases the complexity of modeling and simulation. Similar layout dependence is also reported from shallow trench isolation (STI) stress [15, 16].

To capture such a systematic effect, traditional efforts resort to TCAD simulation, such as the example in Fig. 3.1, to extract the stress level from the entire layout and analyze performance enhancement. This approach usually requires expensive computation, especially when chip size keeps increasing along with technology scaling. Therefore, it is necessary to develop a more effective modeling approach that is able to extract the stress effect for each device and embed it into standard model parameters for circuit simulation. This model should physically capture the impact of circuit layout on transistor performance, rather than empirical fitting [12, 17, 18], such that model scalability is guaranteed for future technology generations.

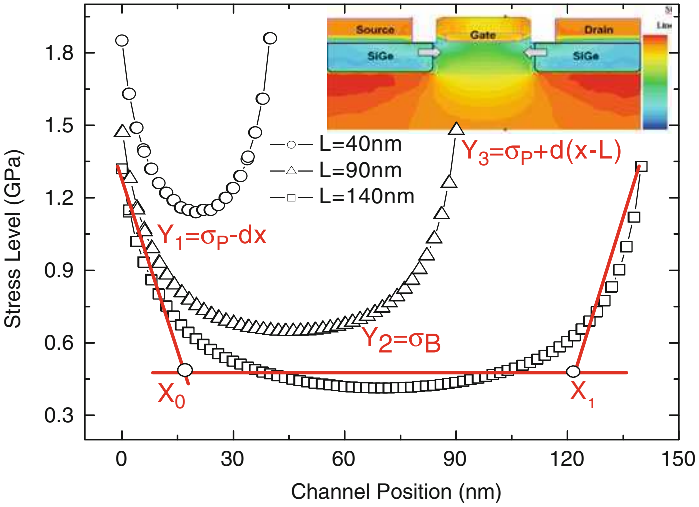


Fig. 3.2 A piecewise linear approximation of the stress distribution in the channel (Adapted from [5])

3.1.1 Modeling of Stress Distribution in the Channel

Figure 3.2 shows the TCAD simulation for stress distribution based on eSiGe technology, and the corresponding non-uniform mechanical stress, which leads to the non-uniform mobility enhancement in the channel. As investigated in [19], the stress magnitude in Si substrate decays sharply from the edge of the channel to the center. It then becomes less dependent on the distance when the location is far from the origin of the applied stress. As the channel length (L) decreases, the overall stress level is elevated, but the stress distribution follows the similar bathtub curve [14]. Without losing generality, a piecewise linear approximation is proposed to capture the stress level as Eqs. 3.1–3.3:

$$Y_1 = \sigma_P - dx \tag{3.1}$$

$$Y_2 = \sigma_B \tag{3.2}$$

$$Y_3 = \sigma_P + d(x - L) \tag{3.3}$$

where σ_P and σ_B denote the peak and bottom stress level in the channel, respectively, and d represents the slope. Y_1 and Y_3 intercept with Y_2 at points of x_0 and x_1, respectively. x_0 and x_1 are expressed as:

$$x_0 = \frac{\sigma_P - \sigma_B}{d} \tag{3.4}$$

$$x_1 = L - \frac{\sigma_P - \sigma_B}{d} \tag{3.5}$$

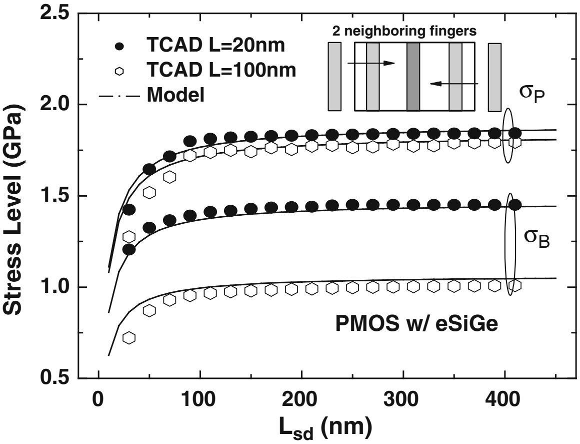


Fig. 3.3 σ_P and σ_B are modeled as functions of channel length (L) and S/D diffusion length (L_{sd}) (Adapted from [5])

Figure 3.3 shows that σ_P and σ_B become higher due to the increased amount of stressor material, when S/D diffusion length (L_{sd}) increases [11]. However, the stress level will become saturated when L_{sd} is large enough. To account for the stress dependence on L and L_{sd}, σ_P is modeled as Eq. 3.6, where σ_m is the saturation stress level and A is a fitting parameter for the dependence on L_{sd}.

$$\sigma_P = \left(1 + \frac{1}{L} + \frac{1}{L + L_{sd}} + \frac{1}{2L + L_{sd}}\right) \cdot \frac{L_{sd}}{A + L_{sd}} \cdot \sigma_m \tag{3.6}$$

$$\sigma_B = \frac{C}{C + L} \cdot \sigma_p \tag{3.7}$$

Each term in the parenthesis represents the contribution by a diffusion region, depending on their separation distance to the channel. Equation 3.6 assumes that all diffusion regions in the neighboring transistors have the same size L_{sd}. If they are different, the exact value should be used to replace the corresponding L_{sd}. On the other hand, as channel length becomes shorter, σ_B grows up and to the limit of σ_P when channel length reaches zero. This channel length dependence can be modeled by Eq. 3.7 with a fitting parameter C. In Fig. 3.3, the model shows good agreement with TCAD simulation.

3.1.2 *Equivalent Mobility Model*

Based on carrier redistribution, the strain-enhanced carrier mobility can be physically modeled in Eq. 3.8, where the coefficient, B, is a physical constant [20].

$$\frac{\mu}{\mu_0} = 1 + B \cdot \left[\exp\left(\frac{\Delta E}{kT}\right) - 1\right] \tag{3.8}$$

ΔE denotes the strain-induced energy splitting of conduction band or valence band and can be calculated by the deformation potential theory, which indicates the applied stress level is linearly proportional to energy splitting [21]. Therefore, energy splitting is modeled by Eq. 3.9.

$$\Delta E = P \cdot \sigma \tag{3.9}$$

Note that P is also temperature-dependent because the temperature alters the bandgap and further affects the energy band splitting. Therefore, the temperature-dependent behavior is modeled as Eq. 3.10, where P_0 denotes its value at room temperature (T_0).

$$P(T) = P_0 \cdot \left(\frac{T}{T_0}\right)^{\alpha} \tag{3.10}$$

Furthermore, since the stress level in the channel is not a constant (Fig. 3.2), the enhancement in carrier mobility is also non-uniform. Based on the principle of current continuity, the non-uniform mobility is captured as an equivalent mobility, μ_e, as shown in Eq. 3.11, where μ_0 denotes the unstrained mobility [22].

$$\frac{\mu_o}{\mu_e} = \frac{1}{L}\int_0^L \frac{\mu_o}{\mu} dx \tag{3.11}$$

Therefore, an analytical solution for mobility can be derived as a function of channel length and L_{sd} to bridge the layout parameters to mobility variation. Equation 3.12 summarizes this result. Figure 3.4 validates the model prediction with TCAD simulations, in which PMOS with eSiGe technology is simulated based on hydrodynamic models.

$$\begin{aligned}\frac{\mu_0}{\mu_e} = &\frac{2kT}{dPL(B-1)} \cdot \left\{\frac{-dPx_0}{kT} + \ln\left[\frac{1 + B\left(\exp\left(\frac{P\sigma_P}{kT}\right) - 1\right)}{1 + B\left(\exp\left(\frac{P\sigma_P - dPx_0}{kT}\right) - 1\right)}\right]\right\} \\ &+ \frac{L - 2x_0}{L \cdot \left[1 - B + B\exp\left(\frac{P\sigma_B}{kT}\right)\right]}\end{aligned} \tag{3.12}$$

While carrier mobility is mainly responsible for the linear operation region, saturation velocity, V_{sat}, is usually used to describe the high E-field behavior in the saturation region. Equation 3.13 shows a simplified solution of the energy balance equation [23], which accounts for the velocity overshoot behavior in a short channel device. This simplified solution considers how mobility influences the high E-field behavior:

$$V_{sat} = V_{sat0} + 0.13\mu_{eff}\sqrt{\tau\mu_{eff}kT/q} \cdot (V_d/L_{eff}^2) \tag{3.13}$$

where μ_{eff} is a linear function of μ_e in Eq. 3.12 [24].

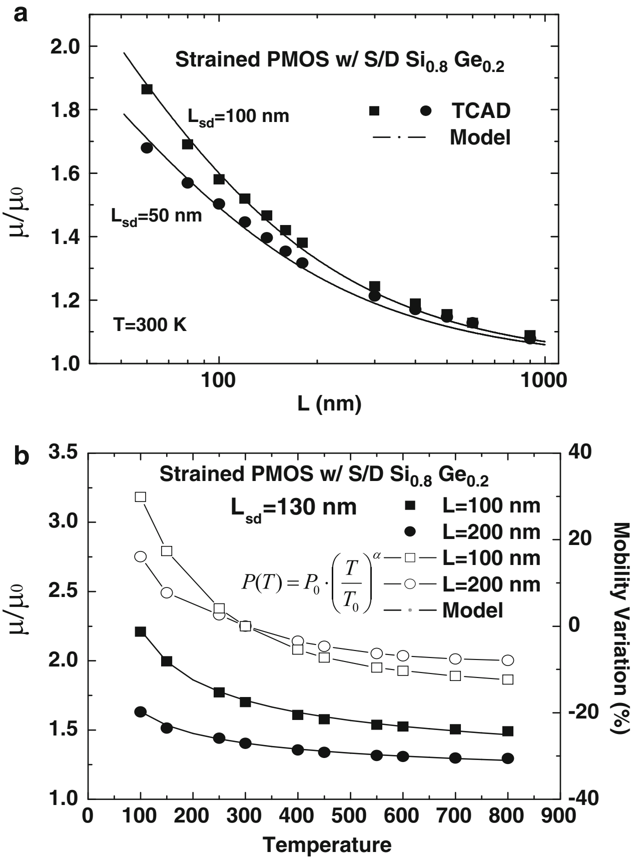


Fig. 3.4 Equivalent mobility enhancement with the dependence on layout parameters and temperature **(a)** Mobility increases as L decreases, and **(b)** Mobility increases as temperature decreases; the device with longer channel length is less vulnerable to the temperature variation (Adapted from [5])

3.1.3 Strain Induced Threshold Voltage Shift

In addition to strain-induced mobility change, threshold voltage reduction is also pronounced in the strained devices. The change in threshold voltage is attributed to strain-induced variation of energy bandgap, electron affinity, and density of states (DOS), where the effect of density of states (DOS) can be ignored due to its insignificant impact [25]. Based on the deformation potential theory [8, 21], the strain-induced change in bandgap and electron affinity is proportional to the applied stress magnitude, so that the threshold voltage change is modeled by Eq. 3.14,

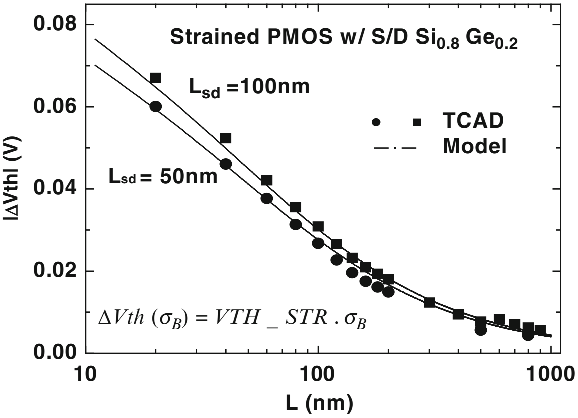


Fig. 3.5 Strained induced threshold voltage shift as a function of channel length and S/D diffusion length (L_{sd}) (Adapted from [5])

where VTH_STR is a fitting parameter to capture the linear relationship between threshold voltage shift and the applied stress magnitude. Note that the bottom stress level (σ_B) is used to calculate threshold voltage shift because the lower the stress level is, the smaller the reduction of the barrier in the channel is.

$$\Delta V_{th}(\sigma_B) = VTH\_STR \cdot \sigma_B \tag{3.14}$$

Figure 3.5 validates the model of strain-induced threshold lowering at various channel lengths. Theoretically, applying stress affects the intrinsic carrier density, which is an exponential function of the bandgap, and changes the barrier between source and substrate, as well as the bulk potential. These effects further influence DIBL, subthreshold swing and the body bias dependence. More careful analysis indicates that these effects are secondary to the change of mobility, velocity, and threshold voltage.

The above models are adequate to predict the performance enhancement by strained silicon technology. They are scalable with device and design parameters. Figure 3.6a evaluates the device performance in both linear and saturation regions. The driving current is significantly improved in the strained device, i.e., 47% and 99% for I_{on} and I_{lin}, respectively. Therefore, strain technology is very promising for future high-performance applications. Figure 3.6b compares the IV characteristics in the sub-threshold region. The off-state current of the strained device is larger than that of an unstrained device. In all operation regions, the new model matches very well with TCAD simulation results.

With the scaling of the device dimension, the strain-induced mobility and threshold voltage shift becomes more significant, as the stress level goes up with channel length scaling (Fig. 3.1). Therefore, it is essential to develop compact models of the layout dependent stress effect for circuit analysis and optimization. The proposed models provide a solution that bridges device and layout parameters with transistor electrical characteristics.

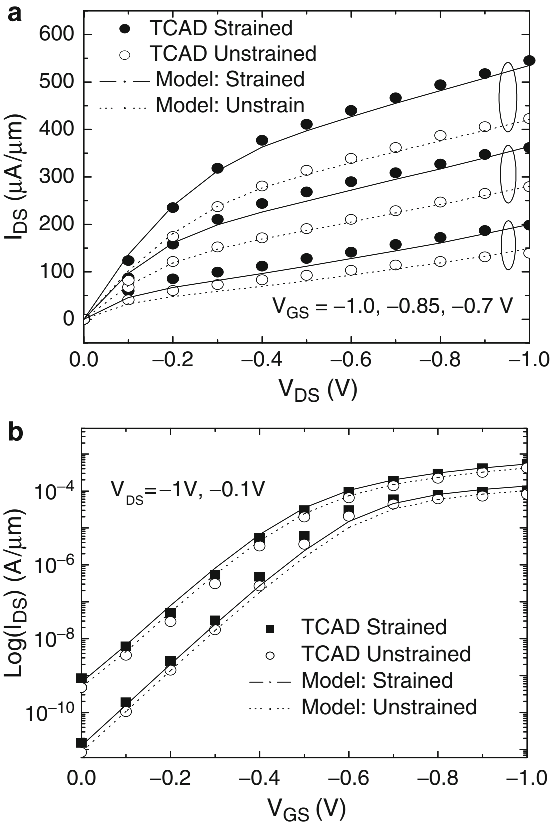


Fig. 3.6 The impact of strain on IV curves in the regions of (**a**) super-threshold and (**b**) sub-threshold (Adapted from [5])

3.2 High-*k*/Metal Gate and Multiple-V_{th} Devices

High-*k*/metal gate (HK/MG) has been adopted into IC production since the 45 nm technology node [26]. High-*k* dielectrics help reduce gate leakage and allow more aggressive scaling of gate dielectrics than classic silicon oxide, while the metal gate is necessary to tune the threshold voltage [2]. However, the implementation of high-*k* dielectrics comes at the expense of transistor reliability. The consequences include a larger amount of negative-bias-temperature-instability (NBTI) and faster degradation of the drain current [27, 28]. Additional compact models need to be developed to account for the instability and to support reliability-aware design (Chapter 5) [29].

Furthermore, as CMOS scales to sub-45 nm nodes, there will be multiple process choices to meet various design requirements, such as high performance and low power [6]. To satisfy this trend and make an impact on leading products, PTM needs to be extended with more diversity and flexibility. This emerging demand requires PTM to cover alternative process choices and some previous secondary

Fig. 3.7 The distribution of transistor numbers at different V_{th} in an industrial 65 nm low power design (Adapted from [30])

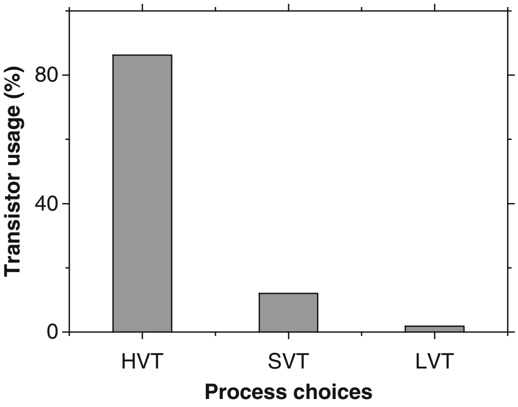
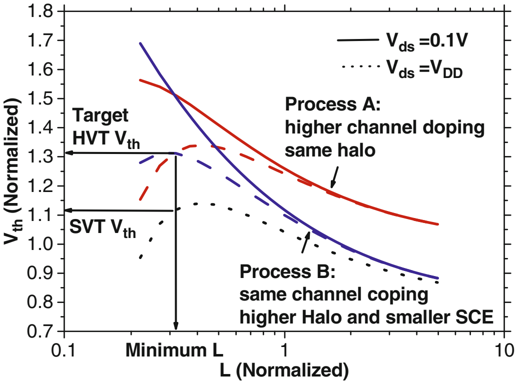


Fig. 3.8 The target threshold voltage can be achieved by tuning either channel doping or halo doping (Adapted from [30])



effects that are prominent in future low power design, especially high-k/metal gate (HK/MG) technology, gate leakage current, multiple-threshold (V_{th}) technology, as well as temperature and body bias effects. Based on realistic technologies at 65 and 45 nm nodes, PTM was extended to address the above issues, enabling early stage design activity for low power applications [30].

To balance the needs for low power and high speed, multiple V_{th} and gate length (L) biasing are commonly adopted. In a typical low power design, high V_{th} (HVT) devices are often dominant, with only a small portion of transistors at standard V_{th} (SVT) and low V_{th} (LVT) to boost the performance of critical paths, as shown in Fig. 3.7 [30]. From SVT to LVT or HVT, different process techniques can be used to tune V_{th}, including the tuning of either channel doping, which only affects long channel V_{th}, or halo doping that controls short channel effects (SCE). For example, to increase V_{th} from SVT to HVT (Fig. 3.8), we can either increase channel doping (Process A) or solely use a higher halo doping (Process B). Both techniques produce the same target HVT at the minimum L, yet they have

Fig. 3.9 Different V_{th} tuning techniques affect the sensitivity of I_{off} under L variation (Adapted from [30])

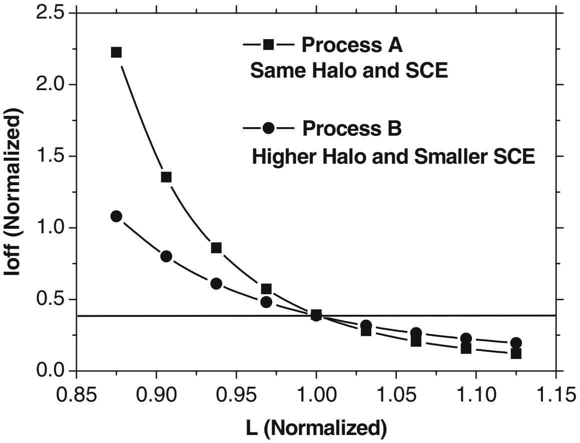
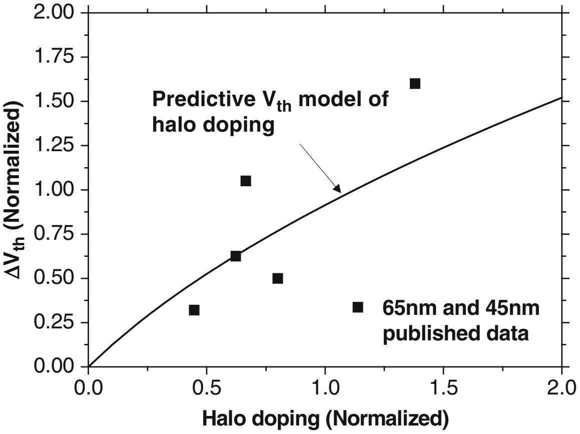


Fig. 3.10 V_{th} change by tuning halo doping (Adapted from [30])



different impact on subthreshold leakage under process variations: as shown in Fig. 3.9, I_{off} of Process B has a lower sensitivity to L variation than that of Process A, while Process A has a better suppression of subthreshold leakage with L biasing.

During the development of PTM, such process options are considered by incorporating physical V_{th} models of channel and halo doping [30]. Figure 3.10 validates the V_{th} change under various halo doping at 65 and 45 nm nodes. By adding these models into the predictive methodology in [1], both IV and CV for multiple V_{th} technologies can be accurately generated. Figure 3.11 verifies the IV of 45 nm SVT devices, which is predicted from previous 65 nm technology node. This predictive methodology is extendable toward the 32 nm node and below. Since the 45 nm node, PTM separate the predictions of high-performance (HP) and low-power (LP) applications. They mainly differ in the values of V_{th}, T_{ox} and V_{dd}.

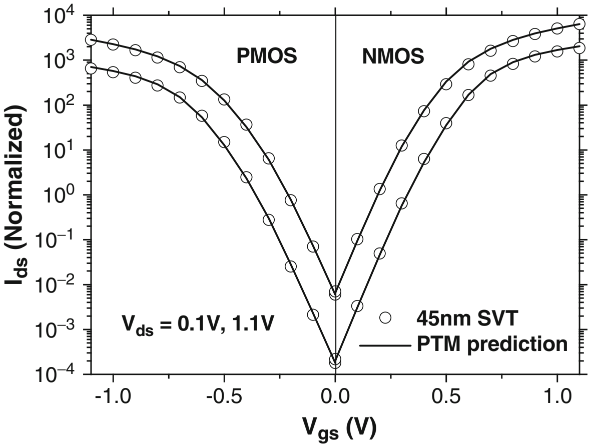


Fig. 3.11 The verification of predicted 45 nm SVT process from a 65 nm technology (Adapted from [30])

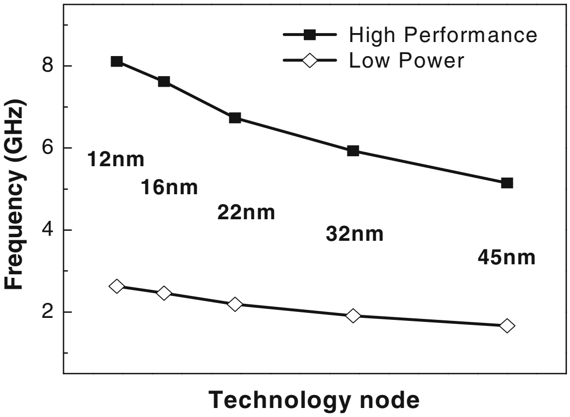


Fig. 3.12 The trends of RO frequency scaling for HP and LP PTM

Figure 3.12 compares the switching frequency between HP and LP predictions, using a 21-stage inverter-based ring oscillator (RO, FO $= 1$).

In addition to diverse process choices, gate leakage current increases exponentially with the scaling of EOT. PTM covers this effect based on scalable models of leakage current and the calibration at 65 and 45 nm nodes (Fig. 3.13). The impact of temperature on mobility, V_{th} and IV is expected to remain the same, as confirmed by the published data [30]. At the 32 nm node and beyond, HK/MG technology will be implemented to control gate tunneling current, which may also boost I_{on}. Figure 3.14 shows the smooth predictions of I_{on} and I_{off} at the 32 nm node with and without HK/MG for three V_{th} processes. I_{off} of HVT slightly deviates from the nominal trend due to the GIDL and tunneling current. Besides the prediction of IV, the scaling trend of gate and parasitic capacitances are included in PTM, since they are important for dynamic circuit performance.

Fig. 3.13 Gate tunneling current (J_g) prediction (Adapted from [30])

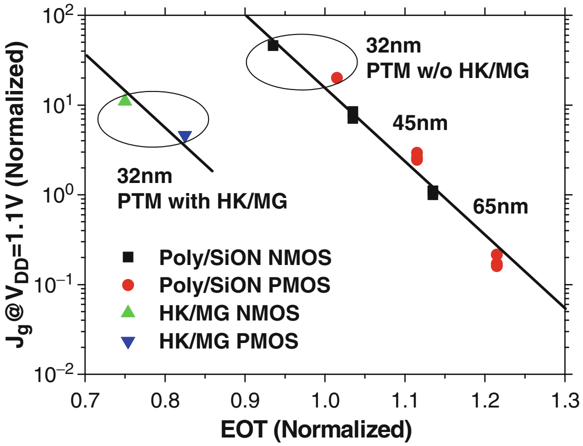
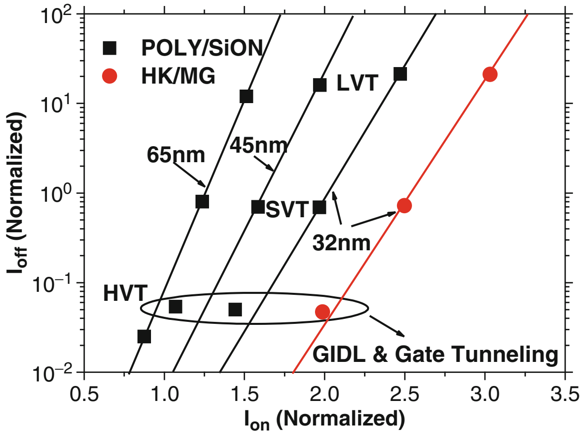


Fig. 3.14 Predictions of 32 nm I_{on} and I_{off} (Adapted from [30])



3.3 Modeling of the FinFET Structure

Beyond the 22 nm technology node, more radical solutions will be necessary to meet the scaling criteria for off-state leakage. The FinFET, or the double-gate device (DG), is a vertical structure that is regarded as the most promising alternative technology because of its improved scalability and the effective suppression of short-channel effects [2, 4]. Figure 3.15 illustrates the structure of a FinFET device. The FinFET device is electrostatically more robust than bulk CMOS since two gates are used to control the channel. When the body silicon thickness (T_{si})

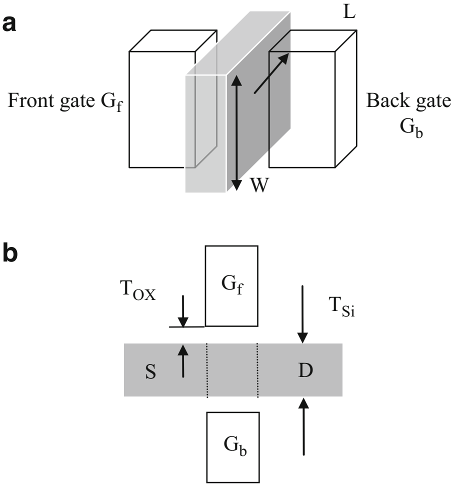


Fig. 3.15 The structure of a FinFET device (**a**) Three dimensional view, and (**b**) The top view of both front and back gates

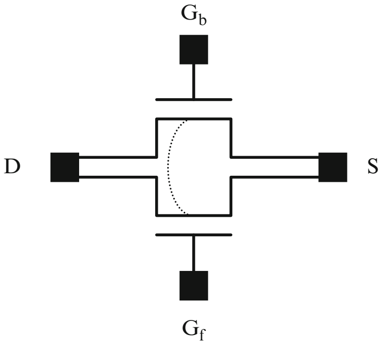


Fig. 3.16 The sub-circuit model of a FinFET device

is sufficiently thinner than the channel length, short-channel effects, such as V_{th} lowering, DIBL, and degraded sub-threshold swing, can be effectively suppressed [31]. With a lightly doped channel, the threshold voltage of a FinFET transistor is weakly affected by random dopant fluctuations [2, 32]. The front and back gates (Fig. 3.15) can be connected together or biased independently, using the front gate to switch the transistor on/off and the back gate as a control signal [2]. At the 32 nm node, it may improve the I_{on}/I_{off} ratio by more than 100% [31].

Extensive research has been conducted to understand the underlying physics [33, 34]. Yet a compact model for DG devices, akin to the BSIM [24] and PSP model [35] for the bulk CMOS transistor, has not been available for the purposes of circuit simulation and technology prediction. Currently early design research with FinFET has to resort to the TCAD simulators (e.g., MEDICI), which are computationally expensive and limit design insights. To overcome these barriers, an equivalent sub-circuit model for a FinFET device is proposed (Fig. 3.16). This circuit model consists of two fully depleted SOI devices for the front and back transistors, respectively. BSIM SOI is used as the model for each device, such that this sub-circuit is compatible with circuit simulators (e.g., SPICE) [36].

Figure 3.16 illustrates the detailed schematics of this equivalent circuit model. Two single gate transistors are used to capture the current conduction controlled by the front and back gate in a FinFET transistor. Each sub-transistor has its own definitions of gate voltage (V_G), V_{th}, and T_{ox}. Their sources and drains are electrically connected to form a four-node circuit. Thus, the drain voltage (V_D) and the source voltage (V_S) are shared. Both sub-transistors have the same gate length (L_{gate}) and W. Since the bottom of a FinFET structure sits on top of a layer of SiO_2, the FinFET is inherently a SOI transistor. Furthermore, in the typical process range of a FinFET, T_{si} is so thin that the silicon body is fully depleted. Therefore, the fully depleted SOI model of BSIM (BSIM FD SOI) is used as the model basis for each sub-transistor in Fig. 3.16.

A unique property of a FinFET device, which is different from a traditional FD SOI transistor, is the electrical coupling between the front and back transistors. Specifically, the threshold voltage of the front transistor (V_{thf}) is governed not only by the process conditions, but also by the back gate voltage V_{Gb}. Such an effect is similar to the body effect in a bulk device; instead of the body contact, V_{Gb} affects V_{thf} through the capacitance partition between the gate oxide capacitance (C_{oxb} and C_{oxf}) and the silicon body capacitance (C_{si}) in a FinFET device [37]:

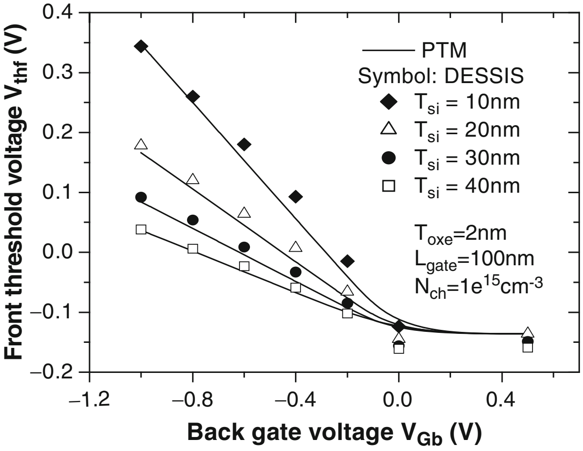
$$\frac{\partial V_{thf}}{\partial V_{Gb}} = -\frac{C_{si}||C_{oxb}}{C_{oxf}} \tag{3.15}$$

$$\frac{\partial V_{thb}}{\partial V_{Gf}} = -\frac{C_{si}||C_{oxf}}{C_{oxb}} \tag{3.16}$$

where $C_{si} = (\varepsilon_{si}/T_{si})$ and $C_{ox} = (\varepsilon_{ox}/T_{ox})$. Note the electrical coupling between V_{thf} and V_{Gb} only exists when the back sub-transistor is in the depletion region. As soon as the back sub-transistor enters the inversion region (i.e., $V_{Gb} > V_{thb}$), the impact of V_{Gb} on V_{thf} is shielded by the inversion layer and rapidly diminishes. These physical relationships are implemented in our sub-circuit model, with an empirical function continuously capturing this effect across the depletion to inversion regions.

Figure 3.17 evaluates the prediction of our equivalent circuit model, which is generated from SPICE, against the results from TCAD simulations (i.e., DESSIS) [38]. For a variety of T_{Si}, the gate coupling behavior is well captured with maximum discrepancy smaller than 10%. For a sub-45 nm FinFET, since the total dopant in the channel is small, V_{th} is independent on the channel doping (N_{ch}) if N_{ch} is smaller than $1e^{17}cm^{-3}$ [38]. Therefore, the FinFET is relatively immune to the variation in V_{th} due to random channel dopant fluctuations, which is a severe concern in the nanoscale bulk CMOS and leads to increase in leakage and SRAM instability (Chapter 4). The coupling between V_{thf} and V_{Gb} is more pronounced when the silicon body becomes thinner, i.e., a relatively larger C_{si}. Based on this

Fig. 3.17 The coupling between V_{thf} and V_{Gb} in a FinFET device



equivalent circuit model, PTM for a FinFET device was developed, following the predictive methodology in Chapter 2 for other primary model parameters.

In summary, PTM introduces scalable models for strained Si, multiple V_{th} and HK/MG processes, and the FinFET structure. Primary parameters under the influence of these technology enhancements include the increase of mobility, the control of SCE and the coupling between front and back gates in a FinFET device. As verified with published data, the thermal effect, particularly that on mobility, V_{th} and IV, is expected to remain the same during the scaling [30]. Predictive modeling of enhanced CMOS devices is applicable toward the 12 nm node, helping illustrate diverse design opportunities and challenges.

References

1. W. Zhao, Y. Cao, "New generation of predictive technology model for sub-45 nm early design exploration," *IEEE Transactions on Electron Devices*, vol. 53, no. 11, pp. 2816–2823, Nov. 2006. (Available at http://ptm.asu.edu).
2. L. Chang, et al., "Extremely scaled silicon nano-CMOS devices," *Proceedings of the IEEE*, vol. 91, no. 11, pp. 1860–1873, Nov. 2003.
3. N. Mohta and S. E. Thompson, "Mobility enhancement: The next vector to extend Moore's law," *IEEE Circuits and Devices Magazine*, vol. 21, no. 5, pp. 18–23, Sept./Oct., 2005.
4. X. Huang, W.C. Lee, C. Kuo, D. Hisamoto, L. Chang, J. Kedzierski., "Sub-50 nm FinFET: PFET," in *Proc. IEEE International Electron Devices Meeting*, pp. 679–682, Dec. 2003.
5. C.-C. Wang, W. Zhao, F. Liu, M. Chen, Y. Cao, "Predictive modeling of layout-dependent stress effect in scaled CMOS design," *International Conference on Computer Aided Design*, pp. 513–520, 2009.
6. The International Technology Roadmap for Semiconductors (ITRS), 2008.
7. A. Khakifirooz and D. Antoniadis, "Transistor performance scaling: The role of virtual source velocity and its mobility dependence," in *Proc. IEEE International Electron Devices Meeting*, pp. 667–670, Dec. 2006.

8. J.-S. Lim, S. E. Thompson, and J. G. Fossum, “Comparison of threshold-voltage shifts for uniaxial and biaxial tensile-stressed n-MOSFETs,” *TED*, vol. 25, no. 11, pp. 731–733, 2004.
9. H. Nii, et al., “A 45 nm High Performance Bulk Logic Platform Technology (CMOS6) using Ultra High NA(1.07) Immersion Lithography with Hybrid Dual-Damascene Structure and Porous Low-k BEOL,” in *IEDM*, pp. 685–688, 2006.
10. H. Aikawa, et al., “Variability aware modeling and characterization in standard cell in 45 nm CMOS with stress enhancement technique,” *VLSI Symp.,* 2008.
11. G. Eneman, et al., “Scalability of the Si1-xGex Source/Drain technology for the 45-nm technology node and beyond,” *TED*, vol. 53, no. 7, Jul. 2006.
12. M. V. Dunga, C.H. Lin, X. Xi, D.D. Lu, A.M. Niknejad, and C. Hu, “Modeling advanced FET technology in a compact model,” *IEEE Transactions on Electron Devices*, vol. 53, no. 9, pp. 1971–1978, Sept. 2006.
13. Taurus Medici, Manual, June 2006. Version Y-2006.06.
14. Taurus Tsuprem4, Manual, Oct. 2005. Version X-2005.10.
15. G. Scott, et al., “NMOS drive current reduction caused by transistor layout and trench isolation induced stress,” in *IEDM Tech. Dig*., pp. 827–830, 1999.
16. R. A. Bianchi, et al., “Accurate modeling of trench isolation induced mechanical stress effects on MOSFET electrical performance,” *IEDM*, pp. 117–120, 2002.
17. K.-W. Su, et al., “A scaleable model for STI mechanical stress effect on layout dependence of MOS electrical characteristics,” *CICC*, pp. 245–248, 2003.
18. X.-W. Lin, “Modeling of Proximity Effects in Nanometer MOSFET’s,” *IEEE/ACM Workshop on Compact Variability Modeling 2008*.
19. C. E. Murray, “Mechanics of edge effects in anisotropic thin film/substrate systems,” *Journal of Applied Physics*, vol. 100, 103532, 2006.
20. J. L. Egley, “Strain effects on device characteristics: implementation in drift-diffusion simulators,” *Solid-State Electronics*, vol. 36, no. 12, pp. 1653–1664, 1993.
21. C. Herring, E. Vogt., “Transport and deformation-potential theory for many-valley semiconductors with anisotropic scattering,” Phys. Rev., vol. 101, pp. 994–961, 1956.
22. F. Payet, F. Buf, C. Ortolland, T. Skotnicki, “Nonuniform mobility-enhancement techniques and their impact on device performance,” *TED*, vol. 55, no. 4, pp. 1050–1057, April 2008.
23. D. Sinitsky, “Physics of future very large-sclae integration (VLSI) MOSFETs,” Ph.D. dissertation, Univ. California, Berkeley, CA, 1997.
24. BSIM4 Manual, University of California Device Group.
25. W. Zhang, J. G. Fossum, “On the threshold voltage of strained-Si-Si$_{1\text{-}x}$Ge$_x$ MOSFETs,” *TED*, vol. 52, no. 2, pp. 263–268, February 2005.
26. K. Mistry, et al., “A 45 nm logic technology with high-k + metal gate transistors, strained Silicon, 9 Cu interconnect layers, 193 nm dry patterning, and 100% Pb-free packaging,” *IEDM*, pp. 247–250, 2007.
27. C. Leroux, et al., “Characterization and modeling of hysteresis phenomena in high-k dielectrics,” in *Proc. IEEE International Electron Devices Meeting*, pp. 737–740, Dec. 2004.
28. A. E. Islam, et al., “Gate leakage vs. NBTI in plasma nitrided oxides: Characterization, physical principles, and optimization,” in *Proc. IEEE International Electron Devices Meeting*, pp. 329–332, Dec. 2006.
29. S. Bhardwaj, W. Wang, R. Vattikonda, Y. Cao, S. Vrudhula, “Predictive modeling of the NBTI effect for reliable design,” *in Proc. IEEE Custom Integrated Circuits Conference*, pp. 189–192, Sept. 2006.
30. W. Zhao, X. Li, M. Nowak, and Y. Cao, “Predictive technology modeling for 32 nm low power design,” *International Semiconductor Device Research Symposium*, TA4-03, 2007.
31. W. Zhao and Y. Cao, “Predictive technology model for nano-CMOS design exploration,” *ACM Journal on Emerging Technologies in Computing Systems*, vol. 3, no. 1, pp. 1–17, April 2007.
32. J. Kedzierski, et al., “High-performance symmetric-gate and CMOS compatible Vt symmetric-gate FinFET devices,” *IEDM*, pp. 497–444, 2001.

33. J. G. Fossum, M.M. Chowdhury, V.P. Trivedi, T.-J. King, Y.-K. Choi, J. An, B. Yu, “Physical insights on design and modeling of nanoscale FinFETs,” in *Proc. IEEE International Electron Devices Meeting*, pp. 679–682, Dec. 2003.
34. B. Iniguez, T. A. Fjeldly, A. Lazaro, F. Danneville, and M. J. Deen, “Compact-modeling solutions for nanoscale double-gate and gate-all-around MOSFETs,” *IEEE Transactions on Electron Devices*, vol. 53, no. 9, pp. 2128–2142, Sept. 2006.
35. G. Gildenblat, et al., “PSP: An advanced surface-potential-based MOSFET model for circuit simulation,” *IEEE Transactions on Electron Devices*, vol. 53, no. 9, pp. 1979–1993, Sept. 2006.
36. H. Wan, X. Xi, A. M. Niknejad, C. Hu, *BSIM SOI Manual*. The Device Group, University of California, Berkeley, CA, 2003.
37. F.-L. Yang, et al., “35 nm CMOS FinFETs,” *Symposium on VLSI Technology*, pp. 104–105, 2002.
38. C.-H. Lin, et al., “Compact modeling of FinFETs featuring in independent-gate operation mode,” in *Proc. IEEE International Symposium on VLSI Technology, Systems, and Applications*, pp. 120–121, 2005.

Chapter 4
Statistical Extraction and Modeling of CMOS Variability

While technology scaling can be extended with alternative materials and structures, CMOS technology will eventually reach the ultimate limits that are defined by both physics and the fabrication process. One of the most profound physical effects will result from the vastly increased parameter variations due to intrinsic randomness, the manufacturing process, and other environmental factors [1–3]. Examples include random dopant fluctuation (RDF), line-edge roughness (LER), and random telegraph noise (RTN) [4–6]. For instance, Fig. 4.1 illustrates the scaling trend of RDF, based on PTM [7]. As the device size scales down, the total number of channel dopants significantly decreases, resulting in a dramatic increase in threshold variation [8].

These effects used to be a design issue primarily for analog circuits, but are now moving to digital circuits as the device dimension is approaching the 10 nm regime. They influence all aspects of circuit performance, especially in the design of SRAM cells that are highly vulnerable to transistor mismatches. Although in tradition, device variability is mostly handled with improvements in the manufacturing process, the semiconductor industry starts to accept the fact that some of the negative effects can be better mitigated during the design stage [9]. To maintain design predictability with those extremely scaled devices, compact models should be extended from the traditional corner-based approach to a suite of research efforts, including in-situ characterization techniques, variation extraction methods, first-principle simulations, modeling of leading variability mechanisms, and highly efficient strategies for the statistical design paradigm [10]. While the characterization and extraction techniques provide realistic data to calibrate the model, compact modeling of device variability is important to understanding the variations, guiding the test chip design, and diagnosing their performance impact [11].

Y. Cao, *Predictive Technology Model for Robust Nanoelectronic Design*,
Integrated Circuits and Systems, DOI 10.1007/978-1-4614-0445-3_4,

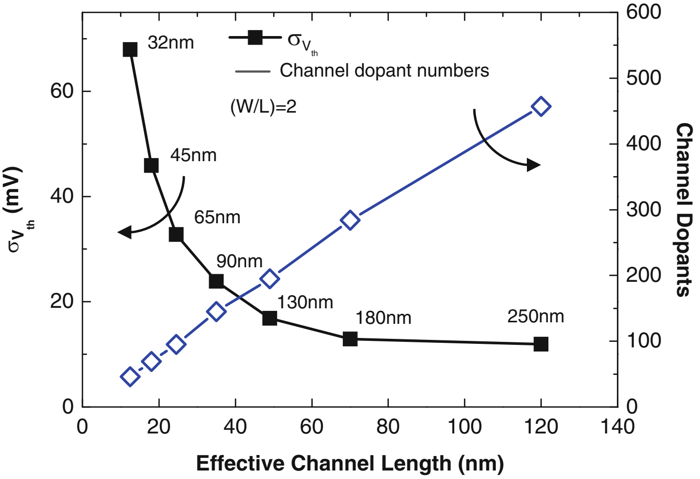


Fig. 4.1 The scaling trend of V_{th} variance due to RDF [7]

4.1 Variability Characterization and Extraction

Process variations usually manifest themselves as parameter fluctuations in nanoscale transistors, such as the channel length, the threshold voltage, and transistor parasitics. The main modeling issue under variations is to identify systematic variation components, develop predictive models for performance analysis, and incorporate them into design tools. By characterizing appropriate test structures, static process variations need to be correctly extracted and embedded into a transistor model file, such that a circuit designer can perform statistical analysis and optimization to mitigate performance variability. A rigorous extraction method further helps understand the variation mechanisms during technology scaling.

Based on compact device models, such as BSIM, EKV, or PSP models, previous works have proposed to extract the statistics of device parameters from measurements [12–14]. Such methods are used by foundries to generate statistical device models. However, existing approaches usually involve empirical fittings of too many model parameters, leading to inaccurate model sensitivity for statistical analysis [15]. Meanwhile, the complexity of the underlying device physics, as well as the manufacturing process, has dramatically increased in scaled CMOS technology [10]. As a result, physical extraction and decomposition of primary variations become even more challenging. The mismatch between the model and the hardware measurement further widens the gap in our understanding of process variations, providing inadequate guidance to the design of on-chip characterization structures [16, 17].

