# Variability-Aware Compact Modeling of Nano-scale Technologies with Customized Test Structure Designs

by

Ying Qiao

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Engineering – Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Costas J. Spanos, Chair

Professor Borivoje Nikolić

Professor Haiyan Huang

Spring 2017

Variability-Aware Compact Modeling of Nano-scale Technologies
with Customized Test Structure Designs

by

Ying Qiao

# Abstract

Variability-Aware Compact Modeling of Nano-scale Technologies
with Customized Test Structure Designs

by
Ying Qiao

Doctor of Philosophy in Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Costas J. Spanos, Chair

It is widely recognized that in nano-scale CMOS technology variation in the manufacturing process has emerged as a fundamental challenge to IC design. While foundries are working hard to mitigate process variability, the design houses are asking for accurate and appropriate models to handle statistical circuit performance evaluation. To accurately represent the process and device variability, it is essential to incorporate the variability during the extraction and calibration phase of compact transistor models. In addition, these compact transistor models require customized test structure designs as well as proper statistical characterization procedures. Conventional statistical compact model characterization methodologies require special single transistor, direct-access test arrays, or virtual measurements from physical simulation data; moreover, these models do not include rigorous statistical model parameter selection criteria.

Our proposed variability-aware compact transistor models can enable statistically optimized designs by capturing device variations in a concise, yet physically accurate way, and they are relatively easy to integrate with existing CAD tool flow. In this work, we have electrical measurements from carefully designed SRAM array test structures with bit transistor access, fabricated using a collaborating foundry's 28nm FDSOI technology. Stepwise parameter selection is combined with sequential extractions of statistical compact model parameters upon foundry-provided nominal compact model cards. These nominal models are trusted as they have been tested extensively and used during the test structure design.

Our characterization methodology selects an optimal statistical model parameter set that can be reliability extracted with the given measurement data. With further data from imaging ROIC test array, we are able to explore the linear spatial propagation of variance method to extract the variability in the compact model parameters with hierarchical models.

We have also built a customized Monte Carlo (MC) simulation platform to utilize these compact transistor models in the statistical IC design flow. Different statistical model parameters can be specified prior to the MC simulation within the scripted wrapper of standard SPICE-based simulators. We further exploit the statistical structure of the extracted parameters in order to capture the nonlinear correlations and the non-Gaussian distributions through mixture of Gaussian distributions. The goal is to demonstrate that significant non-normality in the measured data can be captured by our simplified model. Such non-normality is often evident at the tails of the performance distributions, and capturing that is necessary for the statistical modeling of inherently high-yielding IC designs.

*To my mom and dad*

# Contents

# List of Figures and Tables

## Acknowledgments

First, I want to express my deepest and most sincere gratitude to my research advisor, Prof. Costas Spanos, who has provided me with outstanding support in both personal life and research work throughout my years of doctorate study. Costas has been my research advisor for over six years; and I have always found him dependable, inspirational and unfailingly excited about mentoring. His willingness to take on challenging projects like my early-stage design proposals and see them to successful completion has repeatedly impressed me over the years. I also became acutely aware of his great knowledge and skills in guiding me towards better presentation, as well as the endless hours he worked in a determination to help finish my manuscripts before a top conference submission deadline. I cannot imagine accomplishing so much work without his enthusiastic, yet well-organized mentorship.

I am very grateful to Professor Bora Nikolic for collaboration and support in designing, fabricating and measuring the test patterns I used in this work. I enjoyed our research discussions and I learned a lot from you, who sets the standard for how to be a great researcher, to always try to seek insight and intuition behind a problem. I also owe my gratitude to my thesis committee member, Professor Haiyan Huang, for reading my thesis and providing support suggestions each time I have interacted with her.

The majority part of the thesis project is sponsored by the IMPACT+ funding from Applied Materials, ARM, ASML, Global Foundries, IBM, Intel, KLA-Tencor, Mentor Graphics, Panoramic Tech, Qualcomm, Samsung, SanDisk and Tokyo Electron.

I very much appreciate the collaboration and friendship from all my group mates as well as from Berkeley Wireless Research Center (BWRC), present and past: Brian, Katerina and Amy. Kun, thank you for all your noble help. Zhaoyi and Claire, thanks for all the collaboration during class and all the good times. Yu and Qianying, thank you for all the laughs we shared, and all the advice you gave me, academic or non-academic. You made my life here truly fun and enjoyable.

Finally, my family deserves all the credit for this accomplishment, and thank you for your unconditional love. Cuimei and Jinping, you are the best mom and dad one can ever wish for. Thank you for always being there with me to share the cheers during my success and comfort me during my down times. To my husband, Qikai, you have been extremely loving and caring from the time we met and I cannot thank you enough. Without you, my accomplishment here would mean nothing to me. You deserve every part of this thesis as much as I do. Thank you and I love you so much.

# Chapter 1

# Introduction

## 1.1 Motivation: Process Variation

In 1965, Gordon Moore observed that the number of transistors on a single chip doubled every 18 to 24 months [1-1]; this is an observation now widely known as Moore's Law. The perpetual doubling of transistor density has served as the driving force of an astonishing increase in the functionality and computational capability of electronic devices since then hitherto. Between generations, minimum transistor dimensions undergo scaling by a factor of 0.7, enabling the integration of more transistors with less power dissipation. In recent years, however, several bottlenecks have appeared as we continue to scale down beyond sub 28nm technologies. One of the key issues related to deeply scaled semiconductor manufacturing is the yield, defined as the proportion of manufactured circuits that are functional and meet their performance requirements [1-2]. The overall yield loss falls into two major categories: catastrophic yield loss (due to physical and structural defects, e.g., open, short, etc.) and parametric yield loss (due to parametric variations in process parameters, e.g., threshold voltage, stress, etc.). A large portion of yield loss in circuits now occurs due to process variations, which can be defined as the deviations in the manufactured circuit in comparison to its design [1-3].

With decreasing transistors sizes and increasing transistor densities, the effect of process and manufacturing variabilities are more significant; meeting performance and yield specifications is increasingly challenging. For example, Figure 1-1 shows the general trend in the ratio between the corresponding 3$\sigma$ variation and mean value for some key technology devices and wire parameters from 250*nm* to 45*nm*. Over the time of interest, we see that the proportion of $L_{eff}$ variation increases from 30% to 45%. Wire geometry parameters - width $W$, height $H$ and resistivity $\rho$ - also undergo significant increases. Other parameters such as the threshold voltage $V_{th}$ and oxide thickness $T_{ox}$ increase at a lower rate.

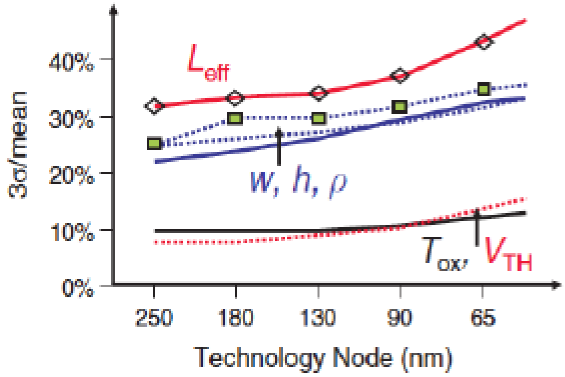


**Figure 1-1 Increase of process variability in conventional scaled MOS technology [1-2]**

Increasing process variations introduce significant uncertainty for both circuit performance and leakage power. It has been shown in that even for the 180*nm* technology, process variation can lead to 1.3X variation in frequency and 20X variation in leakage power [1-4]. In future technology generations, such an impact will become more magnified because the technology is approaching a fundamental randomness regime in the behavior of silicon structures. In recent years, Design for Manufacturability (DFM) methods, including attempts to reduce the systematic sources of variability, statistical modeling, extraction, and optimization for VLSI circuits, have been developed to alleviate the variation effects. For DFM to be meaningful, however, variability needs to be characterized empirically for a specific semiconductor process to obtain a quantitative understanding of variability mechanisms. Such “statistical metrology" methods include measurement techniques for the characterization of variabilities, and statistical modeling and extraction methods for properly interpreting measurement results.

## 1.2 Background

Background knowledge about advanced transistor structures, sources of variability and test structures for variability characterization is necessary to understand the motivation and context behind the following chapters in this work.

### 1.2.1 Advanced Transistor Structures

To achieve a higher transistor density on a chip, the dimensions of a transistor must be scaled. Constant field scaling - where device dimension, doping density, and supply voltages are scaled simultaneously - has worked for a while, but it has started to slow down dramatically when the minimum half-pitch reached 90*nm* [1-5]. Figure 1-2 shows a series of transmission electron microscopy (TEM) images of classical planar bulk MOSFET from the 65*nm* technology node down to the 32*nm* technology node [1-6]. Here, gate leakage due to the thin layer of gate oxide ($SiO_2$) can be mitigated by using a high-permittivity (high-k) dielectric and metal gate (HKMG) stack [1-8]. As the gate length of a transistor is made smaller, non-ideal effects, which are negligible at long gate lengths, can degrade transistor performance. These effects are usually referred to as short-channel effects (SCEs) [1-9].

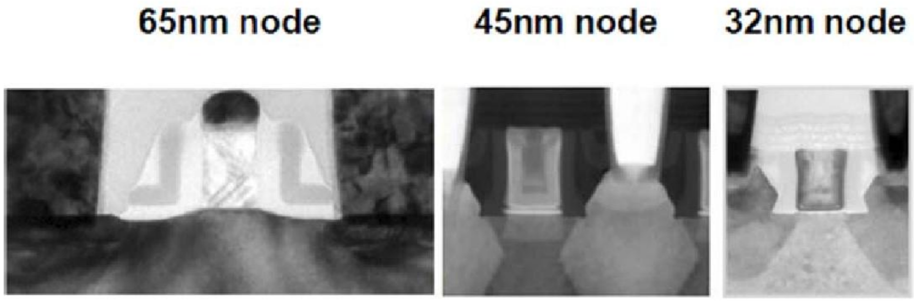


**Figure 1-2 TEM cross-sections of MOSFETs at various technology nodes, with approximately the same scale [1-7]**

Short-channel effects can become very difficult to suppress in a planar bulk transistor for gate lengths below 25*nm*. Therefore, to allow further miniaturization, advanced transistor structures employing thin-body regions such as the fully depleted silicon-on-insulator (FD-SOI) [1-10] and the three-dimensional FinFET have been developed for future generations of CMOS technology [1-11].

A FD-SOI MOSFET is a planar structure that is fabricated in a thin *Si* layer on top of a buried oxide (BOX) layer. An illustration of the FD-SOI MOSFET structure is shown in Figure 1-3(a) and a TEM cross-section is shown in Figure 1-3(b). Electrostatic gate control in the FD-SOI MOSFET is superior to that in a planar bulk MOFSET due to its thin silicon body, since OFF-state leakage current paths far away from the gate are eliminated [1-13]. Because of the minimal differences between FD-SOI and planar bulk MOSFETs, few changes are needed to migrate circuit designs from bulk to FD-SOI.

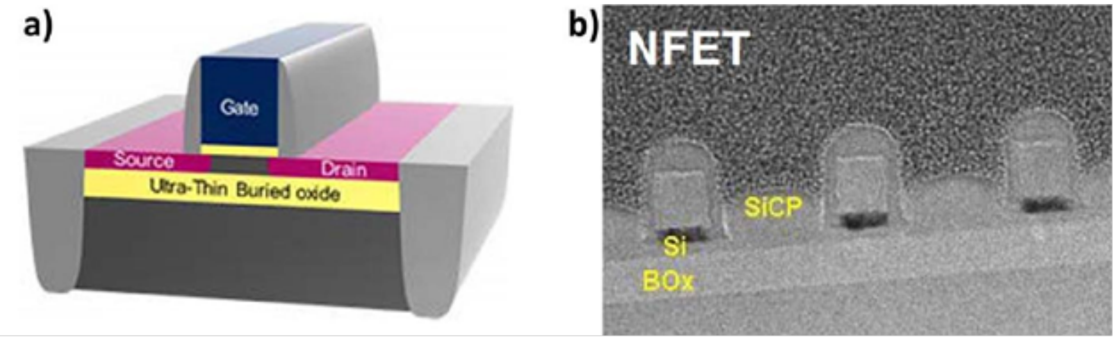


**Figure 1-3 (a) 3D view of a planar FD-SOI transistor [1-12];
(b) A TEM cross-section of N-channel FD-SOI transistors [1-13].**

Another variation of a thin-body MOSFET is the vertical FinFET, where the body has a fin-like shape and a gate electrode straddles it, as depicted in Figure 1-4. Due to gating from all three sides of the channel, SCE can be well suppressed if the fin width is less than half the gate length ($W_{Si} \sim L_g/2$). Experimental results have demonstrated FinFET transfer characteristics with low sub-threshold swing and low drain-induced barrier lowering (DIBL) [1-10]. Recently, the FinFET has supplanted the planar bulk MOSFET in the most advanced microprocessor chips at leading semiconductor foundries.

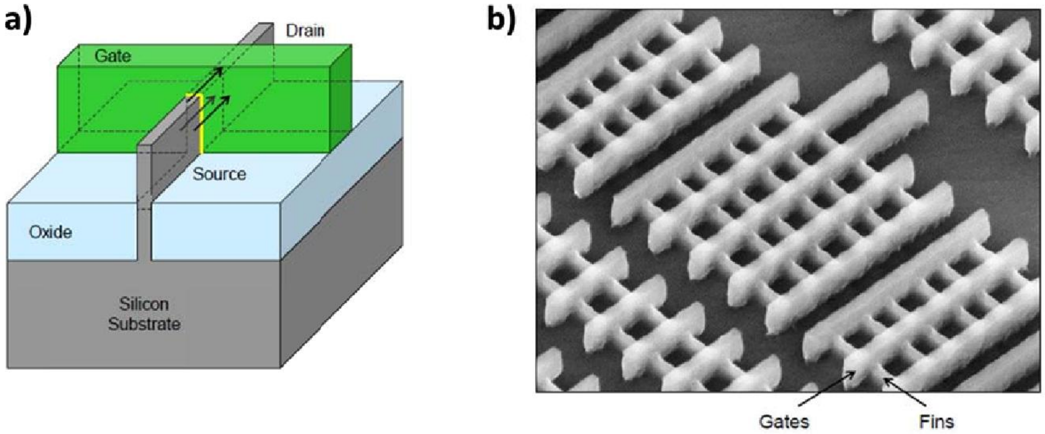


**Figure 1-4 (a) 3D view o f a vertical FinFET transistor;
(b) TEM image of an array of FinFET transistors showing the fin and gate features [1-14].**

## 1.2.2 Sources of Variation

Sources of transistor variation caused by an imperfect manufacturing process are often categorized as either systematic or random. Systematic (or global) variations, dependent on the layout of the transistors and its surroundings, affect all devices on the same die, wafer, or lot; this causes devices of different physical hierarchies or locations to have different drive current versus gate voltage characteristics. Random (or local) variation, on the other hand, can result in differences between identically drawn transistors within the same layout environment, reducing matching in analog differential pairs or proportionally sized devices in SRAM cells. As the critical dimension decreases with technology advancements, the contributions of different physical sources of variation such as random dopant fluctuations (RDF) or line edge roughness (LER) change [1-15].

RDF is caused by variations in the number and placement of dopant atoms in the channel region of the transistor. The small number of dopant atoms makes the threshold voltage susceptible to even the slightest amount of dopant variation. Reducing device dimensions and increasing doping density will result in a larger $V_{th}$ variation. Conversely, by reducing the effective thickness of the gate oxide (e.g. by adopting a high-k dielectric), $\sigma_{V_{th}}$ can be reduced. A more detailed analysis of $V_{th}$ variations due to RDF can be performed through 3D device simulations, which randomly place dopant atoms within the transistor [1-16].

The gate length, aka critical dimension (CD) of the transistor becomes so small that a slight deviation from the nominal value can have a large effect on electrical performance. LER, one of the main contributors to random variation in CD, is caused by the granularity of the photoresist material (used to define the pattern of the gate electrodes) at the molecular level [1-17]. As the CD decreases, LER does not decrease commensurately, which can result in large $V_{th}$ variations [1-18]. LER can affect both the gate length and channel width of a transistor as depicted in Figure 1-5(a). With a more advanced structure like the FinFET, fin width variations caused by LER is also becoming a major concern. Characterization of LER can be done efficiently by analyzing the top view of the scanning electron microscope (SEM) image of the feature as in Figure 1-5(b) [1-19].

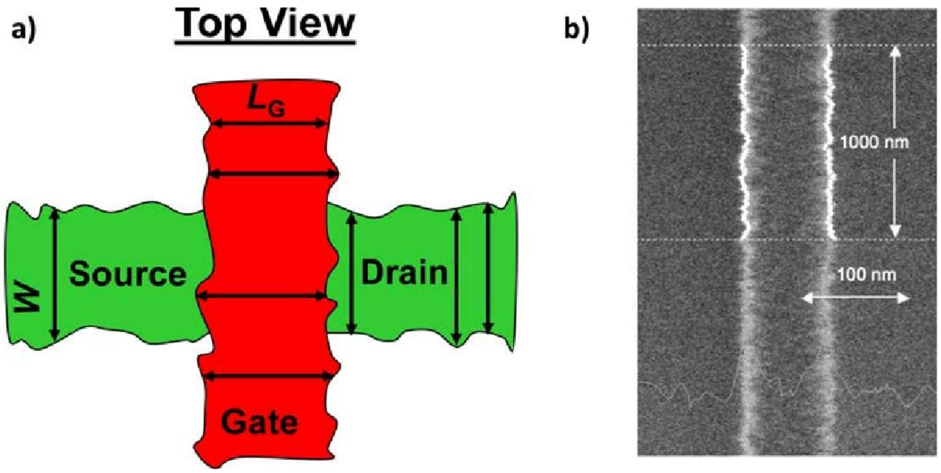


**Figure 1-5 (a) Top view showing how LER can affect the gate length and channel width of a transistor; (b) SEM image of a photoresist line with LER [1-19].**

When we analyze the scaling factors of the recent technology nodes, we come to the conclusion that the electrical characterization of these really challenging technologies becomes an absolute must. However, the extremely small process windows, and the 3-dimensional nature of the FinFET devices and the complicated interconnect schemes, make this characterization very difficult [1-20]. As FinFETs emerging to be the solution for short channel effects, fin shape significantly impacts transistor leakage in bulk tri-gate nFinFETs with thin fins when the fin body doping profile is optimized to minimize leakage [1-21]. With appropriate doping optimization, a 22-nm nFinFET with triangular fin cross section results in an ineligible reduction in leakage current over a rectangular fin with the same base fin width [1-21].

### 1.2.3 Test Structures for Variation Characterization

To improve our understanding of process variation and ultimately reduce this variation to improve yield, process variation needs to be thoroughly characterized with the help of on-chip test structures. The test structures are devices or circuits that are added onto a wafer to help control, understand, and model the behavior of MOSFETs. Per the objective of measurements, test structures fall into two classes: (a) test structures for process control, and (b) test structures for modeling [1-2].

Test structures for process control are used for monitoring and controlling the fabrication line. These are typically small devices or circuits placed in the scribe line on all wafers and therefore can model the history of the line. Monitors often consist of simple test structures that allow the measurements of current-voltage (I~V) characteristics of MOSFETs [1-22], of the resistivity of wires and vias [1-23], and of interconnect capacitance [1-24].

It is important to extract physically meaningful model parameters, as they help to identify possible root causes for process failures, and help drive optimization of the process in early stage process development. The difficulty of this problem comes from the fact that measurements are collected from a limited number of early prototype devices rather than from a full suite of designed test structures. Test structures for modeling are used to generate the fundamental data needed to create models of the fabricated components. These test structures are complex in nature and are typically designed to be sensitive to a specific physical parameter. Therefore, a much richer variety of test structures is needed for modeling purposes.

Prior comprehensive test structures of variability studies have previously been published for 45*nm* [1-25] CMOS, but only limited variability information is available for technology nodes at 28*nm* and below. Moreover, prior experimental work has focused on addressable device arrays with analog switches that require long design times and cumbersome testing setups with separate bench supplies for each device terminal [1-26]. To address this issue, some prior work [1-27] has demonstrated custom technology characterization chips with fully digital interfaces, including a complete current vs. voltage (*I-V*) characterization chip with an integrated digital to analog converter (DAC) and analog to digital converter (ADC).

## 1.3 Thesis Organization

In this thesis we propose accurate and efficient statistical techniques to solve the following problem in DFM: given hands-on knowledge of the test chip design and the measurements of one or several functions (e.g., transistor *I-V* measurements), we need to find the value of transistor compact model parameters (such as parameters sensitive to process variation) to predict circuit performance and to eventually improve product yield.

In Chapter 2, the design of a test chip to study the impact of transistor variation in 28*nm* planar bulk and FD-SOI MOSFETs is discussed. A device characterization array including transistors in mismatch pairs and different layout proximities are used to study the impact of random and systematic variation respectively. Measurement setup is documented in detail, and transistor variation observation and characterization limitations are discussed.

In Chapter 3, we establish the background for statistical compact variability modeling. We propose a general parameter extraction method to enable the extraction of an entire set of MOSFET *I-V* model parameters, even in the face of few or missing *I-V* measurements in the data set. Our improved method for modeling the variability of transistors with compact models works seamlessly with the widely-used PSP model for efficient statistical circuit performance estimation.

In Chapter 4, we apply the statistical compact model extraction methodology to the actual silicon data collected from the SRAM bit cells on the 28*nm* test chip. For the foundry-wrapped PSP model, stepwise parameter selection is applied to find an optimal set of model parameters, from which the measurement data is characterized to create a custom statistical compact model. These statistical compact models provide better accuracies in predicting bit-cell static performance variations than conventional methods.

Chapter 5 demonstrates the full statistical device characterization and circuit performance modeling on the imaging ROIC test array measurement data. With a larger dataset including chip spatial info, we can apply the linear spatial backward propagation of variance method for compact model parameters. Furthermore, a mixture of Gaussian model can be applied onto non-Gaussian correlated model parameters for better circuit performance statistical modeling. Results are shown in detail in this chapter.

Chapter 6 concludes the thesis. We summarize the key contributions of the thesis. With challenges in the analysis of process variation ahead, areas for future research are suggested, including variation prediction within a process development cycle.

## References

[1-1] G. Moore, *"Cramming more components onto integrated circuits,"* Proceedings of the IEEE, vol. 86, no. 2, pp. 82-85, 1998.

[1-2] M. Orshansky, S. Nassif, and D. Boning, *Design for Manufacturability and Statistical Design: A Constructive Approach.* Springer, 2010.

[1-3] W. Zhang, *"IC spatial variation modeling: Algorithms and applications,"* Ph.D. dissertation, Carnegie Mellon University, Department of Electrical Engineering, 2012.

[1-4] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De, *"Parameter variations and impact on circuits and microarchitecture,"* in Design Automation Conference (DAC), June 2003, pp. 338-342.

[1-5] W. Arden, M. Brillouet, P. Cogez, M. Graef, B. Huizing, R. Mahnkopf, *"Morethan-Moore White Paper,"* ITRS 2010. Available:

[1-6] S. Yang, L. G. Lin, M. Han, D. Yang, J. Wang, K. Mahmood, T. Song, D. Yuan, D. Seo, M. Pedrali-Noy, D. Alladi, S. Wadhwa, X. Bai, L. Dai, S. S. Yoon, E. Terzioglu, S. Bazarjani, G. Yeap, *"High Performance Mobile SoC Design and Technology Co-Optimization to Mitigate High-K Metal Gate Process Induced Variations"* in VLSI Symp. Tech. Dig., 2014.

[1-7] V. Moroz, *"Transition from Planar MOSFETs to FinFETs and its Impact on Design and Variability"*, Berkeley Seminar, Oct. 2011.

[1-8] T. Ghani, M. Armstrong, C. Auth, M. Bost, P. Charvat, G. Glass, T. Hoffmann, K. Johnson, C. Kenyon, J. Klaus, B. McIntyre, K. Mistry, A. Murthy, J. Sandford, M. Silberstein, S. Sivakumar, P. Smith, K. Zawadzki, S. Thompson, M. Bohr, *"A 90nm High Volume Manufacturing Logic Technology Featuring Novel 45nm Gate Length Strained Silicon CMOS Transistors,"* IEDM Tech. Dig., 2003.

[1-9] Y. Taur; D. A. Buchanan; W. Chen; D.J. Frank; K.E. Ismail; L. Shih-Hsien; G.A, Sai-Halasz; R.G. Viswanathan; H.-J.C. Wann; S.J. Wind; Hon-Sum Wong, *"CMOS scaling into the nanometer regime,"* Proceedings of the IEEE, vol.85, no.4, pp.486,504, Apr 1997.

[1-10] C. Auth; C. Allen; A. Blattner; D. Bergstrom; M. Brazier; M. Bost; M. Buehler; V. Chikarmane; T. Ghani; T. Glassman; R. Grover; W. Han; D. Hanken; M. Hattendorf; P. Hentges; R. Heussner; J. Hicks; D. Ingerly; P. Jain; S. Jaloviar; R.James; D. Jones; J. Jopling; S. Joshi; C. Kenyon; H. Liu; R. McFadden; B.Mcintyre; J. Neirynck; C. Parker; L. Pipes; I. Post; S.