This section demonstrates a rigorous method in a 65 nm technology that efficiently and physically extract primary variations, including the threshold voltage, gate length, and effective mobility. Based on BSIM4 model, only three critical IV

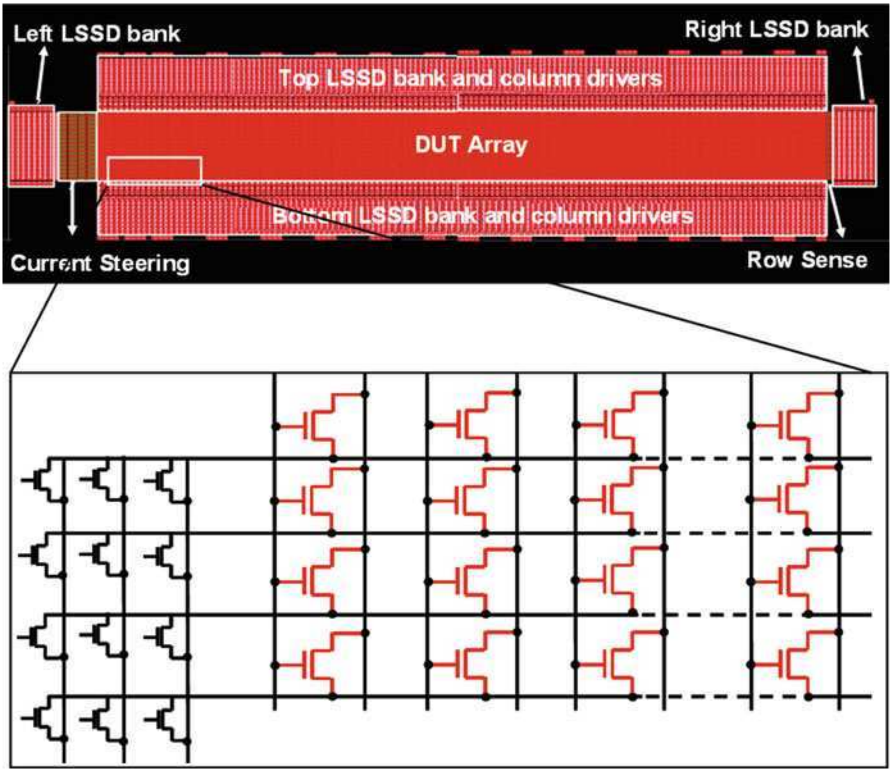

Left LSSD bank
Right LSSD bank
Top LSSD bank and column drivers
DUT Array
Bottom LSSD bank and column drivers
Current Steering
Row Sense

Fig. 4.2 Layout and die photo of the test chip (Adapted from [25])

points are needed to sample the device. The distribution of dominant model parameters is directly identified from these three points. By embedding these parameter variations into the transistor model file, the variability of IV characteristics is accurately predicted in all operation regions.

4.1.1 Test Chip and In-Situ Measurement

Figure 4.2 shows the overall scheme of the test structure that is manufactured in a 65 nm process to evaluate the IV characteristics of each device [18]. It is approximately 1250 μm × 110 μm. The core of the structure is an array containing 96,000 devices densely placed in 1,000 columns, with 96 devices in each column. Level Sensitive Scan Design (LSSD) latch banks are placed on all four sides of the array to enable row-column addressing, calibrating and measuring of each individual device. The array structure significantly improves the efficiency of large-scale measurement. However, the typical array structure is not suitable for measuring leakage current, which is essential to subthreshold IV characterization, since

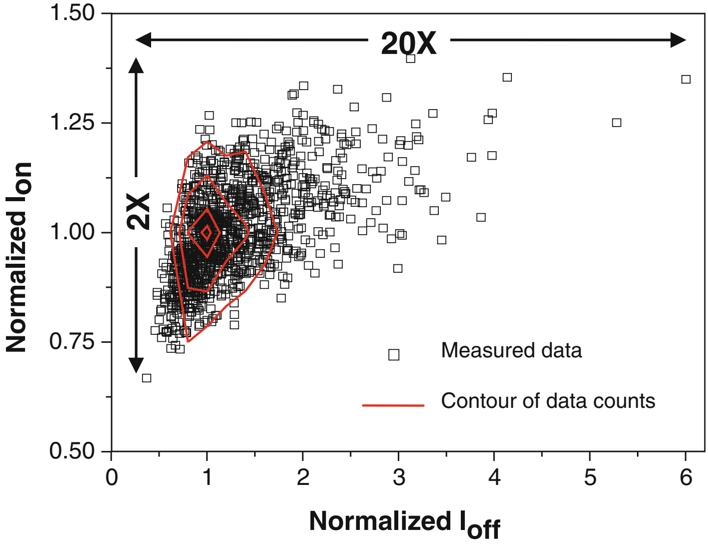


Fig. 4.3 Measured I_{on} and I_{off} data (Adapted from [25])

off-state leakage of unselected devices usually adds to the leakage of the selected device. To address this specific issue, the current steering technique is designed in the test structure [19]. When a target column is activated, only the leakage from the selected device flows to the measurement pin, while the leakage from all other devices in the same column is steered towards the sink pin and thus, does not interfere with the measurement [19]. By implementing current steering and active sensing techniques, the structure allows measuring the IV characteristics of each device with various bias voltages [19]. These techniques also effectively alleviate the requirement on the statistical pre-characterization, which could be challenging for high-volume measurement.

Figure 4.3 illustrates the measurement of the maximum drive current (I_{on}) and leakage (I_{off}), as well as the contours of data distribution. The center of the contours refers to the region with the highest data density. This position indicates the nominal performance of sampled devices. Within this test array, I_{on} varies about 2X, but the distribution of I_{off} is much wider, as a result of its higher sensitivity to threshold variation. Due to its high sensitivity to parameter variations, the leakage region is a better choice for the extraction procedure. This is different from traditional I_{on}-based extraction methods that may not be sufficient to decouple various sources.

4.1.2 Extraction and Decoupling of Variations

The measured IV statistics need to be converted into the variations of transistor parameters in order to support statistical circuit simulation. To begin with, a complete set of primary and independent variation sources needs to be identified [20]. It is well known that channel length (L) and long channel threshold voltage (V_{th}) are the most important variation sources, due to sub-wavelength lithography and etching process steps, and random dopant fluctuations, respectively [21]. In recent years, effective mobility (μ) is also emerging as an additional key variation source due to the local fluctuation of the mechanical stress, either from the strained silicon technology to enhance the current, or from the parasitic stress from shallow-trench-isolation (STI) [22]. In the nanoscale regime, it is increasingly difficult to control the level of stress with different layout patterns. Because of the extreme difficulties in the control of lithography, etching, channel doping, and stress, the variations of L, V_{th} and μ are the dominant sources in our extraction. BSIM4 is used as the model platform to demonstrate this extraction, while the method is general enough for other model templates.

Similar as [23], our extraction method focuses on the subthreshold region instead of the saturation current to determine V_{th}, since the leakage is highly sensitive to V_{th} process variations. The extraction of L variation is traditionally more difficult, because the saturation current is relatively insensitive to gate length due to velocity saturation [24]. It is also coupled with other variation sources, such as mobility. In contrast, the leakage and the value of V_{th} for a short-channel device are significantly different under various V_{ds} and L because of the effect of DIBL [24]. Thus, the difference in V_{th} between high and low V_{ds} (ΔV_{th}) is used to decouple V_{th} and L variation:

$$\Delta V_{th} \propto V_{ds} \exp(-L/l') \tag{4.1}$$

where l' is a DIBL parameter from the nominal model file. At low V_{ds}, I_{ds} is mainly dependent on the V_{th}:

$$I_{ds} \propto e^{\frac{-V_{th}}{kT/q}} \tag{4.2}$$

while at high V_{ds}, I_{ds} has a strong dependence on both V_{th} and L:

$$I_{ds} \propto e^{\frac{-V_{th}}{kT/q}} \cdot e^{\frac{V_{ds} \cdot \exp(-L/l')}{kT/q}} \tag{4.3}$$

Such a difference helps us decouple V_{th} and L variations under different V_{ds}. From Eqs. 4.2 and 4.3, the variations of V_{th} and L are separated, with sufficient accuracy for the prediction of the leakage current. Furthermore, the variation of effective mobility is extracted from the linear region, since the linear current is proportional to effective mobility. Figure 4.4 highlights three critical points for such extraction algorithm: two points from the leakage region under high and low V_{ds}, and the third point is from the linear region: namely Point 1 ($V_{gs} = V_{th}$, V_{ds} ~ 0.1V) and Point 2 ($V_{gs} = V_{th}$, $V_{ds} = V_{dd}$) are selected to extract V_{th} and

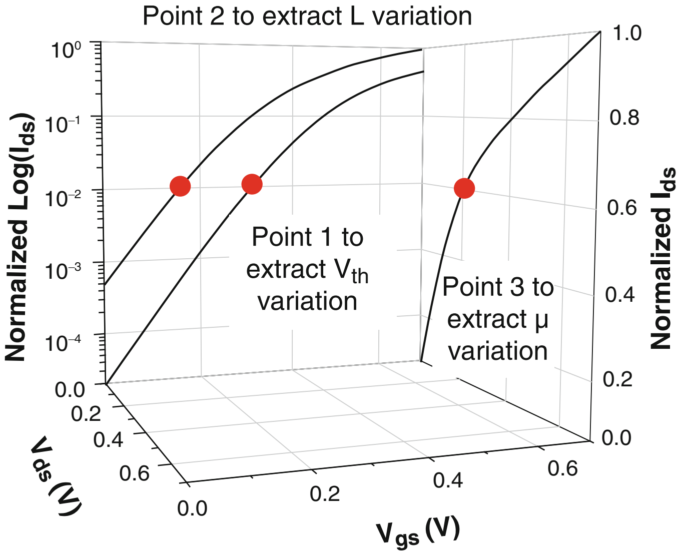


Fig. 4.4 Three critical sampling points on the IV curve for the extraction (Adapted from [25])

L variation, respectively; Point 3 ($V_{gs} = V_{dd}$, $V_{ds} \sim 0.2V$) is to extract mobility variation. The extraction algorithm is:

1. Start from a well-characterized nominal model file: The nominal device is selected from the sampled region with the highest data density (the center of the contours in Fig. 4.3). It provides the basis for further variation study. The nominal values of several important effects, such as DIBL, source/drain resistance (R_{ds}), mobility and velocity, are critical to determine the model sensitivity to parameter changes [25].
2. Extract V_{th} and L variations: V_{th} and L are extracted from the leakage region, relying on the exponential dependence of the leakage current and DIBL on V_{th} and L, respectively. For instance, if the current at Point 1 of a target transistor is lower than that calculated by nominal model file, V_{th} needs to be reduced until they match each other. For a given device, the difference of V_{th} under different V_{ds} values is mainly caused by L variation through DIBL. This effect is used to decouple V_{th} and L variations, as described in Eq. 4.1. Although the leakage is vital to determine V_{th} and L variations, it should be noted that an extremely low value of gate bias, e.g., $V_{gs} = 0$, is not preferred for the extraction. At $V_{gs} = 0$, other leakage components, such as GIDL, may dominant the current over the subthreshold leakage; the change of the subthreshold swing (S) is also pronounced and needs to be considered. To simplify the extraction procedure, a reasonable value of V_{gs}, e.g., 300 mV, is appropriate to exploit the exponential dependence of the leakage on V_{th}, while avoiding other variation sources.
3. Extract μ variation: Effective mobility is extracted from the linear region of IV, assuming R_{ds} is fixed. Note that R_{ds} and mobility are entangled in the linear

region and thus, it is difficult to decouple them from IV measurement only. For the simplicity of model extraction, the fluctuation of linear IV is attributed to mobility variation. A high gate bias is preferred for a larger level of drive current, which reduces the measurement error. The value of μ is used later in the model to calculate the saturation velocity for I_{on} [25].

4. Iterate Step 2 and 3: The steps above provide the initial values of parameter shift. To minimize the overall error in IV matching, two or more iterations for all three variation sources are further introduced in sub-threshold and linear regions for the values of V_{th} and L, and μ, respectively. Usually this final step only requires two to three iterations.

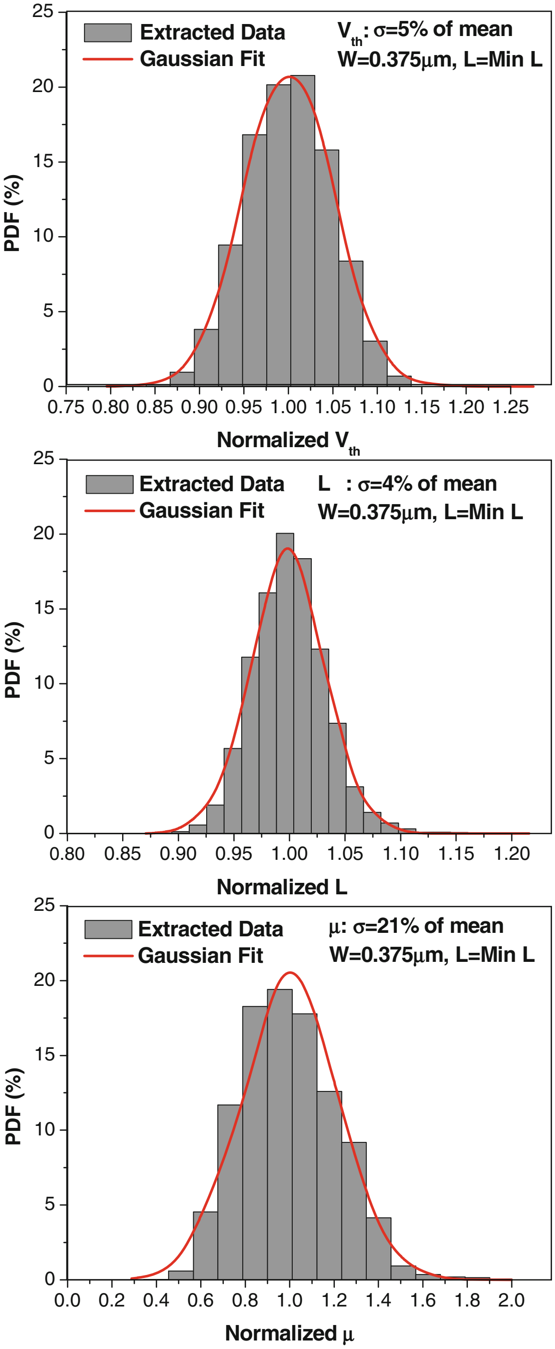
4.1.3 Verification and Statistical Analysis

Based on the new extraction approach, the IV change is translated into parameter variations in the model file. Some important statistics of process variations are further analyzed, such as the dependence on the spatial separation and transistor size, as these characteristics help shed light on robust design strategies.

Figure 4.5 illustrates the distribution of extracted V_{th}, L and μ variations, respectively. The low p-values in all cases indicate very high confidence that they follow the Gaussian distribution. The ratio of standard deviation to mean is 5% for V_{th} variation, 4% for L variation and 21% for μ variation. The wide distribution of μ may be due to the induced stress in this 65 nm technology, while the relatively narrow distribution of L is the benefit from the regular layout pattern of the transistor array (Fig. 4.2). These data illustrate that the variation of mobility has become more pronounced in advanced technology. Furthermore, Figure 4.6 shows the non-correlation between extracted V_{th} and L variations. This behavior proves that these two variations are fully decoupled during this extraction. Similarly, the lack of correlation is also observed between μ and L or V_{th}. The successful decoupling of primary variation sources will further help us understand their statistical properties, as well as the process reasons that lead to the variations.

The incorporation of extracted parameter fluctuations significantly improves the predictability of the nominal model file. For example, I_{on} can be 30% larger or smaller than the nominal I_{on} as shown in Fig. 4.7a, in the absence of the variational parameters. After including the extracted parameter fluctuations into the nominal model file, the IV characteristics can be accurately reproduced for each device. Figure 4.7a shows the strong correlation between measured and modeled I_{on} with an average error of 3.02%. This strong matching in the saturation region is achieved only with three parameter variations that are extracted from the subthreshold and linear regions. I_{on}, together with other three points in Fig. 4.4, captures the most important IV characteristics of a transistor [26]. Excellent model fitting at these points guarantees the accuracy of variation-aware analysis for both DC and AC operations [26]. Figure 4.7b shows the similar correlation between model prediction and the measurement in the subthreshold region. The accuracy in this region is

Fig. 4.5 Extracted parameter variations (Adapted from [25])



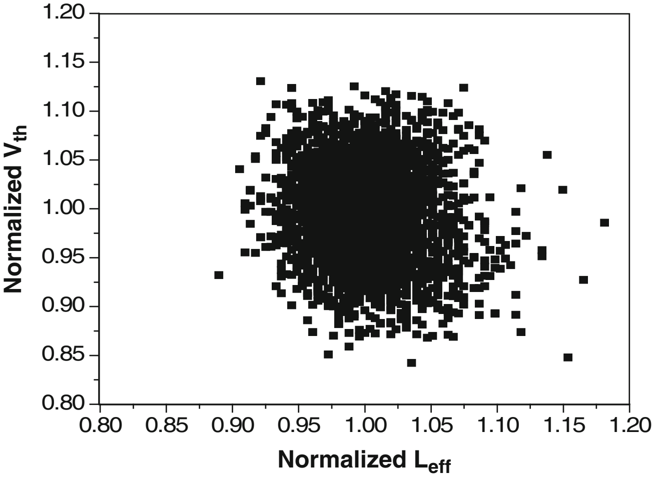


Fig. 4.6 The independence of V_{th} and L variations (Adapted from [25])

slightly less than that of I_{on} prediction because of larger measurement error in I_{off} and the neglect of the variation of the subthreshold swing. Table 4.1 lists the evaluation of the worst case matching error from the test devices at these representative bias conditions. Across different process corners, the model file with extracted parameter variations provides sufficient fidelity to the measurement. The maximum error is smaller than 7% for all operation regions, including the subthreshold, linear, and saturation current.

Besides these particular bias points, more comprehensive evaluation of the variational model is performed. For each test device, model predicted current is compared with measured current for all bias conditions above the threshold. Figure 4.8 shows the error distribution from all the sampling devices. Embedding the variations of L and V_{th}, the matching error in super-threshold region is reduced from 35% to 10% in the worst case. This indicates that L and V_{th} are indeed the dominant components of variations. The consideration of μ variation further reduces the matching error to about 6.5% and achieves more uniform distribution of the error. This observation confirms that the variation in mobility is emerging as a first-order effect and needs to be included into the analysis. Besides the variations in L and V_{th}, it will play an even more important role in the future as the strained silicon technology is widely incorporated into the CMOS structure.

The spatial correlation of variations is further analyzed, which is an important characteristic for statistical analysis. Figure 4.9 reports the variance between two test devices against their physical separation distance. For both V_{th} and L variations, the variance is almost a constant along both column and row directions. Note the dimension of this test array is about 1250 μm × 110 μm [19]. Such a trend indicates that the local spatial correlation is insignificant. The lack of local spatial correlation in V_{th} variation suggests that random parameter fluctuation is the main contributor of local process variation. Our data reveals that

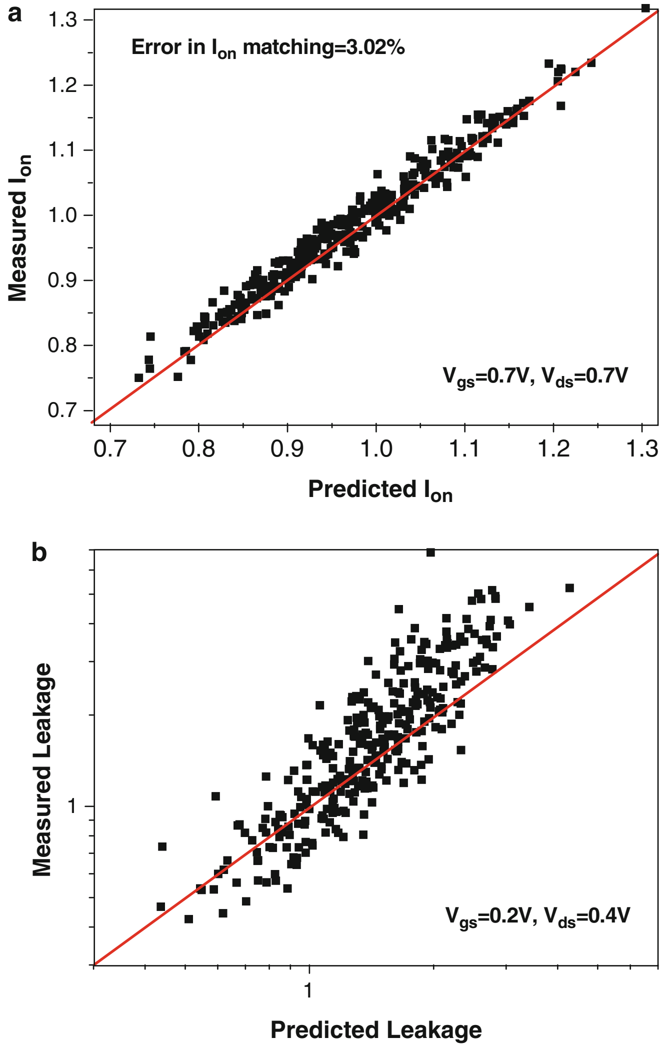


Fig. 4.7 Model predicted currents match well with the measurement. (**a**) I_{on} current matching; (**b**) I_{off} current matching (Adapted from [25])

Table 4.1 The error of IV model at different corners (Adapted from [25])

(V_{gs}, V_{ds}) (V)	Fast	Typical	Slow
(0.7, 0.7)	2.9%	−0.6%	−1.5%
(0.7, 0.4)	−2.5%	−0.6%	−1.6%
(0.4, 0.7)	−2.7%	−0.7%	1.0%
(0.4, 0.4)	−6.5%	−6.7%	1.6%
(0.2, 0.7)	6.2%	−0.3%	−6.5%

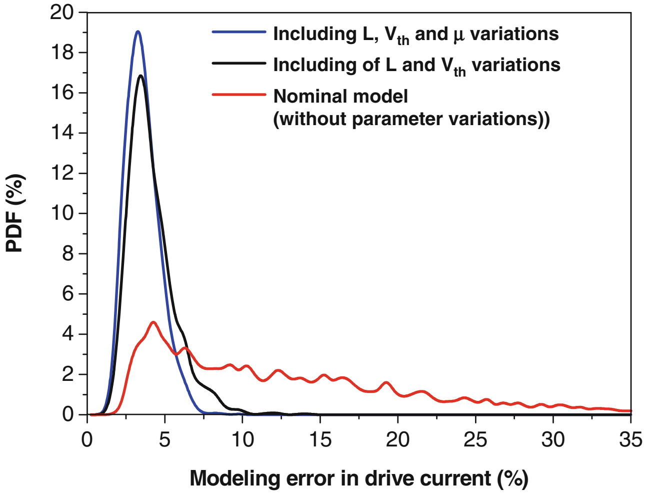


Fig. 4.8 Model error is reduced to $< 6.5\%$ by including extracted variations (Adapted from [25])

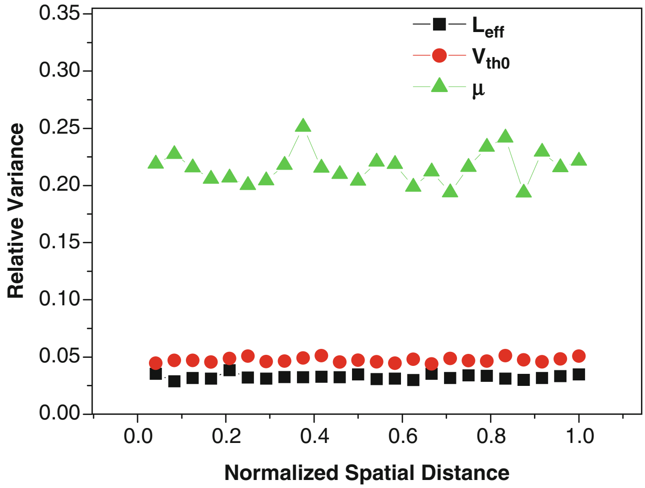


Fig. 4.9 The weak spatial dependence (Adapted from [25])

L variation in this technology also has a negligible level of local spatial correlation. This is different from the strong correlation in L that was published at the 130 nm node [27]. The change of such spatial characteristic may be caused by the regular layout in this test chip. The spatial correlation in effective mobility variation is also negligible, as shown in Fig. 4.9. These facts imply that the impact of process variation can be alleviated in local path timing analysis since propagation delay fluctuations can be averaged out. On the contrary, it indicates challenges in memory cell design since the local mismatch can be dominant.

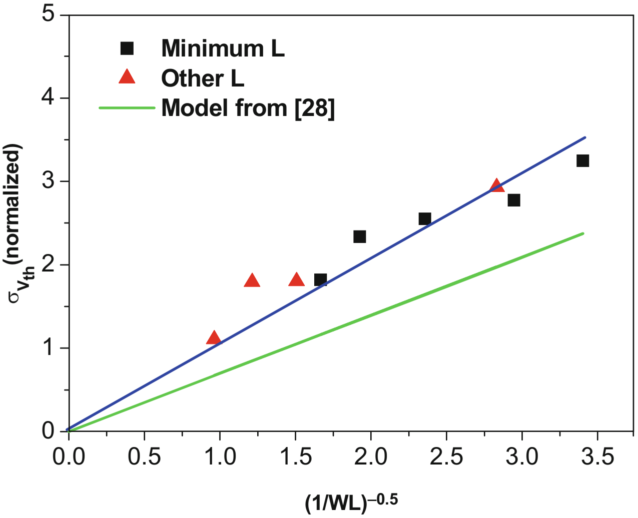


Fig. 4.10 V_{th} variation is linear to $(1/WL)^{-0.5}$ (Adapted from [25])

In addition to the dependence on the spatial distance, the amount of variations is also correlated with the layout and transistor size. V_{th} variation is analyzed for devices with different W and L combinations. Figure 4.10 illustrates that V_{th} variation is inversely proportional to the square root of transistor gate size, across a range of W from 100 to 500 nm. This observation is consistent with other simulation and theoretical results that attribute V_{th} change to random dopant fluctuations. However, the extracted data shows a larger slope than the prediction of dopant fluctuation based model [28, 29]. This implies that additional process factors also have an impact on V_{th} variations besides the RDF effect.

Overall, this extraction method identifies three parameters, L, V_{th} and mobility, as the primary sources due to the uncertainties in lithography, doping and stress. Though this study is based on BSIM4, our approach is general enough for other compact models. The new method will serve as an essential bridge between measured data of process variations and statistical model development [11].

4.2 Predictive Modeling of Threshold Variability

As shown in Sect. 4.1, V_{th} variation in a scaled transistor severely affects device and circuit performance, especially the leakage current. Among multiple variation sources, the effects of RDF and LER represent the primary intrinsic variation sources in the CMOS structure [29, 30], as shown in Fig. 4.11. They stem from atom-level fluctuations, and random in nature. As the device

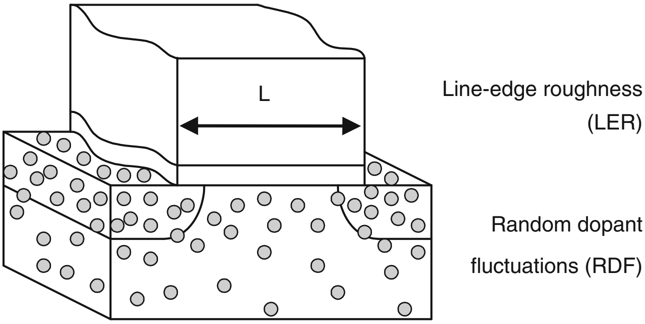


Fig. 4.11 Primary random variations in a nanoscale device

size scales down, their impact rapidly increases (Fig. 4.1), posing one of the ultimate limits on technology scaling [1, 30].

Traditionally, TCAD simulation and compact models are used to quantify these random variations in circuit analysis, but such methods become incorrect as the minimum feature size of a transistor is approaching the characteristic length of these atom-level effects. Instead, 3D Monte-Carlo atomistic simulations become necessary in order to achieve adequate accuracy. For example, [29] and [31] demonstrated the need for and the accuracy of atomistic simulations in the prediction of transistor variations under RDF and LER. However, atomistic simulation is not efficient for statistical circuit analysis, such as the optimization of SRAM cells, since it is too computationally expensive to be incorporated into circuit analysis and statistical optimization. To alleviate this problem, a new methodology is developed, based on the understanding of the underlying physics, particularly the principles of atomistic simulations and short-channel device physics.

RDF is purely a random effect; but LER is induced by both sub-wavelength lithography and the etching process. Lithography usually has a low spatial frequency and causes the so-called non-rectangular gate (NRG) effect [32, 33]. Both RDF and LER change the output current of a transistor by modifying the threshold voltage [34, 35]. In addition to the well-known relationship between V_{th} variation and gate size (W) [34], LER further exacerbates the standard deviation of V_{th} (σ_{Vth}) [30].

4.2.1 Simulation with Gate Slicing Method

To handle the random effects and predict V_{th} variation from a given gate geometry, a non-uniform device is split into slices, which have an appropriate slice width (d) that is larger than the correlation length of RDF in the leakage region, but small enough to track the spatial frequency of LER. Each slice is then modeled as a sub-transistor with correct assignment of narrow-width and short-channel effects, as shown in Fig. 4.12 [30, 33]. Such a representation maps a non-uniform transistor

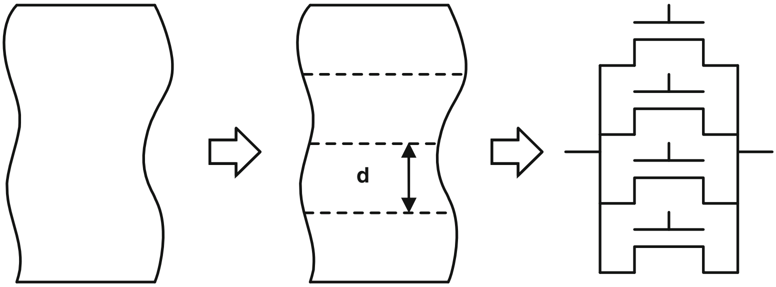


Fig. 4.12 The flow of gate slicing. Each slice has a unique V_{thi} and L_i due to RDF and LER (Adapted from [30])

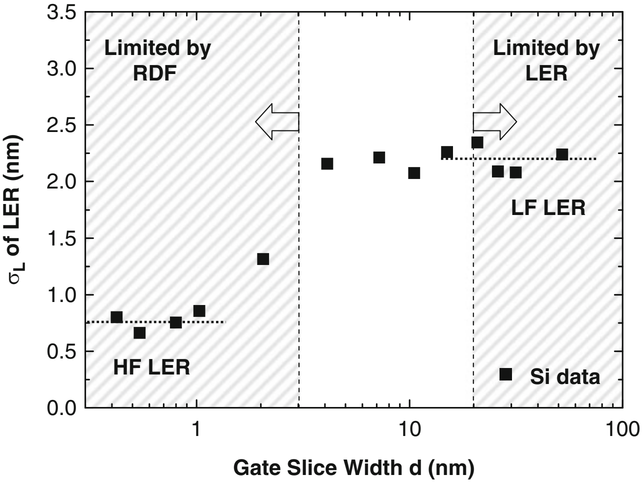


Fig. 4.13 The appropriate selection of slice width under both effects of RDF and LER [32] (Adapted from [30])

into an array of transistors, which can easily be implemented in SPICE. As long as the current in each slice maintains the direction of source-to-drain, i.e., there is no significant distortion of the electrical field along the channel direction, this method is able to provide an accurate prediction on the change of I-V under NRG and LER [33, 36, 37].

On the other hand, there are two fundamental limitations on the slice width, d, especially when the effect of random dopant fluctuations is considered, which requires atomistic simulation to provide sufficient accuracy [29]:

1. Upper bound of d: the spatial frequency of LER. The primary factors to cause LER include sub-wavelength lithography and the etching process. These different factors lead to different spatial frequency and amplitude of the distortion of gate length. Figure 4.13 illustrates the silicon data of gate length change under

LER [32]. The data clearly shows that two regions of LER have distinct spatial frequency: the high-frequency region (HF) that has a characteristic length smaller than 5 nm and a low-frequency one (LF) has a characteristic length larger than 10 nm [32]. While the HF component is usually caused by the property of the photo-resist and the etching condition, the LF component is mainly due to sub-wavelength lithography, i.e., NRG, which can be well predicted from layout by lithography tools [33]. The exact values of their characteristic lengths depend on the fabrication technology. When a non-uniform gate under LER is split, the width of each slice needs to be smaller than the characteristic length in order to track the change in gate lengths with adequate accuracy. For instance, to model a typical LER gate, the slice width should be smaller than 20 nm, as shown in the right side of Fig. 4.13 [32, 33]. This phenomenon defines the upper bound of d during the slicing.

2. Lower bound of d: random dopant fluctuations. Due to the random position of dopants in the channel, V_{th} exhibits an increasing amount of variations with continuous scaling of transistor size [29]. For a relatively long channel device, this behavior is well recorded in the Pelgrom's model [34]. However, as the channel length approaching the length scale of the fluctuation, this type of atom-level randomness can no longer be represented by V_{th} model in the weak-inversion region, which is usually modeled by averaging the potential in the channel. Such an approach hardly tracks the atomistic change [29]. In order to apply the slicing approach with compact V_{th}-based device model, the slice width must be larger than the correlation length of random channel potential near the threshold. A typical value of the length is around several nanometers, depending on the doping concentration [29]. The left side of Fig. 4.13 shows this lower bound of d during the slicing. If d is smaller than the correlation length, then the slicing is not a correct model for the statistical device behavior under RDF, particularly for the weak-inversion current [29].

Considering these two limits, Fig. 4.13 illustrates the appropriate region of d where the slicing approach is applicable. Only when d satisfies both limits that the partition of a single LER transistor becomes meaningful to predict the current in all regions from device physics point of view. Since the L distribution under LER approximately follows the Gaussian function [32, 38], the correlation length of LER (W_c) is selected as the slice width [38]; following the normal distribution, the length for each slice is generated in the experiments [39].

After splitting the original non-uniform transistor into a column of rectangular ones, the gate slicing method assigns different V_{th} values to different slices, and then sum the drive current from each slice to analyze the total output characteristics. In order to perform the linear superposition of currents to understand V_{th} variability, it requires that the drive current should be a linear function of V_{th}. Thus, it is not appropriate to apply the slicing method to the sub-threshold region, since the leakage has an exponential dependence on V_{th}. To solve this problem, V_{th} variation is extracted from the saturation region. Because of the pronounced velocity saturation effect, the output current in the saturation region is a linear function to V_{th} [7].

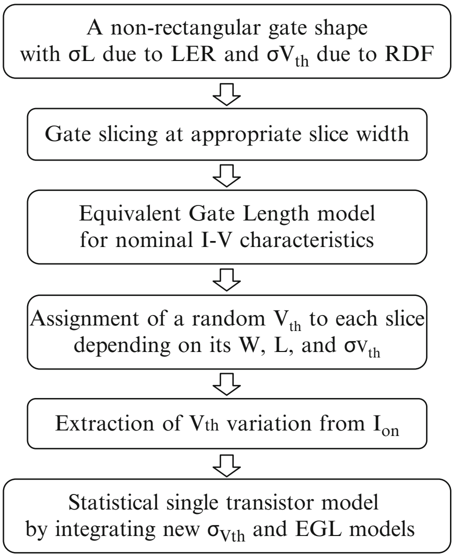

A non-rectangular gate shape with σL due to LER and σV$_{th}$ due to RDF
Gate slicing at appropriate slice width
Equivalent Gate Length model for nominal I-V characteristics
Assignment of a random V$_{th}$ to each slice depending on its W, L, and σv$_{th}$
Extraction of Vth variation from I$_{on}$
Statistical single transistor model by integrating new σ$_{Vth}$ and EGL models

Fig. 4.14 The flow to generate a single device model for statistical analysis of a LER gate

Therefore, it provides a correct mathematical basis to partition the device under RDF and LER, and then linearly superpose the current together to monitor the overall change in V_{th} [30]. Combining this approach with the Equivalent Gate length (EGL) model that describes the nominal device behavior under non-rectangular gate effect [33, 36], the amount of V_{th} variation is predicted under any given transistor characteristics (e.g., non-rectangular gate, reverse narrow-width effect, etc.).

Figure 4.14 summarizes the flow that supports the development of a single device model for statistical analysis under RDF and LER. Given the shape of a LER gate, it is first divided into slices with a suitable width, following the guidance in Fig. 4.13. Then, the model of EGL is produced for the nominal case under NRG [33]. To investigate the interaction of LER and RDF on V_{th} variation, V_{th} is assigned to each slice as a statistical variable. While its mean value is determined by the width and length of the slice (i.e., narrow-width and the DIBL effect [33]), its standard deviation is also dependent on the size of the slice [31, 34, 35]:

$$\sigma_{V_{th}} \propto \frac{1}{\sqrt{WL}} \tag{4.4}$$

The exact value of σ_{Vth} due to RDF is technology dependent [4]. From the summation of I_{on}, the variation of the threshold voltage of the entire transistor is finally obtained under LER and RDF. Since the length of each slice is different under LER, such non-linear relation between σ_{Vth} and L (Eq. 4.4) leads to an increase in V_{th} variation of the entire transistor. The outcome from this procedure is a single device model with EGL and a new σ_{Vth}, which supports efficient statistical performance analysis for any given NRG, LER and RDF.

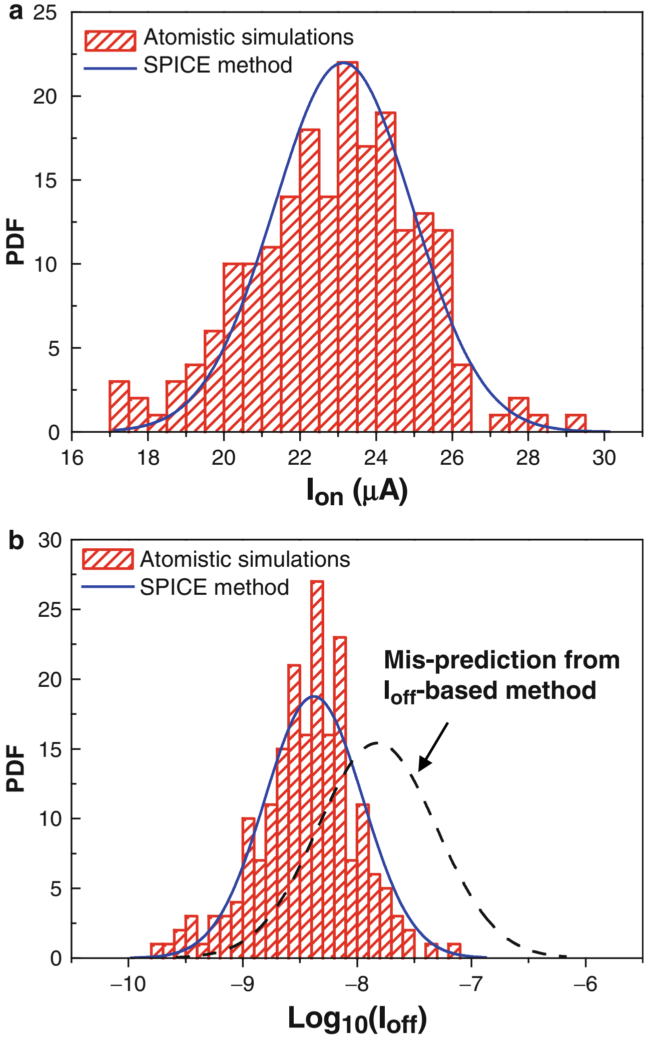
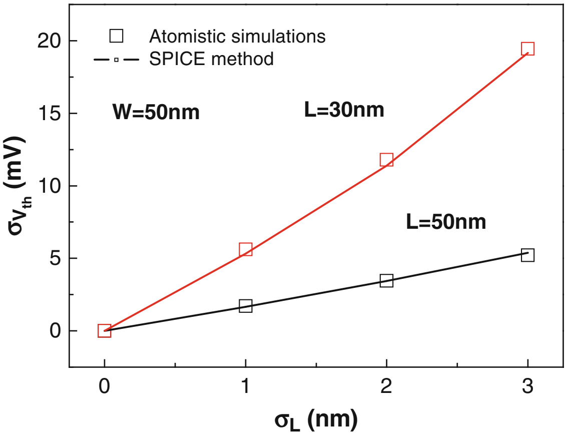


Fig. 4.15 Validation of I_{on} and I_{off} variations under RDF [29]. (**a**) Prediction of I_{on} variation; (**b**) Prediction of I_{off} variation (Adapted from [30])

4.2.2 Validation with Atomistic Simulations

This method is implemented into the SPICE environment to validate its prediction with available 3D Monte-Carlo atomistic simulation results. Figure 4.15 compares the prediction of I_{on} and I_{off} variations under random dopant fluctuations [29]. It indicates that under normally distributed RDF, the variation of I_{on} follows the Gaussian distribution due to its linear dependence on V_{th}. Meanwhile, the variation of I_{off} follows the lognormal distribution because of the exponential dependence of

Fig. 4.16 Validation of σ_{Vth} under LER [29] (Adapted from [30])



$\mathrm{I_{off}}$ on $\mathrm{V_{th}}$. Both mean and sigma of $\mathrm{I_{on}}$ and $\mathrm{I_{off}}$ are well predicted from the $\mathrm{I_{on}}$-based extraction method. Figure 4.15b further shows that if the leakage current is directly summed from every slice to estimate $\mathrm{V_{th}}$ variation, it results in a significant error, as discussed in Sect. 4.2.1.

In addition to the verification of the $\mathrm{I_{on}}$-based method under RDF, Fig. 4.16 evaluates the prediction of σ_{Vth} under different conditions of gate length variations due to LER, assuming a uniform channel doping concentration (i.e., no RDF) [29]. Two devices are studied, with both gate width at 50 nm, and gate length at 30 and 50 nm, respectively. The correlation length of the LER effect ($\mathrm{W_c}$) is 20 nm [29]. For the low-frequency component of LER (NRG), the increase of σ_{L} results in a larger amount of threshold variation, due to the interaction between σ_{Vth} and L, as shown in Eq. 4.4. This interaction is more pronounced when gate length is shorter, in which case the threshold voltage of each slice is more strongly coupled with L through the DIBL effect [33]. Our proposed approach captures this complicated dependence very well, as compared to atomistic simulations.

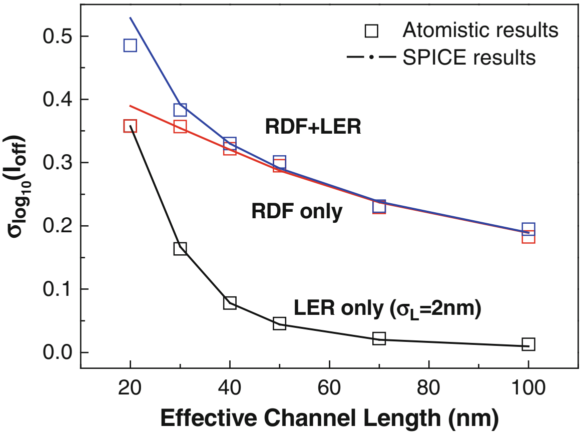
Finally, Fig. 4.17 verifies the prediction of threshold variation in the presence of both RDF and LER. The variation of $\mathrm{V_{th}}$ is evaluated through the distribution of $\mathrm{I_{off}}$, which is very sensitive to $\mathrm{V_{th}}$ change due to its exponential dependence. Three sets of experiments are carried out: LER only with σ_{L} at 2 nm, RDF without LER, and RDF with LER. Again, gate width is 50 nm. Since $\mathrm{V_{th}}$ depends on L through the DIBL effect [24]:

$$V_{th} = V_{th0} - V_{ds} \exp\left(-\frac{L}{l'}\right) \tag{4.5}$$

where $\mathrm{V_{th0}}$ is a function of channel doping, the change of $\mathrm{V_{th}}$ due to L and RDF can be approximated as:

$$\Delta V_{th} = \Delta V_{th0} + V_{ds} \exp\left(-\frac{L}{l'}\right) \cdot \frac{\Delta L}{l'} \tag{4.6}$$

Fig. 4.17 Validation of σ_{Vth} [29] (Adapted from [30])



Therefore, the total variation of V_{th} follows the relationship below, as long as σ_L and RDF are independent and not excessive:

$$\sigma_{total}{}^2 = \sigma_{RDF}{}^2 + \sigma_{LER}{}^2 \tag{4.7}$$

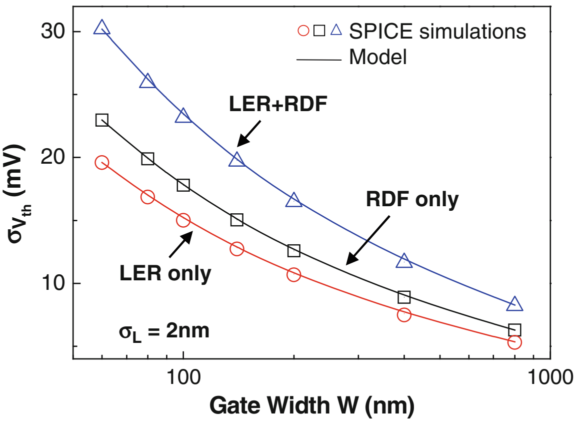
where σ_{RDF}, σ_{NRG}, σ_{total} are V_{th} variations due to RDF only, LER only, and the total amount, respectively. The contributions of LER and RDF are independent to the statistics of V_{th}. The relationship is well verified with atomistic simulations, as shown in Fig. 4.17.

Figure 4.17 indicates that when L is large, RDF is the dominant factor in threshold variation. As gate length decreases, the importance of LER rapidly increases in the calculation of V_{th} variation. Again, the main reason is the strong DIBL effect, which is an exponential function of L, as shown in Eq. 4.5. Overall, our I_{on}-based simulation method provides excellent predictions of V_{th} variation under all situations, as compared to 3D Monte-Carlo atomistic simulations. It significantly enhances the simulation efficiency, with fully compatibility to circuit simulators.