Pradhan; M. Prince; S.Ramey; T. Reynolds; J. Roesler; J. Sandford; J. Seiple; P. Smith; C. Thomas; D.Towner; T. Troeger; C. Weber; P. Yashar; K. Zawadzki; K. Mistry, "A 22nm high performance and low-power CMOS technology featuring fully-depleted trigate transistors, self-aligned contacts and high density MIM capacitors," VLSI Technology (VLSIT), 2012 Symposium on , pp.131,132, 12-14 June 2012.

[1-11] W. Shien-Yang; C,Y, Lin; M.C. Chiang; J.J. Liaw; J.Y. Cheng; S.H. Yang; M. Liang; T. Miyashita; C.H. Tsai; B.C. Hsu; H.Y. Chen; T. Yamamoto; S.Y. Chang; V.S. Chang; C.H. Chang; J.H. Chen; H.F. Chen; K.C. Ting; Y.K. Wu; K.H. Pan; R.F. Tsui; C.H. Yao; P.R. Chang; H.M. Lien; T.L. Lee; H.M. Lee; W. Chang; T. Chang; R. Chen; M. Yeh; C.C. Chen; Y.H. Chiu; Y.H. Chen; H.C. Huang; Y.C. Lu; C.W. Chang; M.H. Tsai; C.C. Liu; K.S. Chen; C.C. Kuo; H.T. Lin; S.M. Jang; Y. Ku, "A 16nm FinFET CMOS technology for mobile SoC and computing applications," Electron Devices Meeting (IEDM), 2013 IEEE International, pp.9.1.1,9.1.4, 9-11 Dec. 2013

[1-12] ST Microelectronics. <http://www.st.com/web/en/about_st/fd-soi.html>

[1-13] L. Le Pailleur, *"28nm FD-SOI Industrial Solution: Overview of Silicon Proven Key Benefits"* FDSOI- Workshop 2013 Kyoto, Japan.

[1-14] M. Bohr, K. Mistry, *"Intel's Revolutionary 22 nm Transistor Technology,"* Intel Presentation, May 2011.

[1-15] K. Kuhn, M. Giles, D. Becher, P. Kolar, A. Kornfeld, R. Kotlyar, S. Ma, A. Maheshwari, and S. Mudanai, *"Process Technology Variation,"* Electron Devices, IEEE Transactions on, vol. 58, pp. 2197-2208, Aug 2011.

[1-16] K. Takeuchi, A. Nishida, T. Hiramoto, *"Random Fluctuations in Scaled MOS Devices,"* Simulation of Semiconductor Processes and Devices, 2009. SISPAD '09. International Conference on, vol., no., pp.1,7, 9-11 Sept. 2009

[1-17] Y. Ma; H. J. Levinson; T. Wallow, *"Line edge roughness impact on critical dimension variation,"* Proc. SPIE 6518, Metrology, Inspection, and Process Control for Microlithography XXI, 651824 (April 05, 2007)

[1-18] A. Asenov, *"Simulation of statistical variability in nano MOSFETs,"* IEEE Symp. VLSI Technol., Dig. Tech. Papers, Jun. 2007, pp. 86–87.

[1-19] A. Hiraiwa, A. Nishida, *"Statistical- and image-noise effects on experimental spectrum of line-edge and line-width roughness,"* J. Micro/Nanolith. MEMS MOEMS, Dec. 2010.

[1-20] Andrzej J. Strojwas; Jacob Abraham; Hong Hao; Max Shulaker, *"Challenges and opportunities in electrical characterization and test for 14nm and below"*, 2016 IEEE 34th VLSI Test Symposium (VTS), 2016.

[1-21] B. D. Gaynor and S. Hassoun, *"Fin Shape Impact on FinFET Leakage With Application to Multithreshold and Ultralow-Leakage FinFET Design,"* in IEEE Transactions on Electron Devices, vol. 61, no. 8, pp. 2738-2744, Aug. 2014.

[1-22] C. Hess, A. Inani, Y. Lin, M. Squicciarini, R. Lindley, and N. Akiya, *"Scribe characterization vehicle test chip for ultra-fast product wafer yield monitoring,"* in Microelectronic Test Structures, 2006. ICMTS 2006. IEEE International Conference on, Mar. 2006, pp. 110-115.

[1-23] L. W. Linholm, R. A. Allen, and M. W. Cresswell, *"Microelectronic test structures for feature placement and electrical linewidth metrology,"* in Handbook of Critical Dimension Metrology and Process Control. New York: SPIE Optical Engineering Express, 1994, vol. CR52, pp. 91 - 118.

[1-24] J. Chen, D. Sylvester, and C. Hu, *"An on-chip, interconnect capacitance characterization method with sub-femto-farad resolution,"* IEEE Transactions on Semiconductor Manufacturing, vol. 11, no. 2, pp. 204-210, May 1998.

[1-25] S. Realov and K. Shepard, *"On-chip combined C-V / I-V characterization system in 45-nm CMOS technology,"* IEEE Journal of Solid-State Circuits, vol. 48, 2013.

[1-26] S. Saxena, *"Variation in transistor performance and leakage in nanometer-scale technologies,"* IEEE Transactions on Electron Devices, vol. 55, 2008.

[1-27] S. Realov, W. McLaughlin, and K. Shepard, *"On-chip transistor characterization arrays with digital interfaces for variability characterization,"* International Symposium on Quality Electronic Design, 2009.

# Chapter 2

# Variability Characterization Test Structures

## 2.1 Introduction

To satisfy Moore’s Law, transistors are miniaturized in each successive technology node so that more of them can be put onto a chip [2-1]. However, such aggressive scaling can also have an adverse effect on the electrostatic integrity of a transistor, causing large off-state current and worsening short-channel effects. One of the root causes of poor electrostatic control is relatively weak capacitive gate coupling to the electric potential in the silicon body region that is further from the gate-oxide interface [2-2]. To tackle this challenge head on, one can think of removing all paths far away from the gate, which is precisely the idea behind the thin-body (fully depleted) MOSFET [2-3]. If the thickness of the silicon body is made much thinner than the gate length, short-channel effects are dramatically reduced. The two most common implementations of a thin-body MOSFET today are the vertical FinFET or planar FDSOI (Fully-Depleted Silicon-On-Insulator) MOSFET [2-4].

The FinFET is a double-gate MOSFET structure, which is more scalable compared to the FDSOI MOSFET due to superior gate control. However, it requires a high aspect ratio Si fin geometry, which presents a major challenge from a fabrication standpoint. Additionally, since the drive strength of a FinFET is adjusted by changing the number of fins, circuit designers must cope with discrete adjustments in the drive current for FinFETs [2-5]. On the other hand, the FD-SOI MOSFET structure, which also uses a thin body like the FinFET, offers improved electrostatic control over the planar bulk MOSFET without adding significant fabrication challenges or imposing new restrictions on circuit design. Instead of a bulk Si wafer, the starting substrate is a Silicon-On-Insulator (SOI) wafer [2-6]. The device fabrication process steps are very similar and less complicated compared to those of a standard planar bulk Si device fabrication process. From a circuit designer's standpoint, the FDSOI design kit is also easier to adapt from that of bulk Si technology: device widths can be adjusted to tune transistor drive strength, and back-biasing can be used to dynamically adjust transistor threshold voltage [2-7].

Given that FDSOI technology is a promising candidate to replace planar bulk Si technology, variability analysis of FDSOI MOSFETs is necessary. This can be achieved by implementing a device characterization array in a test chip.

## 2.2 Transistor Characterization Array Design

To capture and understand the impact of different variability sources on device performance, transistors of different sizes and layout geometries are included in the transistor characterization array. Using built-in circuitry on the chip, each individual transistor can be electrically accessed and characterized through the input/output pads. In general, the variability test structures can be classified as *random* or *systemic* variability test structures. Both NMOS and PMOS transistors of different threshold values are included. Test chips were fabricated by STMicroelectronics using a 28*nm* high-k/metal-gate (HKMG) process, on both bulk-Si and SOI substrates to allow for a direct comparison of planar bulk vs. FDSOI technologies. The layout of the transistor characterization block is shown in Figure 2-1.

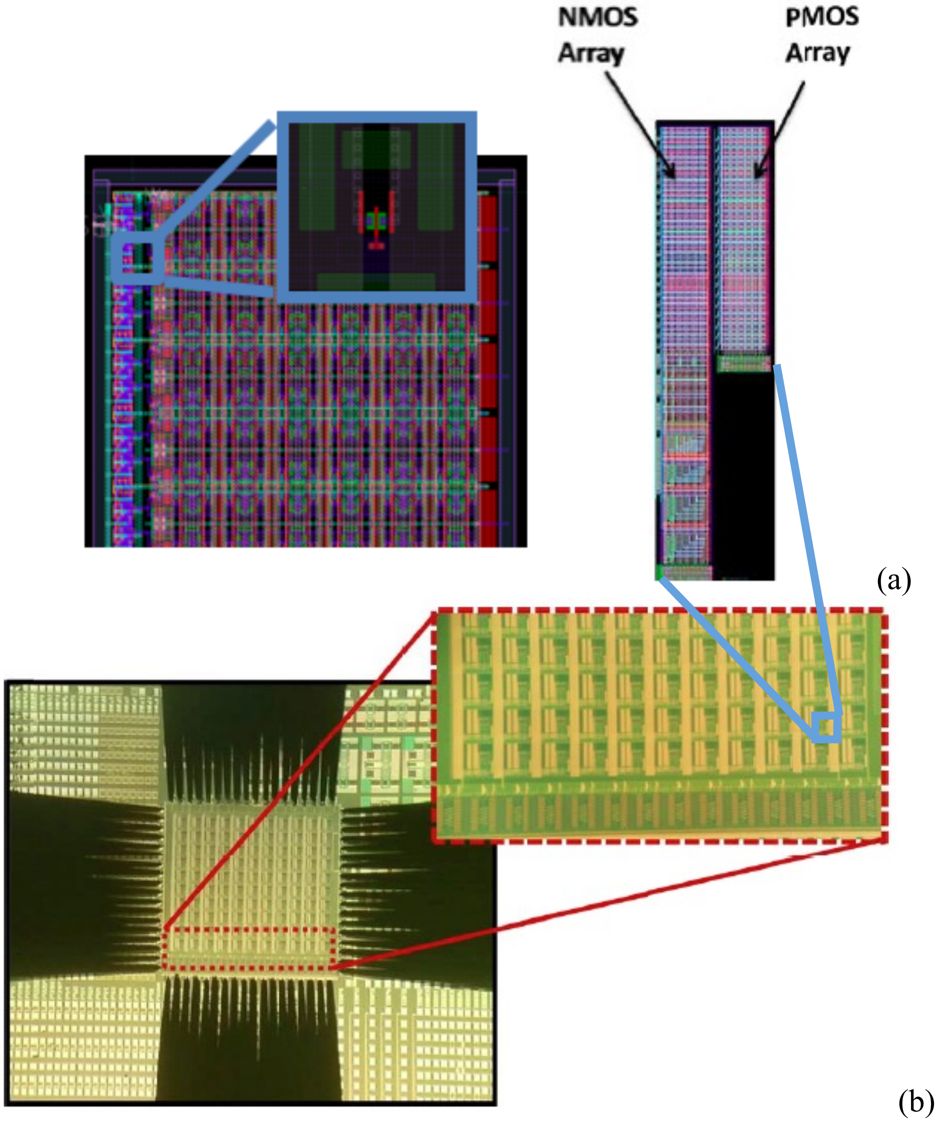


**Figure 2-1 (a) Layout of transistors characterization block and DUT (b) Photo of a die under test with probe tips landed on top; magnified photo of the die showing the macro on the bottom row.**

### 2.2.1 Random Variability Test Structures

Random variability sources such as random dopant fluctuations (RDF), gate work function variation (WFV), and line-edge roughness (LER) can contribute to variations in $V_{th}$, $I_{OFF}$, and $I_{ON}$ between devices with identical layouts. To isolate the impact of random variability from that of systematic variability, transistor pairs (i.e. mismatch test structures) are often used. These test transistors are identically drawn structures that are placed near one another on the chip. If there were a systematic source of variability, its impact would be the same for both devices. Thus, when the difference (as opposed to the absolute value) of the performance parameters between the two transistors in a pair is analyzed, the impact due to systematic variability will cancel out, i.e. the difference will entirely be caused by random variability. To ensure that the transistors in a pair are identical in every possible aspect, it is important to make sure that the surrounding area is the same for both transistors. Figure 2-2 shows a Device-Under-Test (DUT) surrounded by dummy active regions. The other corresponding DUT in the pair is also drawn in a similar manner. Such a layout will help to eliminate variability that might arise from layout-dependent proximity effects such as mechanical stress from Shallow Trench Isolation (STI) or near-by active devices [2-8].

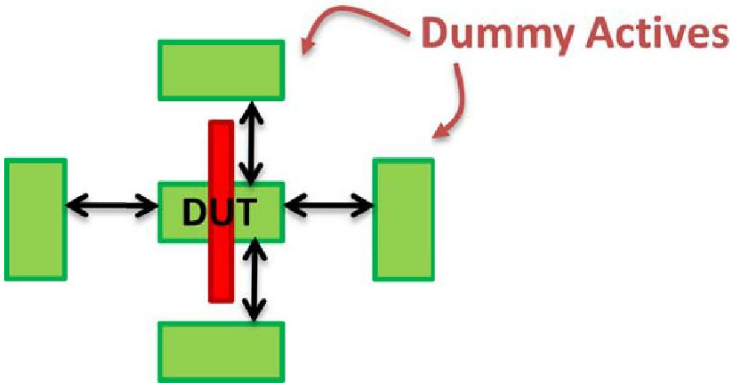


**Figure 2-2 Layout showing DUT surrounded by dummy active regions with equal distance to eliminate any layout-dependent proximity effects**

It has been theoretically derived and experimentally validated that variability in MOSFET threshold voltage increases as the transistor channel dimensions are made smaller. Specifically, the variation of MOSFET threshold voltages is proportional to transistor channel dimensions [2-9]. Thus, devices with different channel area values are included in the array to assess the significance of this trend. To further examine the sensitivity of variability sources to various transistor design parameters (e.g. channel width and gate length), different combinations of $W$ and $L$ corresponding to a fixed channel area are also included. This allows us to decouple different variation sources affecting the transistor threshold voltage.

For System-On-Chip (SOC) products, multiple values of $V_{th}$ must be available to the designers [1-1]. Therefore, it is also important to investigate how variability will affect transistors of different nominal $V_{th}$ values. To this end, three different $V_{th}$ levels (Low, Regular, and High $V_{th}$) are included for each value of width/length combinations.

### 2.2.2 Systematic Variability Test Structures

In addition to transistor structures used to study random variability, several device structures are included to assist with the study of systematic variability associated with layout proximity effects, including mechanical stress induced by STI, Length of Diffusion (LOD), well doping proximity, and segmented channel design. The design and layout of these structures are summarized in the following sections.

#### 2.2.2.1 Shallow Trench Isolation (STI) Effect

Mechanical stress induced by STI can affect carrier mobility and thereby transistor on-state drive currents [2-11]. To quantify the impact of STI-induced stress from different directions, dummy active regions are drawn at different distances ($\lambda$) away from the device under test. Similarly, the effect of STI-induced stress across the channel (along the width direction) can also be captured by placing the dummy active regions at the top and bottom of a DUT at different distances as depicted in Figure 2-3.

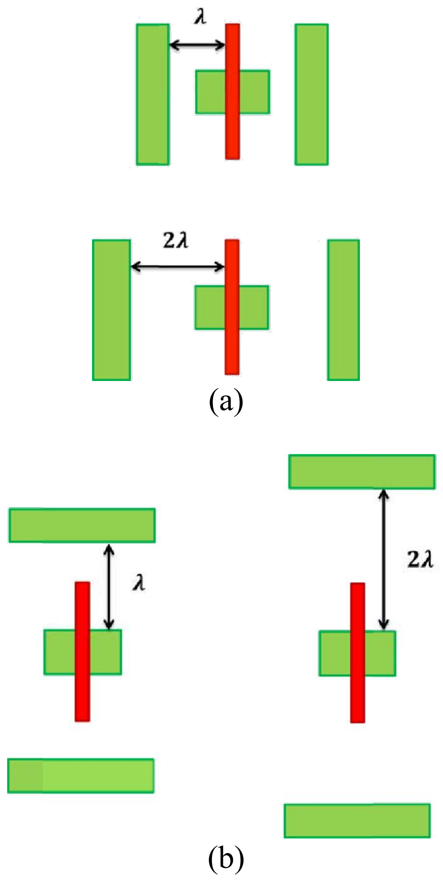
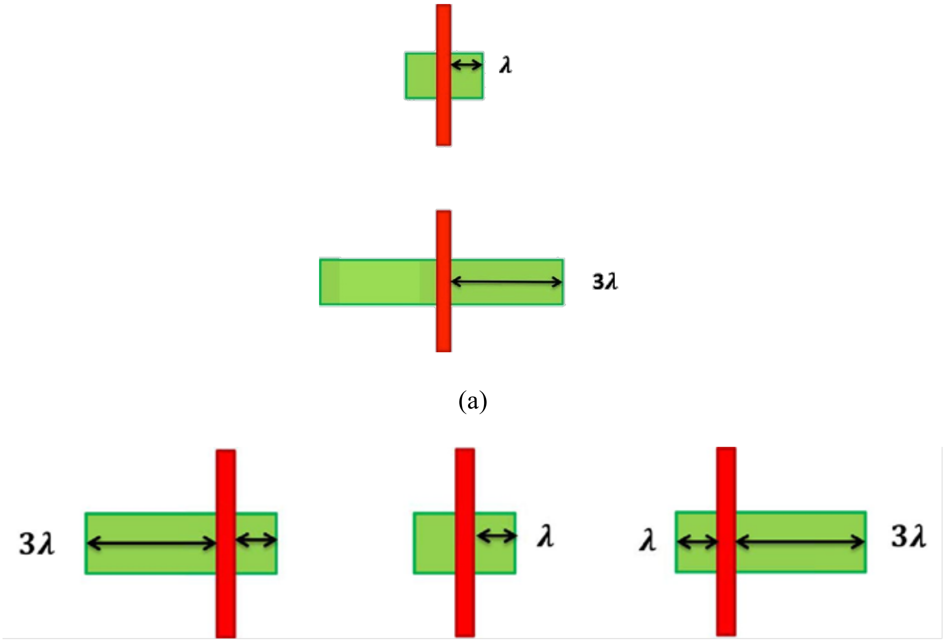


**Figure 2-3 (a) Test structures to monitor the effect of STI-induced stress along the channel direction (lateral) (b) Test structures to monitor the effect of STI-induced stress across the channel direction (vertical)**

### 2.2.2.2 Length of Diffusion Effect

The stress profile within the channel region of the DUT also depends on the length of the diffusion (LOD) or source/drain regions of the transistor. To study the impact of LOD on device performances, transistors with the same gate length $L_g$ and channel width $W$ are drawn with different diffusion lengths at $\lambda$, $3\lambda$, $4\lambda$, and $5\lambda$ for both the source and the drain sides, as shown in Figure 2-4(a).

In addition to devices having equal source/drain diffusion lengths, the LOD of a DUT can also be asymmetric (e.g. source-side LOD is longer than drain-side LOD), as illustrated in Figure 2-4(b). This test structure can be used to decouple the impacts of source-side LOD vs. drain-side LOD, permitting a close examination of parameters that are sensitive to S/D asymmetry such as $V_{TSAT}$ and source-injection velocity.



(a)

(b)

**Figure 2-4 (a) Test structures used to study the impact of length of diffusion on transistor performance (b) Test structures with asymmetric source/drain diffusion lengths**

### 2.2.2.3 Well Proximity Effect

In a planar bulk CMOS technology, NMOS and PMOS transistors are placed inside P-well and N-well regions, respectively [2-12][2-13]. In FD-SOI CMOS technology, the doping type underneath the isolating buried oxide (BOX) layer can be adjusted to achieve the desired $V_{th}$ specification, for both NMOS and PMOS transistors. Due to lateral straggle of implanted dopant atoms, the doping concentration within the well region of the DUT can be affected if it is situated close to the boundary between the N-well and the P-well. To investigate this effect, test structures shown in Figure 2-5 are used, wherein the DUTs are placed at different distances away from the boundary of the well doping, laterally as well as vertically. In order to isolate the well proximity effect from a particular direction, transistors are placed at least 3µm away from that particular well boundary.

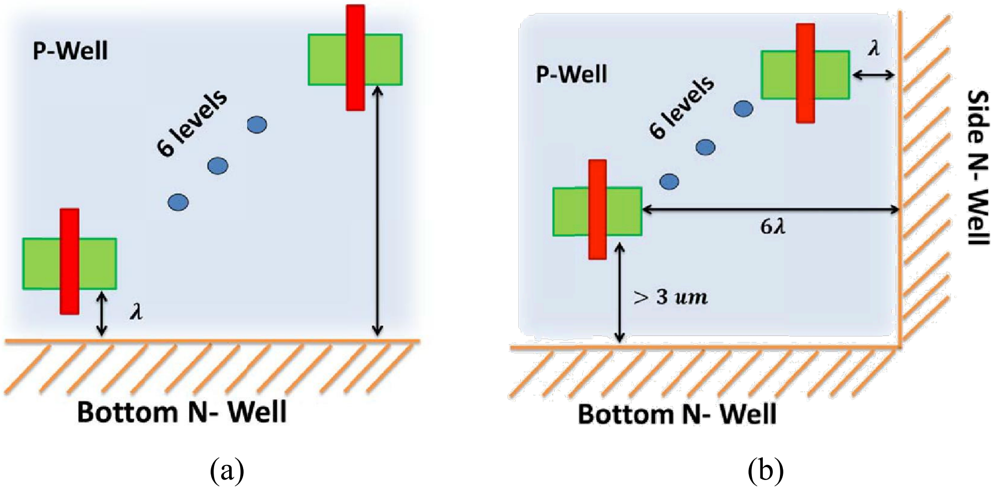


**Figure 2-5 (a) Test structures used to study the impact of proximity to the bottom N-well (b) Test structures to study the impact of proximity to the side N-well, away from bottom N-well**

#### 2.2.2.4 Segmented Channel Transistors

Instead of a transistor having a continuous width, the channel region can be segmented into multiple stripes of equal width as shown in Figure 2-6. From an electrostatic control standpoint, a segmented channel transistor can offer improved short-channel effect due to the slight wrap-around of the gate over the channel and the gate fringing electric field coupling to the channel region through the STI [2-14], if the stripe width is comparable to the channel length. Thus, even though the segmented channel design takes up more layout area as compared to a conventional channel design, the improvement in transistor performance can provide a net benefit when normalized to the same layout area. To observe the greatest benefit of the segmented channel design, the minimum drawn device width is used for each channel segment.

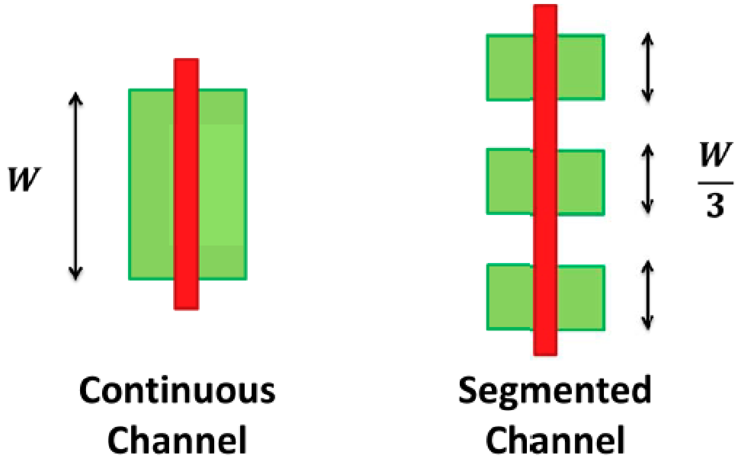


**Figure 2-6 Test structures comparing continuous vs. segmented channel designs**

## 2.3 Transistor Array Test Setup and Measurement

### 2.3.1 Test Circuitry

Transistor characterization arrays allow for many test devices of different designs to be included on a single die. Usually, these test devices are not directly probed due to the limited die area. Therefore, electrical access to individual devices in the array is made through the I/O pads of the test chip. Since the number of I/O pads is limited and must be sufficiently allocated for all the signals in different test blocks, a decoder circuit is used to share some common digital signals such as scan-in (SIN), scan-out (SOUT), and scan-clock (SCLK). The decoder is controlled by a 2-bit signal used to select one of the 4 modules within the array, and only one of the modules is active at any given time. Additionally, an enable signal (EN) for transistor characterization is also included. The floor plan of the characterization array is shown in Figure 2-7. There are 6 columns each in the PMOS and NMOS modules: mismatch pair transistors of RVT, LVT and HVT flavors each take up two columns. The digital signals and the selection circuitry are summarized in Figure 2-8.