4.2.3 Predictive V_{th} Variability Modeling

For traditional long-channel device, V_{th} mismatch is mainly induced by random effects, such as the dopant fluctuation. This consideration is the basis for the well known Pelgrom's model and other V_{th} variation models, in which σ_{Vth} is inversely proportional to the square root of the transistor size [4]. However, as shown in Fig. 4.17, the impact of LER on V_{th} variation becomes pronounced with further scaling of L, and can no longer be ignored in the calculation of threshold mismatch.

Fig. 4.18 Validation of predictive modeling with SPICE simulation using gate slicing method (Adapted from [30])



These two effects superpose each other in the statistical property of V_{th}, as shown in Fig. 4.17 and Eq. 4.7.

As presented in [31, 35], random dopant fluctuations induce the deviation of V_{th} as a linear function of $(WL)^{-0.5}$. For a larger transistor, the random distribution of dopants is averaged out in the modeling of V_{th}. Akin to this effect, the random distribution of gate length under LER also leads to a linear function of $W^{-0.5}$, since the longer gate width is, the more the length distortion is averaged out. On the other hand, due to the DIBL effect, LER induced V_{th} variation has an exponential dependence on L (Eq. 4.5). Therefore, the following formula is derived based on Eqs. 4.6 and 4.7:

$$\sigma_{total}{}^2 = \frac{C_1}{WL} + \frac{C_2 V_{dd}{}^2}{\exp(2L/l')} \cdot \frac{W_c}{W} \cdot \sigma_L{}^2 \tag{4.8}$$

where W_c is the correlation length of LER, and C_1, C_2 and l' are technology dependent coefficients [30]. For example, for 45 nm technology, C_1 is around $10^{-18}V^2{\cdot}m^2$, C_2 is around $1.5 \times 10^{16}m^{-2}$, and l' is around 10 nm. The first term describes conventional Pelgrom's model under RDF. The second term is designated to the variation due to LER. The exponential dependence on L is demonstrated in Fig. 4.17. Figure 4.18 verifies the dependence of threshold variation on gate width. Our model accurately captures the superposition of these two statistical components, as well as the inverse square root dependence on W. Traditional model only considers the RDF effect and thus, significantly underestimates the total amount of V_{th} variation, as shown in Fig. 4.18. Note that due to the exponential dependence on L of the second term in Eq. 4.8, the impact of LER is marginal at long gate length. Yet the second term rapidly affects threshold variation for a device with short gate length and width. For instance, at W = 50 nm, it has a comparable influence as that of RDF. Therefore, its role cannot be neglected, particularly when the minimum sized transistors are used in the design.

Table 4.2 Projection of threshold variation in bulk CMOS devices

LER parameters		Total σ_{Vth} (mV)			
W_c (nm)	σ_L (nm)	65 nm ($V_{ds} = 1.1V$)	45 nm ($V_{ds} = 1V$)	32 nm ($V_{ds} = 0.9V$)	22 nm ($V_{ds} = 0.8V$)
5	0	19.9	23.8	28.1	45.8
	0.5	20.0	24.1	28.7	47.0
	1	20.4	24.9	31.2	53.3
10	0	19.9	23.8	28.1	45.8
	0.5	20.1	24.3	29.3	48.1
	1	20.8	25.9	34.0	59.9

The proposed compact model offers a scalable tool to explore threshold variation under LER and RDF effects. As shown in Figs. 4.17 and 4.18, this approach has the right sensitivity to transistor definitions. Furthermore, these models are extrapolated to future technology generations [7], with the goal to gain early stage insights to robust design under increased variations.

Continuous scaling exacerbates both RDF and LER effects [1]. With the scaling of transistor size, the total number of dopants in the channel significant reduces. Consequently, the amount of random RDF effect becomes more significant (Fig. 4.1). For line-edge roughness, the improvement is limited by the etching process, rather than the lithography process [40–42]. The emerging etching technology may reduce 3σ of LER amplitude down to ~2 nm [43–46] and the correlation length around 10 ~ 20 nm [45, 46]. Yet such improvements still lag behind the scaling rate of nominal channel length. Therefore, the sensitivity of device performance to LER dramatically increases at recent technology nodes. Finally, the situation of NRG is not optimistic due to the difficulty in sub-wavelength photolithography. The distortion in gate length is expected to increase [36], even though lithography recipes and layout techniques, such as regular layout fabrics, may help improve the situation [36, 47].

Using the new method, the amount of threshold variation is projected, under possible scenarios of RDF and LER. The nominal model file is adopted from PTM [7]. In this projection, new technology advances, such as high-*k* and metal gate, are not considered. Other potential variation sources, such as RDF induced mobility variation, have not been included. Therefore, this projection represents the lower bound of threshold variation in future devices. Table 4.2 summarizes the results for various LER parameters of W_c and σ_L. Even under the same amount of LER, the variation of the threshold voltage keeps increasing due to the aggressive scaling of the feature size and the exacerbation of short-channel effects. As the trend goes, future design will suffer a dramatic amount of intrinsic variations. While the improvement of process technology will continue, its effectiveness may be limited by fundamental physics in the future.

Besides intrinsic variations, additional variations are induced by the manufacturing process. Depending on the layout non-uniformity and the specific fabrication technology, these variations may have a spatial correlation length ranging from 1 nm (e.g., lithography effect), to 100 nm (e.g., stress effect), or even millimeter (e.g., rapid thermal annealing). Since they are usually mixed together during data preparation, the modeling challenge is to understand primary components, correlate them with process and design parameters, decompose them from the test data, and embed them into the model file [48]. Predictive modeling of these manufacturing variations requires a coherent cooperation with silicon characterization and parameter extraction.

Increasingly, the consequences of device variability ripple throughout process development, device characterization, physical simulation, compact modeling, and design strategy. At the device and circuit levels, understanding and successfully modeling the leading variation mechanisms is vitally important, not only to current robust design practice, but also to the prediction and management of variation levels for future IC technology.

References

1. International Technology Roadmap of Semiconductors, 2007. (available at http://www.itrs.net).
2. B. H. Calhoun, Y. Cao, X. Li, K. Mai, L. T. Pileggi, R. A. Rutenbar, and K. L. Shepard, “Digital circuit design challenges and opportunities in the era of nanoscale CMOS,” *Proceedings of the IEEE*, vol. 96, no. 2, pp. 343–365, February 2008.
3. S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, V. De, “Parameter variations and impact on circuits and microarchitecture,” *ACM/IEEE Design Automation Conference*, pp. 338–342, June 2003.
4. K. Bernstein et al., “High-performance CMOS variability in the 65-nm regime and beyond,” *IBM J. Res. & Dev.*, vol. 50, no. 4/5, pp. 433–449, Jul./Sep., 2006.
5. R. W. Keyes, “Physical limits in digital electronics,” *Proc. IEEE*, vol. 63, pp. 740–766, 1975.
6. B. Hoeneisen and C. A. Mead, “Fundamental limitations in microelectronics—I. MOS technology,” *Solid-State Electronics*, vol. 15, no. 7, p. 819, 1972.
7. W. Zhao, Y. Cao, “New generation of predictive technology model for sub-45nm early design exploration,” *IEEE Transactions on Electron Devices*, vol. 53, no. 11, pp. 2816–2823, Nov. 2006. (Available at http://ptm.asu.edu).
8. T. Hagivaga, K. Yamaguchi, and S. Asai, “Threshold voltage variation in very small MOS transistors due to local dopant fluctuations,” *Symp. VLSI Technology*, pp. 46–47, 1982.
9. S. R. Nassif, N. Hakim, D. Boning, “The care and feeding of your statistical static timer,” *ACM/IEEE International Conference on Computer-Aided Design*, pp. 138–139, Nov. 2004.
10. S. K. Springer, et al., “Modeling of variation in submicrometer CMOS ULSI technologies,” *IEEE Transactions on Electron Devices*, vol. 53, no. 9, pp. 2168–2006, Sept. 2006.
11. S. R. Nassif, “Model to hardware matching for nanometer scale technologies,” *IEEE International Symposium on Low-Power Electronics and Design*, pp. 203–206, 2006.
12. J. C. Chen, et al., “E-T based statistical modeling and compact statistical circuit simulation methodologies,” *International Electron Devices Meeting*, pp. 635–638, 1996.
13. P. Cox, et al., “Statistical modeling for efficient parametric yield estimation of MOS VLSI circuits,” *IEEE Transactions on Electron Devices*, vol. ED-32, no. 2, pp. 471–478, Feb. 1985.
14. C. C. McAndrew, “Statistical modeling for circuit simulation”, *International Symposium on Quality Electronics Design*, pp. 357–362, 2003.
15. B. Cheng, D. Dideban, N. Moezi, C. Millar, G. Roy, X. Wang, S. Roy, and A. Asenov, “Statistical variability compact modeling strategies for BSIM4 and PSP,” *IEEE Design & Test of Computers*, vol. 27, no. 2, pp. 26–35, March/April 2010.

16. D. Kim, et al, "CMOS mixed-signal circuit process variation sensitivity characterization for yield improvement," *Custom Integrated Circuits Conference*, pp. 365–368, 2006.
17. L. Pang, et al, "Impact of layout on 90nm CMOS process parameter fluctuations," *VLSI Circuits Symposium*, pp. 69–70, 2006.
18. E. Leobandung, et al., "High performance 65nm SOI technology with dual stress liner and low capacitance SRAM cell," *VLSI Tech. Symposium*, pp. 126–127, 2005.
19. K. Agarwal, et al., "A test structure for characterizing local device mismatches," *VLSI Circuits Symposium*, pp. 67–68, 2006.
20. C. McAndrew, et al., "Device correlation: modeling using uncorrelated parameters, characterization using ratios and differences," *Workshop on Compact Modeling*, 2006.
21. S. Nassif, "Modeling and analysis of manufacturing variations," *Custom Integrated Circuits Conference*, pp. 223–228, 2001.
22. G. Scott, et al., "NMOS drive current reduction caused by transistor layout and trench isolation induced stress," *International Electron Devices Meeting*, pp. 827–830, 1999.
23. N. Drego, A. Chandrakasan, and D. Boning, "A test-structure to efficiently study threshold-voltage variation in large MOSFET arrays," *International Symposium on Quality Electronic Design*, pp. 281–286, 2007.
24. BSIM4 Manual, University of California, Berkeley, 2006.
25. W. Zhao, F. Liu, K. Agarwal, D. Acharyya, S. R. Nassif, K. Nowka, Y. Cao, "Rigorous extraction of process variations for 65nm CMOS design," *IEEE Transactions on Semiconductor Manufacturing*, vol. 22, no. 1, pp. 196–203, February 2009.
26. M. Chen, et al., "Fast statistical circuit analysis with finite-point transistor model," *Design, Automation & Test in Europe*, pp. 1391–1396, 2007.
27. P. Friedberg, et al., "Modeling within-die spatial correlation effects for process-design co-optimization," *International Symposium on Quality Electronic Design*, pp. 516–521, 2005.
28. T. Mizuno, J. Okumtura and A. Toriumi, "Experimental study of threshold voltage fluctuation due to statistical variation of channel dopant number in MOSFET's," *IEEE Transactions on Electron Devices*, vol. 41, no.11, pp. 2216–2221, 1994.
29. A. Asenov, A. R. Brown, J. H. Davies, S. Kaya, and G. Slavcheva, "Simulation of intrinsic parameter fluctuations in decananometer and nanometer-scale MOSFETs," *IEEE Transactions on Electron Devices*, vol. 50, pp. 1873, 2003.
30. Y. Ye, F. Liu, S. Nassif, Y. Cao, "Statistical modeling and simulation of threshold variation under dopant fluctuations and line-edge roughness," *Design Automation Conference*, pp. 900–905, 2008.
31. D. J. Frank, Y. Taur, M. Ieong, and H.-S. P. Wong, "Monte Carlo modeling of threshold variation due to dopant fluctuations," *Symp. VLSI Circuits*, pp. 171–172, 1999.
32. J. A. Croon, et al., "Line edge roughness: Characterization, modeling and impact on device behavior," *IEDM*, pp. 307–310, 2002.
33. R. Singhal, A. Balijepalli, A. Subramaniam, F. Liu, S. Nassif, Y. Cao, "Modeling and analysis of non-rectangular gate for post-lithography circuit simulation," *Design Automation Conference*, pp. 823–828, 2007.
34. J. J. M. Pelgrom, A. C. J. Duinmaijer, A. P. G. Welbers, "Matchign properties of MOS transistors," *IEEE Journal of Solid-State Circuits*, vol. 24, no. 5, pp. 1433–1440, Oct. 1989.
35. K. Takeuchi, "Channel size dependence of dopant-induced threshold voltage fluctuation," *Symp. VLSI Technology*, pp. 72–73, 1998.
36. A. Subramaniam, R. Singal, Y. Cao, "Design rule optimization of regular layout for leakage reduction in nanoscale design," *Asia and South Pacific Design Automation Conference*, pp. 474–479, 2008.
37. P. Gupta, A. Kahng, Y. Kim, S. Shah, D. Sylvester, "Modeling of non-uniform device geometries for post-lithography circuit analysis", *SPIE*, vol. 6156, 2006.
38. J. Wu, J. Chen, K. Liu, "Transistor width dependence of LER degradation to CMOS device characteristics," *SISPAD*, pp. 95–98, 2002.

39. S.-D. Kim, H. Wada, J. C. S. Woo, "TCAD-based statistical analysis and modeling of gate line-edge roughness effect on nanoscale MOS transistor performance and scaling." *IEEE Transactions on Semiconductor Manufacturing*, vol. 17, no. 2, pp. 192–200, May, 2004.
40. T. Yamaguchi, H. Namatsu, M. Nagase, K. Kurihara and Y. Kawai, "Line-edge roughness characterized by polymer aggregates in photoresists," *Proc. SPIE*, vol. 3678, no. 1, pp. 617–624, 1999.
41. D. L. Goldfarb, et al., "Effect of thin-film imaging on line edge roughness transfer to underlayers during etch processes," *Journal of Vacuum Science & Technology B*, vol. 22, no.2, pp. 647–653, 2004.
42. H. Namatsu, M. Nagase, T. Yamaguchi, K. Yamazaki, K. Kurihara, "Influence of edge roughness in resist patterns on etched patterns," *Journal of Vacuum Science & Technology B*, vol. 16, no. 6, pp.3315–3321, Nov. 1998.
43. C. Bencher, Y. Chen, H. Dai, W. Montgomery, and L. Huli, "22nm halfpitch patterning by CVD spacer self alignment double patterning (SADP)," *Proc. of SPIE*, vol. 6924, pp. 69244E.1–69244E.7, 2008.
44. S. Sardo, et al., "Line edge roughness (LER) reduction strategy for SOI waveguides fabrication," *Microelectronic Engineering*, vol. 85, no. 5-6, pp. 1210–1213. May–June 2008.
45. P. P. Naulleau, G. Gallatin, "Spatial scaling metrics of mask-induced line-edge roughness," *Journal of Vacuum Science & Technology B*, vol. 26, no. 6, pp. 1903–1910, Nov. 2008.
46. M. Chandhok, et al., "Improvement in linewidth roughness by postprocessing," *Journal of Vacuum Science & Technology B*, vol. 26, no. 6, pp. 2265–2270, Nov. 2008.
47. T. Jhaveri, et al., "Maximization of layout printabili-ty/manufacturability by extreme layout regularity," *J. Micro/Nanolitho., MEMS and MOEMS*, vol. 6, no. 3, 031011, Jul.–Sep. 2007.
48. Y. Cao, F. Liu, "Compact variability modeling in scaled CMOS design," *IEEE Design & Test of Computers*, Special Issue on Compact Variability Modeling in Scaled CMOS Design, vol. 27, no. 2, pp. 6–7, March/April 2010.

Chapter 5
Modeling of Temporal Reliability Degradation

Transistor performance not only depends on static process variations, but also changes over the period of dynamic operation because of the effect of temporal reliability degradation (i.e., aging effect) [1–5]. As CMOS technology is scaling to the 10 nm regime, equivalent oxide thickness will be as thin as 5 Å [1]. Such an aggressive pace inevitably leads to multiple reliability concerns, such as negative-bias-temperature-instability (NBTI), channel-hot-carrier (CHC), and time-dependent-dielectric-breakdown (TDDB). In particular, there has been a recent increase in interest on the reliability impact of PMOS NBTI, and NMOS positive-bias temperature instability (PBTI), which is similar to NBTI and becomes pronounced after high-*k* gate dielectric is adopted [1, 6–9].

NBTI occurs under negative gate voltage (e.g., $V_{gs} = -V_{DD}$ for a PMOS device) and is measured as an increase in the magnitude of threshold voltage [4, 5]. It mostly affects the PMOS transistor and degrades the device driving current, circuit speed, noise margin, the matching property, as well as device and circuit lifetime. Indeed, as gate oxide gets thinner than 4 nm, the threshold voltage change caused by NBTI for the PMOS transistor has become the dominant factor to limit the life time, which is much shorter than that defined by traditional hot-carrier induced degradation of the NMOS transistor [6, 10]. Furthermore, different from CHC that occurs only during dynamic switching, NBTI is induced by static stress on the oxide even without current flow. Consequently, the situation of the NBTI degradation is exacerbated in the nanoscale design, as advanced digital systems tend to have longer standby time for lower power consumption. As the NBTI effect becomes more severe with continuous scaling, it is critical to understand, simulate, and minimize the impact of NBTI in the early design stage, in order to ensure the reliable operation of circuits and systems for a desired period.

To date, research works on NBTI have been active only within the communities of device and reliability physics. Partly due to its complexity and emerging status, design knowledge and CAD tools for managing temporal degradation are not

Y. Cao, *Predictive Technology Model for Robust Nanoelectronic Design*,
Integrated Circuits and Systems, DOI 10.1007/978-1-4614-0445-3_5,

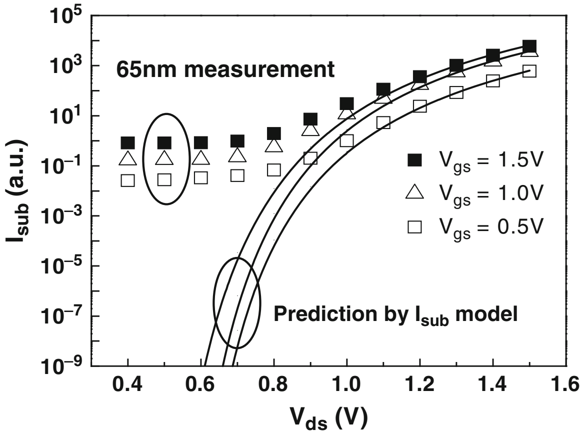


Fig. 5.1 Deviation of traditional I_{sub}-based CHC model (Adapted from [10])

widely available [11]. Leading industrial companies do develop their own models and tools to handle this effect. These tools, however, are usually proprietary and empirical to a specific technology. In this case, a more general and SPICE compatible model that can accurately predict the degradation would be very useful. This predictive model will further serve as a cornerstone to circuit design and optimization in the presence of temporal reliability degradation.

Such a predictive NBTI model is presented in this chapter. It is based on the physical understanding and published aging data for both DC and AC operations. In addition, a new modeling framework is proposed to integrate both NBTI and CHC effects, as CHC is still important for analog and mixed-signal design. Traditionally, CHC is characterized by the substrate current (I_{sub}) that is induced by hot carriers [12]. However, in the nanoscale regime, the I_{sub}-based method is not effective, since the amount of I_{sub} is dominated by other leakage components, such as gate leakage, junction current, and ate-induced drain leakage. Figure 5.1 shows the measured I_{sub} in a 65 nm technology [10]. It exhibits a significant deviation from the traditional hot carrier model [12], particularly when the drain voltage is smaller than 1 V. This phenomenon suggests that continuous usage of I_{sub} would overestimate the degradation and result in an overly pessimistic design.

This chapter unifies the understanding of both NBTI and CHC, based on the general reaction-diffusion mechanism (R-D). Instead of resorting to I_{sub}, the degradation is directly modeled as the shift of key transistor parameters, including the threshold voltage (V_{th}) and mobility (μ). The dependence of NBTI and CHC on process (e.g., L, V_{th}, T_{ox}) and design parameters (e.g., V_{DD}, duty cycle, etc.) are captured in this framework. Representative model coefficients are extracted from silicon data across a wide range of process and stress conditions. Comparisons between published data and model predictions verify the generality and scalability of this approach.

5.1 Static Aging Models

The instability of transistor parameters, e.g., V_{th}, saturation current (I_{on}), etc., under negative bias and high temperature has been known since the 1970s [13]. It is the recent aggressive scaling of CMOS technology that makes NBTI as one of the foremost reliability concerns in nanoscale design [1, 6, 10]. Although there may not be a single physical mechanism that is comprehensive enough to explain all the behaviors, it is arguably believed that NBTI is caused by broken Si-H bonds, which are induced by positive holes from the channel. Then H, in a neutral molecular form (H_2), diffuses away from the interface; positive interface traps (N_{it}) (i.e., from Si^+) are left, which cause the increase of V_{th} [4, 5, 7, 14, 15]:

$$\Delta V_{th} = qN_{it}/C_{ox}, \text{ where } C_{ox} = \varepsilon_{ox}/T_{ox} \tag{5.1}$$

Due to the difference in the flat band voltage, the NMOS transistor has a lower level of holes under the same bias condition and thus, suffers from a smaller amount of PBTI degradation.

For a PMOS transistor, there are two phases of NBTI, depending on its bias condition. These two phases are illustrated in Fig. 5.2, assuming the substrate is biased at V_{DD}. In Phase I, when $V_g = 0$ (i.e., $V_{gs} = -V_{DD}$), positive interface traps are accumulating over the stress time with H diffusing towards the gate. This phase is usually referred as “stress” or “static NBTI”. In Phase II, when $V_g = V_{DD}$ (i.e., $V_{gs} = 0$), holes are not present in the channel and thus, no new interface traps are generated; instead, H diffuses back and anneals the broken Si-H. As a result, the number of interface traps is reduced during this stage and the NBTI degradation is recovered. Phase II is usually referred as “recovery” and has a significant impact on the estimation of NBTI during the dynamic switching. For CHC in a NMOS transistor, its impact cannot be recovered, i.e., only Phase I exists.

There are two critical steps that happen in the static process of NBTI (Phase I) and CHC [16]:

1. Reaction: This is where some Si-H (for NBTI) or Si-O (for CHC) bonds at the substrate/gate dielectric interface are broken under the electrical stress [14, 17].

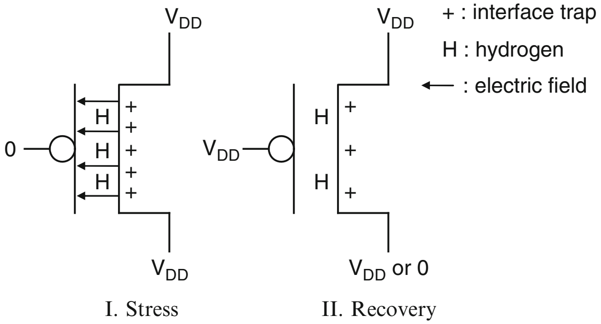


Fig. 5.2 Two phases of NBTI ($V_b = V_{DD}$ for a PMOS device)

The species that trigger such reactions can be positive holes in NBTI or hot electrons in CHC [18]. Consequently, interface charges are induced, which cause the increase of V_{th}. Given the initial concentration of the Si-H bonds, i.e., N_o, and the concentration of the inversion carriers, i.e., P, the generation rate of the interface traps, i.e., N_{it}, is given by [14]:

$$\frac{dN_{it}}{dt} = k_F(N_o - N_{it})P - k_R N_H N_{it} \tag{5.2}$$

where k_F and k_R are the reaction rates of the forward and reverse reactions. Akin to other reactions, the generation rate is an exponential function of the electrical field and temperature. It is also proportional to the density of reaction species, namely holes or hot electrons [14, 17].

2. Diffusion: This is where reaction generated species dif-fuse away from the interface toward the gate, driven by the gradient of the density. While NBTI happens uniformly in the channel, CHC primarily affects the drain end [12]. This process influences the balance of the reaction and is governed by

$$\frac{dN_H}{dt} = D_H \frac{d^2 N_H}{dx^2} \tag{5.3}$$

where D_H is the diffusion constant. The solution of Eq. 5.3 exhibits a power-law dependence on the stress time [14, 17]. The exact value of the power law index indicates the type of diffusion species [17].

The closed-form solutions to the above equations provide such dependence:

$$N_{it} = \sqrt{K^2 \cdot t^{2n} + N_{it0}{}^2} \tag{5.4}$$

where N_{it0} is N_{it} at the starting point; n is about 0.16 for NBTI, which is the signature of neutral H_2 diffusion [10], and n is 0.45 for CHC. Considering the reaction of breaking Si-H or Si-O, the generation rate, K, is linearly proportional to the hole or electron density and exponentially dependent on temperature (T) and the electric field (E_{ox}) [4, 14, 15]. Therefore, for both NBTI and CHC:

$$K \propto \sqrt{C_{ox}\left(V_{gs} - V_{th}\right)} \cdot \exp(E_{ox}/E_0) \cdot \exp(-E_a/kT) \tag{5.5}$$

where $E_{ox} = V_{gs}/T_{ox}$ and k is the Boltzmann constant. K of CHC further depends on the drain current, especially in the saturation region [10].

Using this model and Eq. 5.1, Figs. 5.3 and 5.4 verify the change of V_{th} under static NBTI for 90 and 65 nm technologies at various process and stress conditions [10, 19]. The fitted values of coefficients do converge from the verification [10]. This convergence confirms that E_0 and E_a are technology-independent characteristics of the reaction.

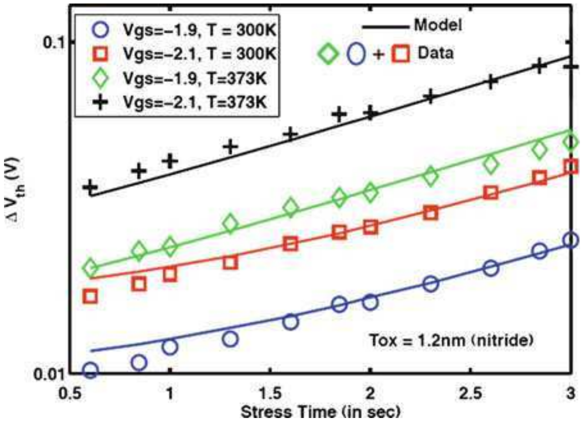


Fig. 5.3 V_{th} degradation under static NBTI for different T and V_{gs} for a 90 nm technology [19]

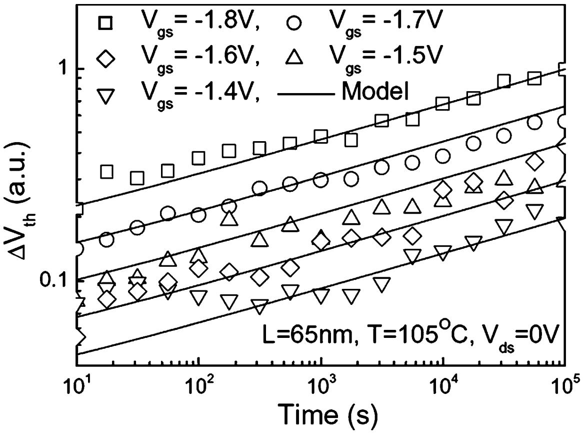


Fig. 5.4 Static NBTI for a 65 nm technology (Adapted from [10])

In addition to the shift of V_{th}, the increase in interface charges further results in the degradation of carrier mobility, due to stronger Coulomb scattering [12, 20, 21]. The mobility degradation as a function of interface trap density can be expressed as:

$$\mu_c = \left(a + \frac{V_{gs} + V_{th}}{V_{th} + b\Delta V_{th}}\right)^{\alpha} \tag{5.6}$$

where μ_c is the Coulomb scattering component in the effective mobility (μ_{eff}) calculation [22]:

$$1/\mu_{eff} = 1/\mu_c + 1/\mu_{surface\ roughness} + 1/\mu_{phonon} \tag{5.7}$$

and ΔV_{th} is the V_{th} change due to aging effects. The degradation mainly happens at low V_{gs}. Figure 5.5 verifies this model with 65 nm data.

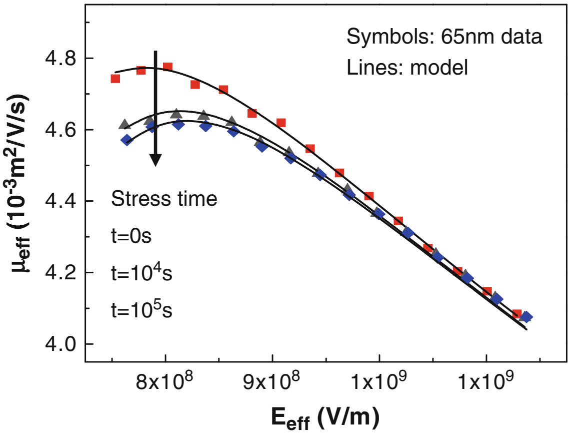


Fig. 5.5 The degradation of mobility under static NBTI stress, where E_{eff} is the effective electric field in the inversion layer [22]

5.2 Dynamic NBTI Models

5.2.1 Cycle-to-Cycle Degradation Model

In a realistic circuit operation, the gate voltage switches between 0 and V_{DD}. For a PMOS transistor, the condition of $V_g = V_{DD}$ removes NBTI stress and anneals interface traps. Such a process solely relies on the diffusion of neutral H_2 and thus, has no field dependence [23]. Assuming the recovery happens at $t = t_0$ with $N_{it} = N_{it0}$, the change of N_{it} can then be modeled as [4]:

$$N_{it} = N_{it0} \cdot \left[1 - \sqrt{\eta(t - t_0)/t}\right] \tag{5.8}$$

Figure 5.6 evaluates this model by verifying the dynamic behavior with data from a 90 nm technology [24]. When the next cycle of stress comes back, the reaction-diffusion process continues as described by Eq. 5.4. V_{th} change during continuous stress is also verified in Fig. 5.6.

In reality, the stress and recovery processes are more complicated. They may involve oxide traps and other charged residues [23, 25–27]. These non-H based mechanisms may have faster response time than the diffusion process. Without losing generality, their impact can be included as a constant of δ:

$$\text{Stress}: \quad N_{it} = \sqrt{K^2 \cdot (t - t_0)^{2n} + {N_{it0}}^2} + \delta \tag{5.9}$$

$$\text{Recovery}: \quad N_{it} = (N_{it0} - \delta) \cdot \left[1 - \sqrt{\eta(t - t_0)/t}\right] \tag{5.10}$$

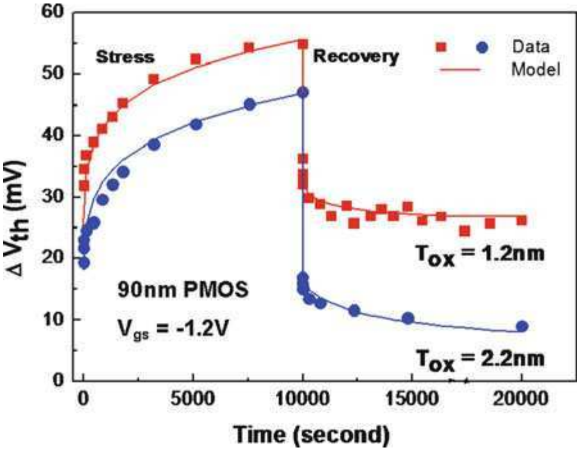


Fig. 5.6 Verification of dynamic NBTI with [24]

The change of V_{th} (ΔV_{th}) is then proportional to N_{it} (Eq. 5.1). A more detailed model of the fast recovery behavior due to δ is presented in [10].

5.2.2 *Long-Term Degradation Model*

In order to predict the long-term threshold voltage degradation due to NBTI at a time t, the stress and recovery cycles given in Eqs. 5.9 and 5.10 can be simulated for $m = t/T_{clk}$ cycles to obtain the long term degradation, where T_{clk} is the clock period. However, for high performance circuits, m can be very large even for $t = 1$ month. Thus, it becomes impractical to perform cycle-to-cycle simulation in order to predict ΔV_{th}. Based on Eqs. 5.9 and 5.10, it is feasible to obtain a closed-form for the upper bound on the long term ΔV_{th} as a function of the duty cycle α, T_{clk} and t [17]:

$$\Delta V_{th} = \left(\frac{\sqrt{K^2 \alpha T_{clk}}}{1 - \beta^{1/2n}}\right)^{2n} \tag{5.11}$$

where β is a function of oxide thickness, T_{clk}, α and t [17].

There is an interesting behavior predicted by these models: the long-term ΔV_{th} is independent on switching frequency as long as the frequency is larger than ~100 Hz. This behavior is confirmed by experimental data, as shown in Fig. 5.7 [17, 25, 28]. With the recovery in dynamic switching, ΔV_{th} due to NBTI may be reduced by two to three times as compared to that purely under static NBTI stress [3, 17].

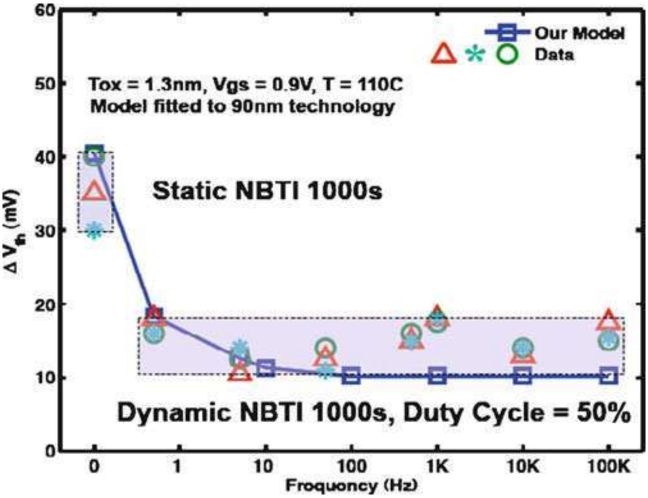


Fig. 5.7 Frequency dependency of the long-term degradation obtained using our model with silicon data [25]

5.3 Model Implementation and Prediction

As the above models are well verified over a wide range of process and design conditions, they provide a solid basis for further simulation studies and tool development. For the direct calculation of V_{th} change under NBTI and CHC, the entire suite of formulas and representative model parameters are summarized in [10]. These models are scalable with key process and design parameters, such as T_{ox}, V_{gs}, V_{th}, V_{ds}, T, L, and time. Even though the gate length (L) is not explicitly expressed in the N_{it} model, L is still able to affect the degradation through its impact on V_{th} (i.e., the DIBL effect).

5.3.1 Sub-circuit for SPICE Simulation

The new model is compatible with standard MOSFET model, such as BSIM and the surface-potential-based PSP. It can be conveniently customized and implemented into the circuit simulation environment to analyze and predict the temporal degradation of circuit performance. Figure 5.8 presents the sub-circuit module for NBTI in the PMOS transistor. The increase in V_{th} was modeled as a voltage-controlled voltage source (VCVS: Egnbti). The VCVS leads to a decrease in V_{gs}, which emulates the V_{th} shift induced by NBTI, and subsequently reduces the drain current. The instantaneous increase in V_{th} is equal to the voltage difference between the VCVS nodes [Egnbti $= \Delta V_{th}(t)$].

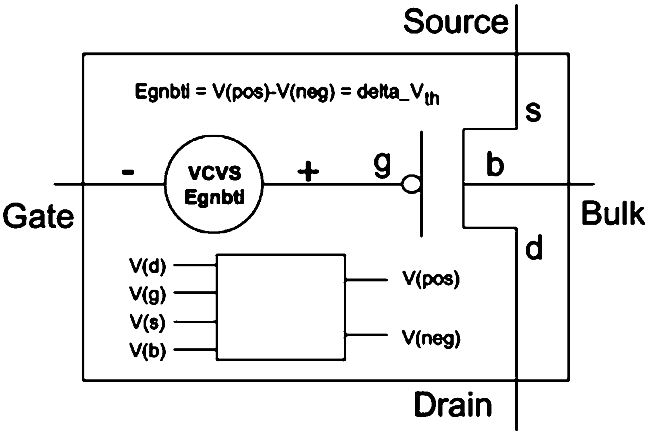


Fig. 5.8 The sub-circuit to simulate aging effects (Adapted from [10])

In complex circuits with a large number of PMOS transistors, a sub-circuit model can be used to accurately estimate the temporal degradation. The V_{th} degradation in a particular PMOS transistor depends on the circuit topology and the bias conditions during the operation [3]. Similarly, the degradation of V_{th} caused by CHC or PBTI is simulated by using the same sub-circuit module for NMOS.

5.3.2 *Device and Circuit Performance Degradation*

Using this implementation method, the impact of temporal device degradation on circuit performance can be conveniently evaluated. Figure 5.9 shows the frequency change of a 65 nm ring oscillator (RO) with 11 stages of inverters. Over the period of 10^5 s, the switching frequency degrades more than 1%. The prediction by device-level aging model well matches the RO measurement data. Note that for this 65 nm technology, the influence of NMOS CHC on RO performance aging is negligible, which indicates the dominance of PMOS NBTI.

Based on the newly developed model, the trend of V_{th} change due to NBTI is extrapolated toward the 12 nm node, as shown in Fig. 5.10. Technology specifications are taken from the nominal Predictive Technology Model [29]. Due to the scaling of V_{DD}, the electric field across gate oxide, $E_{ox} = V_{gs}/T_{ox}$, actually decreases for future technology generations. Consequently, ΔV_{th} due to NBTI is reduced with such a trend of scaling. On the other hand, because of the slow scaling of V_{th} (for leakage control) and T_{ox}, the ratio of V_{DD}/V_{th} is lower and thus, device and circuit performance have increasing sensitivity to V_{th} change.

Such a behavior is illustrated in Fig. 5.11, where the frequency shift (ΔF) of a 65 nm RO is monitored under V_{DD} tuning [30]. Since the amount of the degradation

Fig. 5.9 The frequency degradation of a 65 nm ring oscillator (Adapted from [10])

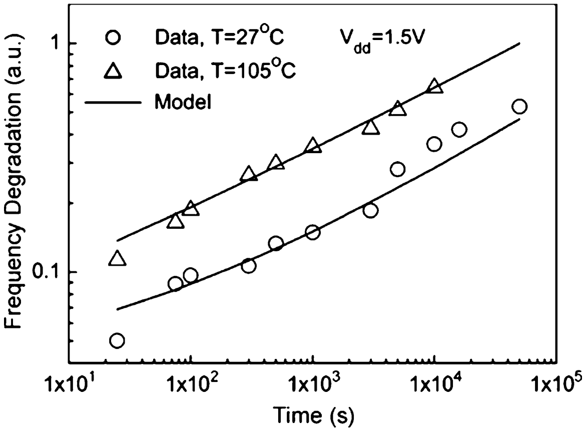


Fig. 5.10 The prediction of V_{th} increase in 7 years under dynamic NBTI effect (50% duty cycle)

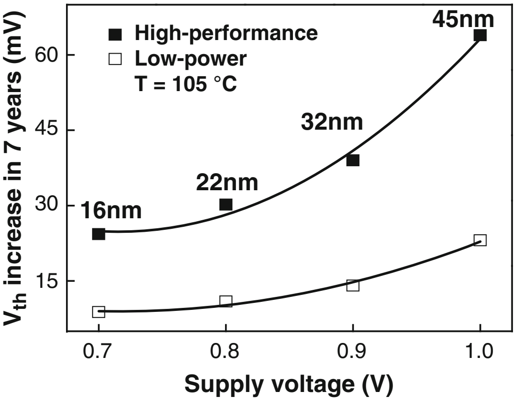
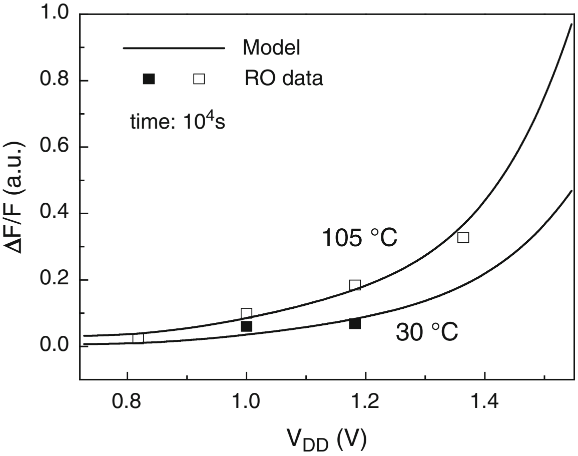


Fig. 5.11 The reduction of aging with lower V_{DD} (Adapted from [30])



is an exponential function of V_{DD} (Eq. 5.5), lower V_{DD} helps reduce the aging. On the other side, if V_{DD} is too low, then circuit performance sensitivity to V_{th} shift is elevated, which eventually cancels the benefit. Figure 5.11 confirms that the reduction rate in ΔF/F is much smaller when V_{DD} is lower than the nominal value.

5.4 Interaction with Process Variations

Since NBTI effect has an exponential dependence on E_{ox}, which is inversely proportional to T_{ox} (Eq. 5.5), device reliability degradation strongly interacts with process variations, significantly shifting both the mean and the variance of the circuit performance. Figure 5.12 shows the measured RO speed degradation from a 65 nm technology [31]. Both static process variations and dynamic operation affect the performance and its variability [5, 31]. Therefore, accurate prediction of the reliability during the lifetime should consider the impact of static variations, primary reliability mechanisms, and more importantly, their interactions. This prediction is essential for designers to safely guardband the circuit for a sufficient lifetime. Otherwise, either an overly pessimistic bound or expensive statistical stress tests need to be used.

A few works have been published in the literature to estimate the statistical variations in temporal NBTI degradation [32–36]. Their assumption is the number of broken bonds in the interface is a Poisson random variable, and correspondingly V_{th} follows the Poisson distribution. With technology scaling, additional V_{th} variations, such as random dopant fluctuation and short channel effects, need to be considered. The measurement data show that the distribution of V_{th} variations follows the Gaussian distribution [36]. In addition, the correlations between process variation and NBTI are ignored in previous work. Starting from the assumption that

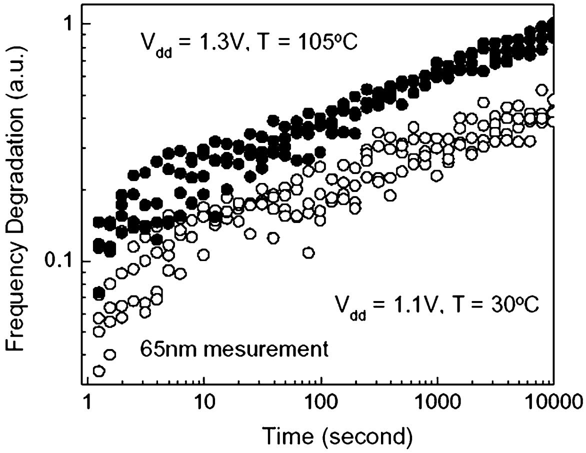


Fig. 5.12 Measured frequency degradation of a 65 nm 11-stage RO under various stress conditions [31]

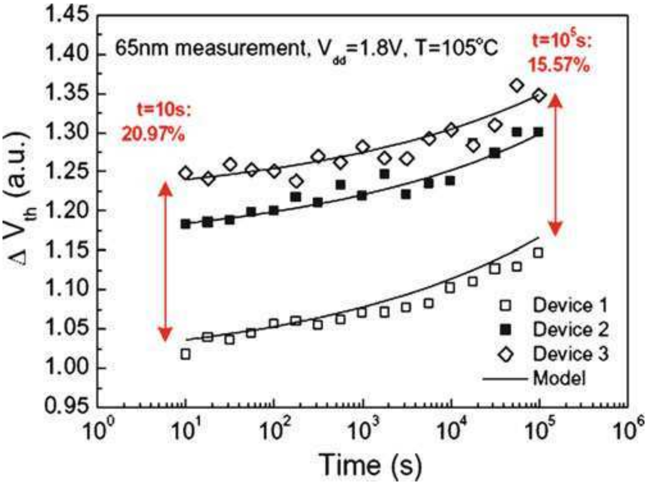


Fig. 5.13 Threshold voltage degradation for different 65 nm devices

process variation induced V_{th} change is a Gaussian random variable, this section analyzes the statistical characteristics of temporal degradation.

NBTI manifests itself as a gradual increase in the magnitude of threshold voltage, resulting in the degradation of circuit performance over time. The model in Sect. 5.2 assumes nominal degradation without considering statistical process variations. If there are global and local process variations, especially those in T_{ox}, E_{ox} in Eq. 5.5 will also become a statistical variable. Due to the fluctuation in T_{ox}, the variations in V_{th} and E_{ox} are correlated: thinner T_{ox} leads to higher E_{ox} and lower V_{th} at the same time [22]. Statistically, V_{th} can be expressed as

$$V_{th} = V_{th0} + \Delta V_{th-g} + \Delta V_{th-l} \tag{5.12}$$

where V_{th0} is the nominal threshold voltage, ΔV_{th-g} and ΔV_{th-l} represent the change of V_{th} due to global and local variations, respectively. Equation 5.12 shows that positive T_{ox} variation (i.e., thicker T_{ox}) results in V_{th} increase, which correspondingly leads to smaller V_{th} degradation due to weaker E_{ox} (Eq. 5.5). Figure 5.13 shows V_{th} degradation over time for three different transistors at the 65 nm node [31]. Due to static process variations, Device 1 starts with a larger V_{th} and Device 3 starts with a smaller V_{th}. Under the same stress conditions, the degradation of V_{th} for these three devices is shown in Fig. 5.13. At the beginning, the difference in V_{th} between Device 1 and Device 3 is 20.97%. With the increase of stress time, the difference becomes smaller and smaller. After 10^5 s stress, it decreases to 15.57%. Such compensation between process variations and reliability degradation is well captured by our models.