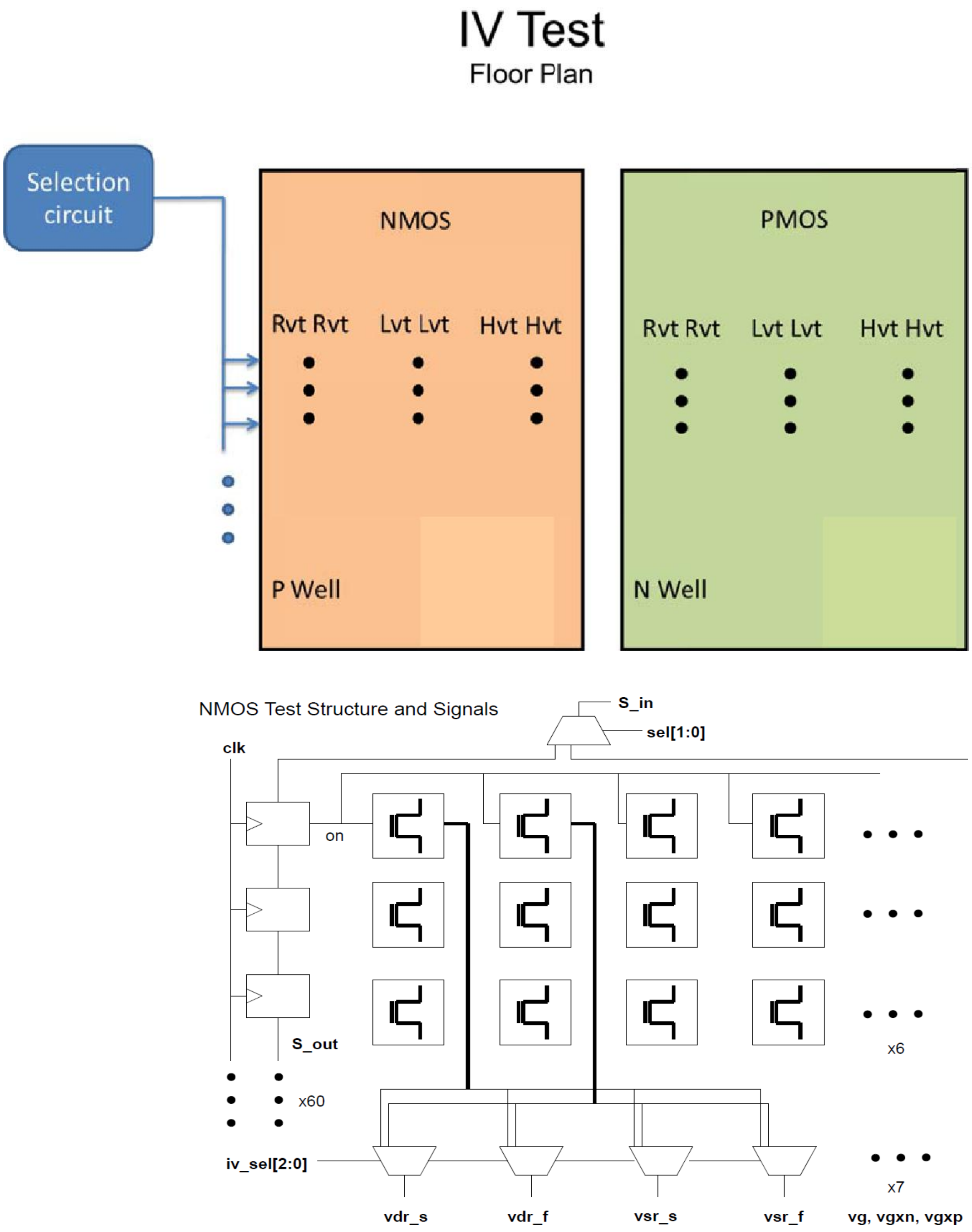


**Figure 2-7 Floor plan for the transistor characterization array consisting of selection circuitry and DUTs**

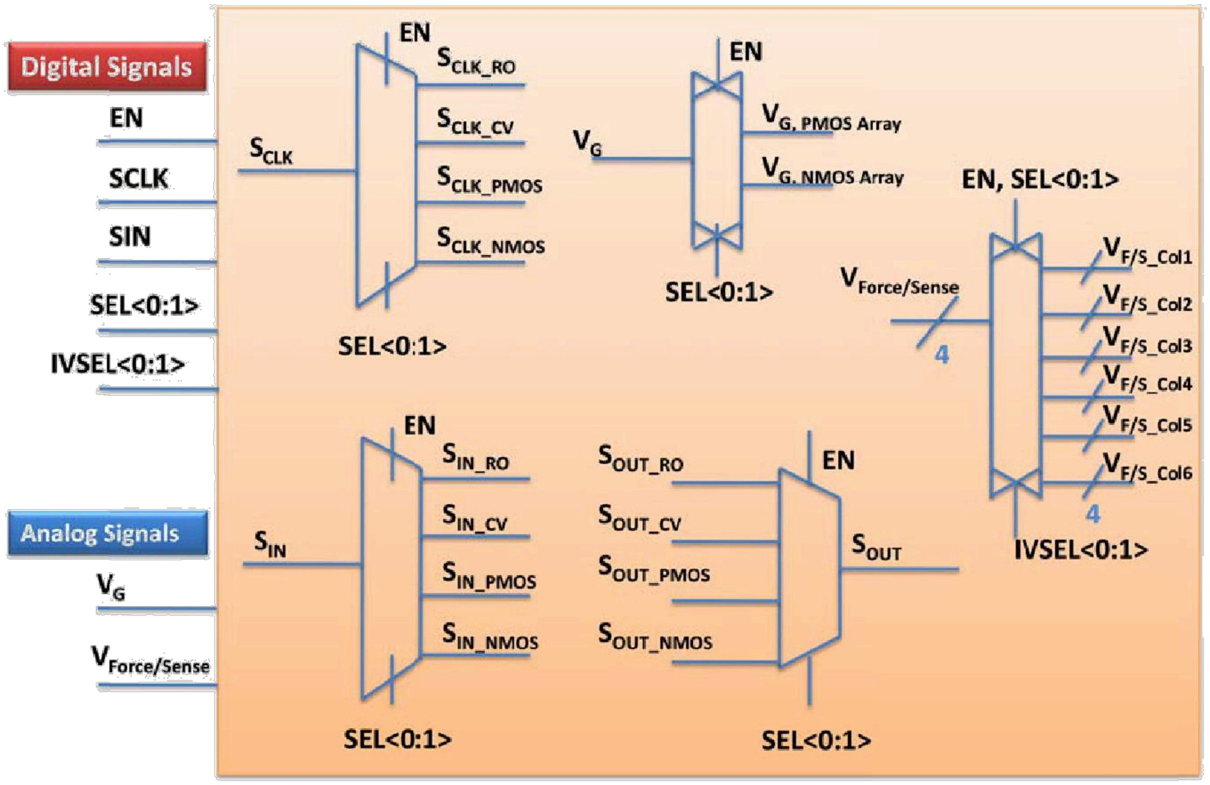


**Figure 2-8 Top-level schematic showing selection circuitry and analog signals**

Transistors within an array are accessed in a serial manner, in a row-wise fashion using a scan-chain circuit. A simple scan-chain circuit comprised of D flip-flops (D-FF) chained together can be used to activate/deactivate the row under test. A column can be selected through a column multiplexer that has a 3-bit control signal, allowing one out of the six columns to be selected at a given time. The source and drain of devices in the same column share electrical lines. With a combination of row select (through scan-chain clocking) and column select (through multiplexer), each individual transistor inside the array can be accessed, allowing full control over the biasing of the source, drain, and gate terminals through the I/O pads.

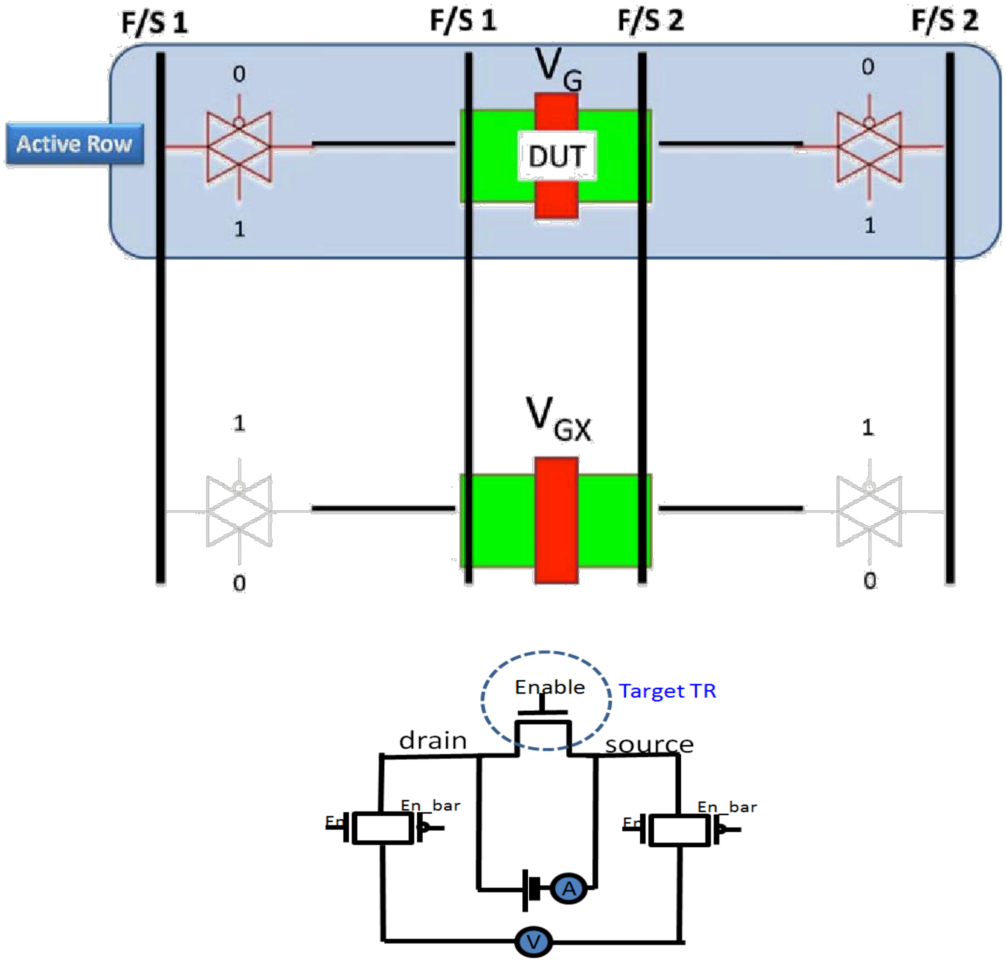


**Figure 2-9 Circuit schematic showing access transistors for row selection and Kelvin Force/Sense configuration**

Analog signal lines used to apply voltage to and measure current from the terminals of the DUT are routed through a series of pass-gates. Due to parasitic resistance along the wire trace from the probe pad to the DUT, the voltage applied to the device's terminals is smaller than the nominal voltage. This voltage difference can be significant if the resistance along the wire is large. To circumvent this problem, the Kelvin measurement technique - utilizing separate sets of Force and Sense lines - is implemented for the source and drain terminals of the DUT [2-15]. Figure 2-9 shows the basic Kelvin measurement configuration for the source and drain terminals. The current is passed through the Force lines and the voltage drop across the DUT is sensed across the Sense lines. The access transistors which are used to connect the force lines to the device in a row must be large enough to support the current level of the DUT, but small enough such that they will not have a large off-state leakage current. Due to the high impedance associated with the sense line (like the impedance of a voltage meter), very low parasitic current can flow through it.

Leakage control is also important in this type of transistor characterization array. Since only one active device is to be characterized at a time, the leakage current from the other devices in the same column should be as small as possible. To this end, a separate gate bias voltage (VGX) is applied to turn off the transistors that are not being tested. Figure 2-10 shows the schematic of the leakage control circuit.