In summary, a set of predictive models for device aging effects are developed. Excellent model scalability and predictability have been verified with experimental data. By implementing these models into the circuit simulator, it enables efficient design practice with emerging reliability concerns. As VLSI design in the late CMOS era is driven by an ever-increasing challenge to cope with unreliable components, these predictive models serve as a solid basis to explore innovative design and test solutions for reliability [37].

References

1. International Technology Roadmap of Semiconductors, 2007. (available at http://www.itrs.net).
2. S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, V. De, "Parameter variations and impact on circuits and microarchitecture," *ACM/IEEE Design Automation Conference*, pp. 338–342, June 2003.
3. W. Wang, S. Yang, S. Bhardwaj, R. Vattikonda, F. Liu, S. Vrudhula, Y. Cao, "The impact of NBTI on the performance of combinational and sequential circuits," *Design Automation Conference*, pp. 364–369, 2007.
4. R. Vattikonda, W. Wang, Y. Cao, "Modeling and minimization of PMOS NBTI effect for robust nanometer design," *Design Automation Conference*, pp. 1047–1052, 2006.
5. D. K. Schroder, J. A. Babcock, "Negative bias temperature instability: road to cross in deep submicron silicon semiconductor manufacturing," *J. of Applied Physics*, vol. 94, no. 1, pp. 1–17, July 2003.
6. N. Kimizuka, T. Yamamoto, T. Mogami, K. Yamaguchi, K. Imai, and T. Horiuchi, "The impact of bias temperature instability for direct-tunneling ultra-thin gate oxide on MOSFET scaling," *VLSI Symp. on Tech.*, pp. 73–74, 1999.
7. V. Reddy, et al., "Impact of negative bias temperature instability on digital circuit reliability," *IRPS*, pp. 248–254, 2002.
8. B. C. Paul, K. Kang, H. Kufluoglu, M. A. Alam, and K. Roy, "Impact of NBTI on the temporal performance degradation of digital circuits," *EDL*, vol. 26, pp. 560–562, 2003.
9. H. Puchner and L. Hinh, "NBTI reliability analysis for a 90 nm CMOS technology," *ESSDERC*, pp. 257–260, 2004.
10. W. Wang, V. Reddy, A. T. Krishnan, R. Vattikonda, S. Krishnan, Y. Cao, "Compact modeling and simulation of circuit reliability for 65 nm CMOS technology," *IEEE Transactions on Device and Materials Reliability*, vol. 7, no. 4, pp. 509–517, December 2007.
11. A. S. Goda, G. Kapila, "Design for degradation: CAD tools for managing transistor degradation mechanisms," *ISQED*, pp. 416–420, 2005.
12. C. Hu, S. C. Tam, F.-C. Hsu, P.-K. Ko, T.-Y. Chan, and K. W. Terrill, "Hot electron induced MOSFET degradation – model, monitor, and improvement," *IEEE Tran. on Electron Devices*, vol. 32, no. 2, pp. 375–385, Feb. 1985.
13. K. O. Jeppson and C. M. Svensson, "Negative bias stress of MOS devices at high electric fields and degradation of MNOS devices," *J. of Applied Physics*, vol. 48, pp. 2004–2014, 1977.
14. M. A. Alam, S. Mahapatra, "A comprehensive model of PMOS NBTI degradation," *Microelectronics Reliability*, vol. 45, pp. 71–81, 2005.
15. S. Chakravarthi, A. T. Krishnan, V. Reddy, C. F. Machala and S. Krishnan, "A comprehensive framework for predictive modeling of negative bias temperature instability," *IRPS*, pp. 273–282, 2004.
16. S. Ogawa and N. Shiono, "Generalized diffusion-reaction model for the low-field charge-buildup instability at the Si-SiO2 interface," *Physical Review B*, vol. 51, no. 7, pp. 4218–4230, February 1995.
17. S. Bhardwaj, W. Wang, R. Vattikonda, Y. Cao, S. Vrudhula, "A scalable model for predicting the effect of NBTI for reliable design," *IET Circuits, Devices & Systems.*, vol. 2, no. 4, pp. 361–371, 2008.

18. S. Mahapatra, D. Saha, D. Varghese, and P. B. Kumar, "On the generation and recovery of interface traps in MOSFETs subjected to NBTI, FN, and HCI stress," *IEEE Trans. Electron Devices*, vol. 53, no. 7, pp. 1583–1592, Jul. 2006.
19. S. Mahapatra, "Electrical characterization and modeling of negative bias temperature instability in p-MOSFET devices," *IRPS*, 2006.
20. F. C. Hsu and S. Tam, "Relationship between MOSFET degradation and hot-electron-induced interface state generation," *IEEE Electron Device Lett.*, vol. EDL-5, no. 2, pp. 50–52, Feb. 1984.
21. S. C. Sun and J. D. Plummer, "Electron mobility in inversion and accumulation layers on thermally oxidized silicon surfaces," *IEEE Trans. Electron Devices*, vol. ED-27, no. 8, pp. 1497–1508, Aug. 1980.
22. "BSIM4 Manual," University of California, Berkeley, 2005.
23. S. Rangan, N. Mielke, E. C. C. Yeh, "Universal recovery behavior of negative bias temperature instability," *IEDM*, pp. 341–344, 2003.
24. A. T. Krishnan, C. Chancellor, S. Chakravarthi, P. E. Nicollian, V. Reddy, and A. Varghese, "Material dependence of hydrogen diffusion: Implication for NBTI degradation," *IEDM*, 2005.
25. G. Chen, et al., "Dynamic NBTI of PMOS transistors and its impact on device lifetime," *IRPS*, pp. 196–202, 2003.
26. V. Huard, M. Denais, "Hole trapping effect on methodology for DC and AC negative bias temperature instability measurements in PMOS transistors," *IRPS*, pp. 40–45, 2004.
27. M. A. Alam, "A critical examination of the mechanics of dynamic NBTI for PMOSFETs," *IEDM*, pp. 345–348, 2003.
28. B. Zhu, J. S. Suehle, J. B. Bernstein, and Y. Chen, "Mechanism of dynamic NBTI of pMOSFETs," *IRW*, pp. 113–117, 2004.
29. W. Zhao and Y. Cao, "New generation of Predictive Technology Model for sub-45 nm design exploration," *ISQED*, pp. 585–590, 2006.
30. R. Zheng, J. Velamala, V. Reddy, V. Balakrishnan, E. Mintarno, S. Mitra, S. Krishnan, Y. Cao, "Circuit aging prediction for low-power operation," Custom Integrated Circuits Conference, pp. 427–430, 2009
31. W. Wang, V. Reddy, B. Yang, V. Balakrishnan, S. Krishnan, Y. Cao, "Statistical prediction of circuit aging under process variations," *Custom Integrated Circuits Conference*, pp. 13–16, 2008.
32. S. E. Rauch, "The statistics of NBTI induced vt and β mismatch shifts in PMOSFETs," *IEEE Trans. on Device Material Reliability*, pp. 89–93, 2002.
33. S. E. Rauch, "Review and reexamination of reliability effects related to NBTI-induced statistical variations," *IEEE Transactions on Device and Materials Reliability*, vol. 7, no. 4, pp. 524–530, 2007.
34. G. L. Rosa, W. L. Ng, S. Rauch, R. Wong, and J. Sudijono, "Impact of NBTI induced statistical variation to SRAM cell stability," *IEEE International Reliability Physics Symposium*, pp. 274–282, Mar. 2006.
35. K. Kang, S. P. Park, K. Roy, and M. A. Alam, "Estimation of statistical variation in temporal NBTI degradation and its impact on lifetime circuit performance," *IEEE/ACM International Conference on Computer-Aided Design*, pp. 730–734, Nov. 2007.
36. W. Zhao, F. Liu, K. Agarwal, D. Acharyya, S. R. Nassif, K. Nowka, Y. Cao, "Rigorous extraction of process variations for 65 nm CMOS design," *IEEE Transactions on Semiconductor Manufacturing*, vol. 22, no. 1, pp. 196–203, February 2009.
37. Y. Cao, J. Tschanz, P. Bose, "Reliability challenges in nano-CMOS design," *IEEE Design & Test of Computers*, Special Issue on Design for Reliability at 32nm and Beyond, vol. 26, no. 6, pp. 6–7, November/December 2009.

Chapter 6
Modeling of Interconnect Parasitics

With continual scaling of CMOS technology, the parasitics of backend-of-the-line (BEOL) interconnect (i.e., wire resistance and capacitance) become increasingly important to circuit performance [1, 2]. In order to match the shrinking pitch of transistors on the silicon substrate, local metal wires need to be narrower and closer, leading to the dramatic increase in the coupling capacitance and RC delay [1]. To overcome this barrier and enhance circuit speed, many technology advances have been made in sub-65 nm CMOS technology. Traditional Al/SiO$_2$ technology is replaced by Cu/low-*k* inter-layer dielectric (ILD) that helps reduce metal resistance and improve the reliability under the electromigration effect (Fig. 6.1). However, the integration of Cu into the CMOS process requires a special diffusion barrier to prevent the rapid diffusion of Cu through ILDs, as shown in Fig. 6.1. This diffusion barrier usually has a higher dielectric constant than that of ILD and thus, increases the capacitance. As the ILD thickness keeps decreasing in CMOS scaling, the impact of the diffusion layer becomes more pronounced. Moreover, to minimize wire capacitance, especially the coupling capacitance between neighboring wires, recent process development focuses on new dielectric material with even lower-*k* value. For instance, the air gap is expected to be integrated into the BEOL structure (Fig. 6.1) [1]. These technology innovations extend the lifetime of current multiple-layer BEOL. Their impact on circuit performance needs to be quantitatively assessed and integrated into design tools.

6.1 Background of Interconnect Models

In today's electrical circuit simulation, a physical metal wire is usually translated into an equivalent RC or RLC model for circuit simulation, where the values of parasitic resistance, capacitance, and inductance (if needed) are extracted from the dimension of wires and ILDs [2]. The accuracy and efficiency of such an extraction are essential to evaluating circuit performance metrics, such as the speed, power consumption and coupling noise.

Y. Cao, *Predictive Technology Model for Robust Nanoelectronic Design*,
Integrated Circuits and Systems, DOI 10.1007/978-1-4614-0445-3_6,

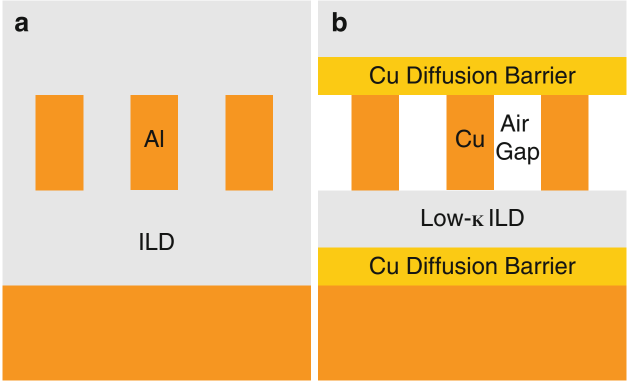


Fig. 6.1 Cross-sectional view of (**a**) Al interconnect in previous technology generations, and (**b**) contemporary Cu interconnect with the diffusion barrier and the air gap (Adapted from [14])

The parasitic capacitances can be accurately calculated by field solvers, such as Raphael and FastCap [3, 4], from the specifications of a BEOL structure. However, these numerical approaches are often computationally expensive, requiring a large amount of CPU time and memory. Therefore, they are inefficient to support large-scale circuit analysis. An alternative approach is based on look-up tables, in which the capacitance values are pre-solved for a specific BEOL technology. Yet the size of the tables limits the flexibility and efficiency. To support general interconnect analysis, compact model, which describes the parasitics as a closed-form function of wire and ILD dimensions, is a desirable solution that achieves excellent scalability and efficiency.

Many works have been devoted to analytical capacitance modeling of basic BEOL structures, such as a single line on the ground plane [5–8]. In [9], Sakurai et al. developed an empirical model for a typical structure of global interconnects: parallel lines above a ground plate with homogeneous ILD (Fig. 6.1a). Wong et al. derived empirical models for a representative structure of local interconnects: parallel lines between two ground plates [10, 11], and improved the fitting accuracy of Sakurai's model. Bansal et al. further developed an analytical model of non-overlapping interconnects in different layers using conformal mapping method [12]. These models provide closed-form solutions that are applicable to a limited range of wire dimensions. Some of them were adopted by Berkeley Predictive Technology Model (BPTM) to estimate the parasitics in scaled BEOL [13].

However, the physical basis of previous models is not adequate to accurately predicting the capacitance value without an intensive fitting process. Furthermore, their empirical nature limits the extension to advanced BEOL structures, including the non-uniform dielectrics, the diffusion barrier and the air gap (Fig. 6.1b). These contemporary features are necessary to meet the scaling criteria of BEOL [1], even though the exact choice varies among different technologies. In this context, compact capacitance models should be sufficiently flexible and accurate to cover a wide range of BEOL parameters.

To achieve these targets and help predict the performance of future metal interconnect, a new field-based compact capacitance model is developed for general 2D on-chip interconnect [14], with the emphasis on those new technology features. Different from wire capacitance, wire resistance and inductance are relatively insensitive to those new technology features. Therefore, previous approach in BPTM is still applicable [13]. Chapter 7 further provides some updates on the calculation of wire resistance in the nanometer regime.

In today's CMOS technology, a general BEOL structure can be decomposed into three types of basic structures [10]:

1. A single line above one plate
2. Parallel lines above one plate, which emulate metal wires in the top layer of BEOL
3. Parallel lines between two plates, which represent metal wires in the intermediate and local layers

Compact capacitance models for the above three structures will help calculate the capacitance in a general layout configuration.

The model derivation is based on the careful analysis of the electrical fields between lines and plates. In this new model, the total capacitance is decomposed into different building components, namely the plate capacitance, the fringe capacitance and the terminal capacitance. Their values are derived from the electrical field for each capacitance component that is independent and localized. Through this partition, model development is greatly simplified; non-uniform dielectric structures, such as copper diffusion barrier and the air gap, can be conveniently modeled by adaptively tuning the corresponding components. Though the importance of the terminal capacitance has long been speculated, its calculation and impact are clarified for the first time in this work. Furthermore, the effects of electrical field shielding and charge sharing are considered and integrated into the model in the case of multiple electrodes.

6.2 Modeling Principles

For each type of three basic structures, the capacitance exists between each pair of conductive surface, such as those shown in Fig. 6.2: C_{bottom} (C_{top}) is the capacitance between metal wire and the lower (upper) plate; C_{couple} is the coupling capacitance between neighboring lines in the same layer. Some commonly used wire dimensions are also denoted in Fig. 6.2: T for wire thickness, W for wire width, S for wire space and H for the distance between the wire and the plate. The full notation of other wire dimensions, dielectric constants and capacitance components are defined in Table 6.1.

The capacitance model can be obtained by conformal transformation [12]. However, this approach often leads to lengthy and complicated solutions [15]. On the other hand, empirical solutions simply use rational functions to fit the

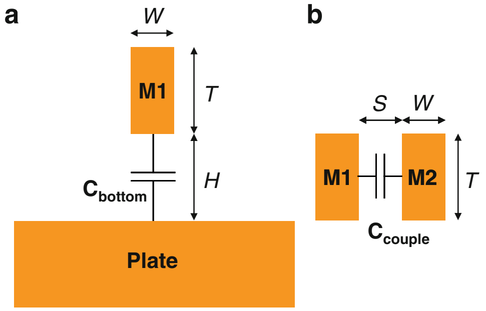


Fig. 6.2 Capacitance components between (**a**) a line and a plate, and (**b**) two parallel lines (Adapted from [14])

Table 6.1 Definitions of model parameters

Symbols	Parameter definitions
W	Wire width
T	Wire thickness
S	Wire space
H_B (H)	Wire to bottom plate distance
H_T	Wire to top plate distance
T_{DB}	Thickness of bottom Cu diffusion barrier
T_{DT}	Thickness of top Cu diffusion barrier
ε	Dielectric constant of low-κ dielectric
ε_D	Dielectric constant of Cu diffusion barrier
C_{bottom}	Wire to lower plate capacitance
C_{top}	Wire to upper plate capacitance
C_{couple}	Coupling capacitance between parallel wires
$C_{terminal}$	Capacitance from wire terminal
C_{fringe}	Fringe capacitance of the wire
C_{plate}	Capacitance between parallel surfaces

nonlinear behavior of the capacitance [9–12]. Given a range of wire dimensions, they require a significant amount of parameter fitting in order to achieve the accuracy. To combine the accuracy of the physics-based solution and the simplicity of the empirical approach, a closed-form capacitance model is proposed by analyzing the electrical field of each component. Such an approach improves model flexibility, providing valuable insights to BEOL design and optimization.

In the BEOL structure, the electrical field distributed among metal wires determines the capacitance value. For the basic structures in Fig. 6.2, Fig. 6.3a shows the equal potential contours simulated by Raphael [3]. The distribution of the electrical field is further derived from the equal potential contours (Fig. 6.3a). From its distribution, the electrical field can be approximately partitioned into different regions, as indicated by the solid lines in Fig. 6.3a. Figure 6.3b illustrates the boundaries of these regions. Such a partition helps us focus the model derivation on each region, and then sum all the regions up to obtain the total capacitance. Furthermore, the impact of latest BEOL advances, such as the air

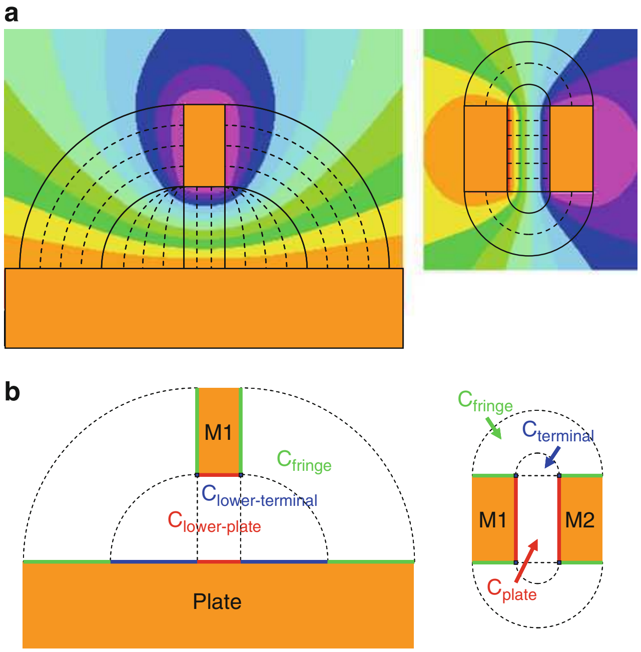


Fig. 6.3 The distribution of electrical fields: (**a**) the equal potential contours from Raphael simulation and the electrical field distribution, (**b**) the decomposition of electrical fields (Adapted from [14])

gap, is mainly on each individual region. Therefore, the partition improves the flexibility of the compact modeling effort.

Based on the partition of the electrical field, the total capacitance is classified into three fundamental cases, as shown in Fig. 6.3b:

1. Plate capacitance: between two parallel metal surfaces
2. Fringe capacitance: from the sidewall of the wire to another perpendicular surface, e.g., the ground plate
3. Terminal capacitance: from the corner of the wire to other metal surfaces

Each component is separately modeled, as described below.

Figure 6.4 illustrates the field in each component. The capacitance between two parallel plates is well known as:

$$\frac{C_{plate}}{\varepsilon} = \frac{W}{H} \tag{6.1}$$

The fringe capacitance between two perpendicular surfaces (Fig. 6.4b) can be derived from the conformal mapping method. A more convenient way is to approximate the electrical field as a circular region from H to H + T on the ground

Fig. 6.4 The fields of three basic components: (**a**) parallel plate capacitance, (**b**) fringe capacitance, (**c**) terminal capacitance (Adapted from [14])

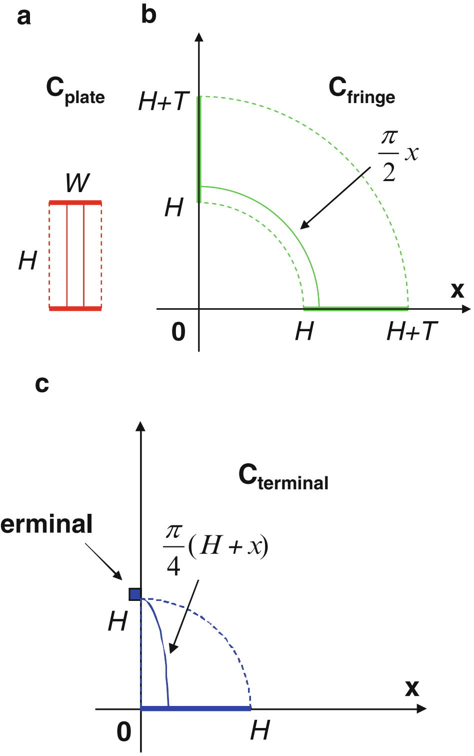


plate (Figs. 6.3b and 6.4b). Thus, the fringe capacitance is integrated from H to H + T along the x direction:

$$\frac{C_{fringe}}{\varepsilon} = \int \frac{width}{dis\tan ce} = \int_{H}^{H+T} \frac{dx}{\frac{\pi}{2}x} = \frac{2}{\pi}\ln\left(1 + \frac{T}{H}\right) \tag{6.2}$$

The last component is the terminal capacitance. Similar as the field from a point charge, the electrical field originated from the terminal spreads toward the plate, but limited to the region as shown in Fig. 6.3b. The range of such a field is approximated from 0 to H along the x direction (Fig. 6.4c). As a result, the terminal capacitance is not negligible. The terminal capacitance is calculated by integrating the ratio of its effective width and distance from 0 to H:

$$\frac{C_{ter\min al}}{\varepsilon} = \int \frac{width}{dis\tan ce} \approx \int_{0}^{H} \frac{dx}{\frac{\pi}{4}(H+x)} = \frac{4}{\pi}\ln 2 \tag{6.3}$$

Note that the terminal capacitance is independent on the dimensions, similar as the capacitance from a point charge.

6.3 Capacitance Modeling of the Basic Patterns

With the model for each component is available, it is ready to combine them together for a practical BEOL structure. This section demonstrates the derivation for the basic patterns, as shown in Fig. 6.2.

6.3.1 Model of the Line-to-Plate Capacitance

The first example is the capacitance of a single line on top of a plate. This capacitance is sometimes named as the ground capacitance. It is important for global on-chip interconnects. A comprehensive comparison of previous developed models is given by Barke [16]. However, those models are either not accurate enough or too empirical. Based on the discussion in Sect. 6.2, an accurate and physical model is presented below. As the electrical fields shown in Fig. 6.3, the total line-to-plate capacitance consists of three main components, i.e., lower-plate, lower-terminal and fringe capacitance. They are independent to each other. The total capacitance, C_{bottom}, is the summation of these three components:

$$C_{bottom} \approx C_{lower-plate} + 2C_{lower-ter\min al} + 2C_{fringe} \tag{6.4}$$

In reality, the electrical field of the three basic components is not exactly as shown in Fig. 6.4. Their boundaries are distorted, leading to some slight differences. Nevertheless, decomposing the electrical field into the basic components maintains the essential scalability to wire dimensions. To account for the charge distribution as compared to the ideal terminal case in Eq. 6.3, the following equation is proposed:

$$\frac{C_{lower-ter\min al}}{\varepsilon} = \frac{2}{\pi} \tag{6.5}$$

This value is a good approximation to compensate the field distortion due to adjacent plate and fringe capacitances. For a single line on top of a plate, we also need to consider the coupling between the upper terminal and the ground plate. By integrating the field, similar as that for Eq. 6.3, it is described as:

$$\frac{C_{upper-ter\min al}}{\varepsilon} = \frac{1}{\pi} \tag{6.6}$$

By combining the upper and lower terminal capacitances with the plate, fringe (Eqs. 6.1 and 6.2) and upper plate capacitances, a physical model for this simple case is completed:

$$\frac{C_{bottom}}{\varepsilon} = \frac{W}{H} + \frac{4}{\pi}\ln\left(1 + \frac{T}{H}\right) + \frac{6}{\pi} + \frac{2}{\pi}\ln\left[1 + \frac{\pi W}{2(1+\pi)(H+T)}\right] \tag{6.7}$$

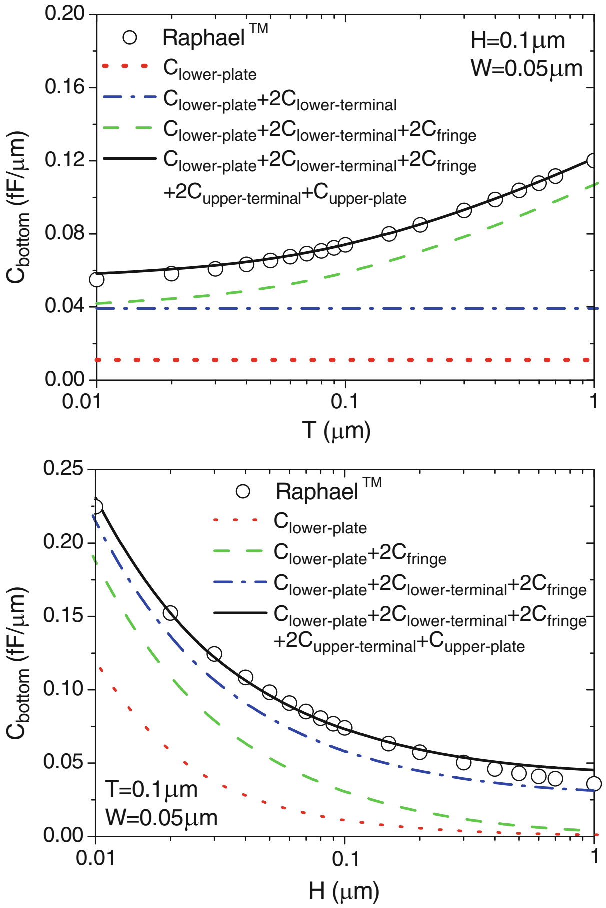


Fig. 6.5 Verification of the line-to-plate capacitance (Adapted from [14])

Note that the upper plate capacitance is calculated based on the same principle as the fringe capacitance in Eq. 6.2.

Figure 6.5 verifies Eq. 6.7 with numerical simulation results from Raphael [3]. The nominal dimensions are for local wires in a 45 nm technology, with $T = 0.1$ μm, $H = 0.1$ μm, and $W = 0.05$ μm, are demonstrated here [1]. Over a wide range of dimensions, the new model matches well with the simulation results. The physical nature of the model guarantees the scalability with all line dimensions.

6.3.2 Role of Terminal Capacitance

Figure 6.5 further illustrates the decomposition of the total capacitance into various components. Different from the traditional understanding, the parallel plate capacitance between the bottom of the wire to the plate, $C_{\text{lower-plate}}$, is actually the smallest component at the nominal dimensions. This is due to the increasing aspect ratio of the metal wire during the scaling [1]. On the other hand, the terminal capacitance, $C_{\text{lower-terminal}}$, is the largest component and contributes approximately half of C_{bottom} as shown in Fig. 6.5. Therefore, an accurate model of the terminal capacitance is important to calculate the total capacitance in a contemporary BEOL structure. While this term is usually ignored in previous models [5–11], our approach physically captures its important role in the calculation.

Equation 6.7 predicts that $C_{\text{lower-terminal}}$ remain as a constant during the scaling of H, while $C_{\text{lower-plate}}$ and C_{fringe} are inversely proportional to H, as shown in Fig. 6.5b. Therefore, the total capacitance, C_{bottom}, does not rapidly approach to zero as H increases (Fig. 6.5b). In principle, $C_{\text{lower-terminal}}$ decreases when H is much larger than W and T, because of the distortion of the electrical field. The neglect of such distortion does not introduce a significant amount of model errors, as shown in Fig. 6.5. Thus, we keep $C_{\text{lower-terminal}}$ as a constant in Eq. 6.7.

6.3.3 Model Comparison and Discussions

We evaluate the model error with several previous models that compute the single line-to-plate capacitance [5–8]. Figure 6.6 summarizes the normalized modeling error as compared to Raphael simulations. In general, the error of the new model is smaller than previous models. With the minimum fitting in the model, the distribution of the error is more stable than other models across the large range of wire dimensions. Previous models in [5] and [6] are also developed based on physical approaches, e.g., the principles in [5] are similar as our new model. However, they ignored both the upper and lower terminal components, which become increasingly important in scaled on-chip interconnect. In addition, the fringe capacitance was miscalculated. Models in [6] are accurate only when W is much larger than T/2. This is no longer the case for sub-65 nm BEOL. Models in [7] and [8] are developed based on an empirical fitting process. They are difficult to adapt to latest structures.

6.3.4 Coupling Capacitance between Parallel Lines

The other basic case in Fig. 6.3 is the coupling capacitance between two identical wires. Based on the model for a single line above one plate, we apply the method of the image charge to the coupling wires, as shown in Fig. 6.7. By inserting a virtual plate in the middle of the wires, the coupling capacitance, C_{couple}, is derived as:

$$\frac{C_{couple}}{\varepsilon} = \frac{T}{S} + \frac{2}{\pi}\ln\left(1 + \frac{2W}{S}\right) + \frac{3}{\pi} + \frac{1}{\pi}\ln\left[1 + \frac{\pi T}{2(1+\pi)(S/2+W)}\right] \tag{6.8}$$

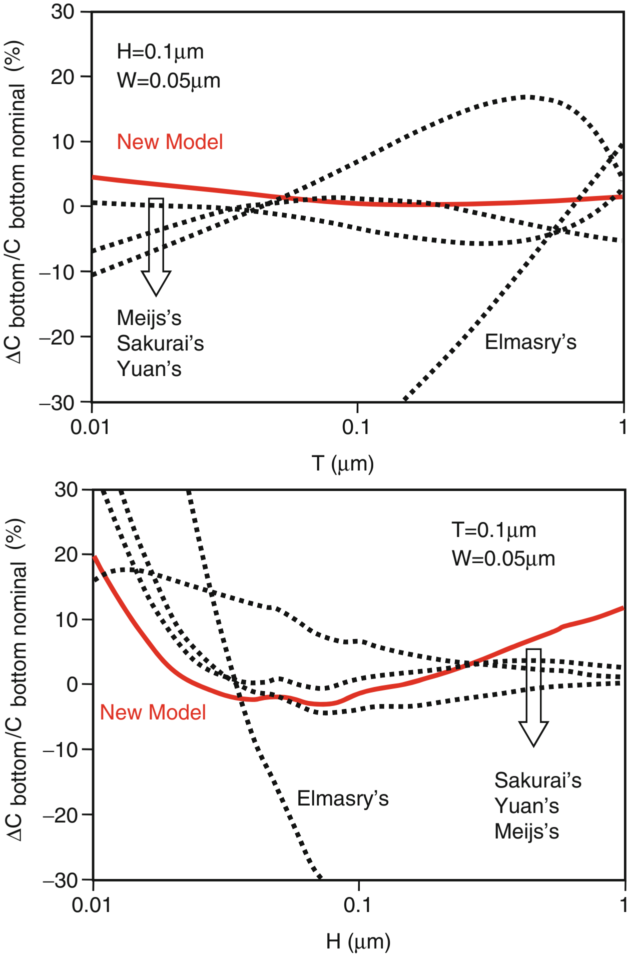


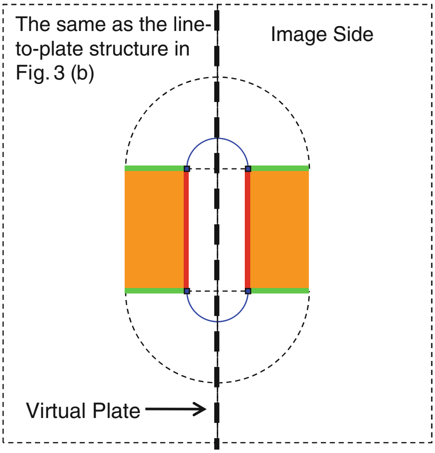
Fig. 6.6 The comparison of model errors (Adapted from [14])

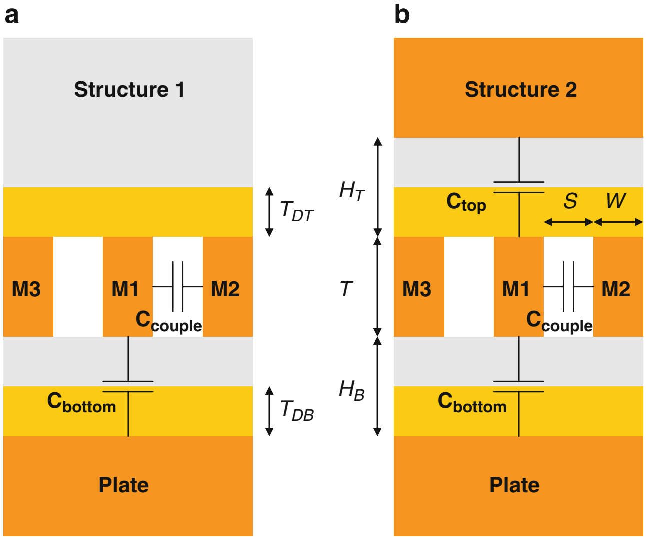
Note that there are also two terminal-to-terminal capacitances between two lines, which are constants.

6.4 Applications to General BEOL Structures

Figure 6.8 shows the contemporary BEOL structure with the air gap and low- *k* dielectrics to reduce the capacitance, and the barrier layer to prevent Cu diffusion. Without losing the generality, two 2D structures are identified that represent global wires on top of the plate (Fig. 6.8a) and local wires between two plates (Fig. 6.8b).

Fig. 6.7 The image method to calculate the coupling capacitance (Adapted from [14])





Structure 1 Structure 2

Fig. 6.8 Two general structures: (**a**) Parallel lines on top of a plate, and (**b**) Parallel lines between two plates (Adapted from [14])

Different from the simple structures in Sect. 6.2, there are multiple electrodes in these structures. Therefore, more complicated physical effects need to be considered, including the shielding effect of the electrical field, as well as the charge sharing effect among different nodes. This section first derives the models without Cu diffusion barrier and the air gap. Then, these advanced technology elements are incorporated into the model.

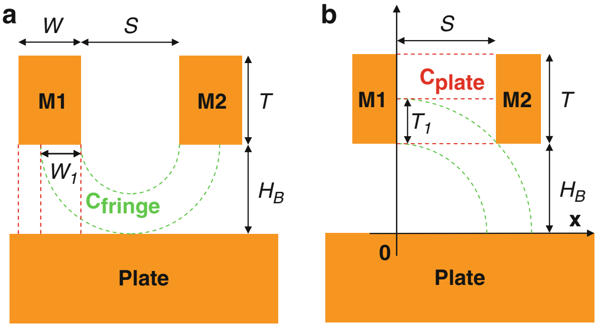


Fig. 6.9 The effects of (**a**) fielding shielding and (**b**) charge sharing (Adapted from [14])

6.4.1 Physical Effects with Multiple Electrodes

The first effect is the shielding effect when there are multiple coupling neighbors: when there are multiple wires, the field lines may not fully end to a single conductive surface; some of them go to other neighboring wire surfaces. For instance, Fig. 6.9a shows the C_{fringe} component of C_{couple}; only part of the electrical field originating from the lower surface of M1, i.e., within W_1, can reach the lower sidewall of M2 (Fig. 6.9a); the rest of the field is shielded by the plate underneath. As a result, C_{fringe} no longer increases with W if W is larger than H_B-S/2. A regional linear function is introduced to handle such a case:

$$f(x,a,b)=\begin{cases}0 & x<a\\ x-a & a\le x\le b\\ b-a & x>b\end{cases}\tag{6.9}$$

The regional dimensions W_1/W_2 are used for the shielding effect in the C_{fringe}:

$$W_1=\begin{cases}f(W,0,H_B-S/2) & H_B\ge S/2\\ 0 & H_B<S/2\end{cases}\tag{6.10}$$

$$W_2=\begin{cases}f(W,0,H_T-S/2) & H_T\ge S/2\\ 0 & H_T<S/2\end{cases}\tag{6.11}$$

Similarly, other regional dimensions under the field shielding effect include:

$$T_1=f\left(T,0,\sqrt{S^2+H_B^2}-H_B\right)\tag{6.12}$$

$$T_2=f\left(T,0,\sqrt{S^2+H_T^2}-H_T\right)\tag{6.13}$$

$$H_{B1} = f(H_B, 0, S/2) \tag{6.14}$$

$$H_{T1} = f(H_T, 0, S/2) \tag{6.15}$$

The other important effect in a multiple electrodes case is the sharing effect, which describes that the field from one conductor may be shared by two or more wire surfaces. An example is shown in Fig. 6.9b: the electrical field originated within T_1 from the right sidewall of M1 can be coupled to both the plate underneath and M2. The total charge within T_1 is therefore shared between the plate and M2. As a result, C_{plate} between M1 and M2 will be smaller than the original value of T/S. We introduce the model below to handle the charge sharing effect:

$$C_1^{'} = C_1 \cdot \frac{C_1}{C_1 + C_2} = \frac{C_1^2}{C_1 + C_2} \tag{6.16}$$

where C_1 and C_2 are the capacitances between two electrodes without considering charge sharing, and C_1' and C_2' are the capacitances with charge sharing. For instance, considering the charge sharing effect, C_{plate} in Fig. 6.9a is calculated as:

$$\begin{aligned}\frac{C_{plate}}{\varepsilon} &= \int\limits_{T_1} \frac{\Delta C_{plate}{}^2}{\Delta C_{plate} + \Delta C_{fringe}} + \frac{T - T_1}{S} \\ &= \int\limits_{H_B}^{H_B+T_1} \frac{\left(\frac{dx}{S}\right)^2}{\frac{dx}{S} + \frac{2dx}{\pi x}} + \frac{T - T_1}{S} = \frac{T}{S} - \frac{2}{\pi}\ln\left(1 + \frac{T_1}{H_B + 2S/\pi}\right)\end{aligned} \tag{6.17}$$

6.4.2 Modeling of the Coupling Capacitance

By including both effects of field shielding and charge sharing, the capacitance components in Structure 1 and 2 (Fig. 6.8) are derived below. Similar as C_{couple} in the simple case (Fig. 6.7), C_{couple} in Structure 1 and 2 has five major components, namely $C_{upper\text{-}fringe}$, $C_{upper\text{-}terminal}$, C_{plate}, $C_{lower\text{-}terminal}$, and $C_{lower\text{-}fringe}$. Their models are obtained from the principles discussed in Sect. 6.2. Table 6.2 summarizes the formulas. Because of the field shielding effect, some components in the capacitance model should reduce to a simpler model depending on line space. For instance, the second term in the denominator of $C_{lower\text{-}terminal}$ model, i.e., ln (1 + 0.3244 S/H_B), should return to Eq. 6.5 when $S = 2H_B$; to satisfy this condition, the coefficient is determined as 0.3244. Other coefficients in Tables 6.2 and 6.3 (i.e., 1.2974 and 0.76) are obtained from similar constraints. In addition, note that Structure 1 is a special case of Structure 2 where H_T is infinite. For the simplicity, only models for Structure 2 are presented. The total C_{couple} is the sum of all five components.

$$C_{couple} = C_{plate} + C_{lower-ter\min al} + C_{upper-ter\min al} + C_{lower-fringe} + C_{upper-fringe} \tag{6.18}$$

Table 6.2 Compact models of C_{couple} (Adapted from [14])

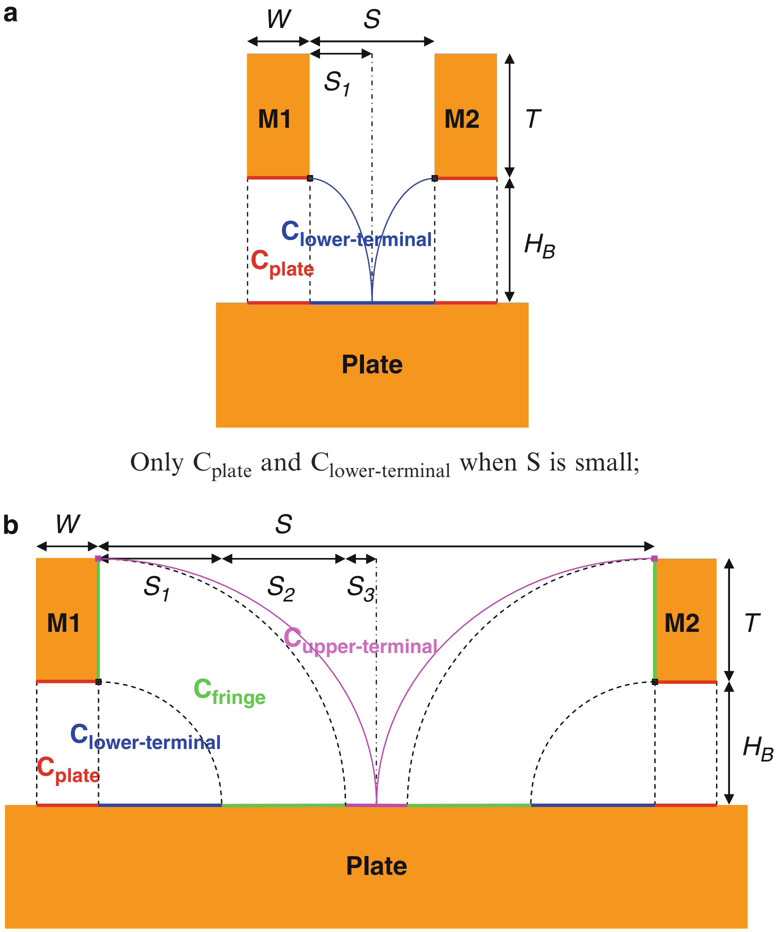
Component	Model
$\frac{C_{plate}}{\varepsilon}$	$\frac{T}{S} - \frac{2}{\pi}\ln\left[\left(\frac{H_B + 2S/\pi + T_1}{H_B + 2S/\pi}\right)\left(\frac{H_T + 2S/\pi + T_2}{H_T + 2S/\pi}\right)\right]$
$\frac{C_{lower-ter\min al}}{\varepsilon}$	$\frac{\left[\frac{2}{\pi}\ln(1 + 1.2974H_{B1}/S)\right]^2}{\frac{2}{\pi}\ln(1 + 1.2974H_{B1}/S) + \frac{4}{\pi}\ln(1 + 0.3244S/H_B)}$
$\frac{C_{lower-fringe}}{\varepsilon}$	$\frac{1}{\pi}\ln\left[\frac{(S + 2W_1)(S + 2H_B/\pi)}{S(S + 2H_B/\pi + 2W_1)}\right]$
$\frac{C_{upper-ter\min al}}{\varepsilon}$	$\frac{\left[\frac{2}{\pi}\ln(1 + 1.2974H_{T1}/S)\right]^2}{\frac{2}{\pi}\ln(1 + 1.2974H_{T1}/S) + \frac{4}{\pi}\ln(1 + 0.3244S/H_T)}$
$\frac{C_{upper-fringe}}{\varepsilon}$	$\frac{1}{\pi}\ln\left[\frac{(S + 2W_2)(S + 2H_T/\pi)}{S(S + 2H_T/\pi + 2W_2)}\right]$

Table 6.3 Compact models of C_{bottom} (Adapted from [14])

Component	Model
$\frac{C_{lower-plate}}{\varepsilon}$	$\frac{W}{H_B}$
$\frac{C_{lower-ter\min al}}{\varepsilon}$	$\frac{\left[\frac{4}{\pi}\ln\left(1 + \frac{0.76S_1}{H_B}\right)\right]^2}{\frac{4}{\pi}\ln\left[\left(1 + \frac{0.76S_1}{H_B}\right)\left(1 + \frac{S_4/4}{H_T + T}\right)\right]}$
$\frac{C_{fringe}}{\varepsilon}$	$\frac{2}{\pi}\ln\left[\left(\frac{H_B + S_2}{H_B}\right)\left(\frac{S + H_B}{S + H_B + S_2}\right)\right]$
$\frac{C_{upper-ter\min al}}{\varepsilon}$ (Structure 1 only)	$\frac{4}{\pi}\ln\left(1 + \frac{S_3/4}{H_B + T}\right)$

6.4.3 Revisiting the Line-to-Plate Capacitance

In Structure 1, C_{bottom} consists of four major components: C_{plate}, $C_{\text{lower-terminal}}$, C_{fringe} and $C_{\text{upper-terminal}}$. Among them, $C_{\text{lower-terminal}}$, C_{fringe} and $C_{\text{upper-terminal}}$ are optional, depending on the space between neighboring lines. Figure 6.10 shows the conditions when these components may not be necessary. When S is smaller than twice of H_B, C_{bottom} will only have two components, i.e., C_{plate} and $C_{\text{lower-terminal}}$ since other fields are shielded out (Fig. 6.10a). When S increases, C_{bottom} has another component, i.e., C_{fringe}; when S is large enough, the field from the top surface will be able to reach the bottom plate and



Only C_{plate} and $C_{lower\text{-}terminal}$ when S is small;

$C_{upper\text{-}terminal}$ is effective only when S is large enough;

Fig. 6.10 The field shielding effect in the line-to-plate capacitance: (**a**) Only C_{plate} and $C_{lower\text{-}terminal}$ when S is small; (**b**) $C_{upper\text{-}terminal}$ is effective only when S is large enough (Adapted from [14])

thus $C_{upper\text{-}terminal}$ shows up (Fig. 6.10b). To account for such a field shielding effect, three regional dimensions related to S are introduced:

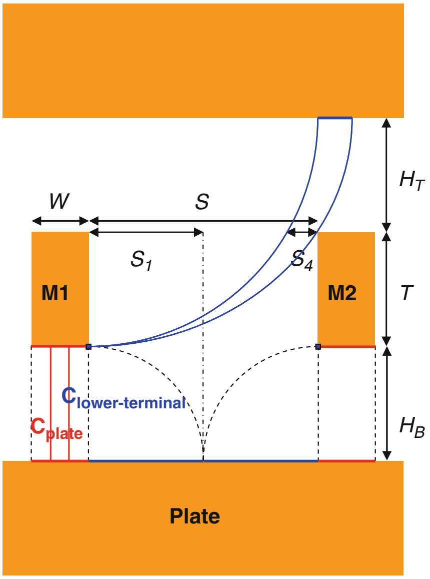
$$S_1 = f(S/2, 0, H_B) \tag{6.19}$$

$$S_2 = f(S/2, H_B, H_B + T) \tag{6.20}$$

$$S_3 = f(S/2, H_B + T, 2H_B + 2T) \tag{6.21}$$

Note that $C_{lower\text{-}terminal}$, C_{fringe} and $C_{upper\text{-}terminal}$ are further divided as the right and left ones if S at different sides are different.