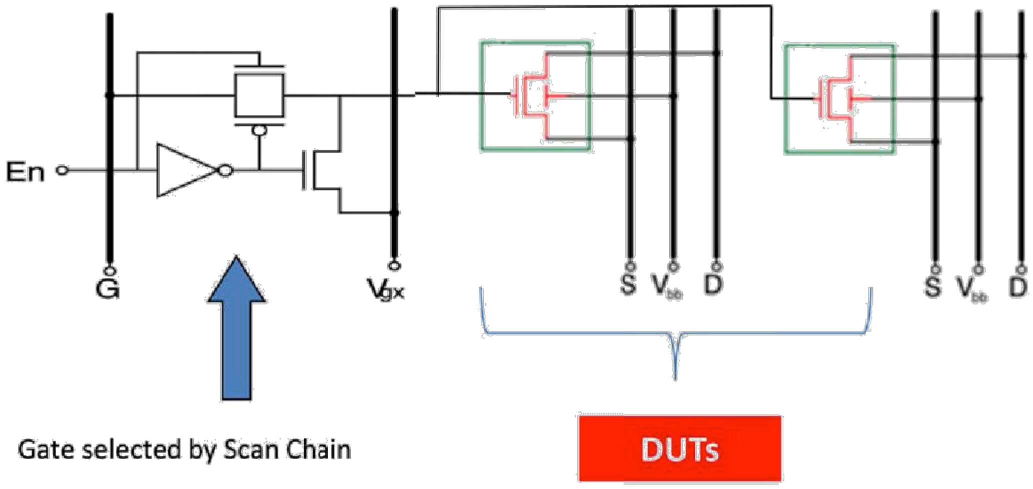
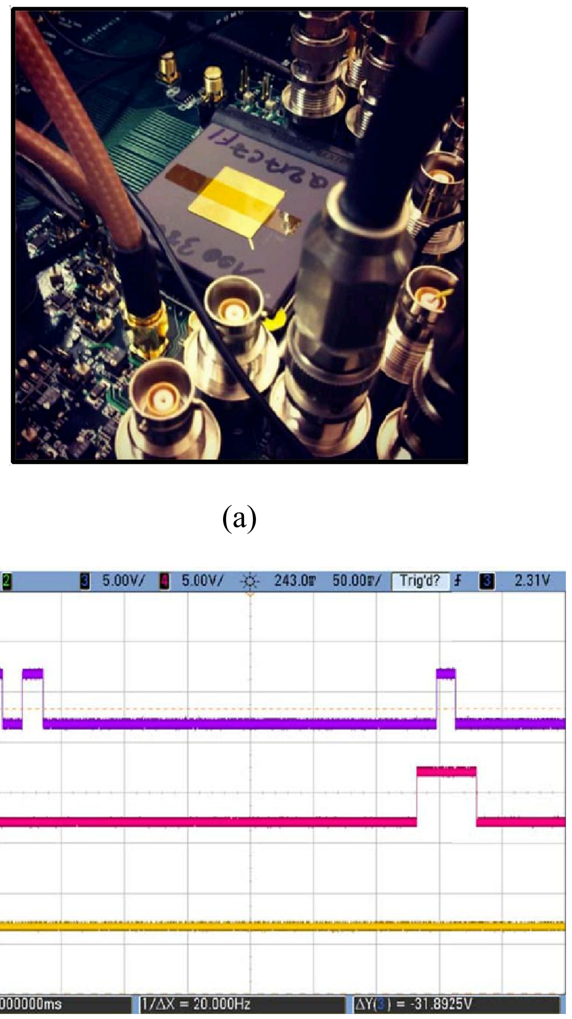


**Figure 2-10 Row selection circuit with $V_{GX}$ biasing for minimizing off-state leakage for an NMOS array**

### 2.3.2 Measurement Setup

The fabricated chip was packaged in a pin grid array (PGA) mountable to a printed circuit board (PCB) using a standard chip socket. The PCB was designed in-house and it contains various test pins, decoupling capacitors, voltage level shifter, and connectors for digital and analog signals. Triaxial connectors are used to bias the source and drain terminals as they help to prevent leakage current through the insulator of the cable, which is necessary for measuring small levels of current flowing through the source and drain of a transistor. A photograph of the test chip mounted on the PCB is shown in Figure 2-11(a).

A semiconductor parameter analyzer (SPA) is used to control the digital and analog signals. The SPA outputs the signal per the binary representation of the programmed decimal value. For this specific instrument, a digital high is represented as a '0' and a digital low is represented as a '1'. A level shifter is needed on the PCB board to convert the digital output from the SPA ($5V$) to the acceptable range used by the test chip ($0.9V$- $1V$). An example of a scan-chain test with signals coming out is shown in Figure 2-11(b).

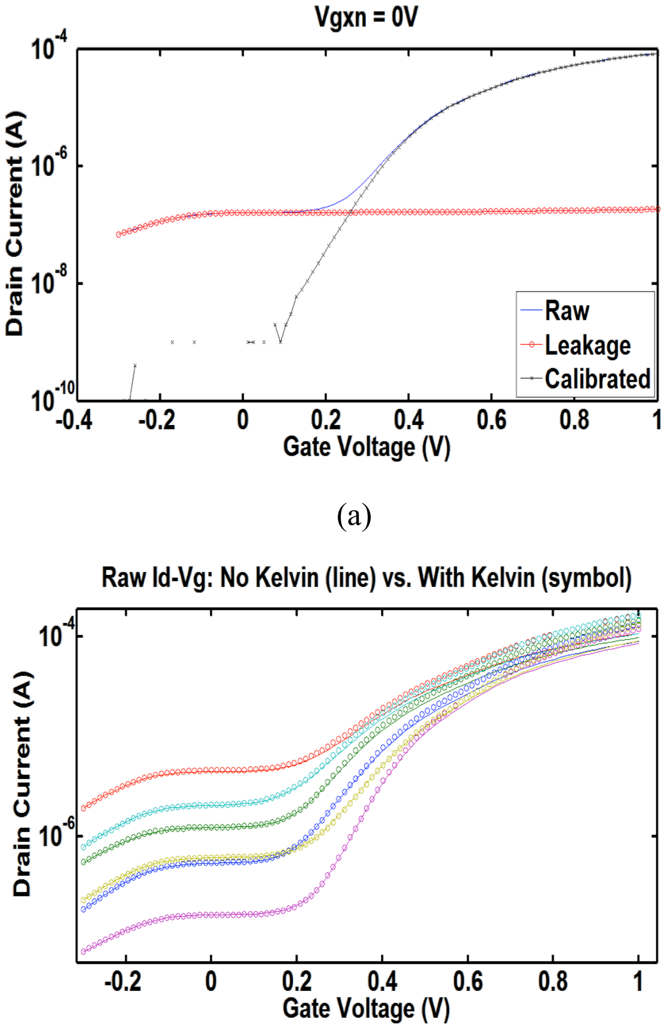


(a)

(b)

**Figure 2-11 (a) Packaged test chip mounted on the custom-made PCB; coaxial and tri-axial cables are used for delivering analog signals to the chip; (b) Digital signals outputted from DB25 port used to control the scan-chain.**

Once the transistor in the array is selected, its source, drain and gate terminals are multiplexed to the I/O pads that are now connected to the SPA. This allows us to perform any basic electrical measurements such as $I_D$-$V_G$ and $I_D$-$V_D$. However, the $I$-$V$ characteristic tends to suffer from a high leakage floor due to the off-state leakage currents flowing through the other transistors within the same column of the characterization array. As mentioned in the previous section, the built-in leakage control circuitry allows the gate terminal of the idle transistors to be set at $V_{GX}$ to make them strongly off. In conjunction with this method, one can also try to calibrate out the leakage current. This is accomplished by first performing a parametric $I$-$V$ measurement sweep with none of the devices in the array selected (i.e. the scan-chain is filled with zeros). Let's call the resulting current $I_{Leakage}$. Then, one can perform the I-V measurement for the DUT to get $I_{DUT}$. To calibrate the leakage-current out, the two current quantities are subtracted from one another, i.e., the quality of the calibrated current depends on the current sensitivity level of the SPA that can be set for optimal operation, as shown in Figure 2-12.



(a)

(b)

**Figure 2-12 (a) Leakage calibration: leakage floor subtracted out from the raw data. (b) Higher $I_{on}$ current is observed with Kelvin measurement.**

### 2.3.3 Variability Observation

Transistors are characterized in order to understand the underlying variability components. Threshold voltages in both the linear and saturation regimes of operation are extracted using a constant-current definition: $300nA \cdot W/L$ for NMOS. Once experimental data is acquired, quantities such as DIBL (the difference between $V_{T\,LIN}$ and $V_{T\,SAT}$) can be derived with respect to different modes of measurements. The box plot of $V_{T\,SAT}$ of the minimum-sized ($L_g$=30nm) transistors across different measured FDSOI dies is shown in Figure 2-13. Different threshold voltage flavors are shown separately across 7 dies with a total of 180 devices in each subcategory. Die mean values are later excluded and modeled as global variations.

The histograms of $V_{T\,SAT}$ for the two minimum-sized ($L_g$=30$nm$, 60$nm$) transistors are shown in Figure 2-14. The number of mismatched pairs used in this study is 60. Experimental distributions are shown in different colors for various $V_{th}$ flavors. The distributions of $V_{T\,SAT}$ are found to be non-Gaussian due to a limited data size and a premature manufacturing phase. The key point to note here is that non-Gaussian behavior should be captured for compact modeling used in the latter statistical circuit simulation.

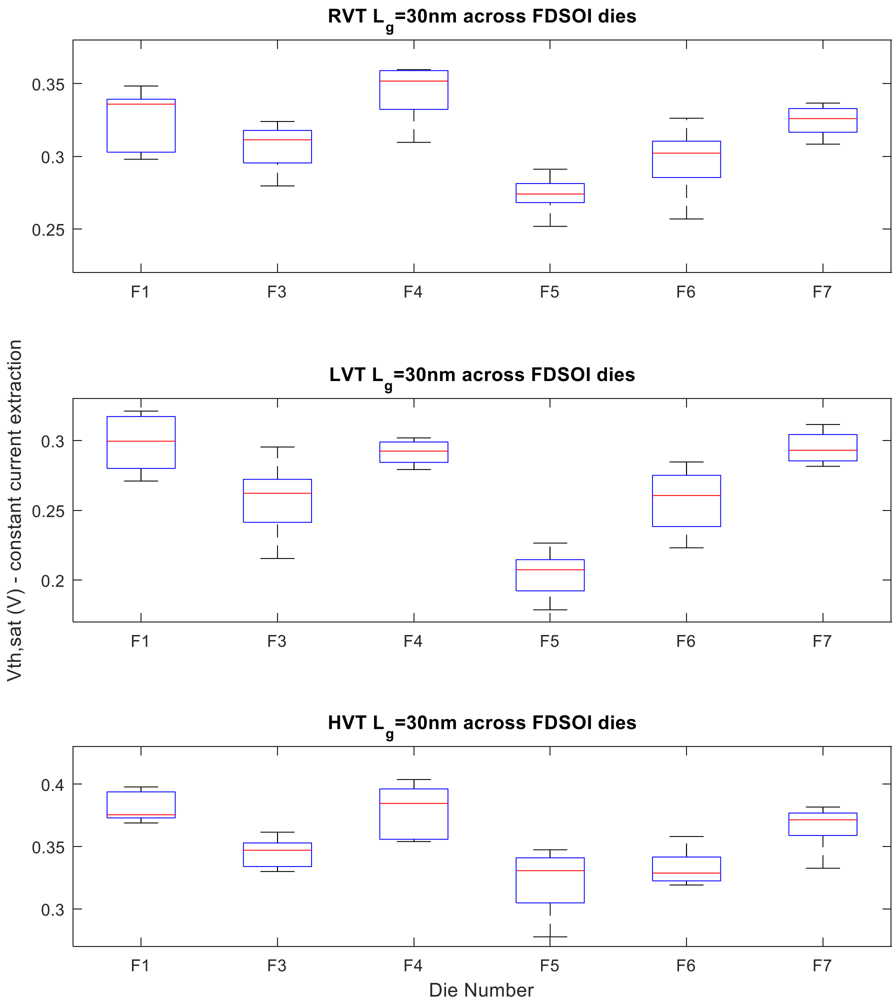


**Figure 2-13 Boxplot of $V_{T\ SAT}$ of the minimum-sized (*Lg*=30*nm*) transistors across various measured FDSOI dies**

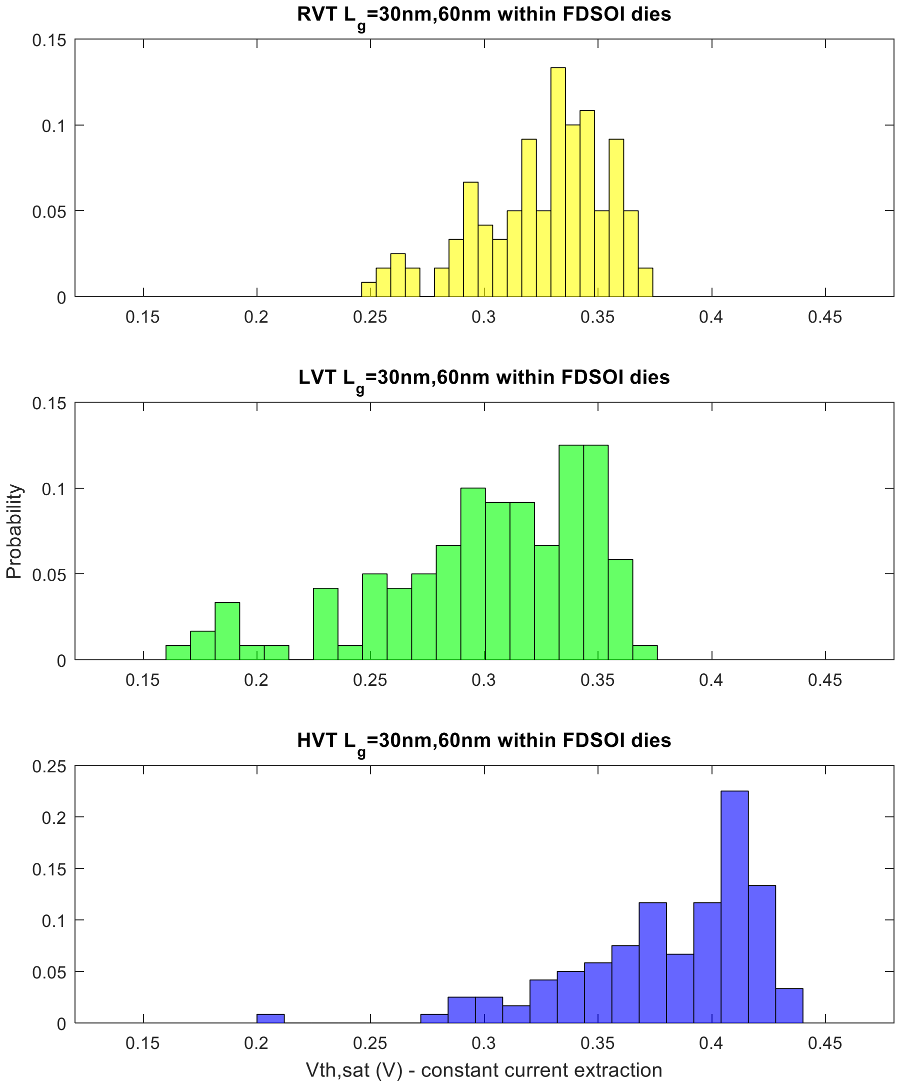


**Figure 2-14 Distributions of $V_{T\,SAT}$ [@$V_{DS}$=1$V$]. RVT have nearly Gaussian distributions, but LVT and HVT do not follow a strict normal distribution.**