Fig. 6.11 The effects of charge sharing and field shielding in Structure 2 (Fig. 6.8) (Adapted from [14])



In Structure 2, when S is larger than $(T^2 + 2TH_T)^{1/2}$, the electrical field from the lower terminal is shared between the lower and the upper plate, as shown in Fig. 6.11. This charge sharing effect reduces $C_{\text{lower-terminal}}$. Thus, we introduce another regional dimension:

$$S_4 = f\left(S, \sqrt{T^2 + 2TH_T}, 2H_B + 2T\right) \tag{6.22}$$

Table 6.3 summarizes the models of C_{bottom}. It is the sum of all four components in Structure 1 and C_{plate}, $C_{\text{lower-terminal}}$, C_{fringe} in Structure 2:

$$C_{bottom} = C_{plate} + 2C_{lower-ter\min al} + 2C_{fringe} + 2C_{upper-ter\min al} \tag{6.23}$$

To calculate C_{top}, H_B is switched to H_T in Eqs. 6.19–6.22 and Table 6.3. The total capacitance of line M1 is $2C_{\text{couple}} + C_{\text{bottom}}$ in Structure 1, and $2C_{\text{couple}} + C_{\text{bottom}} + C_{\text{top}}$ in Structure 2.

6.4.4 Cu Diffusion Barrier

In today’s BEOL technology, the impact of Cu diffusion barrier on the capacitance becomes more pronounced since its thickness scales much more slowly that ILD thickness. With our field-based method, it is convenient to incorporate it into the

Table 6.4 C_{couple} model parameters with Cu diffusion layer (Adapted from [14])

Component	Region	F	Dimension
$\frac{C_{plate}}{\varepsilon}$	Entire	1	$H_B \to H_B{}'$ $H_T \to H_T{}'$
$\frac{C_{lower-ter\min al}}{\varepsilon}$	$H_B - T_{DB} \geq S/2$	1	$H_B \to H_B{}'$
	$H_B - T_{DB} < S/2$	$\frac{\varepsilon_D}{\varepsilon}\left[1+\frac{(\varepsilon/\varepsilon_D - 1)(H_B - T_{DB})}{S/2}\right]$	
$\frac{C_{lower-fringe}}{\varepsilon}$	$H_B - T_{DB} \geq S/2 + W_1$	1	
	$H_B - T_{DB} < S/2 + W_1$	$\frac{\varepsilon_D}{\varepsilon}\left[1+\frac{(\varepsilon/\varepsilon_D - 1)(H_B - T_{DB})}{S/2 + W_1}\right]$	
$\frac{C_{upper-ter\min al}}{\varepsilon}$	$T_{DT} \geq S/2$	$\frac{\varepsilon_D}{\varepsilon}$	$H_T \to \varepsilon_D H_T{}'/\varepsilon$
	$T_{DT} < S/2$	$1+\frac{(\varepsilon_D/\varepsilon - 1)T_{DT}}{S/2}$	
$\frac{C_{upper-fringe}}{\varepsilon}$	$T_{DT} \geq S/2 + W_2$	$\frac{\varepsilon_D}{\varepsilon}$	$H_T \to \varepsilon_D H_T{}'/\varepsilon$
	$T_{DT} < S/2 + W_2$	$1+\frac{(\varepsilon_D/\varepsilon - 1)T_{DT}}{S/2 + W_2}$	

appropriate component. For C_{bottom} or C_{top} in Table 6.3, this is achieved by replacing H_B/H_T with $H_B{}'/\ H_T{}'$:

$$H_B^{'} = H_B + \left(\frac{\varepsilon}{\varepsilon_D} - 1\right)T_{DB} \tag{6.24}$$

$$H_T^{'} = H_T + \left(\frac{\varepsilon}{\varepsilon_D} - 1\right)T_{DT} \tag{6.25}$$

For the coupling capacitance, it is not sufficient only by replacing H_B/H_T with $H_B{}'/\,H_T{}'$ since the electrical field is not uniformly partitioned among different layers. For instance, if T_{DT} is larger than S/2, $C_{upper\text{-}terminal}$ is only in Cu diffusion barrier; however, if T_{DT} is smaller than S/2, part of $C_{lower\text{-}terminal}$ is in the low-κ dielectric layer. Therefore, models of C_{couple} are regional. The regional function, F, is listed in Table 6.4. F approximates the linear combination of the field distribution in non-uniform dielectrics. In presence of Cu diffusion layer, the capacitance component needs to be corrected by the regional function F and the dimension (Table 6.4):

$$C_{component} \to F \cdot C_{component} \tag{6.26}$$

For instance, if T_{DT} is larger than S/2, $C_{upper\text{-}terminal}$ is:

$$\frac{C_{upper-ter\min al}}{\varepsilon} = \frac{\varepsilon_D}{\varepsilon}\frac{\left[\frac{2}{\pi}\ln(1+1.2974H_{T1}/S)\right]^2}{\frac{2}{\pi}\ln(1+1.2974H_{T1}/S)+\frac{4}{\pi}\ln\left(1+0.3244S/H_T^{'}\right)} \tag{6.27}$$

Table 6.5 C_{bottom} model parameters with the air gap (Adapted from [14])

Component	F	Dimensions
$\frac{C_{lower-plate}}{\varepsilon}$	1	$H_B \rightarrow H_B{}'$
$\frac{C_{lower-ter\min al}}{\varepsilon}$	Eq. 6.29	$H_B \rightarrow H_B{}'$ $H_T \rightarrow H_T{}'$
$\frac{C_{fringe}}{\varepsilon}$	$1+\frac{\varepsilon_0(1-\varepsilon/\varepsilon_0)T}{\varepsilon(T+H_B{}')}$	$H_B \rightarrow H_B{}'$
$\frac{C_{upper-ter\min al}}{\varepsilon}$ (Structure 1 only)	$1+\frac{\varepsilon_0(1-\varepsilon/\varepsilon_0)T}{\varepsilon(T+H_B{}')}$	$H_B \rightarrow H_B{}'$

6.4.5 *Air Gap*

The adoption of the air gap successfully reduces the coupling capacitance between coplanar wires. By changing the effective T/S in the model of C_{plate}, C_{couple} with the air gap is handled:

$$\begin{aligned}\frac{C_{plate}}{\varepsilon} = &\frac{\varepsilon_0}{\varepsilon}\frac{T}{S} - \frac{2}{\pi}\left[1+\frac{\varepsilon_0(1-\varepsilon/\varepsilon_0)T_1}{\varepsilon(T_1+H_B^{\prime})}\right]\ln\left(\frac{H_B+2S/\pi+T_1}{H_B+2S/\pi}\right)\\ &-\frac{2}{\pi}\left[1+\frac{\varepsilon_0(1-\varepsilon/\varepsilon_0)T_2}{\varepsilon(T_2+H_T^{\prime}}\right]\ln\left(\frac{H_T+2S/\pi+T_2}{H_T+2S/\pi}\right)\end{aligned} \tag{6.28}$$

Similar as the treatment in the case of the diffusion layer, C_{bottom} or C_{top} with the air gap is calculated by replacing the dimensions and multiplying the F term in Table 6.5. To simplify the model of $C_{lower\text{-}terminal}$ with air gap, a fitting parameter β is introduced:

$$\frac{C_{lower-ter\min al}}{\varepsilon} = \frac{\left[\exp\left(\frac{\beta S_1}{H_B^{\prime}}\right)\frac{4}{\pi}\ln\left(1+\frac{0.76S_1}{H_B^{\prime}}\right)\right]^2}{\exp\left(\frac{\beta S_1}{H_B^{\prime}}\right)\frac{4}{\pi}\ln\left(1+\frac{0.76S_1}{H_B^{\prime}}\right)+\frac{4}{\pi}\ln\left(1+\frac{S_4/4}{H_T^{\prime}+T\varepsilon/\varepsilon_0}\right)} \tag{6.29}$$

The value of β is −0.2 in our validation in Sect. 6.5.

6.5 Model Validation and Comparison

The new models are comprehensively validated with Raphael simulation results. The nominal conditions are from a 45 nm technology: T = H_B = H_T = 0.1 μm, W = S=0.05 μm, T_{DB} = T_{DT} = 0.04 μm, ε = 2.5ε_0 and ε_D = 4ε_0.

These ratios are similar as both ITRS prediction [1] and those in a realistic industry process. Note that the capacitances only depend on the ratio of line dimensions, not the absolute value. During CMOS technology scaling, the ratios of dimensions are relatively stable. On the other hand, the model has the minimum error when the dielectrics are uniform. When dielectric constant of local layers keeps reducing and becomes non-uniform during the scaling (e.g., by using the air gap) [1], the fitting parameter β may need to be slightly modified. The new model is scalable with these features and supports more efficient development of BEOL models.

Figures 6.12 and 6.13 demonstrate the comprehensive model validation with numerical simulations of Raphael, as well as the empirical model developed by Wong et al. [10]. To conduct a fair comparison, H_B' and H_T' are used in the empirical model instead of H_B and H_T, when Cu diffusion barrier exists. Overall, the new compact model is more accurate than previous empirical results over a wide range of dimensions. Furthermore, it offers an excellent flexibility to incorporate various new structures. Since the new model is based on the analysis of the electrical field, its physical nature helps provide valuable insights on the capacitance scaling. It facilitates us to identify and improve the weakness of previous empirical models. One example is the model of C_{bottom} with large line space. In the empirical model [10], C_{bottom} dependences on H_T and T are not considered, which become important when S is large (Fig. 6.13). The new model well predicts such dependences through the fringe and upper terminal capacitances.

Table 6.6 further evaluates the model with simulation results at different dimension corners, assuming $W = S$, $H_B = H_T$, $T_{DB} = T_{DT}$ and other variables remain the same as their nominal values. The maximum error at the corners is only 5.0% and 1.5% for C_{couple} and C_{bottom}, respectively. The mean square root (RMS) error of the model is 2.0% and 1.2% for C_{couple} and C_{bottom}, respectively, which is much smaller than the results from Wong's model (17.6% and 18.3%, respectively). Since the new model is physics based, it is applicable to a wide range of practical geometrical and material values with reasonable accuracy. Finally, as shown in Table 6.6, Cu diffusion barrier leads to 8% increase in C_{couple}, while the air gap reduce C_{couple} by 38% at nominal dimensions of this 45 nm technology.

In summary, this chapter presents a new physical model for the parasitic calculations of scaled BEOL interconnect. Different from previous empirical approaches, the new model is derived from an in-depth analysis of the electrical field distribution between multiple electrodes. The terminal capacitance is identified as an important component in the capacitance modeling. The new model is conveniently customized to incorporate advanced CMOS interconnect structures, such as Cu diffusion layer and the air gap. As demonstrated with Raphael simulations at the 45 nm node, the new model achieves excellent accuracy and scalability in the capacitance calculation over a wide range of interconnect definitions.

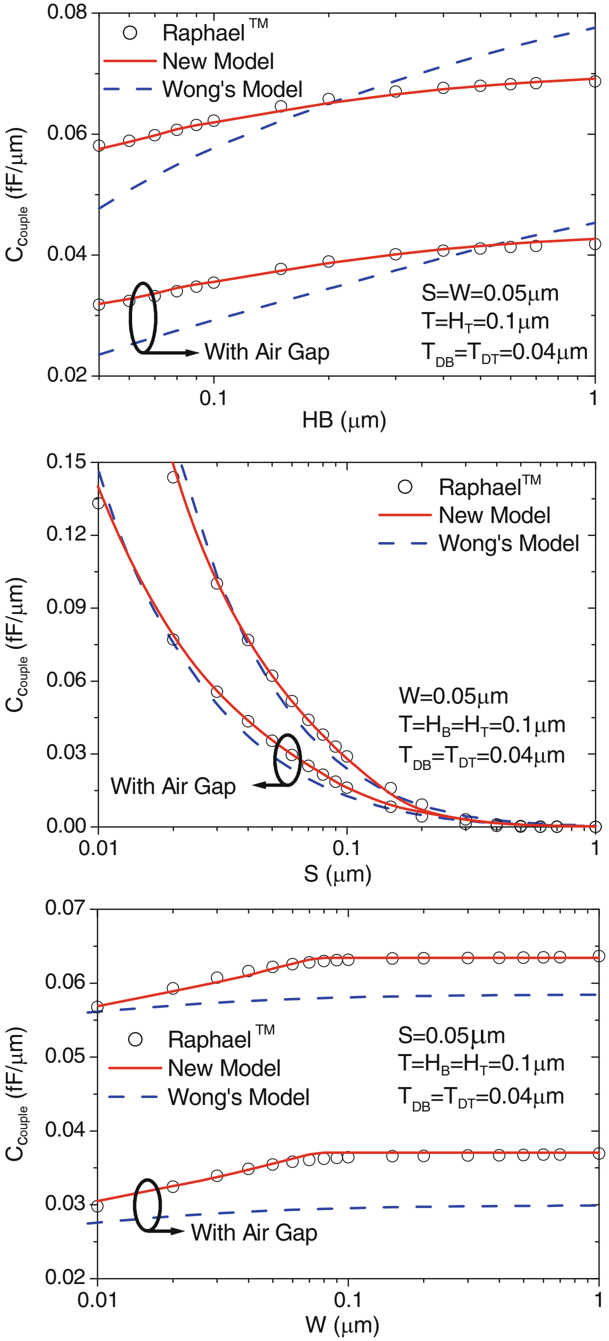


Fig. 6.12 Validation of C_{couple} in Structure 2 (Fig. 6.8) (Adapted from [14])

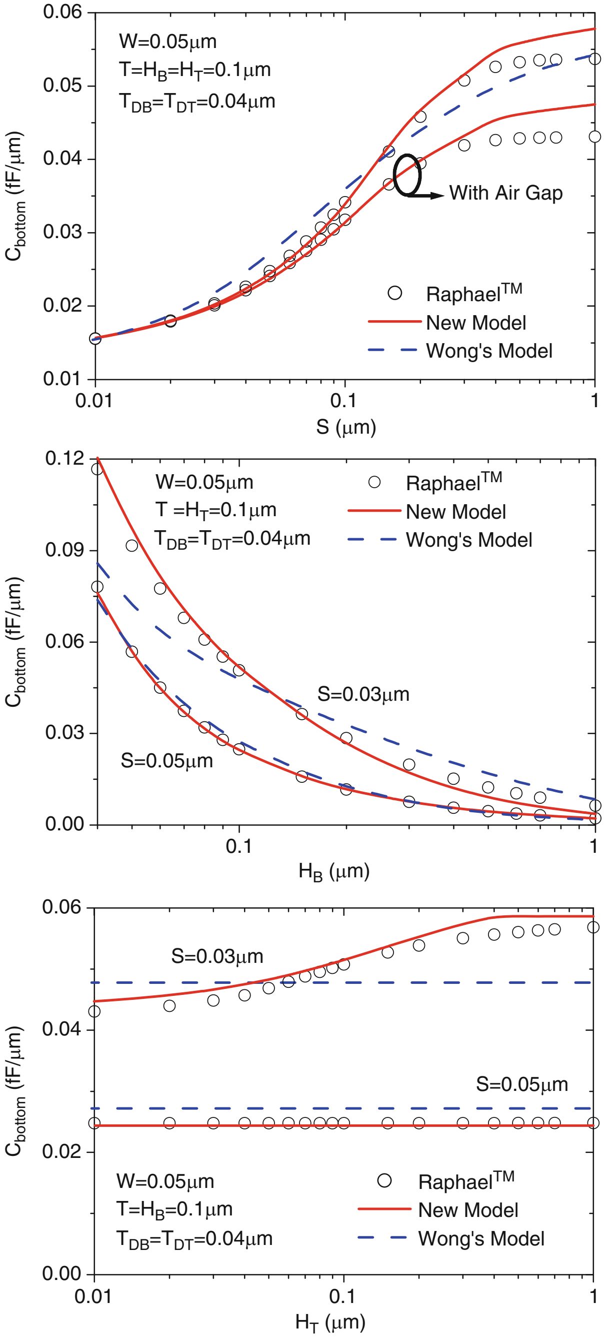


Fig. 6.13 Validation of C_{bottom} in Structure 2 (Fig. 6.8) (Adapted from [14])

Table 6.6 Comparison of model prediction with Raphael simulation (Adapted from [14])

Dimensions (nm)	C_{couple} (aF/μm)			C_{bottom} (aF/μm)		
{W, H_B, T_{DB}, air gap}	Model	Simulated	Error (%)	Model	Simulated	Error (%)
{50, 100, 0, w/o}	57.24	57.58	−0.6	20.87	21.18	−1.5
{50, 100, 40, w/o}	61.94	62.19	−0.4	24.39	24.77	−1.5
{50, 100, 40, w/}	35.58	35.43	0.4	23.74	24.08	−1.4
{50, 200, 40, w/o}	69.45	69.62	−0.2	11.49	11.63	−1.2
{50, 200, 40, w/}	42.91	42.67	0.6	11.35	11.44	−0.8
{100, 100, 40, w/o}	27.89	29.35	−5.0	46.87	47.30	−0.9
{100, 100, 40, w/}	15.97	16.50	−3.2	44.56	45.00	−1.0
{100, 200, 40, w/o}	39.05	39.32	−0.7	22.50	22.80	−1.3
{100, 200, 40, w/}	25.94	25.84	0.4	21.94	22.21	−1.2

References

1. International Technology Roadmap of Semiconductors, 2007. (available at http://www.itrs.net).
2. B. Wong, A. Mittal, Y. Cao, G. Starr, *Nano-CMOS Circuit and Physical Design*, John Wiley & Sons, Inc., 2004.
3. RAPHAEL Users' Manual, Technology Modeling Associate.
4. K. Nabors and J. White, "Fastcap: a multiple accelerated 3-D capacitance extraction program," *IEEE Transactions on Computer Aided Design*, vol. 10, no. 11, pp. 1447–1459, Nov. 1991.
5. M. I. Elmasry, "Capacitance calculation in MOSFET VLSI," *IEEE Electron Device Letters*, vol. EDL-3, no. 1, pp.6-7, Jan. 1982.
6. C. P. Yuan and T. N. Trick, "A simple formula for the estimation of the capacitance of two-dimensional interconnects in VLSI circuits," *IEEE Electron Device Letters*, vol. EDL-3, no. 12, pp. 391–393, Dec. 1982.
7. T. Sakurai and K. Tamaru, "Simple formulas for two- and three- dimensional capacitances," *IEEE Transactions on Electron Devices*, vol. ED-30, no. 2, pp. 183–185, Feb. 1983.
8. N. V. D. Meijs and J. T. Fokkema, "VLSI circuit reconstruction from mask topology," *Integration*, vol. 2, no. 2, pp. 85–119, 1984.
9. T. Sakurai, "Closed-form expressions for interconnection delay, coupling, and crosstalk in VLSI's," *IEEE Transactions on Electron Devices*, vol. 40, no. 1, pp. 118–124, Jan. 1993.
10. S.-C. Wong, G.-Y. Lee, and D.-J. Ma, "Modeling of interconnect capacitance, delay, and crosstalk in VLSI", *IEEE Transactions on Semiconductor Manufacturing*, vol. 13, no. 1, pp. 108–111, Feb. 2000.
11. S.-C. Wong, T. G.-Y. Lee, D.-J. Ma, and C.-J. Chao, "An empirical three-dimensional crossover capacitance model for multilevel interconnect VLSI circuits", *IEEE Transactions on Semiconductor Manufacturing*, vol. 13, no. 2, pp. 219–227, May. 2000.
12. A. Bansal, B. Paul, and K. Roy, "An analytical fringe capacitance model for interconnects using conformal mapping", *IEEE Transactions on Computer-Aided Design of Integrated Circuit and Systems*, vol. 25, no. 12, pp. 2765–2774, Dec. 2006.
13. Y. Cao, T. Sato, M. Orshansky, D. Sylvester, and C. Hu, "New paradigm of predictive MOSFET and interconnect modeling for early circuit simulation," *CICC*, pp. 201–204, 2000.

14. W. Zhao, X. Li, S. Gu, S. H. Kang, M. Nowak, Y. Cao, “Field-based capacitance modeling for sub-65 nm on-chip interconnect,” *IEEE Transactions on Electron Devices*, vol. 56, no. 9, pp. 1862–1872, September 2009.
15. W. H. Chang, “Analytic IC-metal-line capacitance formulas,” *IEEE Transactions on Microwave Theory Techniques*, vol. MTT-24, pp. 608–611, Sept. 1976.
16. E. Barke, “Line-to-ground capacitance calculation for VLSI: a comparison”, *IEEE Transactions on Computer-Aided Design*, vol. 7, no. 2, pp. 295–298, Feb. 1988.

Chapter 7
Design Benchmark with Predictive Technology Model

CMOS technology scaling is increasingly challenged by fundamental physics and manufacturing limits at the 22 nm node and beyond [1]. High-*k*/metal gate devices and strained silicon techniques help extend the lifetime of CMOS technology, but also complicate the fabrication process and increase the amount of variations. The situation is compounded by low power process, which has different device and design requirements from high performance process, and RC parasitics of scaled backend-of-the-line (BEOL) interconnect. During the pathway of scaling, process and design tradeoffs, such as those between power consumption and circuit performance, become much more complex, due to the issues in aggressively scaled CMOS technology and the implementation of new circuit design techniques. These challenges reduce the predictability of circuit performance and increase the development cycle for new products. In order to continue the design success with nanoscale CMOS, it requires an early comprehension of the technology impacts and adaptively making design decisions up front. Such a predictive capability helps identify potential issues, enables early design research, and guarantees the time to market. To accomplish this new design paradigm, it requires Predictive Technology Models (PTM) to assess performance trends, and to evaluate key modules before silicon is ready [2, 3].

There have been many successful examples using PTM to benchmark various design techniques and expose potential design problems, including those in low power design, on-chip memory, and circuit robustness under variations [4–8]. This chapter demonstrates a predictive strategy to enable simultaneous exploration of low power CMOS process and design concepts at the 22 nm node, based on silicon data at 90-45 nm nodes [9]. The general PTM methodology is customized with specific enhancements of previously secondary physical effects [2, 3, 10], which are now significant for transistor and interconnect performance. Specific examples include high-*k*/metal gate, gate fringe capacitance, temperature effects, parasitic capacitances, high-*k* cap layer and etch damage layer in metal wires, metal grain scattering effects, and contact/via resistance [1]. These customized low power PTM models are systematically calibrated with 90-45 nm Poly/SiON data and published high-*k* /metal gate (HK/MG) information. PTM with multiple threshold voltage

Y. Cao, *Predictive Technology Model for Robust Nanoelectronic Design*, Integrated Circuits and Systems, DOI 10.1007/978-1-4614-0445-3_7,

(V_{th}) is successfully generated down to the 22 nm node for design assessment. It facilitates the projection of various behaviors of transistors, interconnect, and representative circuit modules, such as ring oscillator (RO), standard cell and SRAM down to the 22 nm node.

Furthermore, this chapter examines the roadmap of circuit resilience, recognizing the increasing impact of technology scaling on both the amount of and the performance sensitivity to process variations and reliability degradation. Leveraging predictive models of variability and reliability (Chaps. 4 and 5), failure rates in representative circuit units are evaluated.

7.1 Customization of PTM

The PTM was first introduced in 2000 based on BSIM3 model [1]. It was further improved in 2006, by identifying the scaling trend of key parameters and incorporating physical models [2]. It covers both frontend-of-the-line (FEOL) devices and backend-of-the-line (BEOL) metal interconnects. Predictions of FEOL technology rely on a set of simplified equations that capture the essential behavior of charge and carrier transport, rather than the full set of BSIM equations [2]. The electrostatic models emphasize the dependence of V_{th} on channel length (e.g., DIBL), channel doping, HALO, etc. The transport model adopts the velocity saturation model with overshoot behavior [2]. Such simplification allows easier extraction of critical model parameters from published data [2, 3, 10], capturing major device characteristics and their scaling trends [2]. In addition, the layout dependent stress effects are embedded into mobility and V_{th} models, and HK/MG transistor models are adopted for sub-45 nm devices [3]. The general PTM models from 180 to 16 nm are available at http://ptm.asu.edu.

In this chapter, the generic PTM is customized for an industrial low power process with multiple V_{th} choices [11]. Benefiting from the continuity of process scaling, we are able handle secondary device effects with better confidence, including the body effect, temperature dependence, and parasitic capacitances. For example, the trend of gate fringe capacitance (C_f) is calculated based on a physical equation [12], while source/drain resistance (R_{dsw}) remains constant, as shown in Fig. 7.1. Based on the predictive methodology explained above, Fig. 7.2 shows predicted I-V characteristics from 65 to 22 nm for both NMOS and PMOS devices. Strain effect has been included in PMOS devices at 32 and 22 nm nodes. Figure 7.3 shows predicted I_{on} and I_{off} of Poly/SiON and HK/MG devices from 45 to 22 nm node, as compared with published data [13, 14].

In addition to the transistor, parasitic BEOL resistance and capacitance play an increasingly important role in determining circuit performance. A predictive model for conventional BEOL structures was presented in [10]. Table 7.1 shows the scaling of interconnect geometries [13, 14, 16]. More complicated BEOL structures and new physical effects, such as high-*k* cap layer, etch damage layer, and metal

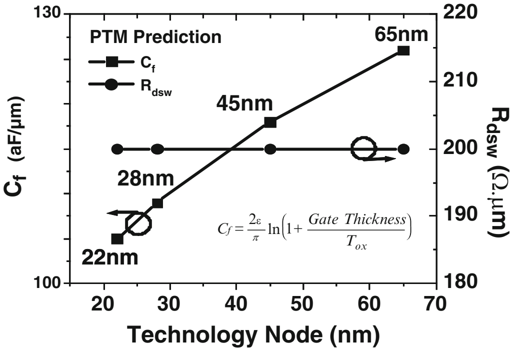


Fig. 7.1 The scaling trends of C_f and R_{dsw} (Adapted from [9])

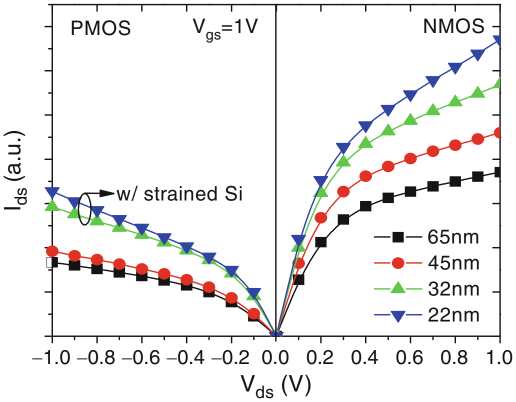


Fig. 7.2 The nominal I-V curves of scaled CMOS (Adapted from [9])

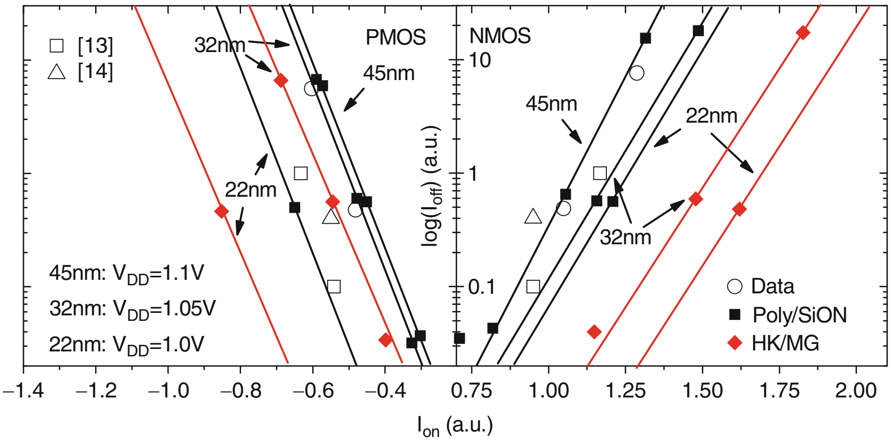


Fig. 7.3 I_{on}/I_{off} are predicted for Poly/SiON CMOS and HK/MG devices from 45 to 22 nm nodes (Adapted from [9])

Table 7.1 The scaling trend of interconnect parameters (Adapted from [9])

Technology (nm)	65 [13]	45 [13]	32 [15]	22
Gate pitch w/ contact (nm)	260	162	130	90 [16]
Contact pitch (nm)	200	126	110	80*
M1 pitch (nm)	180	126	100	70*
Intermediate metal pitch (nm)	200	126	100	70*
IMD *k* value	2.9	2.5	2.4	2.2*

*denotes predicted values

Table 7.2 Models of wire and contact/via resistance (Adapted from [9])

Metal resistivity	$\rho_m = \rho_{bulk}(1 + d/w)$
Metal resistance	$R_m = \left(\frac{\rho_m \cdot L}{A_{metal}}\right) \Vert \left(\frac{\rho_l \cdot L}{A_{liner}}\right)$
Contact/via resistance	$R_c = \frac{\rho_m \cdot H/\pi}{r^2 + (2r + T_t)T_t(\rho_m/\rho_l)} + \frac{\rho_l \cdot T_t/\pi}{(r + T_t)^2}$ $\propto (\rho_m \cdot H + \rho_l \cdot T_t)/W^2$

grain scattering effects exist in advanced process nodes [1]. To address these advanced features, a new field-based physical capacitance model is proposed in Chap. 6 [12]. The new model decomposes the electrical field into various regions and solves each basic component into a closed-form solution. Such a physical approach is convenient to incorporate new structures and materials, minimizing the complexity and the error in the model fitting process. Metal wire and contact/via resistance models are also developed, as listed in Table 7.2. In Table 7.2, the effect of electron scattering is considered: d is electron scattering coefficient; ρ_m and ρ_{bulk} are metal resistivity and metal bulk resistivity, respectively; A_{metal} and A_{liner} are the area of metal and the barrier metal liner, respectively; H is the height of contact/via; T_t and ρ_l are barrier metal liner thickness and resistivity; r is the radius of contact/via; and W is the structure width. Based on Tables 7.1 and 7.2, the PTM BEOL model projects the scaling trend of contact/via and metal resistances, and their variations. Figure 7.4 presents the comparison between model and silicon data for resistance of via, contact, metal 1 and metal 2 layers, with $\pm 3\sigma$ variation of metal resistance and $+3\sigma$ variation of contact/via resistance. The predictive model exhibits a close correlation with silicon data.

7.2 Exploratory Design of 22 nm CMOS Circuits

Through SPICE simulations, PTM offers an insightful pathway to evaluate the trends and tradeoffs of circuit performance metrics, under given low power design constraints. This section presents the benchmark study of representative

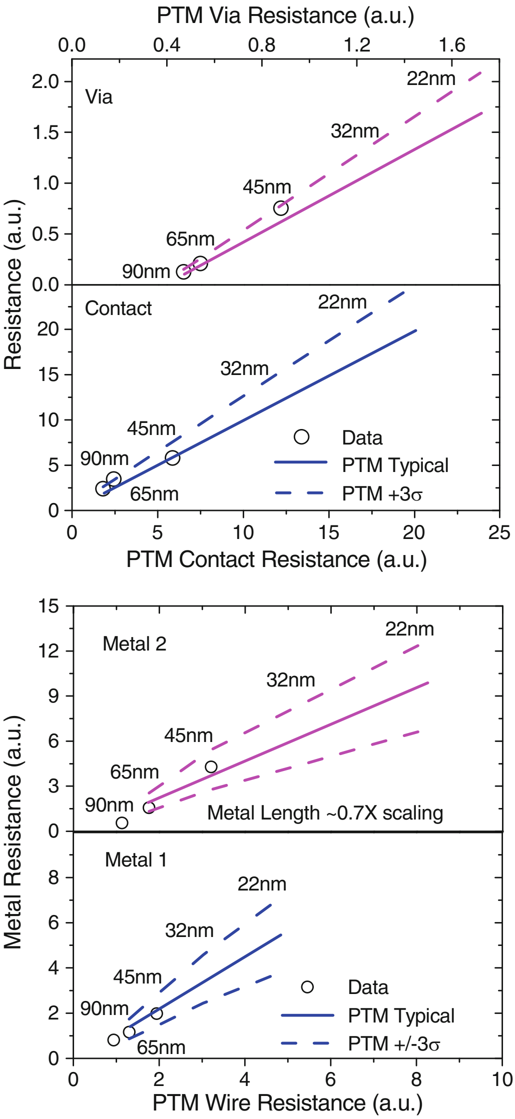


Fig. 7.4 Scaling trends of contact/via and metal wire resistance, for typical and +3σ values (Adapted from [9])

combinational and sequential circuit elements, as well as the impact of BEOL scaling. The customized PTM (Sect. 7.1) of both FEOL and BEOL serves as the basis for the simulation study. This exploratory approach allows designers to evaluate critical performance metrics with various technological components, and to start competitive design research before silicon data is mature.

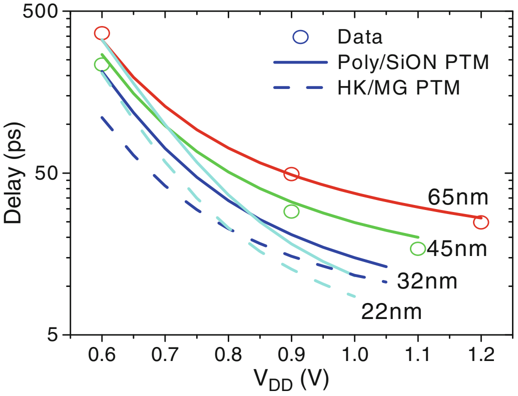


Fig. 7.5 The trend of RO delay under voltage scaling (Adapted from [9])

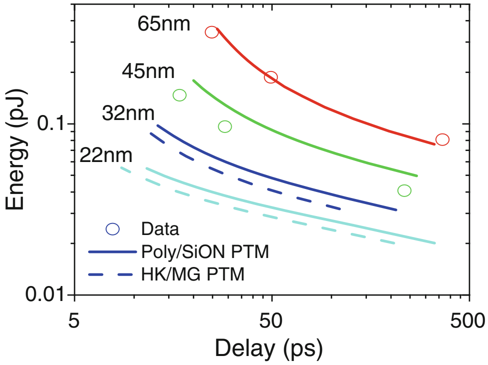


Fig. 7.6 The trend of RO energy versus delay under voltage scaling (Adapted from [9])

7.2.1 Ring Oscillator Delay and Energy

The first study is on self-loading ring oscillator (RO) that evaluates the driving capability of frontend-of-the-line (FEOL) transistors. Figure 7.5 shows a smooth reduction in the delay of an inverter-based FO = 4 RO. Note that the delay of 22 nm Poly/SiON RO is longer than that of 32 nm HK/MG device. For all generations, delay of RO rises rapidly as V_{DD} is reduced. During the V_{DD} scaling, the RO delay and dynamic energy trends are similar for all generations (Figs. 7.5 and 7.6). However, it is observed that the RO delay at 22 nm increases rapidly at lower supply voltage, possibly due to the strain effect. Dynamic energy of HK/MG RO is lower than that of Poly/SiON RO at a given delay, because of a lower V_{DD} of HK/MG RO.

Figure 7.7 evaluates the prediction of total power consumption for each RO stage at 10% duty cycle. A similar trend is predicted by PTM, as compared to available data. Under the same I_{off} target, HK/MG RO may not save total energy (i.e., standby and active energy) at the same voltage at 22 nm. However, a design

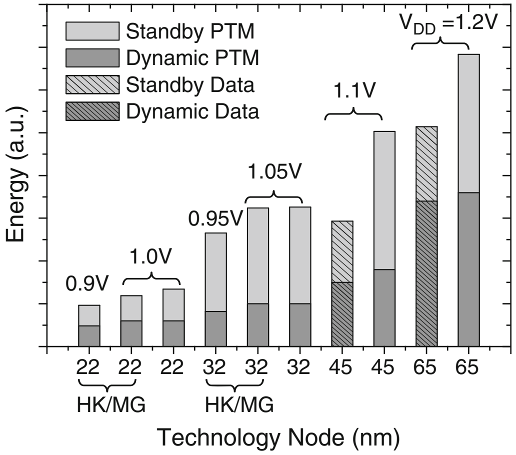


Fig. 7.7 The trend of total energy consumption at 10% duty cycle (Adapted from [9])

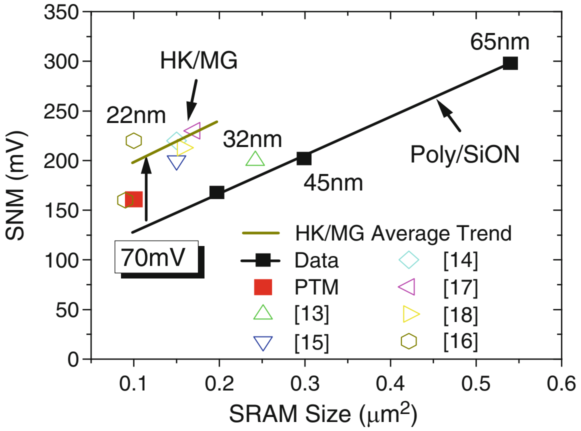


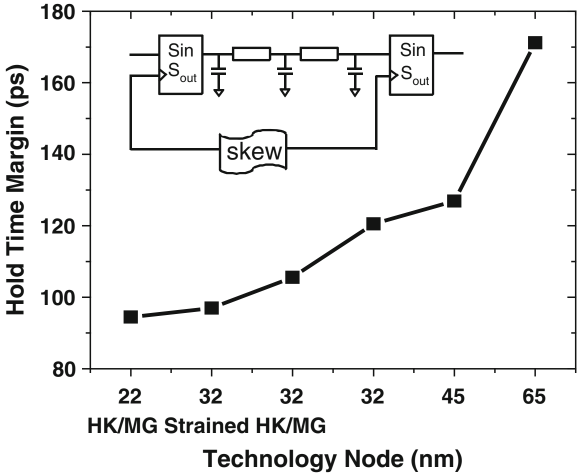
Fig. 7.8 The trend of SRAM static noise margin (Adapted from [9])

with HK/MG devices allows further V_{DD} reduction at the same I_{on} target, since HK/MG effectively boosts the drive current compared with Poly/SiON. Thus, it helps to reduce I_{off} and dynamic energy at lower V_{DD}. If duty cycle is below 10%, then total energy reduction is marginal at lower frequency because standby power will be dominant in that situation.

7.2.2 *Performance of Sequential Elements*

During technology scaling, one of the fundamental problems is the reduction of transistor switching characteristics, such as the I_{on}/I_{off} ratio (Chap. 1). Such degradation raises a considerable concern to sustain acceptable data storage capability in sequential elements. Figure 7.8 presents the scaling trend of static noise margin

Fig. 7.9 Hold time margin of flip-flops continuously declines with technology scaling



(SNM) of a 6-T SRAM cell. It illustrates that a 22 nm HK/MG device may be still able to provide adequate SNM [1, 13, 16–18]. This trend also illustrates that HK/MG devices produce better performance than Poly/SiON ones for the same SRAM size, benefiting from their enhanced drive current. SNM predicted from HK/MG PTM is slightly below the average of published data. The reason may be that the PTM model is generic for both logic and SRAM design, not specialized for a SRAM cell; the cell layout is not optimized for a scaled SRAM design either.

In today's synchronous sequential design, the margin of circuit timing changes with technology scaling and operation conditions. One important metric is the hold time margin of a flip-flop (FF) scan path, which is defined as the maximum clock skew between two FFs before hold failure happens (Fig. 7.9). As Fig. 7.9 demonstrates, hold time margin in the test of scan chain integrity continuously decreases with the scaling of FEOL device, BEOL interconnect, and V_{DD}. Such a trend induces lower design margin, posing an increasing challenge on robust synchronous design. The situation is further exacerbated by process variability and signal integrity issues in low power design [19]: hold time failure in a scan chain occurs even when there is zero clock skew. This phenomenon severally affects product yield, demanding new circuit techniques to improve the reliability of sequential circuits.

7.2.3 *Impact of BEOL Scaling*

As transistor delay is reduced, parasitic RC delay becomes relatively significant in total path delay. In addition, V_{DD} drop due to local wiring resistance increases as wire resistance keeps increasing with technology scaling. Because of these reasons, BEOL is increasingly important at 22 nm and beyond. For instance, Fig. 7.10 shows

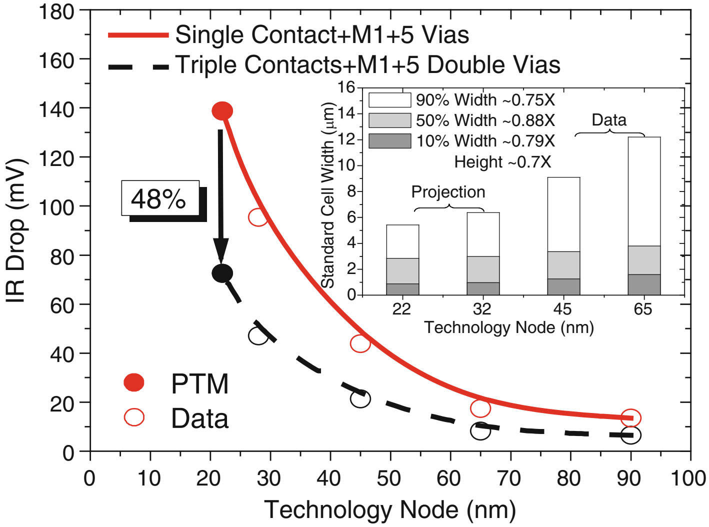


Fig. 7.10 V_{DD} drop at 500 μA (M6 to M1 by a stack via). The inset is the trend at 10/50/90% of cumulative standard cell width (Adapted from [9])

the IR drop at various technology nodes. Assuming M1 length is 50% of cumulative standard cell width of each generation, local V_{DD} drop from M6 to active region dramatically increases as metal resistance and, more importantly, contact resistance become larger during technology scaling. As contact resistance becomes more dominant, the adoption of triple-contacts effectively reduces the IR drop by 48% in a 22 nm design (Fig. 7.10); meanwhile, adding two single contacts, the RC delay of M1 wire increases by 79% at 22 nm. The adoption of double- and triple-contact reduces RC delay of M1 with two contacts by 50% and 62%, respectively (Fig. 7.11).

BEOL RC delay is increased significantly when FEOL delay, which is represented by RO delay in Sect. 7.2.1, is scaling down from 65 to 22 nm. RC delay of M2 plus two vias is about 80% lower than RC delay of M1 plus two contacts (Fig. 7.11). Therefore, M1 with two contacts may be more dominant in local routing delay if M2 has the same length as M1. As shown in Fig. 7.12, the delay gap between FEOL RO and BEOL interconnect reduces significantly: it is about only 10× and 20× at the 22 nm node for high performance (HP) and low power (LP) applications, respectively, assuming M1 and M2 is 20% and 80% length of 90% of cumulative standard cell width of each generation, respectively.