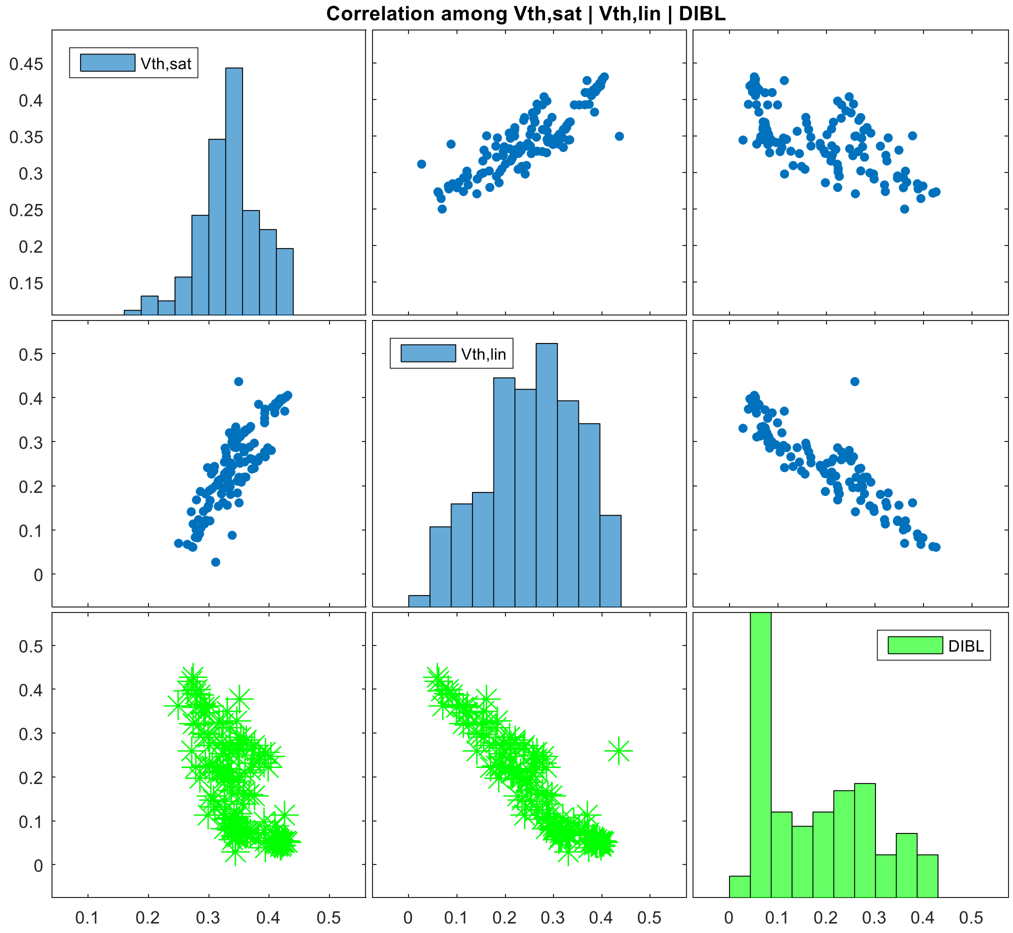


**Figure 2-15 Strong positive correlation shows between $V_{T\,LIN}$ and $V_{T\,SAT}$; weak correlation between $V_{T\,SAT}$ and DIBL, negative correlation between $V_{T\,LIN}$ and DIBL.**

Correlation between linear and saturation $V_{th}$ of the same device measured is plotted in Figure 2-15 (first two rows/columns). Overall, there is a positive correlation between $V_{T\,SAT}$ and $V_{T\,LIN}$. The relationship between DIBL and $V_{th}$ can also be understood by comparing the experimental data in Figure 2-15 (last green row). DIBL (i.e. $V_{T\,LIN} - V_{T\,SAT}$) is plotted against the linear threshold voltage $V_{T\,LIN}$ as well as the saturation threshold voltage $V_{T\,SAT}$. The medians of $V_{th}$ of the different operation regimes are essentially the same (i.e. this is seen by the same horizontal spread in the scatter plot). However, there is a large difference in the corresponding DIBL value. Here, the $V_{T\,LIN}$ shows a higher correlation with DIBL and a much larger variability. This is because these DUTs are prone to components of RDF [2-16].

## 2.4 Summary

The impact of transistor variability can be efficiently studied using a test chip vehicle. Characterization of an array of test devices including mismatch pairs, different combinations of gate length and channel width dimensions, and different layout proximity allowing for the collection of data, which can be used to analyze random and systematic variability. Since the number of I/O pads is quite limited, many signals must be shared among test blocks either through a decoder or a multiplexer. Care must be taken when designing the selection circuitry to ensure correct operation when accessing a device in an array and to minimize its impact on the measured device characteristics. Leakage minimization circuits can be designed to ensure that the devices that are not under test are turned strongly off to minimize their contributions to the measured current. The variation sources identified in this experiment will help build variability-aware compact models of transistors with physically meaningful parameters sensitive to variability observed in the manufacturing process.

## References

[2-1] K. J. Kuhn, A. Murthy, R. Kotlyar, and M. Kuhn, *"Past, Present and Future: SiGe and CMOS Transistor Scaling (Invited)"* ECS Trans. 2010 33(6): 3-17;

[2-2] C. Hu, *"FinFET 3D Transistor & the Concept Behind it"*, Solid State Techonology & Devices Seminar, UC Berkeley, Aug 26, 2011.

[2-3] C. Auth; C. Allen; A. Blattner; D. Bergstrom; M. Brazier; etc., "*A 22nm high performance and low-power CMOS technology featuring fully-depleted trigate transistors, self-aligned contacts and high density MIM capacitors*," VLSI Technology (VLSIT), 2012 Symposium on, pp.131,132, 12-14 June 2012.

[2-4] Q. Liu; M. Vinet; J. Gimbert; N. Loubet; R. Wacquez; L. Grenouillet; and etc, "*High performance UTBB FDSOI devices featuring 20nm gate length for 14nm node and beyond,*" Electron Devices Meeting (IEDM), 2013 IEEE International, pp.9.2.1-9.2.4, 9-11 Dec. 2013.

[2-5] S.H. Rasouli; H.F. Dadgour; K. Endo; H. Koike; K. Banerjee, "*Design Optimization of FinFET Domino Logic Considering the Width Quantization Property*," Electron Devices, IEEE Transactions on, vol.57, no.11, pp.2934,2943, Nov. 2010.

[2-6] M. Bruel; B. Aspar; B. Charlet; C. Maleville; etc*, "'Smart cut': a promising new SOI material technology,"* SOI Conference, 1995. Proceedings, 1995 IEEE International, pp.178-179, 3-5 Oct 1995.

[2-7] C. Fenouillet-Beranger; P. Perreau; T. Benoist; C. Richier; and etc, "*Impact of local back biasing on performance in hybrid FDSOI/bulk high-k/metal gate low power (LP) technology*," Ultimate Integration on Silicon (ULIS), 2012 13th International Conference on, pp.165,168, 6-7 March 2012.

[2-8] J.V. Faricelli, "*Layout-dependent proximity effects in deep nanoscale CMOS*," Custom Integrated Circuits Conference (CICC), 2010 IEEE, pp.18, 19-22 Sept. 2010.

[2-9] K. J. Kuhn, M. D. Giles, D. Becher, P. Koler, and etc, "*Process technology variation*," IEEE Trans. Electron Devices, vol. 58, no. 8, pp. 2197–2208, Aug. 2011.

[2-10] B. Pelloux-Prayer; A. Valentian; B. Giraud; Y. Thonnart; J.-P. Noel; P. Flatresse; E. Beigne, "*Fine grain multi-VT co-integration methodology in UTBB FD-SOI technology*," Very Large Scale Integration

(VLSI-SoC), 2013 IFIP/IEEE 21st International Conference on, pp.168,173, 7-9 Oct. 2013.

[2-11] M.G. Bardon; V. Moroz; G. Eneman; P. Schuddinck; and etc, "*Layoutinduced stress effects in 14nm & 10nm FinFETs and their impact on performance*," VLSI Technology (VLSIT), 2013 Symposium on, pp.T114, 11-13 June 2013.

[2-12] J.V. Faricelli, *"Layout-dependent proximity effects in deep nanoscale CMOS,"* Custom Integrated Circuits Conference (CICC), 2010 IEEE, pp.1-8, 19-22 Sept. 2010.

[2-13] P.G. Drennan; M.L. Kniffin; D.R. Locascio, *"Implications of Proximity Effects for Analog Design,"* Custom Integrated Circuits Conference, 2006. CICC '06. IEEE, pp.169-176, 10-13 Sept. 2006.

[2-14] R.A. Vega; T.-J. King, "*Comparative Study of FinFET Versus Quasi-Planar HTI MOSFET for Ultimate Scalability*," Electron Devices, IEEE Transactions on, vol.57, no.12, pp.3250-3256, Dec. 2010.

[2-15] N. Damrongplasit, *"Study of Variability in Advanced Transistor Technologies",* Ph.D. dissertation, UC Berkeley, May 2015

[2-16] K. J. Kuhn, M. D. Giles, D. Becher, P. Koler, A. Kornfeld, R. Kotlyar, S. T. Ma, A. Maheshwari, and S. Mudanai, *"Process technology variation,"* IEEE Trans. Electron Devices, vol. 58, no. 8, pp. 2197–2208, Aug. 2011.

# Chapter 3

# Statistical Compact Model Characterization and Statistical Circuit Verification

## 3.1 Introduction

It is widely recognized that in nanoscale CMOS technology, variation in the manufacturing process has emerged as a fundamental challenge to IC design. While foundries are working hard to mitigate process variability, the design houses are asking for accurate and appropriate models to handle statistical circuit performance evaluation. To accurately represent the process and device variability, it is essential to incorporate the variability during the extraction and calibration phase of compact transistor models. In addition, these compact transistor models require customized test structure designs as well as proper statistical characterization procedures. Conventional statistical compact model characterization methodologies [3-1][3-2] require special single transistor, direct-access test arrays, or virtual measurements from physical simulation data;

moreover, these models do not include rigorous statistical model parameter selection criteria.

After obtaining measurement results from the test structures described in Chapter 2, the next step is to apply statistical analysis techniques to interpret these measurement data. In this chapter, we propose a methodology consisting of a complete flow of statistical compact model characterization (analysis of variability in transistor behavior) and statistical circuit performance evaluation (translation of transistor variability to circuit performance variability).

Our proposed variability-aware compact transistor models can enable statistically optimized designs by capturing device variations in a concise, yet physically accurate way, and they are relatively easy to integrate with existing CAD tool flow. Stepwise parameter selection is combined with sequential extractions of statistical compact model parameters upon foundry-provided nominal compact model cards. These nominal models are trusted as they have been tested extensively and used during the test structure design. Our characterization algorithm selects an optimal statistical model parameter set that can be reliability extracted with the given measurement data.

In addition, we have built a customized Monte Carlo (MC) simulation platform to utilize these compact transistor models in the statistical IC design flow. Different statistical model parameters can be specified prior to the MC simulation within the scripted wrapper of standard SPICE-based simulators. With Gaussian Mixture Models (GMM), non-Gaussian tails in the circuit performances can be more accurately estimated compared to estimation with conventional methods.