Finally, the impact of technology scaling on RO performance is examined, by integrating both transistors and metal wires. By decomposing delay and power into various components, we identify key factors that limit the performance and adaptively search technological or design solutions. Figures 7.13 and 7.14 show the decomposition of RO delay and dynamic energy, respectively. Assuming that M1 length in one RO stage is 90% of cumulative standard cell width of each generation,

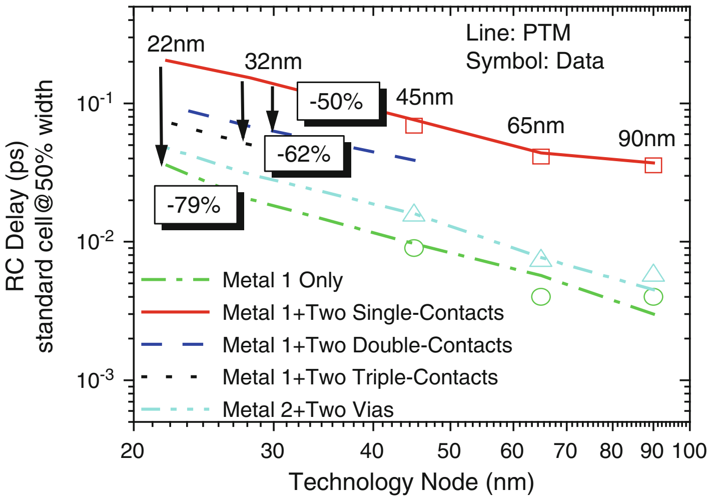


Fig. 7.11 Interconnect RC delay at 50% of cumulative standard cell width (Adapted from [9])

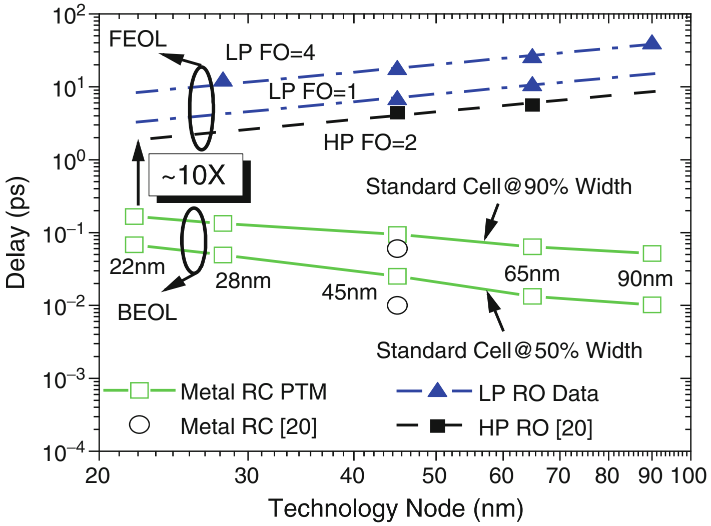


Fig. 7.12 The comparison between FEOL only delay and BEOL RC delay (Adapted from [9])

we observe that a larger impact on RO (FO = 4) delay comes from device intrinsic channel and gate fringe capacitance C_f (Fig. 7.1). Intrinsic channel delay component is reduced consistently. Nevertheless, scaling of gate delay due to gate fringe capacitance slows down. RO delay of HK/MG FEOL is smaller than that of Poly/SiON FEOL as expected. For dynamic power of RO (FO = 4), BEOL parasitics is the second largest component of dynamic energy at the 22 nm node. The intrinsic component of total dynamic power consumption is consistently reduced but the

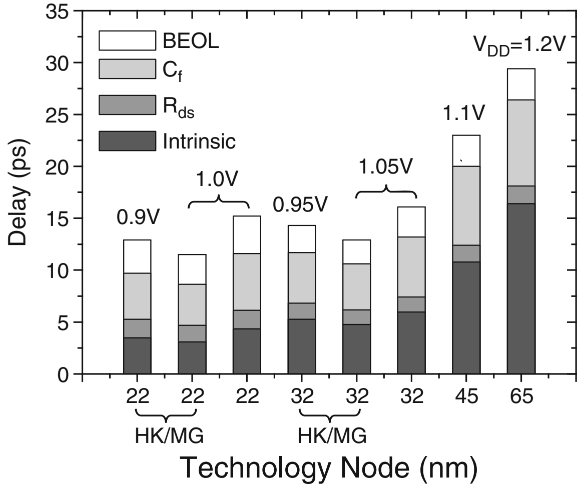


Fig. 7.13 The decomposition of RO delay (FO = 4), with two contacts per stage and M1 length at 90% of cumulative standard cell width (Adapted from [9])

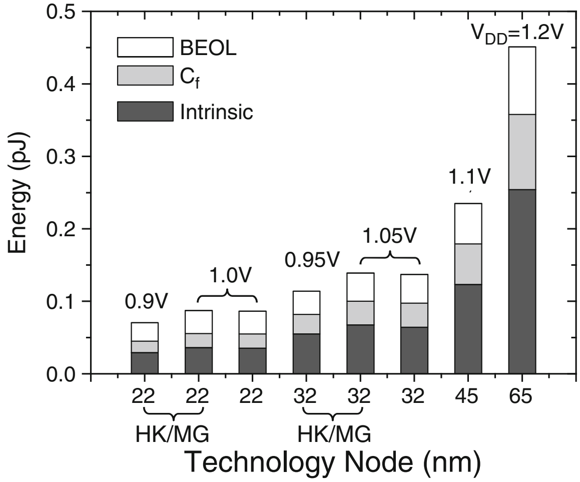


Fig. 7.14 The decomposition of RO dynamic energy with the same interconnect loading as that in Fig. 7.13 (Adapted from [9])

scaling of power due to BEOL parasitics slows down. As a result, the impact of BEOL parasitics becomes more significant on dynamic energy. In a brief summary, gate fringe capacitance becomes increasingly significant on RO delay, while BEOL parasitics play a more important role on power consumption.

7.3 Scaling Trend of Circuit Resilience

Technology scaling has an increasing impact on the resilience of CMOS circuits. This is a result of the escalation in both the amount of parametric variability and the sensitivity of circuit performance to various intrinsic and extrinsic variation sources

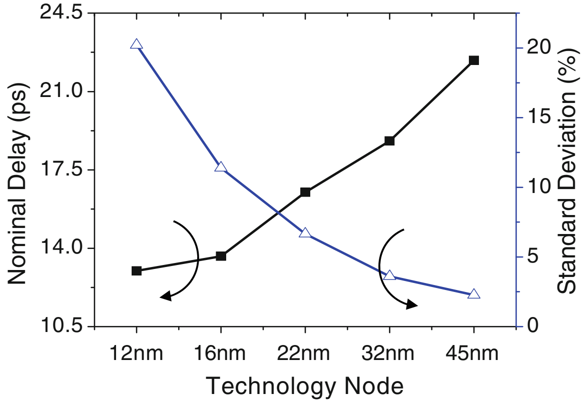


Fig. 7.15 The trends of nominal path delay and its variance in a seven-stage inverter chain (FO = 1)

[1, 20, 21]. Besides traditional manufacturing defects (e.g., via shorts or opens), the emerge of process variations and reliability degradation further exacerbates the failure rate of circuit operation, such as path delay in a synchronous design, data stability on-chip memory, power and other similar metrics. One canonical example is SRAM, where the need for cell density leads to using the smallest device feature size. These extremely small devices are highly susceptible to variations, such as random dopant fluctuations (RDF), line edge roughness (LER), and oxide thickness fluctuations (OTF). For a SRAM cell, excessive device mismatch may lead to scenarios where a particular bit cannot be reliably read or written, or where the data cannot be safely stored under low supply voltage. Another example is the widening delay distribution in logic paths. Depending on the clock frequency, it may cause incorrect logical value in the output register.

Based on nominal PTM and predictive models of intrinsic random variations (Chap. 4) [2, 22], this section targets to illustrate how continual technology scaling will cause current circuit failures to become much more pervasive, and to demonstrate the trends for future technology generations. Similar as that in previous sections, two representative circuits are benchmarked, including a seven-stage inverter chain (FO = 1) and a 6-T SRAM cell. Their performance variability is quantified through SPICE simulations.

Figure 7.15 presents the scaling trends of the nominal delay and its variance in the inverter chain, under random variations of RDF, LER, and OTF [22]. The P/N ratio is adjusted for each technology generation in order to achieve equal rise and fall times through the path. While the nominal path delay decreases with technology scaling, the standard deviation as a percentage of the mean value increases rapidly. Such a trend indicates the increasing importance of random variations on logic circuit performance. Furthermore, the path delay variability is decomposed into each individual factor (i.e., RDF, LER and OTF), as shown in Fig. 7.16. LER and OTF become more significant in advanced technology nodes. As gate length and oxide thickness are aggressively reduced, their variations due to atom-level randomness do not scale. Therefore, the impact of LER and OTF on device and

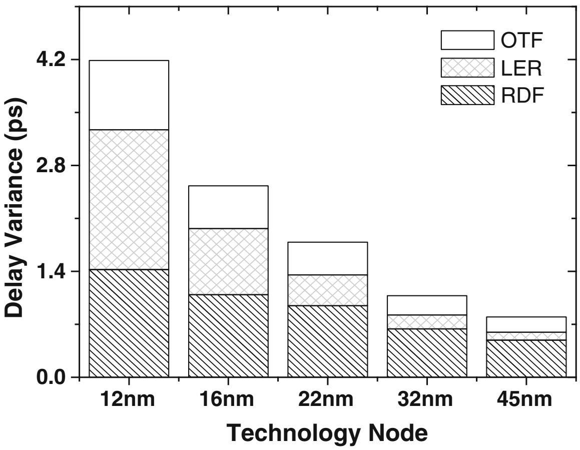


Fig. 7.16 The decomposition of delay variability to each variation source

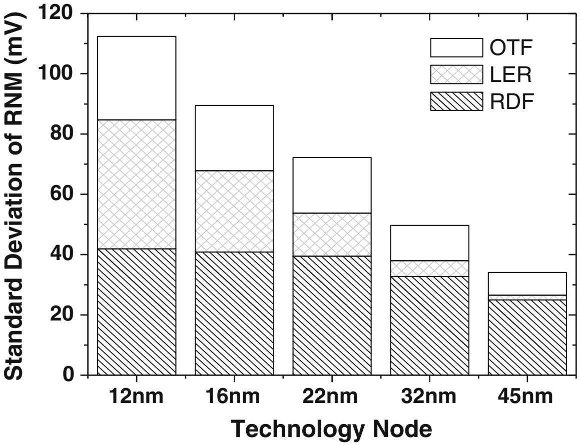
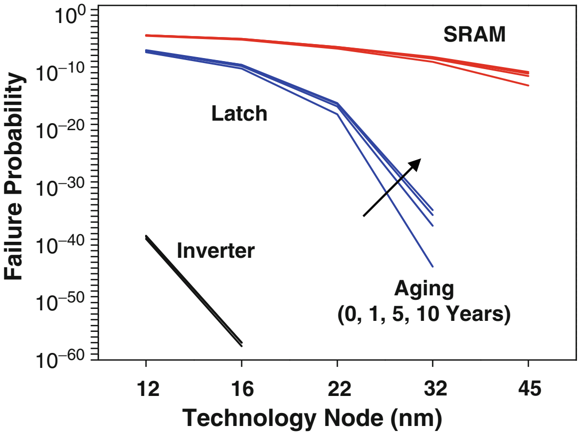


Fig. 7.17 The decomposition of RNM variability to each variation source

circuit parameters, especially V_{th}, is much more pronounced in a short-channel device. The other benchmark circuit in this study is a 6-T SRAM cell, which is well known for its higher failure rate than other circuit elements in the same technology. An SRAM cell may fail in many different ways, ranging from the readability, writability, data retention, to cell access time. For simplicity, static noise margins, such as read noise margin (RNM) and write noise margin (WNM), are monitored in the simulation. Since the SRAM cell uses the smallest device in the fabrication and is extremely sensitive to device mismatches, it exhibits a much higher failure rate than that of the inverter chain. Figure 7.17 illustrates the scaling trend of RNM variability and the contribution by each random variation source. While RDF dominates the variability in current technologies, the randomness in device geometry (LER and OTF) becomes the major contributor since the 22 nm node.

Fig. 7.18 Impact of NBTI on circuit failure probability (Adapted from [20])



Finally, Fig. 7.18 plots the impact of NBTI, which is the dominant aging mechanism [22, 23], on failure probabilities of an inverter, a D-type latch, and a 6-T SRAM cell [20]. In this case, the failure in the inverter is defined as the point where the inverter can no longer switches from one to zero (i.e., it appears to be stuck at one). This phenomenon happens if the PMOS device is too leaky or the NMOS device is too weak. The latch fails when there is a write latency violation, i.e., the D to Q delay is too long as compared to the clock cycle. for the SRAM cell, the write failure is considered. V_{th} shift due to NBTI over a desired time span is estimated assuming nominal V_{DD}, temperature, and 50% duty cycle [22]. The result illustrates a rapid increase in failure probability toward the end of the lifetime, especially in SRAM and the latch [20].

In summary, these design benchmarks focus on the manner with which technology scaling affects circuit performance, assuming constant circuit implementation styles and topologies. They help understand upcoming design issues and attempt to guide innovations at the device, circuit, and architecture levels. The benchmark infrastructure is simple and open to incorporate other studies, with the hope to promote research in the area of robust nanoscale design.

References

1. The International Technology Roadmap for Semiconductors (ITRS), 2008.
2. W. Zhao and Y. Cao, "New generation of predictive technology modeling for sub-45 nm early design exploration," *IEEE Trans. Electron Devices*, vol. 53, pp. 2816–2823, Nov. 2006.
3. W. Zhao, X. Li, M. Nowak, Y. Cao, "Predictive technology modeling for 32 nm low power design," *International Semiconductor Device Research Symposium*, TA4-03, 2007.
4. D. Greenhill, "Design for reliability in CMOS VLSI," Tutorial, *International Solid-State Circuits Conference*, 2002.
5. R. Heald, P. Wang, "Variability in sub-100 nm SRAM design," *ICCAD*, pp. 347–352, 2004.

6. T. Sakurai, “Perspectives of low-power VLSI’s,” *IEICE Trans. Electron*, vol. E 87-C, no. 4, pp. 429–436, April 2004.
7. K. Roy, et al., “Gate leakage reduction for scaled devices using transistor stacking,” *Trans. on VLSI*, vol. 11, no. 4, pp. 716–730, August 2003.
8. S. Rodriguez and B. Jacob, “Energy/Power breakdown of pipelined nanometer caches (90 nm/ 65 nm/45 nm/32 nm),” *ISLPED*, pp. 25–30, 2006.
9. X. Li, W. Zhao, Y. Cao, Z. Zhu, J. Song, D. Bang, C.-C. Wang, S. H. Kang, J. Wang, M. Nowak, N. Yu, “Pathfinding for 22 nm CMOS designs using predictive technology models,” *Custom Integrated Circuits Conference*, pp. 227–230, 2009.
10. Y. Cao, T. Sato, M. Orshansky, D. Sylvester, and C. Hu, “New paradigm of predictive MOSFET and interconnect modeling for early circuit simulation,” *Custom Integrated Circuits Conf.*, pp. 201–204, 2000.
11. W. Zhao, F. Liu, K. Agarwal, D. Acharyya, S. Nassif, K. J. Nowka and Y. Cao, “Rigorous extraction of process variations for 65-nm CMOS Design,” *IEEE Trans. Semiconductor Manufacturing*, vol. 22, pp. 196–203, Feb. 2009.
12. W. Zhao, X. Li, S. Gu, S. H. Kang, M. Nowak and Y. Cao, “Field-based capacitance modeling for sub-65 nm on-chip interconnect,” *IEEE Transactions on Electron Devices*, vol. 56, no. 9, pp. 1862–1872, September 2009.
13. K.-L. Cheng, et al., “A Highly scaled, high performance 45 nm bulk logic CMOS technology with 0.242μm^2 SRAM Cell,” *IEEE International Electron Devices Meeting*, pp. 243–246, 2007.
14. S.-Y. Wu, et al., “A 32 nm CMOS low power SoC platform technology for foundry applications with functional high density SRAM,” *IEEE International Electron Devices Meeting*, pp. 263–266, 2007.
15. C. H. Diaz, et al., “32 nm gate-first high-k/metal-gate technology for high performance low power applications,” *IEEE International Electron Devices Meeting*, pp. 629–632, 2008.
16. B. S. Haran, et al., “22 nm technology compatible fully functional 0.1 μm2 6 T-SRAM cell,” *IEEE International Electron Devices Meeting*, pp. 625–628, 2008.
17. S. Natarajan, et al., “A 32 nm logic technology featuring 2nd-generation high-k + metal-gate transistors, enhanced channel strain and 0.171μm^2 SRAM cell size in a 291 Mb array,” *IEEE International Electron Devices Meeting*, pp. 941–943, 2008.
18. F. Arnaud, et al., “32 nm general purpose bulk CMOS technology for high performance applications at low voltage,” *IEEE International Electron Devices Meeting*, pp. 633–636, 2008.
19. X. Zhang, et al., “On zero clock skew hold time failure in scan test,” *Solid-State and Integrated-Circuit Technology*, pp. 2070-2074, 2008.
20. S. Nassif, N. Mehta, Y. Cao, “A resilience roadmap,” *Design, Automation and Test in Europe*, pp. 1011-1016, 2010.
21. K. Bernstein, et al., “High-performance CMOS variability in the 65-nm regime and beyond,” *IBM J. Res. & Dev.*, vol. 50, no. 4/5, pp. 433–449, Jul./Sep., 2006.
22. W. Wang, S. Yang, S. Bhardwaj, R. Vattikonda, F. Liu, S. Vrudhula, Y. Cao, “The impact of NBTI on the performance of combinational and sequential circuits,” *Design Automation Conference*, pp. 364–369, 2007.
23. R. Vattikonda, W. Wang, Y. Cao, “Modeling and minimization of PMOS NBTI effect for robust nanometer design,” *Design Automation Conference*, pp. 1047-1052, 2006.
24. K. Mistry, et al., “A 45 nm logic technology with high-k + metal gate transistors, strained silicon, 9 Cu interconnect layers, 193 nm dry patterning, and 100% Pb-free packaging,” *IEEE International Electron Devices Meeting*, pp. 247–250, 2007.
25. Y. Ye, F. Liu, S. Nassif, Y. Cao, “Statistical modeling and simulation of threshold variation under dopant fluctuations and line-edge roughness,” *Design Automation Conference*, pp. 900–905, 2008.

Chapter 8
Predictive Process Design Kits

W. Rhett Davis and Harun Demircioglu

8.1 Introduction

For nearly half a century, semiconductor technology has continued to deliver exponential growth in the number of transistors on a chip. Even in the 22 nm processes of today, with exponentially increasing costs of research and development, masks, and design, transistors are still cheaper and denser than in previous process nodes. However, the cracks are showing in the industry's armor. Prior to 2005, each technology generation brought not only lower cost, but also more speed and less power consumption. Today, designers must be much more creative to balance the competing customer needs of cost, speed, and power. One size no longer fits all.

Semiconductor manufacturers have responded to this problem by offering a dizzying array of options to the designer: first multiple threshold voltages, then multiple supply voltages and gate-oxide thicknesses, then multiple standard-cell heights. At the same time, transistor and wire variation continues to increase, leading to larger and more complicated design rule decks and corner simulations. Finally, with the new emphasis on greater system level integration (e.g. "More than Moore"), the number of options will only continue to increase.

Such complexity creates a tremendous barrier to innovation. Global Foundries reports that the number of design starts in the first 5 years of development dropped from 1,012 in its 65 nm process to just 156 in its 22 nm process [1]. It is clear that removing barriers to innovation is necessary for the continuing health of the industry. Simpler process design kits are one solution to the complexity problem. The range of options must be reduced in order to make the design process more approachable. Reducing the number of options, however, is a risky move for a foundry eager to fill its fab lines.

W.R. Davis • H. Demircioglu
Department of Electrical and Computer Engineering,
North Carolina State University, Raleigh, NC 27695

Y. Cao, *Predictive Technology Model for Robust Nanoelectronic Design*,
Integrated Circuits and Systems, DOI 10.1007/978-1-4614-0445-3_8,

Simplified, predictive process design kits are the key to reducing that risk. A process design kit (PDK) is a collection of rules, models, and scripts for electronic design automation tools: everything needed a designer to complete his or her work before sending it to a foundry. A predictive PDK targets a predictive technology. Simplified, predictive PDKs can be used for market research, allowing a foundry to propose a set of possible options to potential customers to see which options will lead to the best products. Simplified, predictive PDKs can also promote innovative new electronic design automation (EDA) tools by providing a platform for research and development that does not disclose trade secrets.

This chapter presents the FreePDK™ [2], a simplified, predictive PDK for universities, targeting the least expensive CMOS process options at the 45 nm node. The FreePDK project began as a predictive PDK for teaching VLSI design, because the venerable scalable CMOS rules [3] and NCSU Cadence Design Kit (CDK) [4] that are typically used for teaching have not been used for fabrication in any technology node smaller than 180 nm. Since then it has grown to be used extensively by computer architecture researchers to create virtual prototypes and EDA companies to create virtual demonstrations. Version 1.3 of the FreePDK45™ has been downloaded by more than 900 individuals from its primary distribution site since its release in March of 2009.

The rest of this chapter presents the set of extensions beyond the scalable CMOS rules and NCSU CDK that were chosen for the FreePDK. It also presents an analysis of how successful those choices were in simplifying design exploration in newer technologies. Section 8.2 presents the choice of transistor types and models for the kit. Section 8.3 presents the transistor or front-end-of-line (FEOL) design rules. Section 8.4 presents the metallization or back-end-of-line (BEOL) design rules. Section 8.5 presents the lithography simulation rules. Lastly, section 8.6 presents the conclusions and a perspective for the future of process design kits.

8.2 Transistor Types and Models

The most important change in PDKs that was not included in the scalable CMOS rules is the multiplicity of transistor types. As supply voltages dropped below 1.8 V at the 180 nm node, threshold voltages and gate oxide thicknesses scaled to keep a constant electric field strength in the saturation region. Lower thresholds (V_T) brought the problem of higher channel off-currents (I_{off}), while thinner gate oxides (t_{ox}) brought the problems of increased gate-leakage currents (I_{gate}) and greater vulnerability to electro-static discharge (ESD) in the off-chip interfaces. A choice of V_T and t_{ox} that was well optimized for a 5 GHz processor was poorly optimized for a 500 MHz processor. Manufacturers responded to this problem by offering a range of V_T and t_{ox} values. Two thresholds were common for 180 nm technologies, and it is common to see 5 thresholds and 2 oxide thicknesses in 45 nm processes.

We chose to offer three V_T options to target the three options presented in the 2005 technology roadmap [5]: high-performance, low operating-power, and low standby-power. These devices were assigned the identifiers VTL, VTG, and VTH to

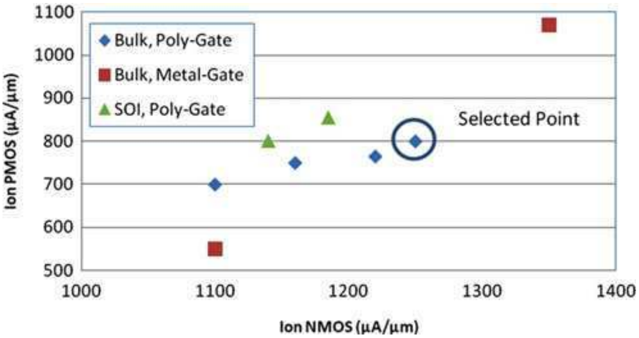


Fig. 8.1 On-current (Ion) values for commercial CMOS processes

indicate low, general, and high threshold voltages. Furthermore, a thick-oxide is necessary for off-chip interfaces. This device was given the identifier ThkOx. To indicate these options in the layout, four new layers were created, using the identifiers as layer names. These layers were considered modifiers to the well-shapes. Therefore, all shapes in threshold-adjustment and oxide-adjustment layers were required to be coincident with well shapes to avoid design-rule violations.

The next step was to choose a set of simulation models for these transistors. Published data on 45 nm transistors [6–13] show a wide range of options as well as differing choices of body-style (bulk vs. SOI) and gate-style (polysilicon vs. metal). Figure 8.1 shows the on-current (Ion) for these technologies for both NMOS and PMOS transistors. Clear trends for specific device types are not clear, and so we chose a point in the relative center of the range (labeled "Selected Point") for the FreePDK. This point is close to the best performance reported for a poly-gate, bulk technology, and so it was decided to assume this type of transistor for the FreePDK.

We also needed to choose values for leakage currents. Published technologies tend to report Ion values for specific value of subhreshold leakage, also called off-current (Ioff). These values are around 100–200 nA/µm for high-performance technologies, 20–30 nA/µm for general technologies, and 1–5 nA/µm for low-power technologies [6–13]. We chose to target 100 nA/µm for our high-performance transistors. Finally, maximum gate-leakage current-density (Jgate) in the range of 15–20 A/cm^2 is reported for bulk, poly-gate technologies with the high-K gate dielectrics [11, 14]. Because the published technology closest to the selected point in Fig. 8.1 used a high-K nitride oxide (SiON) dielectric [11], a value of 15 A/cm^2 was targeted for the FreePDK45.

The last step was to develop a set of simulation models for these devices. The Predictive Technology Model (PTM) [15] for 45 nm poly-gate bulk CMOS (V1.0) provided a starting point. This model needed to be tuned to match our target technology. Model parameters were adjusted as shown in Table 8.1. We began this process by pulling values for NGATE, NDEP, and XJ from the 2005 ITRS [5] for the 2007 technology node. Next, we lowered the values of long-chanel threshold (VTH0) and electrical gate oxide thickness (TOXE) until the target Ion was reached, searching for justifications in the literature. Although the PTM V1.0 value for

Table 8.1 Model parameters for the FreePDK45 highpeformance transistors

BSIM4 card	Description	NMOS	PMOS
NGATE (1/cm^3)	PolySi gate doping	3.0×10^{20}	2.0×10^{20}
NDEP	Channel doping	3.4×10^{18}	2.4×10^{18}
XJ (nm)	S/D junction depth	19.8	19.8
VTH0 (V)	Long channel threshold	0.322	−0.302
TOXE (nm)	Electrical gate ox. thick	1.14	1.26
ETA0	DIBL coefficient	0.006	0.0055
AIGC (Fs^2/g)$^{0.5}$m^{-1}	Parameter for I_{gcs} & I_{gcd}	0.02	0.0107
AIGSD (Fs^2/g)$^{0.5}$m^{-1}	Parameter for I_{gs} & I_{gd}	0.02	0.0107
Ion (μA/μm)	On-current	1,246	−801
Ioff (nA/μm)	Off-current	100	−100
Jgate (A/cm^2)	Gate current-density	15.3	−14.4

threshold is 0.466 V, a number of papers show threshold voltages of under 0.3 V for the longest channels [16–18], so a smaller value seemed justified. Also, though the PTM V1.0 model uses a low-K dielectric oxide thickness of 1.75 nm, several papers state equivalent oxide thicknesses in the range of 1.05–1.25 nm [14, 19] for the high-K dielectrics targeted for this work.

These parameter choices brought Ion into the target range, but Ioff was still too low. Therefore, the DIBL coefficient ETA0 was tweaked larger until Ioff matched the target of 100 nA/μm, bringing it closer to the value found in the PTM V2.1 models for high-K, metal-gate transistors. Finally, in order to raise Jgate into the target range, the gate current parameters were tweaked upwards to match the values for the newer PTM V2.1 models. The complete set of parameter changes, along with simulated values of Ion, Ioff, and Jgate, are given in Table 8.1.

8.3 Front-End Design Rules

The Front-End-of-Line design rules, which govern the semiconductor devices, are the primary determinant of the density of a process. Because the transistor-density of a process is the primary determinant of cost, it is important to understand how these rules have changed in advanced processes. The most important point for designers to remember is that manufacturers have pursued density first and foremost. Even though many front-end width and spacing rules have increased in advanced processes, transistor pitch has not increased. The minimum area of a standard-cell or other macro-cell in an advanced process can be accurately predicted by simply scaling the area from an old design by the feature-size of the new technology. However, such high density can yield transistors with poor characteristics. Designers may want to increase certain transistor dimensions in order to get better performance. For this reason, advanced processes tend to have a large number of recommended rules. This section presents a simplified approach to understand how these rules have changed and how they affect transistor behavior.

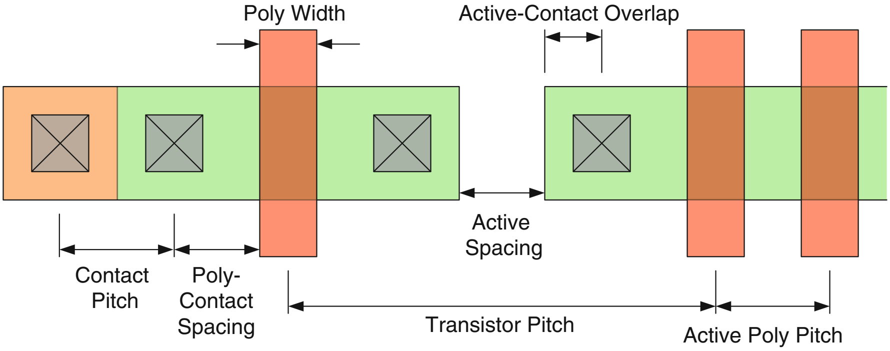


Fig. 8.2 Design rules that determine standard-cell width

To understand the trends, we present the design-rule differences between our 45 nm technology and the three flavors of MOSIS scalable CMOS (SCMOS) rules [20]. The original SCMOS rules were developed for 1–3 μm processes and were based largely on the work of Mead and Conway [3]. These rules were modified for technologies below 1 μm as the "submicron" (SUBM) rules. For 250 nm and 180 nm technologies, these rules were again updated as the "deep submicron" (DEEP) rules. The rules have not been used for smaller technologies, for reasons that will be described here.

8.3.1 Width-Affecting Rules

The simplest way to understand advanced front-end design rules is to recognize that transistor pitch is largely unchanged. The width of a standard-cell can be determined by six primary values, as illustrated in Fig. 8.2. For ease of comparison to the SCMOS rules, these values are listed in Table 8.2 in units of lambda (λ), which is one-half of the minimum poly width[1]. The poly width defines the transistor length, and all other design rules can be understood by their relation to this minimum value. The active spacing reflects the ability to isolate transistors. Table 8.2 shows these two values as largely unchanged across all technologies.

The differences in advanced technologies begin with the contact rules. Because wire resistances have become the limiting factor in interconnect, wire thicknesses have increased while contact area has decreased. These factors lead to ever-increasing contact resistance. Table 8.2 shows that contact pitch has increased in advanced technologies, in order to reduce this resistance. In order to prevent increasing contact

[1] Even though many 45 nm and smaller technologies use metal for transistor gates instead of polysilicon, design rules have not been affected by this change. Here we use the term "poly" for the gate conductor, regardless of whether it is manufactured as metal or semiconductor.

Table 8.2 Trends in width-affecting design rules, in units of lambda (λ)

	SCMOS	SUBM	DEEP	FreePDK45
Poly Width	2	2	2	2
Active Spacing	3	3	3	3.2
Contact Pitch	4	4	4	5.6
Active-Contact Overlap	2	2	2	1.5
Poly-Contact Spacing	2	3	3	2.7
Transistor Pitch	15	15	17	13.6
Active Poly Pitch	4	5	6	7.6

size from adversely affecting transistor density, the active-contact overlap shrank in 45 nm, even though it had held steady for so many process generations. The poly-contact spacing increased in submicron technologies, because it greatly affects the variability of transistor properties. This value has held relatively steady in advanced technologies, as Table 8.2 shows.

The sum of these effects is that transistor pitch began to increase as the SCMOS rules scaled to the 180 nm DEEP rules. Such an increase meant that these rules were no longer useful, since using them would mean that designers would be wasting area. Our 45 nm design rules, however, show a density that is higher than the SCMOS rules, which is much more in line with published transistor density in 45 nm technologies. This makes the FreePDK45 suitable for architecture studies at the 45 nm node.

Another important note is how the poly-over-active spacing has increased. Table 8.2 shows how this rule began increasing below 1 μm and continues to increase today. This rule has little to do with the ability to print these features and more to do with the increase in transistor variability caused by this proximity. Until lithographic techniques are developed with reduced variation, this trend is likely to continue. Luckily, this tends to have little effect on overall transistor density, because it only occurs on transistors with shared source-drain regions without a contact. Our comparisons show that these cases are rare enough that they have little effect on standard-cell width.

8.3.2 Height-Affecting Rules

Although standard-cell width tends to be easily predictable, standard-cell height does not. Figure 8.3 shows the three rules that have the greatest effect. The first to consider is the minimum transistor width. In all variants of the SCMOS rules, minimum transistor width held steady at 3 λ. In 45 nm technologies, we see a wide variation between foundries in the this value, some as low as 4 λ, some as high as 8 λ. This variation seems to have less to do with a foundry's ability to print active areas and more to do with how much variation they are willing to permit in their transistor characteristics. Foundries less willing to permit variations will have wider transistors. Our approach with the FreePDK45 has been to choose a low value for

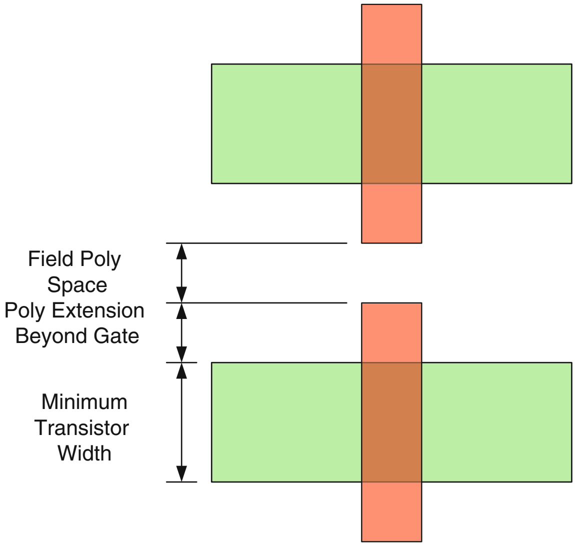


Fig. 8.3 Design rules that determine standard-cell height

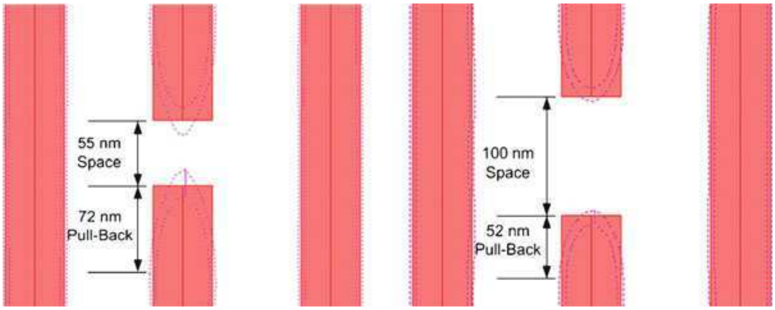


Fig. 8.4 Simulated poly trace pull-back for varying space

active width (3.6 λ) and to pursue techniques for prediction of device variation. The choice for a minimum transistor width will have a huge effect on standard-cell height, because the widths of all transistors tend to be chosen based on multiples of the minimum-sized transistor's dimensions [21]. Standard-cell heights also vary depending on whether a library is targeted for high performance or low power. Some libraries are based on a transistor that is wider than the minimum, because it reduces delays at the expense of increased power.

The poly-extension and field poly space rules also have a significant effect on standard-cell height. In all variants of the SCMOS rules, the poly extension was identical to the poly width. In technologies smaller than 180 nm, however, there tends to be large variations at the end of poly traces. Figure 8.4 shows an example of this variation, simulated with the lithographic model distributed with the FreePDK45. The dotted-lines illustrate the process-variation bands (PV-bands) indicating the range of shapes that are likely to be printed. Because the poly

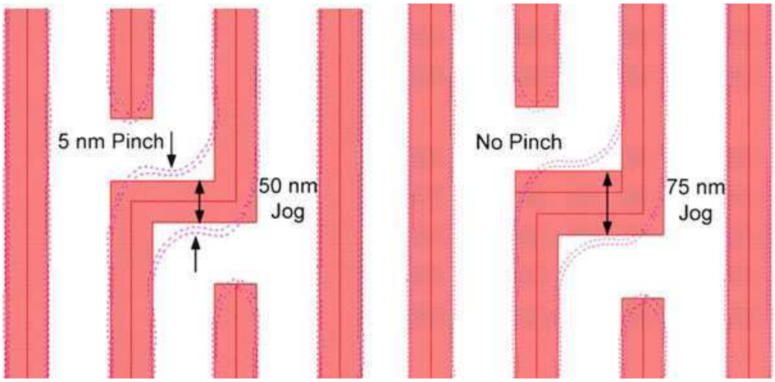


Fig. 8.5 Simulated pinching of a poly jog for varying width

space represents an irregularity in the pattern of repeating poly traces, it is difficult to create a mask that prints the space. The sharper the irregularity, the less the PV-bands conform to the desired shape. If this line-end were an extension beyond the edge of a gate, then there would be significant variation in the length of the transistor. For this reason, the poly extension has increased beyond 2 λ in advanced technologies. In our simulation, widening the space from 55 nm (2.2 λ) to 100 nm (3 λ) decreases the pull-back from 72 nm (2.9 λ) to 52 nm (2.2 λ). The FreePDK45 rules use 3 λ for the poly space and 2.2 λ for the poly extension, but other technologies go as high as 4 λ and 3 λ for these rules. The value of this rule in the long term is dependent on how much variation there is in the poly patterns.

Finally, the irregular poly width also plays a role in standard-cell height. As stated above, regular patterns are the easiest to print. An irregular segment in a poly-line, such as a jog, bend, or branch, is difficult to print. A wider segment represents an irregularity that is less sharp and easier to print. Figure 8.5 shows a simulation in which a jog was widened from 50 nm (2 λ) to 75 nm (3 λ), which eliminated pinching of 5 nm (0.2 λ). At the moment, the FreePDK contains no irregular poly width or space rules, because deciding on values and coding the rules for a rule-checker are non-trivial. Commercial kits contain tens of special rules governing the many cases of possible poly bends, jogs, and branches. Fortunately, this rule has only a minor effect on standard-cell height and can be safely ignored in computer architecture studies.

8.3.3 Antenna Rules

The last density-affecting rule that needs to be included in advanced PDKs is the antenna rule. The antenna rule is needed to prevent gate-oxide breakdown during manufacturing. Because wires are taller (thicker) than they are wide in advanced processes, directional plasma etching must be used to fabricate them. Plasma etching involves the ionization of an etching gas and the creation of an electric field around the wafer, causing the ions to impact the metal and remove it in the undesired locations. On impact, the ionized molecules transfer their charge to the metal. In many cases, this charge can build up to the point that the voltage on a net exceeds the breakdown voltage of a transistor gate.

Fortunately, there is a simple solution to this problem. Source and drain junctions are engineered to reach non-catastrophic Zener breakdown at a lower voltage than the transistor gates. As long as each gate is connected to a reverse-biased junction diode, any excess charge left during etching safely flows through the diode into the substrate. This is the same approach used to create electro-static discharge (ESD) protection on pads. The drawback to this approach is that space must be made for these diodes between the transistors, which can reduce density. These diodes also add to the capacitance of a net and contribute a small amount of leakage power. Therefore, the antenna rule is needed to discover which nets require these protection diodes.

The basic theory behind the antenna rule is that tunneling current density through the oxide must be kept below a certain threshold. Collisions between electrons and impurities in the gate oxide during tunneling can create low resistance paths through the oxide. Direct tunneling is generally assumed to dominate over Fowler-Nordheim tunneling. The direct-tunneling current density can be calculated from the oxide thickness t_{ox} and voltage V_{ox} as follows [22]:

$$J_{gate} = \frac{q^2(\phi_B - V_{ox}/2)}{2\pi h t_{ox}^2} exp\left[-4\pi t_{ox}(2qm*)^{1/2}(\phi_B - V_{ox}/2)^{1/2}/h\right] \qquad (8.1)$$

where q is the electron charge, h is Planck's constant, ϕ_B is the barrier height of the metal-oxide interface, and $m*$ is the tunneling electron effective mass. If the amount of incident charge per second on the metal during etch is known along with C_{ox}, then the incident charge on the metal can be related to J_{gate} through this equation to compute the maximum amount of charge that can be collected during the etch. This calculation simplifies to a simple ratio of exposed metal area to transistor gate area. If the calculated ratio for a node is below the limit, then no diode is needed. Otherwise, a diode must be connected to the node.

One confusing aspect of the antenna rule is that exposed metal area is calculated for every layer of metal and includes only the metal connected during that processing step. Figure 8.6 shows an example of the shapes considered for a sample layout during the antenna checks for poly, metal1, and metal2. Charge collects on exposed metal during each step, and is not released until connection is made to a junction diode. The metal1 check must include both the area for metal1 and poly. The metal2 check includes all three layers as well as a small strip of metal1 that was not previously connected. In this final check, the added strip of metal1 is not a problem, because it connects the node to a diode, eliminating the need to observe the antenna ratio. For the first two checks, however, the maximum allowed ratio must be met.

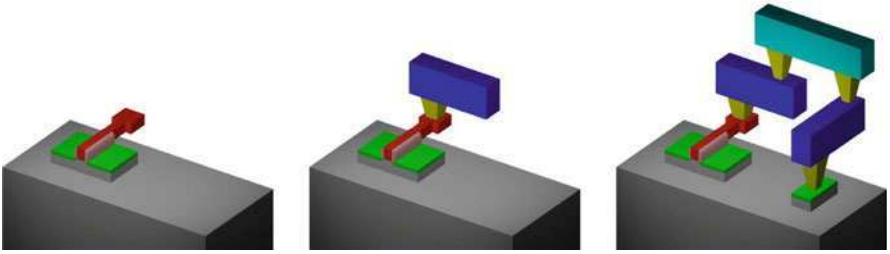


Fig. 8.6 Example of shapes considered during antenna rule checks for poly, poly-metal1, and poly-metal1-metal2

Another confusing aspect of the antenna rule is the manner in which exposed metal area is calculated. During the etching, charge is collected on the sides of a shape in addition to the top but at a rate that varies as the etch progresses. Some antenna rules therefore calculate the "exposed area" of a shape with a function that includes both area and perimeter, effectively the surface area of the metal trace (not including the bottom). For simplicity, the current version of the FreePDK ignores this complication and uses a simple area ratio.

Finding a value for the FreePDK antenna ratio has been problematic. There was a great deal of published research on the topic when the 180 nm technology node was introduced, but little details have emerged since then. A commonly accepted maximum ratio of 1000:1 (exposed metal area to gate area) was common for 180 nm [23], but this ratio assumed a maximum allowed J_{gate} during etch of 0.02 A/cm^2. As shown in Table 8.1, J_{gate} for a typical 45 nm transistor during normal operation is 1,000 times this value. It is possible that this rule will diminish in importance as gate tunneling currents become more common. Bang et al. [24] claimed that oxide charging currents were unlikely to increase for t_{ox} below 1.5 nm, which is thinner than the typical t_{ox} for a 45 nm process. Weng et al. [25] later concluded that plasma damage is negligible for t_{ox} below 1.5 nm. However, antenna ratios have dropped below 1000:1 in commercial PDKs. One possibility is that increased electric fields are needed to make wires with taller aspect ratios, but this is supposition. The reason for these decreasing ratios appears to be unknown. We chose a maximum ratio of 300:1 for the FreePDK to be 1/3 of the maximum ratio for the 180 nm node, but this is arbitrary. We further chose a 1/3 smaller ratio of 100:1 for the poly antenna check, because foundry rules typically allow a smaller ratio for that check only. These rules provide a valuable learning tool for users of the FreePDK, but their accuracy is questionable.

Fortunately, antenna protection diodes are uncommon enough that their impact on density is minimal. In custom designs, they tend to be placed systematically to keep antenna ratios much smaller than the maximum, lest the layout need to be reworked and much time lost. In standard-cell designs, there tends to be enough vacant area between cells to create these diodes as needed in the gaps when routed. We therefore include the rule mostly to inform users of the FreePDK of this hazard in advanced processes.

8.4 Back-End Design Rules

Back-end-of-line design rules govern the metallization for a process. These rules determine density for designs that are wire-limited. They also govern the way that global signals and power are distributed on a chip, and so it is important to understand how these rules have changed in advanced processes. Here we present the most significant changes to vias and spacing. We also present a typical metal stack in an advanced process and changing capacitance models.

8.4.1 Via Rules

As wire widths have decreased, wire thicknesses have increased to keep resistance as low as possible. Chemical-mechanical polishing (CMP) has allowed the stacking of an arbitrary number of metal layers, but uniform layer thicknesses are very difficult to control. Therefore, the inter-layer dielectrics (ILD) have also increased in thickness. This increase makes the manufacturing of a reliable via hole more difficult. Via areas have increased relative to wire width to accommodate this change.

Via rules have changed most significantly in that metal enclosure of a via is no longer required. Extension of metal on two opposite sides tends to be required. Figure 8.7 shows an illustration of this rule in the λ-based SCMOS rules and an advanced process. Metal enclosure of a via on all sides used to be required in order to handle the worst-case overlay misalignment. This resulted in a via enclosure that was 1 λ wider than the typical 3 λ width and space rules for metal1, but this had minimal effect on wire density. Observing this rule in an advanced process would lead to a blockage of the two adjacent wire tracks. Therefore, an extension is required on two opposite edges only. The disadvantage of this approach is a dramatic increase in worst-case via resistance, which is roughly 5 times higher for contact and low-level via layers in 45 nm processes (10–50 Ω) compared to 180 nm processes (2–10 Ω). For higher levels of metal, worst-case via resistances

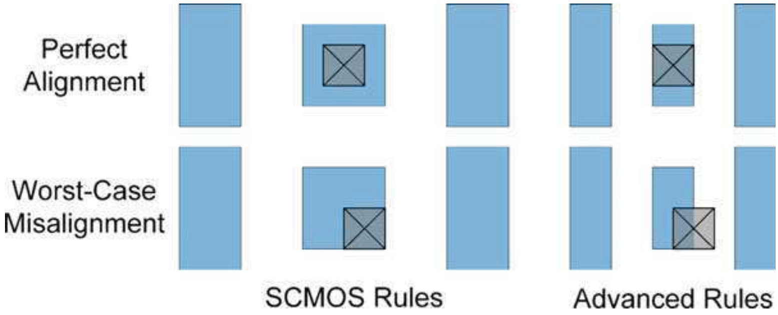


Fig. 8.7 Via rule changes in advanced technology nodes

drop sharply, because wire widths are larger, and the misalignment is a smaller proportion of the total via area.

To chose a new extension rule for the FreePDK, we turned to the ITRS [5], which publishes a “3σ overlay” tolerance for alignment at each node. Foundry rules typically require extension on two opposite sides that is around three times this value. This leads to an opposite-side-extension rule that is 1.4 λ, slightly larger than the 1 λ enclosure required by the SCMOS rules. For higher levels of metal, where the minimum width is more than 4 times this extension, the rule is simply dropped, and no extension is required. The approach allows the metal rules in the FreePDK to minimize blockages to adjacent wire tracks and target the maximum wire density possible.

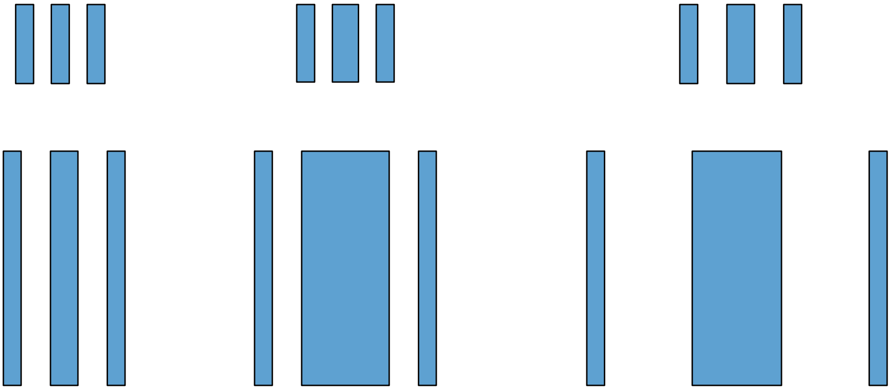
8.4.2 Variable Spacing and Density Rules

The last rule that designers need to be familiar with in advanced technologies are metal spacing rules that vary with shape width and length. The large number of these rules is very confusing and difficult for most designers to track. Of the 82 individual rules defined for the FreePDK, for example, 28 of them (just over 1/3) are variable spacing rules. Commercial PDKs have see a similar mulitplication of the number of rules for each metal layer.

The reason for the increase in the number of rules is that the simple rule is even more difficult for designers to follow. Variable spacing rules arise from a need for uniform metal and dielectric thicknesses. The CMP techniques used to fabricate each layer depend on the assumption that roughly the same amount of material must be removed at every location during processing. If there is great variation in the density of a metal layer, then there is also great variation in the resistance and capacitance of every metal trace, making it more difficult to guarantee that delay constraints will be met. The typical way to express a density rule is to pass a window over the entire design and check the density of metal for that window to ensure that it is within a certain range (25–75%, for example).

When CMP techniques were first introduced (around the 250 nm node), density rules were expressed as a window size and density range. Unfortunately, these rules tended to cause an unnecessary reduction in productivity for custom designs. Most custom designers create shapes that are much smaller than the window, which meant that the rule could never be met and was ignored. Upon assembly of the larger design, however, if the density rule was not met, then a tremendous amount of time would be required for unexpected re-design. Variable spacing rules are a way to impose a set of constraints at design-time to ensure that density rules can be met.

These variable spacing rules themselves tend to vary greatly from one foundry to the next, making it difficult to come up with a rule for the FreePDK. These rules tend to be added late in the development of a process, after the metal stack has been finalized. Foundries try to make them as simple as possible by providing a limited



Rule	Value	Description
METAL1.1/2	65 nm	Minimum width and space of metal1
METAL1.5	90 nm	Minimum spacing of metal wider than 90 nm and longer than 300 nm
METAL1.6	270 nm	Minimum spacing of metal wider than 270 nm and longer than 900 nm

Fig. 8.8 Illustration of variable spacing breakpoint layouts for 3 rules

number of breakpoints that are based on integer multiples of the minimum metal width and space. We followed the same approach with the FreePDK. The simple way for designers to understand these rules is to visualize the set of breakpoints and relate them to the density constraint.

Figure 8.8 shows the breakpoint layouts defined by three of 28 variable spacing rules in the FreePDK45. Examination of the density of these layouts shows that it is never below 25% (excluding the adjacent metal shapes) or above 63% (excluding the metal shapes). Furthermore, the aspect ratio of the windows with and without metal are always between 0.75 and 1.6. Examination of the complete set of breakpoint shapes for the FreePDK45 across would show that the density is constrained between 25% and 90% with aspect ratios in the range of 0.30–3.3. In these respects, there is little variation from one foundry to another. Some foundries omit the length constraint from the spacing rule, which effectively removes the aspect ratio constraint from the window. Also, the 90% maximum density observed with the FreePDK rules are higher than generally observed in commercial rules, in which the maximum density tends to be closer to 80%.

8.4.3 Metal Stack

The lambda-based SCMOS rules were never intended to support more than three metal layers. The 250 nm and 180 nm variants of the SCMOS rules show a doubling of the number of metal layers with an equal doubling of the number of rules.

Table 8.3 Metal stack in the FreePDK45

Name	Pitch (width/space) (nm)	Thickness (nm)
ILD 9		2,000
Global (9–10)	1,600 (800/800)	2,000
ILD 7-8		820
ThinGlobal (7–8)	800 (400/400)	800
ILD 4-6		290
Semi-global	280 (140/140)	280
ILD 2-3		120
Intermediate (2–3)	140 (70/70)	140
ILD 1		120
Metal 1	130 (65/65)	130
Poly-Dielectric		85
Poly	125 (50/75)	85

A 45 nm process can have a further doubling of the number of metal layers. This exponentially increasing number of rules can be hard to for designers to track. Fortunately, foundries do tend to use a simple approach to definition of these metal layers. The poly, metal1 and metal2 rules are made as tight as possible to guarantee high-density local connections between transistors. Higher levels of metal generally increase in width and thickness based on an integer multiple of metal1 or metal2. Because varying metal widths complicate the problem of routing, these metal layers are organized in groups of similar width (such as intermediate, semi-global, global or 1X, 2X, 4X). It is expected that each set of interconnect layers will be used at a different level of design hierarchy. Some foundries also offer a “Thin-Global” metal layer, which is a variant of the global layer that is much thinner. Global layers have much less resistance, making them better suited for delivering power. Thin-Global layers have much less coupling capacitance to adjacent wires, making them better suited for signal wires.

Table 8.3 shows the metal stack assumed by the FreePDK45. This stack was derived from a merging of stacks offered by Toshiba [10] and IBM [12], using thickness information from the ITRS [5]. It is important to note that there is little variation among foundries up to and including the semi-global level of interconnect. The differences for higher levels of metal are primarily due to the differing wiring requirements of the products each foundry makes. Computer architects looking for guidance may want to assume complete freedom of choice for the width, space, and thickness of higher metal layers, provided that the density requirement are met (as described in the previous section) and that the aspect ratio of a wire (height/thickness) never rises above two. Foundries often grant customers’ requests to tweak the metal stack, provided that the volume of requested chips is high enough. This is naturally not an option for designers participating on multi-project wafer (MPW) runs, such as the ones organized by MOSIS [20].

8.5 Lithography Simulation Model

The rules documented so far in this chapter are still not sufficient to capture the nuances of lithographic variation. A large number of rules in commercial PDKs are devoted to constraining designers in various ways, depending the yield that they hope to achieve. Rather than attempt to re-create this complexity, we chose with the FreePDK to document a typical lithographic simulation model for an advanced technology, in order to popularize the use of lithographic simulation as a tool for understanding advanced design rules.

The goal of lithographic simulation is to create process variability bands (PV-bands) which show how much design objects may vary due to focus and dose imperfections of an exposure system. Lithographic simulation assumes the use of a set of resolution enhancement techniques (RET), such as optical-proximity correction, phase-shift masks, *etc.*, to allow printing of features smaller than the wavelength of the light used for exposure. The resolution enhancement flow is very costly and time-consuming and is typically performed by the foundry after the complete layout is finalized. Therefore, it is impossible to know at design-time exactly what recipe will be used. Litho-Friendly Design (LFD) refers to the estimation of the RET flow during the design phase. Figure 8.9 illustrates the LFD flow. After the RET recipe is estimated, process variation experiments are simulated, which include a set of off-focus and off-dose conditions. Based on these experiments, as set of PV-bands such as the ones shown in Figs. 8.4 and 8.5 can be determined and superimposed on the layout.

Here we document the FreePDK lithography model. The first group of parameters is related to the optical models, which include the physical properties of the illumination system used in photolithography. According to the Rayleigh criterion, resolution, in other words achievable half pitch of a lens is given by Eq. 8.2 [26].

$$R = k1\frac{\lambda}{n\sin\alpha} = k1\frac{\lambda}{NA} \tag{8.2}$$

where n is the index of refraction of the medium between the lens and the mask, α is the acceptance angle of the lens, which is the measure of the ability of the lens to collect the diffracted light, λ is the wavelength of the light source and NA is the numerical aperture of the lens. In addition, k1 is an experimental parameter, which

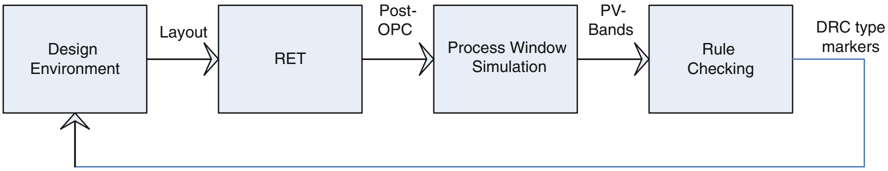


Fig. 8.9 A typical litho-friendly design (LFD) flow

depends on the lithography system and resist properties [26]. Its value is around 0.25–0.5 in modern lithography systems [10]. The nominal values of these parameters are taken from the ITRS [5]. In advanced processes, ArF lithography is used which has a wavelength of 193 nm. Since the minimum half pitch of the FreePDK 45 nm technology is much lower than this wavelength, a numerical aperture greater than 1 is needed. Therefore, immersion in water is assumed, which gives an index of refraction of 1.44. In the FreePDK rules, the minimum half pitch is 65 nm, which requires a numerical aperture of 1.2, which is common for modern 45 nm lithography systems. The geometry of the exposure system also affects the printing ability dramatically. Through trail-and-error simulations with Mentor Graphics® Calibre LFD™ simulations, we eventually found that an annular illumination system using 4X reduction matched well with published images [10, 12, 27].

The second group of parameters is related to the photoresist films. In photolithography systems, a wafer is coated with a photoresist material so that mask objects can be transferred to it. The thickness of the photoresist and the refraction and absorption indexes of the photoresist material highly affect the resolution of the process. In addition, to minimize the reflection from the wafer surface, bottom anti-reflective coating (BARC) material is also employed below the photoresist material, to further improve the resolution. The material properties were determined by literature survey [28] and tuned again by trial-and-error simulations. The thickness of the photoresist material is 90 nm with an index of refraction (n) of 1.71 and index of absorption (k) of -0.015. For the BARC material, the thickness is 40 nm, $\mathrm{n} = 1.82$ and $\mathrm{k} = -0.034$, which shows that it has more refraction and absorption than the photoresist material, hence enhancing the mitigation of reflections from a wafer surface. In addition, the process models include minimum light intensity for wafer printing at the surface of the resist. In other words, the light intensity below this threshold value does not change the photoresist properties, so it cannot print an image. The normalized intensity threshold value is found to be 0.25 after simulations.

Our model also assumes the use of attenuated phase shift masks (ATT-PSM), which are widely used to improve resolution [5]. An attenuation factor of 0.06 is assumed.

In order to simulate the process window, reasonable limits of variation in focus and dose must be known. Depth of focus for an exposure system can be estimated by the Eq. 8.3 [26].

$$DOF = \pm k2 \frac{\lambda}{(NA)^2} \tag{8.3}$$

where k2 is an experimental parameter of around 0.5. For the defined exposure system, the depth of field is in the range ±70–120, and so the worst-case defocus for the flow was estimated to be ±75 nm. The worst-case dose variation was determined to be $\pm 5\%$.

This model provides sufficient information for designers to create the technology files for a variety of lithographic simulation tools. The FreePDK45 includes a set of lithographic simulation rules for the Calibre LFD tool. This tool allows the definition of design rules based on the generated PV-bands, rather than the user-defined shapes. It is tempting to think that LFD simulation can replace traditional DRC. However, there is a limitation to such an approach. Since the RET changes with the layout, any errors in the layout will influence the LFD results. It is easy for minor errors (such as tiny notches) in the layout to cause generation of a RET that widely diverges from what the foundry would likely use. Our experience shows that these errors can lead to such wide variations as shapes that completely disappear or merge at different process corners. Current LFD tools are not capable of detecting such errors, but they are easy to detect with traditional DRC. Still, the combined use of DRC and LFD checks may eventually prove to be an effective way to reduce the complexity of design rules.

8.6 The Future of Process Design Kits

With the complexity of design rules increasing and the number of design starts falling, there has been increasing pressure to reduce or somehow manage this complexity. Recently, there have been three significant efforts aimed at standardizing PDKs, in order to bring the semiconductor industry together on common solutions. The first of these efforts is the PDK checklist, published by the Global Semiconductor Alliance (GSA, formerly the Fabless Semiconductor Association or FSA) since 2004 [29]. This checklist is more of a minimum list of ingredients for documentation, however, rather than an interface standard.

The second effort is the Interoperable Process Design Kit (iPDK™), introduced by Taiwan Semiconductor Manufacturing Company (TSMC®) in 2009 [30]. The release of the iPDK was viewed by many as the most aggressive attempt to date by a semiconductor company to impose a standard for automation interfaces that was not tied to a particular EDA vendor. TSMC subsequently released the iPDK trademark and organization to the Interoperable Process Design Kit Library (IPL) Alliance, a consortium of companies that includes TSMC and every large EDA vendor except Cadence Design Systems [31]. Cadence refused to join the alliance, because it viewed the iPDK as an attempt to erode its dominance of the custom design tool market. Other foundries, such as IBM®, had little interest in adopting a PDK standard from competitor TSMC.

The third effort is the OpenPDK effort from Si2 [32] in 2010. Si2 has brought together the IPL Alliance along with IBM and Cadence. Because this effort includes two of the largest foundries along with the largest four EDA companies, this effort has the potential to create a significant standard with a broad impact. This effort differs from the iPDK in that it does not aim to produce a PDK, but rather a PDK compiler. This compiler will impose more of a standard structure on PDKs, simply because foundries would rather use a compiler to create their PDKs, rather than continue to throw more manpower at the problem.

8.7 Conclusion

The complexity of design rules and PDKs has increased significantly in the last decade. The most significant changes have been presented in this chapter, including more devices with multiple threshold and gate oxide options, more complex rules for vias, variable spacing rules, antenna rules, and more metal layers. Lithographic variation leads to a further explosion of design rules and the need for lithographic simulation. The FreePDK aims to collect these issues into an easily distributable package to help inform educators, computer architects, and EDA developers. This effort has led to the creation of standard-cell libraries based on these rules, including the library from Oklahoma State, packaged with the FreePDK [33, 34], and the Nangate™ Open Cell Library. The industry appears to be taking the first steps toward a standard for process design kit interfaces. The FreePDK will likely to follow this emerging standard. The Predictive Technology Model and FreePDK provide the free, realistic basis for this standard to take root and succeed in the electronic design marketplace.

Acknowledgements We would like to thank and recognize the students whose hard work contributed the NCSU Basekit portion of the FreePDK, including Michael Bucher who developed the first set of design rules and LFD flow, Sunil Basavarajaiah for the LVS and extraction rules, and Shivam Priyadarshi and Suhas Satish for developing the Antenna rules. We would also like to thank SRC, NSF (grant 643700), C2S2, Mentor Graphics, Si2, and Synopsys for their generous support of this work. Finally, thanks to Paul Franzon and the developers of the NCSU CDK for their inspiration and continuing technical support in this work.

References

1. D. Grose, "From Contract to Collaboration: Delivering a New Approach to Foundry," *Design Automation Conference*, June 13–18, 2010.
2. The FreePDK45™ Process Design Kit, version 1.3, Mar. 4, 2009, (available at http://www.eda.ncsu.edu).
3. C. A. Mead and L. A. Conway, *Introduction to VLSI Systems*, Addison-Wesley, 1980.
4. "The NCSU Design Kit for IC Fabrication through MOSIS," *International Cadence User Group Conference*, Austin, Texas, 1998.
5. International Technology Roadmap of Semiconductors, 2009. (available at http://www.itrs.net).
6. K. Mistry, et al., "A 45 nm Logic Technology with High-k + Metal Gate, Strained Silicon, 9 Cu Interconnect Layers, 193 nm Dry Patterning, and 100% Pb-Free Packaging," *IEEE Intl. Electron Devices Meeting (IEDM)*, pp. 247–250, 2007.
7. Kuan-Lun Cheng, et al., "A highly scaled, high performance 45 nm bulk logic CMOS technology with 0.242 μm^2 SRAM cell," *IEEE Intl. Electron Devices Meeting (IEDM)*, pp. 243–246, 2007.
8. Samuel K.H. Fung, et al., "45 nm SOI CMOS Technology with 3X hole mobility enhancement and Asymmetric transistor for high performance CPU application," *IEEE Intl. Electron Devices Meeting (IEDM)*, pp. 1035–1037, 2007.
9. T.Sanuki, et al., "High-Performance 45 nm node CMOS Transistors Featuring Flash Lamp Annealing (FLA)," *IEEE Intl. Electron Devices Meeting (IEDM)*, pp. 281–284, 2007.

10. H.Nii, et al., “A 45 nm High Performance Bulk Logic Platform Technology (CMOS6) using ultra High NA(1.07) Immersion Lithography with Hybrid Dual-Damascene Structure and Porous Low-k BEOL,” *IEEE Intl. Electron Devices Meeting (IEDM)*, pp. 1–4, 2006.
11. T. Miyashita, et al., “High Performance Low Power Bulk Logic Platform Utilizing FET Specific Multiple-Stressors with Highly Enhanced Strain and Full-Porous Low-k Interconnects for 45-nm CMOS Technology,” *IEEE Intl. Electron Devices Meeting (IEDM)*, pp. 251–254, 2007.
12. S. Narasimha, et al., “High Performance 45-nm SOI Technology with Enhanced Strain, Porous Low-k BEOL, and Immersion Lithography,” *IEEE Intl. Electron Devices Meeting (IEDM)*, pp. 1–4, 2006.
13. M.A. Quevedo-Lopez, et al., “High Performance Gate First HfSiON Dielectric Satisfying 45 nm Node Requirements,” *IEEE Intl. Electron Devices Meeting (IEDM)*, pp. 424–428, 2005.
14. A. Oishi, et al., “High Performance CMOSFET Technology for 45 nm Generation and Scalability of Stress-induced Mobility Enhancement Technique,” *IEEE Intl. Electron Devices Meeting (IEDM)*, pp. 229–232, 2005.
15. W. Zhao and Y. Cao, “New Generation of Predictive Technology Model for sub-45 nm early design exploration,” *IEEE Trans. on Electron Devices*, vol. 53 no. 11, pp. 2816–2823, Nov. 2006.
16. T. Komoda, et al., “Mobility Iimprovement for 45 nm Node by Combination of Optimized Stress and Channel Orientation Design,” *IEEE Intl. Electron Devices Meeting (IEDM)*, pp. 217–220, 2004.
17. T. Yamamoto, et al., “Junction Profile Engineering with a Novel Multiple Laser Spike Annealing Scheme for 45-nm Node High Performance and Low Leakage CMOS Technology,” *IEEE Intl. Electron Devices Meeting (IEDM)*, pp. 143–146, 2007.
18. H. Ohta, et al., “High Performance 30 nm gate Bulk CMOS for 45 nm node with Σ-shaped SiGe-SD,” *IEEE Intl. Electron Devices Meeting (IEDM)*, 2005.
19. M. A. Quevedo-Lopez, et al., “High Performance Gate-First HfSiON Dielectric Satisfying 45 nm Node Requirements,” *IEEE Intl. Electron Devices Meeting (IEDM)*, 2005.
20. MOSIS, *MOSIS Scalable CMOS Design Rules*, revision 8.00, May 11, 2009, (available at http://www.mosis.com/Technical/Designrules/scmos)
21. I. Sutherland, B. Sproull, and D. Harris, *Logical Effort: Designing Fast CMOS Circuits*, Morgan Kaufmann, 1999.
22. M. Hiroshima, T. Yasaka, S. Miyazaki, and M. Hirose, “Electron Tunneling through Ultrathin Gate Oxide Formed on Hydrogen-Terminated Si(100) Surfaces,” *Japan J. of Applied Physiscs*, vol. 33, part 1, no. 1B, Jan. 1994, pp. 395–398.
23. C. T. Gabriel and E. de Muizon, “Quantifying a Simple Antenna Design Rule,” *Intl. Symp. on Plasma Process-Induced Damage*, 22–24 May, 2000, pp. 153–156.
24. D. S. Bang, *et al*, “Effect of Cu Damascene Metalization on gate SiO_2 Plasma Damage,” *Intl. Symp. on Plasma Process-Induced Damage*, 4–5 Jun., 1998, pp. 64–67.
25. W. T. Weng, et al, “A Comprehensive Model for Plasma Damage Enhanced Transistor Reliability Degradation,” *Intl. Reliability Physics Symp.*, 15–19 April, 2007, pp. 364–369.
26. J. D. Plummer, M. D. Deal, and P. B. Griffin, *Silicon VLSI Technology: Fundamentals, Practice, and Modeling*, Prentice Hall, 2000.
27. E. Josse, et al., “A Cost-Effective Low Power Platform for the 45-nm Technology Node,” *IEEE Intl. Electron Devices Meeting (IEDM)*, pp. 1–4, 2006.
28. A. Bourov, *et al*, “Experimental Measurement of Photoresist Modulation Curves,” Proc. of SPIE, vol. 6154, pp. 1003–1008, Apr. 2006.
29. Global Semiconductor Association, *GSA Mixed-Signal/RF PDK Checklist Users Guide version 3.0*, 2008.
30. Taiwan Semiconductor Manufacturing Company, “TSMC Launches First Advanced Technology Interoperable Process Design Kit”, Press Release, Jul. 21, 2009.

31. M. LaPedus, “IPL group releases PDK standard,” *EE Times*, Feb. 24, 2010.
32. M. LaPedus, “Si2 rolls open PDK effort,” *EE Times*, Feb. 2, 2010.
33. J. E. Stine, I. Castellanos, M. Wood, J. Henson, F. Love, W. R. Davis, P. D. Franzon, M. Bucher, S. Basavarajaiah, J. Oh, and R. Jenkal, “FreePDK: An Open-Source Variation-Aware Design Kit,” *Intl. Conf. on Microelectronic Systems Education (MSE)*, 2007, pp. 173–174.
34. J. Stine, *Oklahoma State University System on Chip Design Flows*, available online at http://vcag.ecen.okstate.edu/projects/scells.

Chapter 9
Predictive Modeling of Carbon Nanotube Devices

Silicon based devices have been the forerunner in mainstream computing for the last 40 years. Their success relies on simultaneously achieving sustainable scaling of physical dimensions and device performance [1]. However, such a scaling trend has been significantly slowing down in recent years due to fundamental physics, materials, and manufacturing limits. Examples of major bottlenecks for continual scaling include short channel effects, high leakage currents, large process variations and reliability issues [2–4]. These pitfalls are rendering design and fabrication of integrated circuits increasingly difficult with scaled silicon devices. As we approach these fundamental limits in planar CMOS process, it becomes imperative to search for alternative materials, structures, and devices to replace silicon transistor as the building block of future nanoelectronics.

These needs drive the innovation of alternative structures like FinFET and trigate device [5, 6], strained channel to enhance carrier mobility and high-k/metal gate to reduce gate leakage current [7, 8]. Though these implementations promise to mitigate some of the problems, their potential is limited and only able to extend the scaling by a generation or two. Amongst more radical search for new devices and materials, carbon nanotube electronics has attracted significant attention owing to the high intrinsic carrier mobility of carbon nanotubes.

Carbon nanotube (CNT) can be simplistically defined as a hollow cylinder made up of one (single-walled) or more (multi-walled) concentric layers of carbon atoms arranged in a hexagonal lattice structure, which is similar to a rolled-up sheet of graphene. With diameters of 1–4 nm and the length extending to several micrometers, carbon nanotube is essentially a one-dimensional object with unique properties attributed to low dimensional structures, such as 1-D density of state for electrons [9]. This allows reduced phase space for scattering and near ballistic transport of carriers when the device dimensions are less than the mean-free path for scattering. Depending on the detailed arrangement of atoms in the nanotube, or

Y. Cao, *Predictive Technology Model for Robust Nanoelectronic Design*,
Integrated Circuits and Systems, DOI 10.1007/978-1-4614-0445-3_9,

the direction in which the graphene sheet is rolled up, single-walled carbon nanotubes is either metallic or semiconducting. Hence CNT transistor and interconnect can be made out of semiconducting and metallic nanotubes, respectively.

Theoretically, with CNTs in parallel, it is possible to get current densities much higher than that of silicon devices with the similar dimension [10]. Various research groups have fabricated and demonstrated functional field effect transistors with semiconducting carbon nanotube channel and metallic nanotubes as interconnects [11–13]. All the transistors reported use metal as the source and drain junctions with direct contact with the CNT channel. This forms Schottky barriers (SB) at the source-drain junctions, which severely restrict the intrinsic current-carrying capability of CNTs, reducing the on current. These SB-CNT transistors further show ambipolar behavior, i.e., an increasing current for negative gate bias. This is an unwanted characteristic for digital applications. In addition to the Schottky junction, some other hurdles that prevent the integration of CNT into the IC industry include lack of process control to separate semiconducting and metallic nanotubes, the alignment of nanotubes, the definition of diameter and junctions, and stable doping methods to develop complementary CNT channels.

To speed up the evolution of this novel alternative technology, parallel efforts in circuit design are essential. For this purpose, the development of predictive compact model is a vitally important step that enables circuit simulation and exploration. Currently most of the models developed for carbon nanotube transistors and interconnects employ numerical or semi-numerical approaches to get the I-V and C-V characteristics [14, 15]. Though highly physical, these models rely on the solution of 1-D differential equations for the solutions. Such numerical approach degrades the computation efficiency and is not suitable for large-scale circuit simulations. Other compact modeling efforts so far include threshold voltage based models and models that resort to SPICE simulator to solve iterative differential equations and compute the surface potential [16, 17].

In this chapter, we propose an integrated compact model for carbon nanotube transistors and interconnects that is non-iterative and SPICE compatible. Initial models concentrated on modeling the channel part of the transistor alone, which is a ballistic transport model. However, since the Schottky barrier effect cannot be decoupled from the channel region, we have developed a non-iterative triangular approximation model to calculate the carrier tunneling probability at the source-drain region for Schottky barrier CNT devices. The implemented model has been systematically verified with TCAD simulations and published measurement data. Leveraging the new CNT model and direct measurements, we further decompose a dramatic range of I-V variability (e.g., 100X in I_{on} and $>10^4$X in I_{off}) into a set of key device parameters, including the Schottky barrier height (Φ_{SB}), CNT diameter (d), the length (L), etc. Such a statistical extraction procedure helps gain insight into physical and process causes of variations. Finally, using the new model, we benchmark digital and analog performance metrics and compare them with 22 nm CMOS process to explore design potentials with CNTs [18, 19].

9.1 Predictive Transistor Model Development

9.1.1 Device Structure

The CNT based device is a strong contender for FET and interconnect applications due to its inherent ballistic transport properties. The cross-sectional view of a typical carbon nanotube transistor is shown in Fig. 9.1. The basic structure is similar to a conventional FET with the channel replaced by a semiconducting carbon nanotube. The top-gated region is defined as the gate length (L_g) and highly doped ungated portion is defined as the access length (L_a). The similarity to the structure of CMOS device improves the compatibility with today's process and design infrastructure, reducing the overhead to incorporate a new type of technology.

With a similar structure of the CNT transistor, metallic carbon nanotubes can be integrated for the interconnect application. Figure 9.2 shows the basic schematic of carbon nanotube interconnects. The structure comprises of metallic nanotubes aligned together over an oxide (in this case SiO_2) of height h with spacing s between the tubes and metal reservoirs at the two lateral ends. This facilitates high-density integration during large-scale manufacturing.

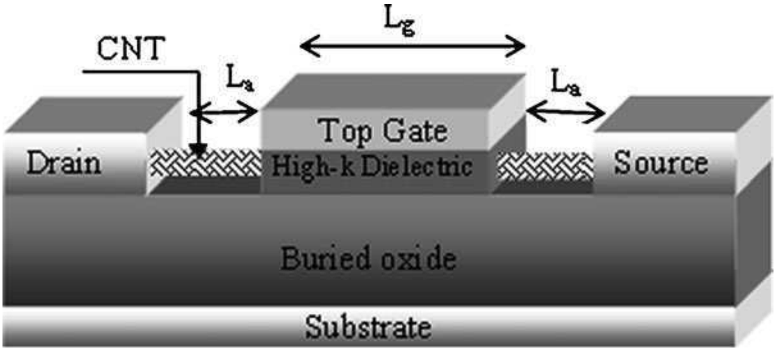


Fig. 9.1 Cross-section of a CNT-FET structure with top gated region as the intrinsic transistor of length L_g and highly doped undated access region of length L_a as the extrinsic part (Adapted from [19])

Fig. 9.2 Cross-section of a generic interconnect structure using carbon nanotubes (Adapted from [19])

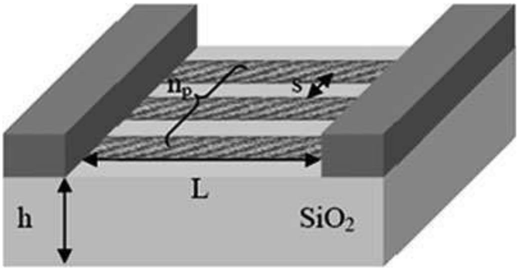
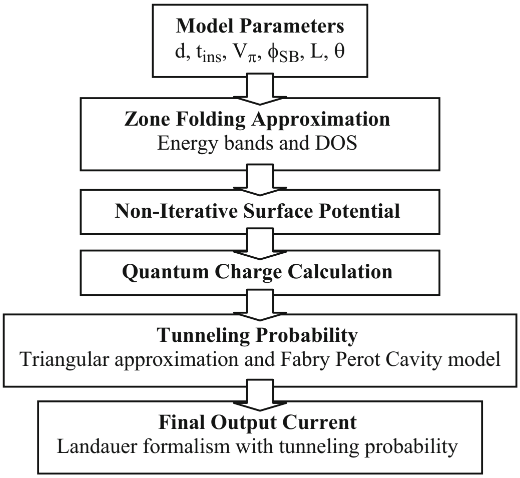


Fig. 9.3 The flow chart describing elements in the model development


Model Parameters
d, t_ins, V_π, ϕ_SB, L, θ
Zone Folding Approximation
Energy bands and DOS
Non-Iterative Surface Potential
Quantum Charge Calculation
Tunneling Probability
Triangular approximation and Fabry Perot Cavity model
Final Output Current
Landauer formalism with tunneling probability

In the ideal case for ballistic transport, the source and drain electrodes would behave as reservoirs that supply and sink unlimited carriers without any reflection at the source and drain. This is true only when there are ideal source and drain contacts, i.e., no significant energy gap between the channel and the contact. However, such an ideal case is difficult to implement in reality. There has been extensive work on finding the appropriate contact material for the CNT-FET and they all have a finite energy gap when contacting the carbon nanotube [20]. Due to Fermi pinning at the contacts, the device behave like a Schottky barrier one where the gate has less control of the channel than that of the ideal case. The device performance is primarily limited by the Schottky contact, depending on the properties of the contact material and the nanotube. The energy gap is sensitive to the work function of the contact, the diameter of the nanotube, as well as the chirality. Therefore, a compact model needs to capture these variations in materials and the fabrication process. Figure 9.3 shows a flowchart of CNT-FET modeling.

9.1.2 Zone-folding Approximation

We begin with characterizing the structure of single-walled carbon nanotubes (SW-CNTs) and defining its basic electronic properties like band-gap, density of states etc. A SW-CNT is essentially a one-dimensional nanowire formed by rolling a two-dimensional graphene sheet. The $2s$, $2p_x$ and $2p_y$ orbitals form σ bonds in graphene. Since the σ bonds are weakly coupled to the $2p_z$ orbitals, they form π bonds, which give rise to the electronic properties of graphene. The E-k values for graphene can be obtained from the tight-binding model given by Eq. 9.1 [21]:

$$E_{g2D}\left(k_x, k_y\right) = \pm t\left\{1 + 4\cos\left(\frac{\sqrt{3}k_x a}{2}\right)\cos\left(\frac{k_y a}{2}\right) + 4\cos^2\left(\frac{k_y a}{2}\right)\right\}^{1/2} \qquad (9.1)$$

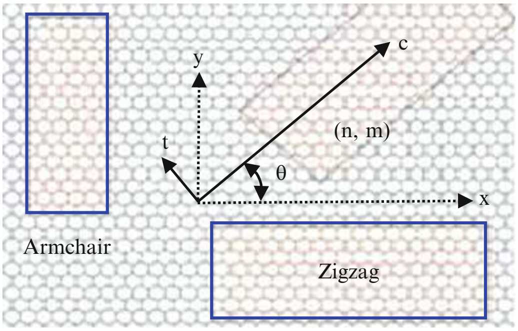


Fig. 9.4 Honeycomb lattice graphene sheet showing the chiral vectors (n, m). The corresponding E–k and DOS are calculated using Eqs. 9.1 and 9.2, respectively

To get the band structure of carbon nanotubes, we begin with the band structure of graphene given in Eq. 9.1, and then apply periodic boundary conditions along the circumference of the nanotube. The rolling-up of the honeycomb lattice of the graphene sheet along a specific direction, known as the chiral vector (shown in Fig. 9.4) causes the quantization of the wave-vector space along its direction. A chiral vector can be denoted by the coordinates (n, m): if (n-m) is a multiple of 3, the carbon nanotube is metallic, else it is semi-conducting; when n = m, the carbon nanotube is known as 'zigzag', and when m = 0, it is known as 'armchair'. The energy gap (E_g) of a semiconducting nanotube is dependent on its diameter (d), which is dependent on the chiral vector (n, m). Hence, E_g is effectively a function of the chiral vector or the chiral angle. To calculate the current, the electron density of states (DOS) near the Fermi level is required. Classical tight-binding models are used to accurately compute the DOS but at low bias, the DOS D(E) at energy E can be approximated as expressed in Eq. 9.2 [21]:

$$D(E) = \frac{D_0|E|}{\sqrt{E^2 - E_n^2}}, \text{ where } D_0 = \frac{8}{3\pi V_\pi a} \tag{9.2}$$

All variables used in the above equations are defined in Table 9.1.

9.1.3 Surface-potential Based Modeling

When a gate voltage V_G is applied, the surface potential (ϕ_s) is modulated. Figure 9.5 illustrates the concept of the surface potential. The expressions for surface potential and the total charge are as follows:

$$\phi_s = V_G - \frac{|Q_{CNT}|}{C_{ins}} \tag{9.3}$$

Table 9.1 Constants and parameters used in the model (Adapted from [19])

Physical constants			
V_π	C-C bonding energy		2.97 eV
a	C-C bonding length		0.142 nm
q	Electron charge		$1.6e^{-19}C$
V_t	Thermal voltage		26 mV
Model parameters			
d	Diameter (m)	θ	Chiral angle (degree)
L	Nanotube length (m)	t_{ins}	Insulator thickness (m)
ϕ_{sb}	Barrier height (eV)	ε_{ins}	Insulator dielectric constant
Derived parameters			
Energy gap (eV)		$E_g = 2V_{pi}a/d$	
Sub-band energy levels (eV)		$E_n = (E_g/8)(6n - 3 - (-1)^n)$	
Intrinsic carrier concentration		$N_0 = 4q/(3\pi V_\pi a)$	
Insulator capacitance		$C_{ins} = 2\pi\varepsilon_r\varepsilon_0/\log[(t_{ins} + d/2)/(d/2)]$	

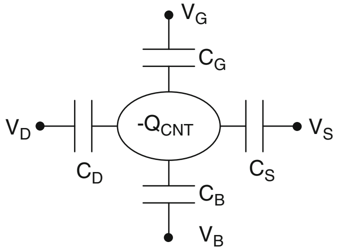


Fig. 9.5 Surface potential plays a central role to determine the channel charge (Adapted from [19])

$$Q_{CNT} = N_0 \sum_n \int_{E_n} \left[F\left(\sqrt{E^2 - E_n^2}, \mu_s\right) + F\left(\sqrt{E^2 - E_n^2}, \mu_s - V_{DS}\right) \right] dE \quad (9.4)$$

where

$$F(E, \mu) = \frac{1}{1 + e^{(E-\mu)}} (\text{Fermi} - \text{Dirac Integral}) \quad (9.5)$$

The classical method to compute ϕ_s (using the conduction-band minima and DOS calculated from Table 9.1 and Eq. 9.2, respectively) involves numerically solving the 1-D Poisson equation and the total charge equation with self-consistency. In spite of being accurate, this method is not a good choice for compact modeling since it is computationally inefficient; in addition, SPICE solvers often encounter convergence errors when loaded with the task of solving complicated numerical functions. Hence, in our model, we derive a linear equation for ϕ_s. By eliminating the iterations involved, the simulation speed is considerably improved making the model suitable for large-scale circuit simulation.

To derive the surface potential, we first condition the bias voltages at the source and drain into intrinsic potentials ξ_s and ξ_d with respect to the source Fermi level E_f and sub-band energy, E_0,p. The non-iterative compact equation for the surface potential at zero-bias is obtained by the first order approximation of charge in the CNT (Eq. 9.4) and is given by:

$$\phi_s = \sum_n \left(\frac{V_t \gamma (\xi_s |\xi_s| + \xi_d |\xi_d|)}{2(1 + 2\gamma)} \right) - V_G \tag{9.6}$$

where $\gamma = N_0/C_{ins}$, and

$$\xi_{s,d} = \frac{\left(E_f - V_{s,d} - E_{o,p} + V_{gs}\right)}{V_t}; \left|\xi_{s,d}\right| = \begin{cases} 1, \text{ if } \xi_{s,d} > 0 \\ 0, \text{ if } \xi_{s,d} < 0 \end{cases}$$

This expression forms the basis of our compact model. All existing models use self-consistent numerical methods to solve for ϕ_s. Figure 9.6 shows the variation of surface potential as a function of V_{GS} and V_{DS}, for different diameters. At low voltages, the model is in good agreement with the numerical simulations and no regional approximations are required in the expression. The surface potential is a function of the diameter, temperature and gate dielectrics to the first order. At higher voltages, higher sub-bands are filled and therefore the slope of the line in Fig. 9.6 (top) changes and is modeled by Eq. 9.6.

9.1.4 Schottky Barrier Modeling

Due to the work function difference between carbon nanotubes and the source/drain metals, a Schottky barrier is formed at the junction. The barrier height ϕ_{SB} depends on the work function difference while the barrier width depends on the thickness of the insulator between the gate and the nanotube channel. The total current at the junction is the sum of thermionic emission and the tunneling current through the barrier. The worst case is when $\phi_{SB} = E_g$, the Fermi level is pinned to the valence band and ambipolar behavior is severe. As the insulator thickness reduces, the barrier at the source and drain become more transparent and the thermionic emission over the barrier dominates. Hence, the tunneling model is important to accurately model carrier conduction in a CNT-FET. Tunneling probability through a Schottky barrier is given by the WKB approximation:

$$T(E) = \exp\left[- \int_{z_i}^{z_f} k(z) dz \right]$$

An exponential barrier profile has been approximated by a triangular barrier, which gives a closed form solution for tunneling probability [22], thus significantly

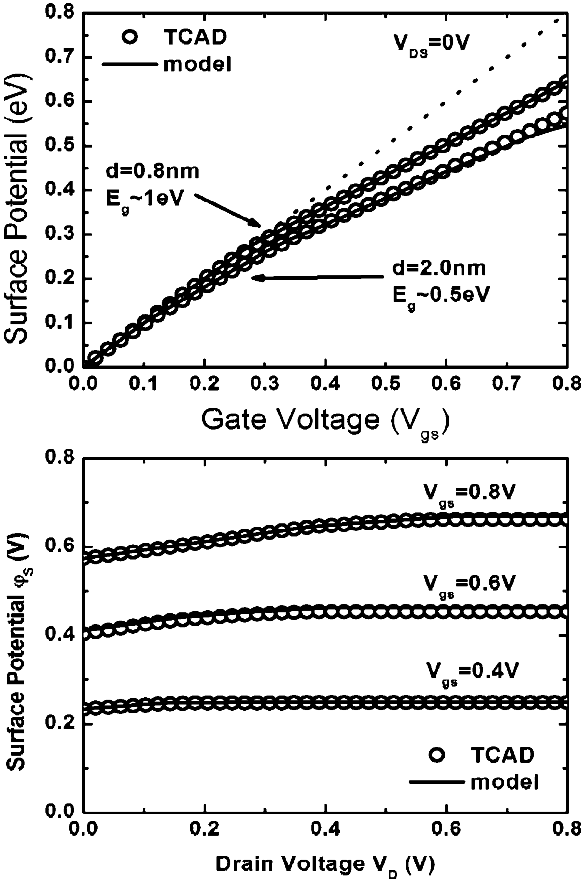


Fig. 9.6 ϕ_s as a function of V_{gs} and V_{ds} for d = 0.8 nm and 2 nm. The voltage range is the region where there is good gate control and FET. type behavior (Adapted from [19])

enhancing the computational efficiency of the model. The non-iterative tunneling probability as a function of energy is given by:

$$T(E) = \exp\left[\frac{-t_{ins}k_n}{\phi_{sb'}}\left(E'\sqrt{1-K'^2} + (E-\phi_{sb'})E_t\right)\right] \tag{9.7}$$

where

$$E_t = \sqrt{(1-K(E-\phi_{sb'}))^2} - \sin^{-1}\left(-E'\sqrt{1-KE'^2}\right) + \sin^{-1}(\phi_{sb'} - E)$$

$$K = \frac{q\pi}{4k_nN_0}$$

$$E' = (E-\phi_{sb'}) + \phi_{sb'}\ln\left(\frac{E}{\phi_{sb'}}\right)$$

$$\phi_{sb'} = \mu_{s,d} + \phi_{sb} \tag{9.8}$$

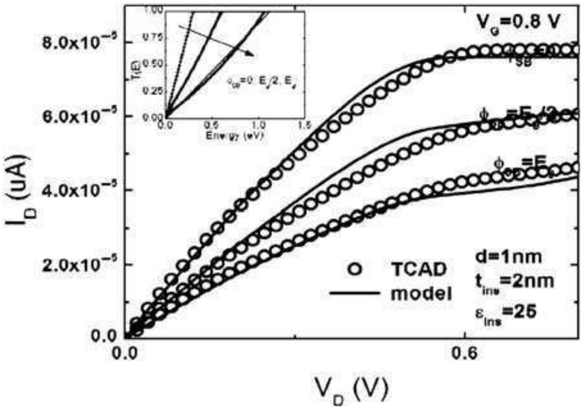


Fig. 9.7 I_{ds} vs. V_{ds} at $V_{gs} = 0.8$ V for three different barrier heights (Adapted from [19])

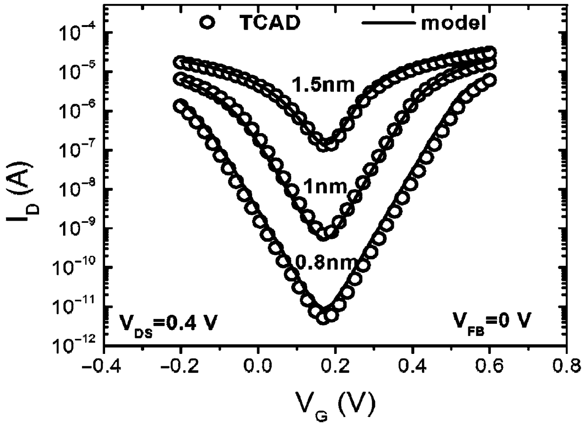


Fig. 9.8 I_{ds} as a function of V_{gs} for d = 0.8 nm, 1 nm and 1.5 nm. $V_{FB} = 0$ V, $t_{ins} = 2$ nm, $\varepsilon_r = 25$, and $L = 10$ nm (Adapted from [19])

Figure 9.7 demonstrates a good agreement between the triangular approximation model and the numerical model for the contact part. The tunneling probability equation given by Eq. 9.7 is solved at the source and drain junctions and Eq. 9.9 is used to compute the final current:

$$I = \frac{4q}{h}\mathrm{sgn}(E)T(E)\sum_n \int_{E_n} [F(\mathrm{sgn}(E)(E,\mu_s)) + F(\mathrm{sgn}(E)(E,\mu_s - V_{ds}))]dE \quad (9.9)$$

where $sgn(E) = 1$ or -1 for conduction and valence band respectively and $F(\mu,E)$ is as defined in Eq. 9.5. Using the equations and results discussed as summarized in Table 9.1, a physics based compact model of CNT-FET was implemented in Verilog-A which is computationally efficient and is useful to run transient simulations. The I-V characteristics are presented in Fig. 9.8. These results prove

Table 9.2 Parameters in the SPICE model file (Adapted from [19])

Parameter	Description	Default value
Instance parameters		
d	Diameter	2 nm
θ	Chiral angle ($0 \leq \theta < 30^\circ$)	0
t_{ins}	Insulator thickness	10 nm
ε_{ins}	Dielectric constant of insulator	9
t_{back}	Backgate insulator thickness	130 nm
ε_{back}	Dielectric constant of substrate	3.9 (SiO_2)
L	Gate length	100 nm
type	n-type = 1, p-type = −1	1
Model parameters		
phisb	Schottky barrier height	0 eV
mob	Mobility parameter	1
R_s	Parasitic source access resistance	0 ohm
R_d	Parasitic drain access resistance	0 ohm
β	Coupling coefficient	1
C_C	Coupling capacitance	7aF
C_p	Parasitic capacitance	120aF

that the model is suitable for the different diameters and bias conditions without the need for any empirical parameters from numerical simulations, thus making this the first compact model for the CNT. This model does not include scattering effects that may further affect the I-V characteristics. However, since we use the surface potential approach, they can be easily incorporated in future.

9.1.5 Transistor Model Extraction and Validation

The parameters enlisted in Table 9.2 comprise the SPICE based circuit model for CNT-FET developed in Verilog-A. Running simulations by varying each parameter enables us to gain detailed insight on the effect of each parameter on performance of the CNT-FET.

Our compact model can be used to fit measurement data to gain process-related insight such as parasitics, variations etc. This is achieved by properly tuning the model parameters enlisted in Table 9.2. The main fitting steps are:

1. Define instance parameters; calculate physical parasitics (C_C is set to a very small value, which is about 1/10 of the insulator capacitance);
2. C_{subfit}: tuned to fit I_{DS} vs. V_{GS} at low V_{DS} (0.1 V) and V_{BS} fixed. This is to match the flat bland voltage;
3. β: tuned to fit I_{DS} vs. V_{DS} at a high V_{GS} to match the saturation region (basically the shape of the I_{DS} vs. V_{DS} curve);
4. C_p: tuned to match I_{DS} vs. V_{GS} in the subthreshold region, at high V_{DS}; sometimes, phisb also needs to be tuned to match I_{DS} vs. V_{GS} in the saturation region;

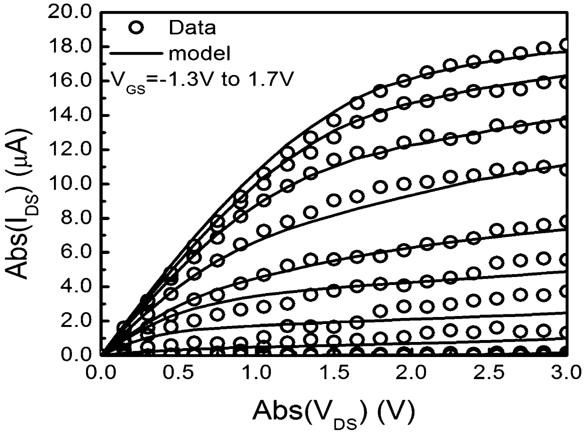


Fig. 9.9 Model validation with experimental data [23] (Adapted from [19])

5. $R_{D,S}$: tuned to match I_{DS} vs. V_{DS} in the linear region;
6. mob: used to match the saturated drain current;

Using the fitting procedure described in the previous section, the model has been validated with published measurement data (Fig. 9.9). An interesting feature of the fitting is the exact replication of the gap in the I-V plot, which is due to the multiple band conduction in carbon nanotubes. The I-V characteristics distinctly show the following trends: (1) the off current varies exponentially with diameter and barrier height, and (2) the on current degrades with barrier height and increases linearly with diameter. These conclusions have been observed even in other models [14–17]. The new model now helps us run SPICE simulations fast enough to benchmark circuit performance metrics. All the results in the following sections are generated using the Verilog-A model that supports AC and DC analysis that is several times faster than numerical simulations in matlab. The model can be extended in the future for high-field effects and other non-idealities.

9.2 Interconnect Modeling

Metallic CNT interconnects have recently gained a lot of interest due to their properties of high mechanical and thermal stability, thermal conductivity and high current carrying capabilities [24]. Ideally, metallic SW-CNTs have a Fermi velocity of 8×10^5 m/s. However, in reality the ballistic motion is mitigated by several scattering mechanisms, such as acoustic phonon scattering, zone boundary scattering and optical-phonon scattering. These mechanisms have been explained by several models [25, 26]. In this section, we present a continuous expression for the resistance of the interconnect and the resistance of the contact. The circuit model for the interconnect is shown in Fig. 9.10. At high frequencies, the inductance and the capacitance determine the total impedance of the interconnect.

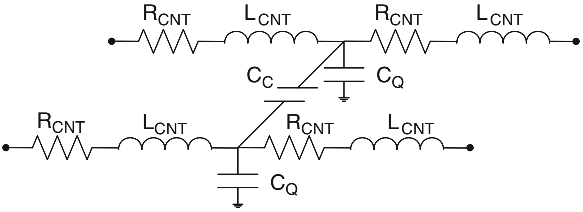


Fig. 9.10 Circuit model for CNT interconnect (Adapted from [19])

The following subsections present the DC and small-signal parameters of the CNT interconnect.

9.2.1 CNT Interconnect Resistance

Due to the nature of the band structure, in an ideal ballistic motion regime, the resistance is constant:

$$R_{ballistic} = \frac{h}{4e^2} = \frac{1}{\mu_0} \tag{9.10}$$

However, when the length of the interconnect is much longer that the mean free path (MFP), several scattering mechanisms dominate. At low bias, the predominant mechanism is the acoustic phonon scattering with a MFP of 1 µm-1.6 µm [25]. As the bias voltage increases, the electrons can scatter from band to band and within the same band. This leads to optical phonon scattering and zone-boundary scattering. These scattering mechanisms are well known and have been modeled in the past. In this compact model, we have derived a single equation to model the conductance under all these effects:

$$G(V,L) = G_{op\_zo} + \frac{V_{eff}[G_{acc} - G_{op\_zo}]}{V} \tag{9.11}$$

Where

$$V_{eff} = V_{cr} - \frac{1}{2}\left[(V_{cr} - V - \delta) + \sqrt{(V_{cr} - V - \delta)^2 + 4V_{cr}\delta}\right]$$

Equation 9.11 combines the effect of acoustic phonon scattering and optical phonon scattering in a single equation. Below the critical voltage G_{acc} dominates. Using the expression for V_{eff}, G has a smooth transition to G_{op_zo}. This allows better convergence in circuit simulation tools as compared to piecewise linear equations for the two scattering regions. Figure 9.11 illustrates the resistance at various lengths.

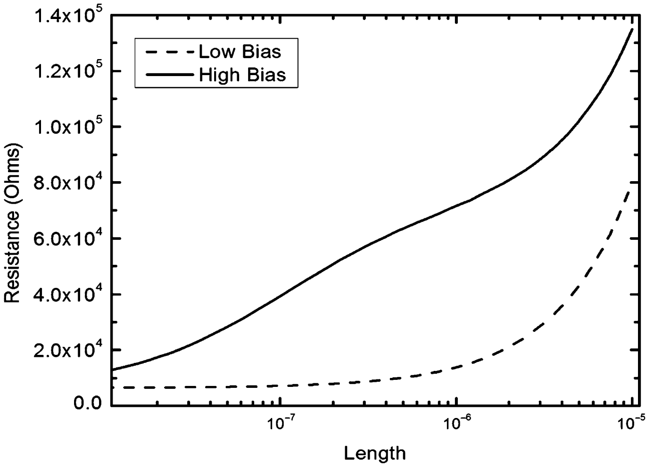


Fig. 9.11 Resistance of a CNT interconnect with varying length for high and low bias across the terminals (Adapted from [19])

9.2.2 Capacitance and Inductance of CNT Interconnect

As shown in Fig. 9.2, carbon nanotube interconnects are formed by arranging arrays of nanotubes aligned next to each other with the terminals at the ends of the two tubes. Two capacitances become important due to this structure, the coupling capacitance between two adjacent nanotubes C_C, and the quantum capacitance within the nanotube C_q. The coupling capacitance has the form

$$C_c = \frac{\pi \varepsilon L}{\log\left(d/s + \sqrt{(d/s)^2 + 1}\right)} \tag{9.12}$$

and the quantum capacitance is given by

$$C_Q = \frac{4e^2 L}{\pi h v_f} \tag{9.13}$$

Theoretically, there are two kinds of inductances that need to be modeled for metallic carbon nanotubes, the magnetic or mutual inductance and the kinetic or self-inductance. As discussed in [26], it can be shown that for a one-dimensional structure like carbon nanotubes, kinetic inductance dominates mutual inductance and hence our model only considers on kinetic inductance. It is given by the following expression:

$$L_e = \frac{h}{2e^2 v_f} \tag{9.14}$$

Table 9.3 CNT interconnect model parameters (Adapted from [19])

Parameter	Description	Default value
Instance parameters		
d	Diameter	1 nm
np	Number of CNTs in parallel	1
s	Spacing between CNTs	10 nm
ε_{ins}	Dielectric constant of insulator	25
C_C	Coupling capacitance	0
L	Gate length	100 nm
h	Substrate insulator thickness	100 nm
Model parameters		
phisb	Schottky barrier height	0 eV
V_{crit}	Optical-phonon scattering parameter	0.16 eV
R_p, R_n	Parasitic access resistance	0 ohm
l_{acc}	MFP for acoustic phonon scattering	1.0 μm
l_{zb}	MFP for zone boundary phonon scattering	20 nm

Due to their multiple band structure, carbon nanotubes have two modes of propagation. In each mode, it is also possible to have two electrons (spin up and spin down). Hence, CNT has four modes of propagation, thus resulting in one-fourth of the total inductance calculated above and four times the quantum capacitance as given in Eq. 9.13.

9.2.3 *Interconnect Model Extraction and Validation*

The resistance of CNT interconnect is controlled by the effective mobility due to several scattering mechanisms. Therefore, we use the three model parameters V_{crit}, l_{acc} and l_{zb} to model the optical phonon scattering, acoustic phonon scattering and zone boundary phonon scattering, respectively. The SPICE circuit parameters for the interconnect model are enlisted in Table 9.3.

The instance parameters are geometry dependent parameters. The coupling capacitance is either calculated by external 2D or 3D solvers, e.g., Raphael [27], or can be calculated internally by Eq. 9.12. If the length ranges between 10 nm and 1 μm, V_{crit} is tuned in the range of 0.08–0.16 to decrease the resistance; if the length is longer than 1 μm, acoustic phonon scattering dominates and therefore l_{acc} will affect the slope of the curve. When the contacts are short and Ohmic, R_n and R_p can be ignored. At high current values, the phisb value can be extracted. The model has been validated against measured data in Fig. 9.12.

9.3 Statistical Extraction of Process Variability

Theoretical calculations [10] and experimental results [28] have shown that CNT device has superior performance with respect to conventional silicon devices. Yet, the challenges in precise process control are still tremendous, especially in the

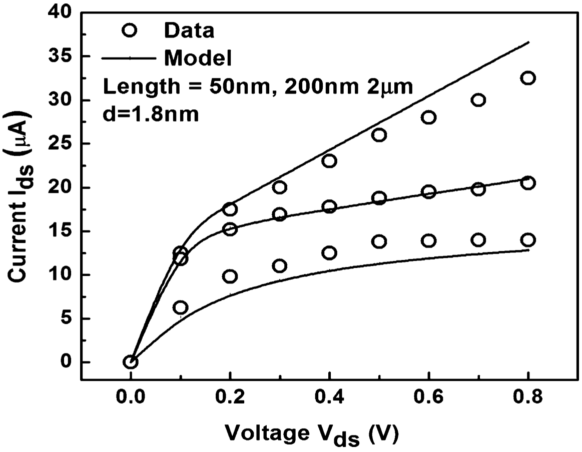


Fig. 9.12 Interconnect model validation with measured data for varying length [25] (Adapted from [19])

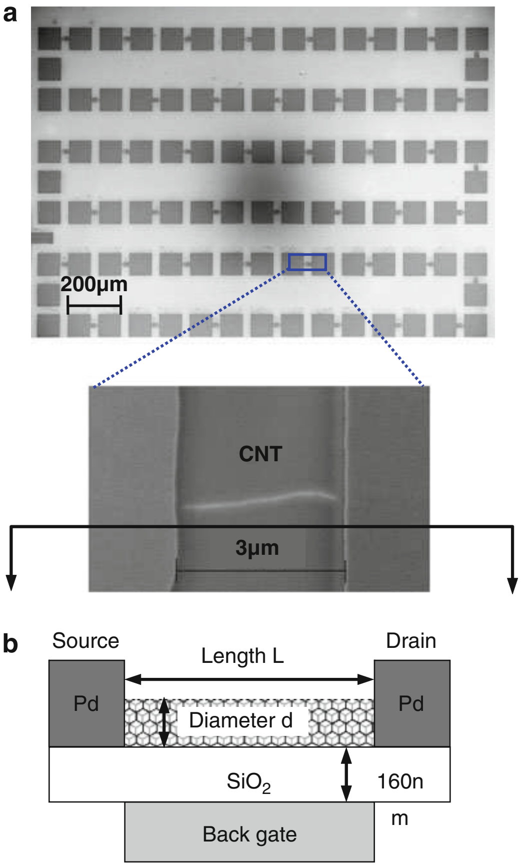
definitions of diameter, chirality, alignment and contact [29]. Structurally, carbon nanotube transistor is a three-terminal device similar to Si-based devices, as shown in Fig. 9.13. However, metals have to be used as source and drain, instead of doped nanotubes. The metal-semiconducting CNT channel forms Schottky barrier which limits drain to source current (I_{DS}). Since I_{DS} is a combination of thermionic and trans-mission emission through the Schottky barrier, the Schottky barrier height (Φ_{SB}) plays a crucial role in determining the performance of a carbon nanotube device. In addition, the diameter and length of a carbon nanotube have a significant influence on the current of the CNT transistor. The diameter determines the band structure of the nanotube while the length affects various scattering effects that reduce the current from the ballistic limit.

Previous experimental work has shown that the Schottky barrier height is dictated by the work function of the metal used to form the contact [30], fabrication method [31] and the diameter [32] of the nanotube. A recent study shows that the chemical nature of the atomic species of the electrode also plays an important role in determining the transmission characteristics [33]. It is imperative to systematically characterize the impact of these variation sources to improve fabrication quality and facilitate large-scale circuit implementation with carbon nanotube transistors. This section develops a model-based statistical method to assess major variation sources in CNT-FET devices.

9.3.1 Device Fabrication and Measurement

Figure 9.13 illustrates the regular array of CNT devices. The fabrication process starts from a highly doped wafer, which operates as the back gate to modulate the conductivity of the CNT device. The wafer is covered with 160 nm of SiO_2 (Fig. 9.13c). Using chemical vapor deposition (CVD), the P-type carbon nanotubes

Fig. 9.13 (**a**) Die photo of the test array consisting of p-type CNT devices; (**b**) SEM image of a single CNT transistor with Pd-contacts; and (**c**) Schematic of the structure used for the extraction



are grown from patterned catalyst islands. The targets of CNT diameter and length are 1.5–2 nm and 2 μm, respectively. Pd metal contacts are added to both ends of the CNT device using standard photolithography and e-beam evaporation. Figure 9.13 shows the die photos and the cross-sectional structure.

From this regular array, I-V characteristics are conveniently measured. Due to the variation of the chirality, ~30% of the CNT devices are metallic ones. The rest of 97 semiconducting CNT transistors are collected to study other variational parameters. For these P-type CNT-FETs, I_{ON} and I_{OFF} exhibit 100X and 10^4X variability, respectively (Fig. 9.14), because of process variation in the channel and contact regions. I_{ON} is the maximum drive current at $V_{GS} = -15$ V and $V_{DS} = -3$ V, while I_{OFF} is the minimum current in the V_{GS} range of ±15 V.

Furthermore, the diameter of each CNT-FET, which is the height of the CNT from SiO_2 (Fig. 9.13c), is determined by tapping mode atomic force microscopy (AFM). The exact length of each nanotube is measured by scanning electron microscopy (SEM), considering their nonlinear alignment (Fig. 9.13b). Figure 9.15 shows the variations of both parameters. The variation of coupling capacitance (C_c)

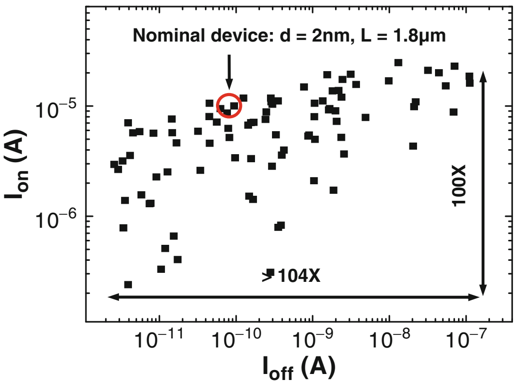


Fig. 9.14 Measured I_{ON} vs. I_{OFF} variability. I_{ON} and I_{OFF} exhibit 100X and 10^4X variability, respectively

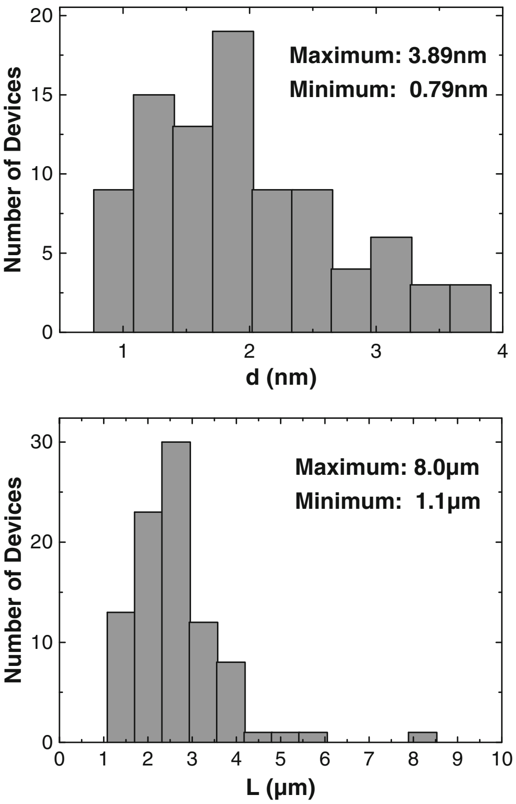


Fig. 9.15 Histograms of measured variations of the diameter and CNT length. *d* significantly affects I_{ON} and I_{OFF}, while *L* mainly affects I_{ON} due to mobility degradation

between S/D and the channel is further calibrated by the capacitance bridge technique. Even though *d* and *L* variations account for a large portion of I-V fluctuations, they are still not sufficient to explain the entire range of variability: for these 97 CNT-FETs, the error in I_{on} is still higher than 100% if

Table 9.4 Variational parameters and summary of the extraction method

Parameters	Unit	I-V Sensitivity	Extraction Method	Range of values
d	nm	High	AFM	0.7–3.9
L	µm	High	SEM	1.1–8.5
C_c	aF	Low	Capacitance bridge	10–30
Φ_{SB}	eV	High	Model based	0.01–0.7
R_{ds}	kΩ	Medium	Model based	3–8

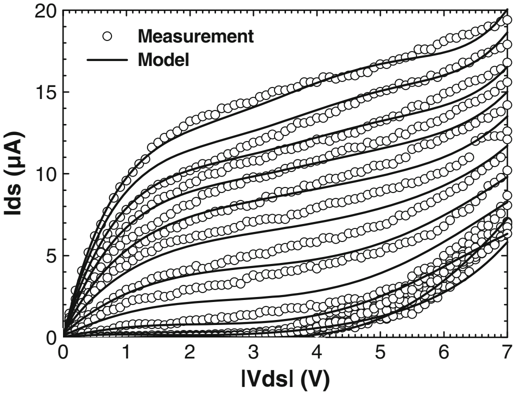


Fig. 9.16 Fitting of the nominal model to I-V measurement (V_{GS}: −15 V to 15 V)

only d and L variations are included. In this case, the newly developed compact model of CNT-FET (Sect. 9.1) is adopted to decompose other physical parameters, which cannot be extracted from direct measurements.

9.3.2 Model Based Extraction of Variations

Table 9.4 summarizes the main process parameters and the extract methods for CNT variations. The model developed in Sect. 9.1 well captures intrinsic variations, especially in the Schottky barrier height, and serves as the cornerstone of the extraction method. Primary variational parameters in this step include Φ_{SB} and S/D parasitic resistance (R_{DS}), with parasitic capacitance (C_p) fixed at 1.2 pF for this fabrication process. The values of Φ_{SB} and R_{DS} are extracted by iteratively fitting I_{OFF} and I_{ON}, respectively, as shown in the flowchart (Fig. 9.16). To be specific, the model based extraction procedure starts from the calibration of nominal model parameters. This is achieved by fitting the full I-V characteristics of the nominal device (Fig. 9.14), as shown in Fig. 9.16. The ambipolar behavior is observed at high V_{DS} because of the presence of the Schottky contacts.

Based on the nominal model, Φ_{SB} and R_{DS} values are tuned to match I_{OFF} and I_{ON} of each individual CNT-FET. The impact of d and L variations on I-V is incorporated by tuning the bandgap and the mobility due to the scattering.

Fig. 9.17 The variation in I_{ON} is most sensitive to Φ_{SB} and moderate to R_{DS}, but negligible with respect to change in the coupling capacitance C_c

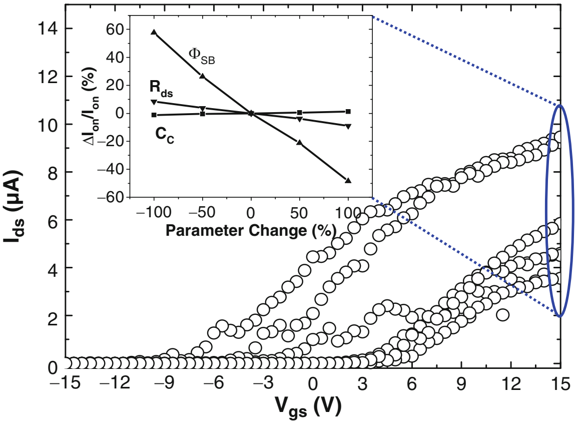


Fig. 9.18 Extracted variation f Φ_{SB} with respect to the bandgap E_g. The trend matches theoretical expectation [35]

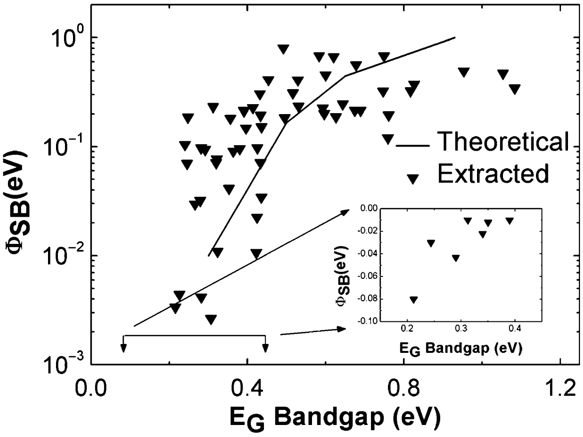


Figure 9.17 confirms the high sensitivity of I_{ON} to Φ_{SB} and R_{DS}. From this procedure, the statistics of Φ_{SB} and R_{DS} are obtained. Figure 9.18 illustrates a distinct behavior where smaller Φ_{SB} is formed for a larger diameter (i.e., a smaller bandgap). In addition, there is a clear trend that Φ_{SB} variation is much higher in CNT-FETs with smaller bandgaps ($d > 1.5$ nm); the variation reduces significantly when d is smaller than 1.5 nm. These results match the theoretical expectation since Fermi level pinning is not observed at the metal-CNT junction in carbon nanotube devices [34, 35]. It is concluded that there is a trade-off between low Schottky barrier height (i.e., near Ohmic contact) and Φ_{SB} variation. The inset in Fig. 9.18 shows extracted data with negative values of Φ_{SB} resulting in Ohmic contacts. Devices with Ohmic contacts are desirable since they have higher I_{ON}. However, the amount of variation in Φ_{SB} is considerably larger for these diameters.

By including the variations of d, L, Φ_{SB} and R_{DS} into the nominal model, the dramatic I-V fluctuations are captured. Fig. 9.19 shows the correlation between model predictions and the measurement data. Both I_{ON} and I_{OFF} are well matched,

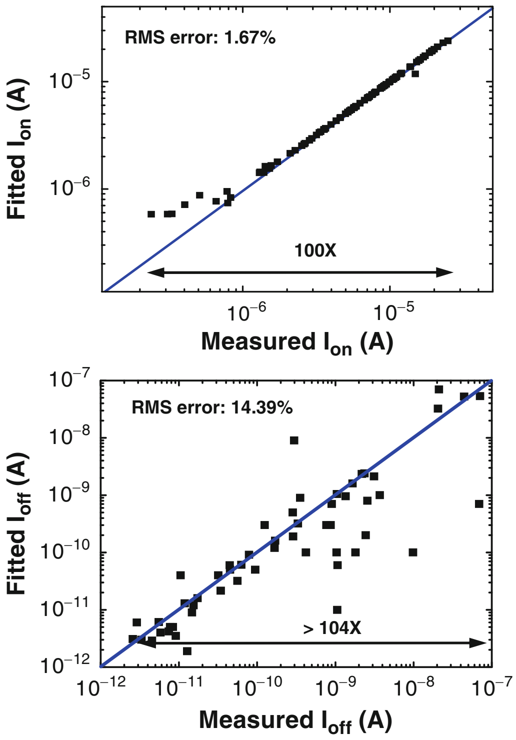


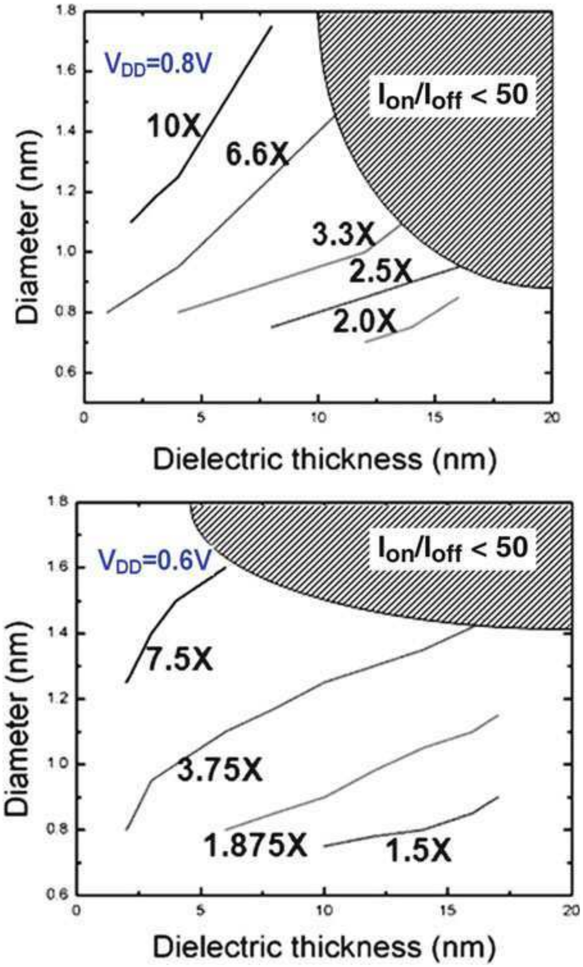
Fig. 9.19 Excellent correlation between model predictions and measured variations. I_{ON} is sampled at $V_{GS} = -15$ V while the bias value for I_{OFF} varies with the diameter

with the root-mean-square (RMS) error in $I_{ON} < 5\%$. By combining direct measurement with the compact model, the extraction of primary variations in CNT devices provides new insight into the source of variations, guiding further investigation on optimizing the fabrication.

9.4 Design Insights with CNT Devices

Based on the concept of the surface potential, the new compact model of CNT accurately predicts I-V and C-V characteristics, as well as the variability. It is scalable to key process and design parameters, including the diameter, chirality, gate dielectrics, and bias voltages. Using this model, we explore design possibilities in order to extract the optimum design space. CNT with L = 100 nm has been compared with 22 nm bulk CMOS from PTM for both analog and digital applications [18]. For consistency in the analysis below, we have used $V_{FB} = V_{DD}/2$ for N-type CNT and $-V_{DD}/2$ for P-type CNT. The dielectric material used has $\varepsilon_r = 25$. Parasitic capacitances have been lumped into a single parameter based on published values [36]. Since all the characteristics are dependent on the diameter of the nanotube, our analysis is for varying diameters. Above 1.8 nm, the SB-FET has I_{ON}/I_{OFF} less than 50, which is not practical for design applications, and thus, it is not included in this study.

Fig. 9.20 Speed contours for varying diameters and t_{ins}



To benchmark digital design, SPICE simulations of FO4 inverter comparing CNT-FETs with 22 nm bulk CMOS have been performed to study the effect of Schottky barrier height (Source/Drain contact material), gate dielectric thickness, leakage power, supply voltage scaling and process variations on digital design. It is found that for smaller diameters of the range of 1–1.5 nm and optimum contact materials, up to 10X improvement in speed, power and energy consumption can be achieved as compared to 22 nm bulk CMOS. High-k dielectrics are undoubtedly the best choice for CNT transistors.

The speed contours have been plotted for adequate scaling in dielectric thickness to ensure the same performance. It can be clearly seen that up to 10X increase in speed can be achieved when compared to 22 nm CMOS. The contours shown in Fig. 9.20 can be followed by varying the diameter. The reason for diameters of 1–1.5 nm being optimal is depicted by the shaded region in the Fig. 9.20. Since larger diameters have higher leakage, it is more difficult to switch them off.

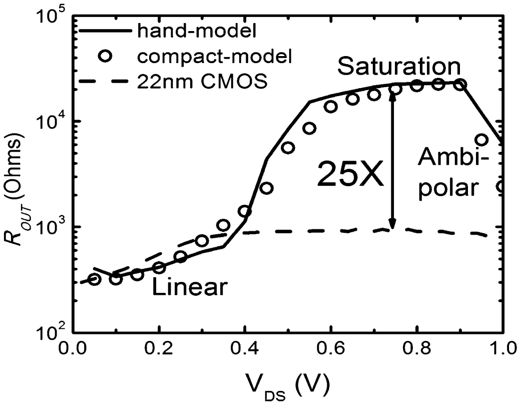


Fig. 9.21 R_{out} as a function of drain voltage, compared with 22 nm CMOS with the same saturation current

Smaller diameters have a 5X decrease in speed as compared to CNTs with a larger diameter. There is a trade-off between speed and power in using CNT-FET for digital applications.

Carbon nanotubes have a multiple band structure. Hence, CNT FET has a much higher current density with comparable bulk semiconductors. If parasitic capacitance is reduced, CNTs have another advantage in low quantum capacitance. Therefore, the device can have very high cut-off frequency, which is given by Eq. 9.15 [37]:

$$f_T = \frac{g_m}{2\pi C_g} \tag{9.15}$$

Efficient measurement technique to characterize analog performance and reducing the parasitic capacitance during the fabrication are the two major hurdles facing the industry. The AC gain and frequency response are mainly controlled by the transconductance (g_m) and output impedance (R_{out}). Figure 9.12 plots the variation of output impedance of CNT-FET compared to 22 nm bulk CMOS. For a fair comparison, R_{out} is calculated for the same saturation current for both devices. For CMOS, R_{out} vs. V_{DS} is mainly influenced by the triode region, channel length modulation, drain induced barrier lowering (DIBL) and finally substrate current induced body effect (SCBE) with increasing V_{DS} [38]. Contrarily for CNTs, R_{out} is affected by the linear, saturation and ambipolar characteristics of the CNT device. As can be seen from Fig. 9.21, due to better saturation characteristics in CNTs, a CNT-FET can have up to 25X higher R_{out} as compared to 22 nm CMOS for the same saturation current.

In conclusion, CNTs possess the capacity to surpass CMOS transistors in both analog and digital domains assuming high-level integration and process-related challenges are solved. This new predictive model serves as one of the most important bridges between process and design giving key insights into the development of carbon based nanoelectronics.

References

1. G. E. Moore, “Cramming more components onto integrated circuits,” *Electronics*, vol. 38, no. 8, 04/19 1965.
2. C. H. Wann, K. Noda, T. Tanaka, M. Yoshida, and C. Hu, “A Comparative Study of Advanced MOSFET Concepts,” *IEEE Transactions on Electron Devices*, vol. Vol. 43, no. No. 10, pp. 1742–1753, October 1996.
3. K. A. Bowman, S. G. Duvall, and J. D. Meindl, “Impact of die-to-die and within-die parameter fluctuations on the maximum clock frequency distribution for gigascale integration,” *Solid-State Circuits, IEEE Journal of*, vol. 37, no. 2, pp. 183–190, Feb 2002.
4. I. C. Chen, S. Holland, and C. Hu, “Electrical breakdown in thin gate and tunneling oxides,” *IEEE Trans. Electron Devices*, vol. 32, pp. 413–422, February 1985.
5. D. Hisamoto, et al., “FinFET-a self-aligned double-gate MOSFET scalable to 20 nm,” *IEEE Trans. Electron Devices*, vol. 47, no. 12, pp. 2320–2325, 2000.
6. B. Doyle, et al., “Tri-gate fully-depleted CMOS transistors: Fabrication, design and layout,” *VLSI Symp. Tech. Dig.*, pp. 133–134, 2003.
7. J. Welser, J. L. Hoyt, and J. F. Gibbons, “Electron mobility enhancement in strained-Si n-type metal-oxide semiconductor field-effect transistors,” *IEEE Electron Device Lett.*, vol. 15, no. 3, pp. 100–102, 1994.
8. R. Chau, S. Datta, M. Doczy, B. Doyle, J. Kavalieros, and M. Metz, “High-k/metal gate stack and its MOSFET characteristics,” *IEEE Electron Device Lett.*, vol. 25, no. 2004, pp. 408–410, 2004.
9. P. L. McEuen, F. M. S, and H. Park, “Single-walled carbon nanotube electronics,” *Nanotechnology, IEEE Transactions on*, vol. 1, no. 1, pp. 78–85, Mar 2002.
10. A. Raychowdhury, A. Keshavarzi, J. Kurtin, V. De, and K. Roy, “Carbon nanotube field-effect transistors for high-performance digital circuits-DC analysis and modeling toward optimum transistor structure,” *Electron Devices, IEEE Transactions on*, vol. 53, no. 11, pp. 2711–2717, Nov. 2006.
11. Y.-M. Lin, J. Appenzeller, Z. Chen, Z.-G. Chen, H.-M. Cheng, and P. Avouris, “High performance dual-gate carbon nanotube FETs with 40-nm gate length,” *IEEE Electron Device Lett.*, vol. 26, pp. 823–825, 2005.
12. G. Zhang, X. Wang, X. Li, Y. Lu, A. Javey, and H. Dai, “Carbon nanotubes: From growth, placement and assembly control to 60 mV/decade and Sub-60 mV/decade tunnel transistors,” *IEDM*, pp. 1–4, Dec. 2006.
13. G. F. Close and H.-S. P. Wong, “Fabrication and Characterization of Carbon Nanotube Interconnects,” *IEDM*, pp. 203–206, Dec. 2007.
14. J. Guo, S. Datta, M. Lundstrom, “A numerical study of scaling issues for Schottky-barrier carbon nanotube transistors,” *Electron Devices, IEEE Transactions on*, vol. 51, no. 2, pp. 172–177, Feb. 2004.
15. H.-S. P. Wong, et al., “Carbon nanotube transistor circuits: Models and tools for design and performance optimization,” *ICCAD*, pp. 651–654, Nov. 2006.
16. A. Raychowdhury, S. Mukhopadhyay, and K. Roy, “A circuit-compatible model of ballistic carbon nanotube field-effect transistors,” *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 23, no. 10, pp. 1411–1420, Oct. 2004.
17. J. Deng and H.-S. P. Wong, “A compact SPICE model for carbon-nanotube field-effect transistors including nonidealities and its application Part I: Model of the intrinsic channel region,” *Electron Devices, IEEE Transactions on*, vol. 54, no. 12, pp. 3186–3194, Dec. 2007.
18. W. Zhao and Y. Cao, “New generation of Predictive Technology Model for sub-45 nm design exploration,” *ISQED*, pp. 585–590, 2006.
19. S. Sinha, A. Balijepalli, Y. Cao, “Compact model of carbon nanotube transistor and interconnect,” *IEEE Transactions on Electron Devices*, vol. 56, no. 10, pp. 2232–2242, October 2009.

20. Z. Chen, J. Appenzeller, J. Knoch, Y.-M. Lin, and P. Avouris, "The role of metal-nanotube contact in the performance of carbon nanotube field-effect transistors," *Nano Lett.*, vol. 5, pp. 1497–1502, 2005.
21. J. Guo and M. Lundstrom, *Nanoscale Transistors: Device Physics, Modeling and Simulation*, Springer, 2006.
22. D. Jimenez, X. Cartoix, E. Miranda, J. Su, F. A. Chaves, and S. Roche, "A simple drain current model for Schottky-barrier carbon nanotube field effect transistors," *Nanotechnology*, vol. 18, no. 2, p. 025201, 2006.
23. I. Amlani, J. Lewis, K. Lee, R. Zhang, J. Deng, H.-S. P. Wong, "First demonstration of AC gain from a single-walled carbon nanotube common-source amplifier," *IEDM*, pp.1–4, Dec. 2006.
24. A. Naeemi, R. Sarvari, and J. D. Meindl, "Performance comparison between carbon nanotube and copper interconnects for gigascale integration (GSI)," *IEEE Electron Device Lett.*, vol. 26, no. 2, pp. 84–86, Feb. 2005.
25. J.-Y. Park, et al., "Electron-phonon scattering in metallic single-walled carbon nanotubes," *Nano Lett.*, vol. 4, no. 3, pp. 517–520, 2004.
26. A. Raychowdhury and K. Roy, "Modeling of metallic carbon-nanotube interconnects for circuit simulations and a comparison with Cu interconnects for scaled technologies," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 25, no. 1, pp. 58–65, Jan. 2006.
27. *Raphael Interconnect Analysis Program Reference Manual*, Synopsys Inc.
28. G. Zhang, X. Wang, X. Li, Y. Lu, A. Javey, and H. Dai, "Carbon nanotubes: From growth, placement and assembly control to 60 mV/decade and sub-60 mV/decade tunnel transistors," *IEDM*, pp. 1–4, 2006.
29. J. Appenzeller, "Carbon nanotubes for high performance electronics – progress and prospect," *Proc. of the IEEE*, vol. 96, pp. 201–211, 2008.
30. Z. Chen, J. Appenzeller, J. Knoch, Y.-M. Lin, and P. Avouris, "The role of metal nanotube contact in the performance of carbon nanotube field effect transistors," *Nano Letters*, vol.5(7), pp.1497–1502, 2005.
31. V. Derycke, R. Martel, J. Appenzeller, and P. Avouris, "Controlling doping and carrier injection in carbon nanotube transistors," *Applied Physics Letters*, vol. 80, no. 15, pp. 2773–2775, 2002.
32. W. Kim, A. Javey, R. Tu, J. Cao, Q. Wang, and H. Dai, "Electrical contacts to carbon nanotubes down to 1 nm in diameter," *Applied Physics Letters*, vol. 87, no.17, p. 173101, 2005.
33. P. Tarakeshwar, J. Palacios, and D. Kim, "Interface study of metal electrode and semiconducting carbon nanotubes: Effects of electrode atomic species," *IEEE Transactions on Nanotechnology*, vol.7, no.2, pp. 124–127, March 2008.
34. F. L´eonard and J. Tersoff, "Role of fermi-level pinning in nanotube Schottky diodes," *Physical Review Letters*, vol. 84, no. 20, pp. 4693–4696, May 2000.
35. F. L´eonard and A. A. Talin, "Size-dependent effects on electrical contacts to nanotubes and nanowires," *Physical Review Letters*, vol. 97, no. 2, p. 026804, 2006.
36. H.-S. Wong, J. Deng, *A circuit-compatible SPICE model for enhancement mode carbon nanotube field effect transistors*, Synopsys Inc., Sept. 2006.
37. D. Akinwande, G. F. Close, and H.-S. P. Wong, "Analysis of the Frequency Response of Carbon Nanotube Transistors," *Nanotechnology, IEEE Transactions on*, vol. 5, no. 5, pp. 599–605, Sept. 2006.
38. J. H. Huang, Z. H. Liu, M. C. Jeng, P. K. Ko, and C. Hu, "A physical model for MOSFET output resistance," *Electron Devices Meeting,* pp. 569–572, Dec 1992.

Chapter 10
Predictive Technology Model for Future Nanoelectronic Design

Beyond that 10 nm benchmark, the present scaling approach may have to take a different route. The grand challenge to the integrated circuit design community is to identify unconventional materials and structures, such as carbon-based electronics, spintronics, nano-electromechanical relays, and steep subthreshold devices, integrate them into the circuit architecture, and enable continuous growth of chip scale and performance. The predictive technology model (PTM), which bridges the process/material development and circuit simulation through device modeling, is essential in assessing potentials and limits of new technology and in supporting early design prototyping. Figure 10.1 illustrates the roadmap of PTM development, from nominal prediction, to variational behaviors, and to heterogeneous integration beyond the Silicon.

Current PTM focuses on predictive modeling of CMOS devices down to the 12 nm node, with results validated by available TCAD simulation and silicon measurement data. Approaching the end of the silicon roadmap and going beyond, compact modeling without interface to novel materials and structures will not be adequate for advanced technological predictions. Such capabilities as first-principles calculation of bandstructure and carrier transport are must to capture the physical property of emerging materials, structures, and devices. In addition, innovative methodology for compact modeling needs to be developed since multi-dimensional effects become more significant in nanoscale devices. These exploratory models should be further implemented into realistic design environment in order to evaluate their design potential, and to construct the optimal circuit architecture.

Toward this goal, extensive research efforts are needed to cover material/structure simulation, device modeling, and design tools.

- *TCAD simulation for novel materials and structures*: CMOS will arguably be the technology of choice for the next 10 years. Besides traditional scaling efforts, novel materials and structures are necessary to enhance the performance and scalability of transistors. Nanoscale devices usually feature large ratio of surface area to device volume. The material properties, such as bandstructure, may differ from bulk ones within devices. Therefore, it becomes essential to have in-situ

Y. Cao, *Predictive Technology Model for Robust Nanoelectronic Design*,
Integrated Circuits and Systems, DOI 10.1007/978-1-4614-0445-3_10,

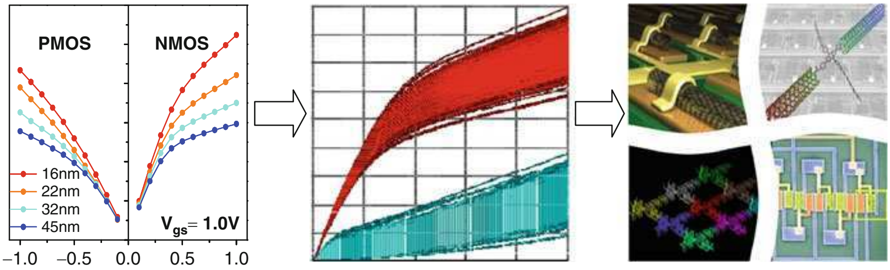


Fig. 10.1 Future development of predictive technology model

material characterization capabilities in device simulation. The efficiency and accuracy of current calculation methods, such as those for the bandstructure and quantum transport, need to be significantly improved and integrated into device simulation.

- *Compact modeling and design kits for early design research*: Predictive device models are the critical interface between technology innovation and exploratory circuit design. They should be scalable with latest technology advances, accurate across a wide range of process and operation conditions, and efficient for large-scale computation. In the nanometer regime, these demands are tremendously challenged by the introduction of alternative materials and structures that boost CMOS performance, as well as more radical device experiments beyond CMOS. These technological solutions extend the scaling, but also result in new physical effects that are not well captured in today's compact models, such as the layout dependence, carrier transport, and 2D or even 3D channel. Novel compact modeling approach will be crucial to describe these effects in device operation. In addition to modeling of intrinsic components, parasitic effects, especially the contact, become increasingly important realistic design evaluation.

  In nanoelectronic design, the modeling task is compounded with ever-increasing process variations and reliability degradation, when technology scaling eventually reaches the ultimate limits that are defined by physics and manufacturability. The exact amount of variations further depends on layout and operation conditions. PTM will continuously provide not only nominal model files for scaled CMOS and post-Si devices, but also analytical models to account for systematic and random variations. These models will help shed light on robust design solutions, generating physical insights into process and design choices. They will be implemented into circuit simulators and further lead to the development of statistical process design kits.

Overall, future development of PTM seeks general and flexible models that are able to efficiently bridge emerging device research with circuit design infrastructure. With an integral set of TCAD simulation, compact modeling and design kits, PTM aims to achieve a coherent environment of technological prediction and exploratory design research. Such a predictive capability will ensure a timely and smooth transition from CMOS-based design to robust integration with post-silicon technologies.

Index

Y. Cao, *Predictive Technology Model for Robust Nanoelectronic Design*,
Integrated Circuits and Systems, DOI 10.1007/978-1-4614-0445-3,