## 3.2 Statistical Compact Model Characterization

### 3.2.1 MOSFET Device Models and Extraction

Interconnect and MOSFET device models are the critical interface between the manufacturing technology and integrated circuit design. *Compact models* include key equations that describe the current of a device as a function of its terminal voltages, to enable circuit simulation. To meet accuracy requirements on device models, existing BSIM [3-4] and PSP [3-5] models are being constantly upgraded to account for the emerging physical phenomena in the nanometer regime. The BSIM models illustrate an example of the evolution of the parametric complexity of transistor models in industrial design kits. For the 0.5*μm* technology in the early nineties, this model had 99 parameters, 7% of which were physical [3-4]. Here, physical quantities directly describe the physical attributes of the system. In the deep sub-micron era (65*nm*), the BSIM4 generation of this model has 355 parameters, 2.5% of which are physical. The PSP compact model has a similar parametric complexity [3-5].

Once the required transistor measurement data are acquired, one can perform either analytical regression or numerical optimization to estimate the compact model parameters. This procedure and the full set of compact model parameters are commonly referred to as *compact model parameter extraction* and *model card*, respectively. The increasing number of parameters and complexity of equations of compact transistor models drive the need to accurately determine all that many model parameters to reproduce the behavior of a specific observed device. In practice, a complete compact model, for example in this work, a foundry-provided nominal model card, is usually generated using a combination of pre-known technological process data, analytic methods, and numerical optimization methods.

The most widely used numerical optimization methods are based on the deterministic minimization of a nonlinear least squares error function between model output and measurement data. Nearly all optimization methods are iterative, and defining an appropriate starting point as well as parameter bounds is of crucial importance, and often relies on a deep understanding of the model to guide the minimization process.

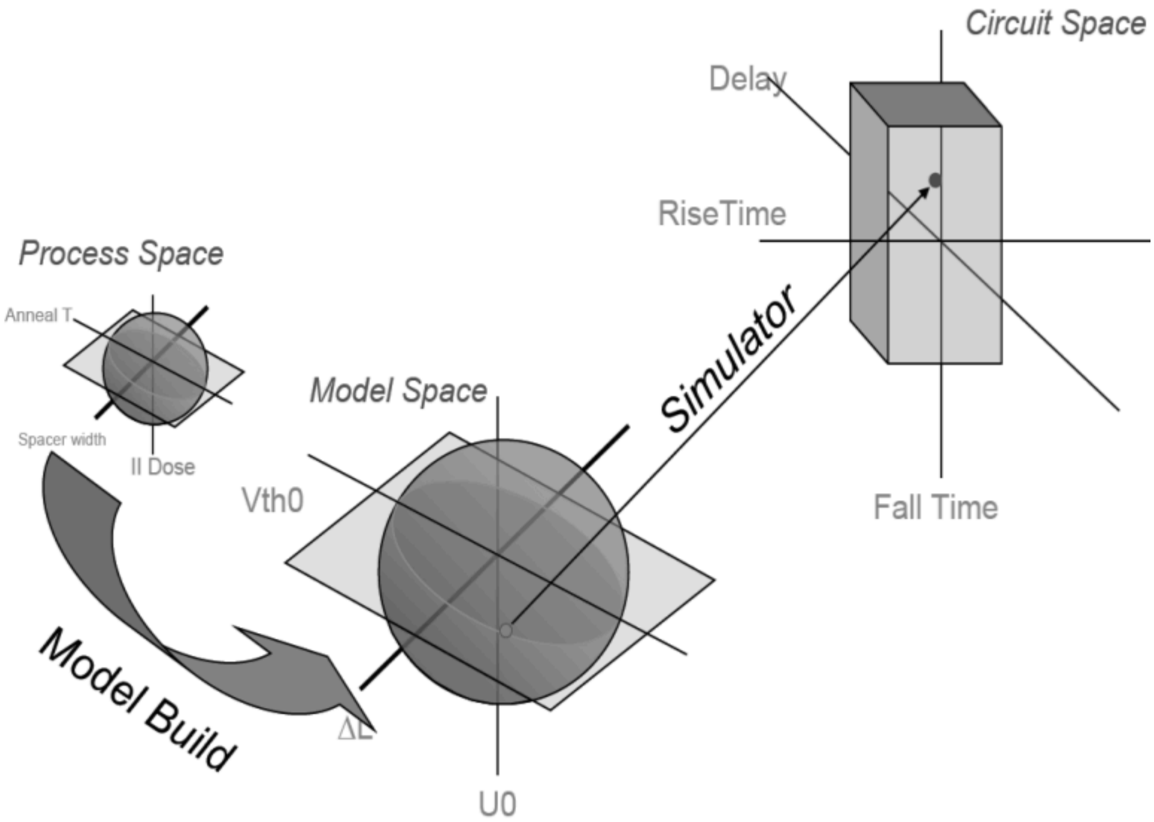


**Figure 3-1 Compact Modeling for Statistical Circuit Design Scheme [3-7]**

### 3.2.2 Statistical Device Characterization

Process variations usually manifest themselves as parameter fluctuations in nanoscale transistor physical dimensions or material/electronic properties, such as channel length, threshold voltage, and transistor parasitic [3-6]. All challenges with nominal device characterization become more serious for statistical device characterization, since statistical extraction procedures rely on an accurate extraction of nominal parameters.

After designing appropriate test structures as described in Chapter 2, variation measurements need to be correctly mapped and embedded into a statistically capable design kit, such that circuit designers can perform statistical circuit analysis and optimization to improve yield. This task is referred to as *statistical extraction,* as illustrated in Figure 3-1. One major problem in statistical estimation is to determine the appropriate distribution of the parameters (e.g., to determine the distribution of $V_{th}$ follows a normal distribution or a log-normal distribution). The next key question for statistical extraction is to determine the parameters of a specific distribution (e.g., determine mean and variance of a parameter which follows a normal distribution, or to find the mean and covariance structure of a correlated set of multivariate normally distributed parameters).

We focus here on the statistical sub-model of the transistor compact model, which consists of a set of model parameters assigned with a Gaussian variability model. The mean of the distribution is the nominal value given in the full model, while the standard deviation can be determined by various methods, including direct extractions from atomistic device simulation data [3-2] or backward propagation of variance from unit circuit performance measurement data [3-8][3-9].

### 3.2.2.1 Stepwise Compact Model Parameter Selection Algorithm

Our proposed methodology, using *stepwise selection of compact model parameters with sequential extractions from transistor measurement data*, can provide statistical model parameters that reflect true variability while preserving reduced complexity in subsequent statistical circuit simulations.

The statistical compact model parameter extraction is normally done using numerical nonlinear least-squares optimization methods. Given a reasonable set of initial guesses, (i.e. the *nominal model card*), a set of model parameter values can be estimated by minimizing the error between the model and the measurement data. Industrial standard transistor compact models, as well as open-source compact models under development, have many parameters. However, due to the high computational cost involved in optimization problems with large numbers of variables, it is helpful to reduce the number of model parameters to be extracted so that only essential parameters are fitted. Furthermore, imposing constraints on the optimization problem can ensure that the results are physically realistic and can reduce the indeterminacy or numerical instabilities during the optimization.

The nonlinear least-squares regression problem can be formulated as follows:

$$\begin{aligned}\boldsymbol{p}^* &= \underset{\boldsymbol{p}}{\operatorname{argmin}}\, F(\boldsymbol{p}) \\ &= \operatorname{argmin}_{\boldsymbol{p}} \textstyle\sum_{i=1}^{m} [I_i - G_i(p_1, \dots, p_n)]^2 \end{aligned} \tag{3-1}$$

Here $\{I_i\}$ are the electrical measurement data, $\{p_j\}$ are the compact model parameters, of which $\boldsymbol{p}$ is the $n$-dimensional parameter vector, and $\{G_i(\boldsymbol{p})\}$ are the evaluated model equation values with parameter $\boldsymbol{p}$, where $\boldsymbol{G}\,(\boldsymbol{p})$ is an $m$-dimensional function (Transistor compact model equation is normally a continuous one, which can be mapped onto space of any dimensions). This sum of squared residuals cost function is nonnegative, real-valued and continuously differentiable, which leads to least square estimation of the optimal value $\boldsymbol{p*}$. We rewrite the fitting residual or error function as

$$\Delta\boldsymbol{I} = [\Delta I_1, \Delta I_2, \dots, \Delta I_m\,]^T, \qquad \Delta I_i(\boldsymbol{p}) = I_i - G_i(\boldsymbol{p}) \tag{3-2}$$

We can approximate the error function stepping with the following Taylor series expansion, where $\boldsymbol{J}$ is the Jacobian matrix of the objective function $\boldsymbol{F}$, if the derivatives exist for $F\ (\boldsymbol{p})$:

$$\Delta F = F(\boldsymbol{p} + \Delta\boldsymbol{p}) - F(\boldsymbol{p})$$

$$\cong \left\{\frac{\partial F}{\partial p_1}\Delta p_1 + \cdots + \frac{\partial F}{\partial p_n}\ \Delta p_n\right\} + \frac{1}{2}\sum_{j=1}^{n}\sum_{k=1}^{n}\frac{\partial^2 F}{\partial p_j \partial p_k}\Delta p_j \Delta p_k \tag{3-3}$$

$$= \nabla F(\boldsymbol{p})^T \cdot \Delta\boldsymbol{p} + \frac{\mathbf{1}}{\mathbf{2}}\ \Delta\boldsymbol{p}^T\boldsymbol{H}(\boldsymbol{p})\Delta\boldsymbol{p}$$

Here, the gradient and *Hessian* matrix of the objective function can be estimated through the *Jacobian* matrix as:

$$\nabla F(\boldsymbol{p}) = \boldsymbol{J}(\boldsymbol{p})^{\mathrm{T}}\Delta\boldsymbol{I},\ \ \boldsymbol{H}(\boldsymbol{p}) \cong \boldsymbol{J}(\boldsymbol{p})^T\boldsymbol{J}(\boldsymbol{p})$$

$$\boldsymbol{J}(\boldsymbol{p}) = \begin{bmatrix} \frac{\partial \Delta I_1}{\partial p_1} & \cdots & \frac{\partial \Delta I_1}{\partial p_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial \Delta I_m}{\partial p_1} & \cdots & \frac{\partial \Delta I_m}{\partial p_n} \end{bmatrix} \tag{3-4}$$

This approximation of the Hessian matrix is acceptable because we assume that, at least locally, the objective function is linear.

Various optimization algorithms use slightly different methods in determining the size of each iterative step [3-10]. Gradient descent method will simply compute the tangent line of an objective function, and take a fixed step size in that direction; alternatively, the traditional Newton's method takes a step of

$$\Delta \boldsymbol{p} = -\boldsymbol{H}(\boldsymbol{p})^{-1} \nabla F(\boldsymbol{p}) \tag{3-5}$$

If the model parameters can be described as a relatively tight statistical distribution centered on their respective typical values, we can locally approximate the compact model equations with the following linear equation, where $\boldsymbol{J}$ is the Jacobian matrix of the original non-linear problem and $\Delta \boldsymbol{p}$ is the optimal step size:

$$\Delta \boldsymbol{I} = \boldsymbol{J}(\boldsymbol{p})^{\boldsymbol{T}} \cdot \Delta \boldsymbol{p} + \boldsymbol{\epsilon} \tag{3-6}$$

As in the context of linear regression, the estimated variance of a model parameter is given by the diagonal elements of the covariance matrix, and a common way to express it is via the normalized confidence interval (CI) of the estimated value [3-11]:

$$\sigma^2(\boldsymbol{p}^*) = \sigma^2(\boldsymbol{\epsilon}^*) \cdot diag[(\boldsymbol{J}(\boldsymbol{p}^*)^T \boldsymbol{J}(\boldsymbol{p}^*))^{-1}]$$

$$CI_j = t_{1-\frac{\alpha}{2}} \sigma(p_j) \quad \rightarrow \quad 95\% \text{ normalized } CI_j: \pm 1.96 \frac{\sigma(p_j)}{\mu\sqrt{N}} \tag{3-7}$$

The objective now is to identify the subset of model parameters that can be reliably extracted based on the available transistor *I-V* measurements. A stepwise parameter selection scheme is therefore developed not only upon fitting quality, but also upon extraction quality [3-3].

Starting with $n$ parameters, we fit the measurement data by nonlinear least squares optimization. Suppose we have a criterion function that represents the "goodness" of each extracted parameter, or the so-called extraction quality. If the current round of extraction provides an acceptable fitting quality, measured by the overall extraction/simulation fitting relative residual error, the "worst" parameter will be removed from the extraction and be set to a proper constant value, i.e. a nominal value from design kit model card or a median of previously extracted parameter values. With the reduced parameter set containing $n-1$ parameters, we repeat the same procedure until the fitting error begins to increase significantly or above a cutoff value, as shown in Figure 3-2.

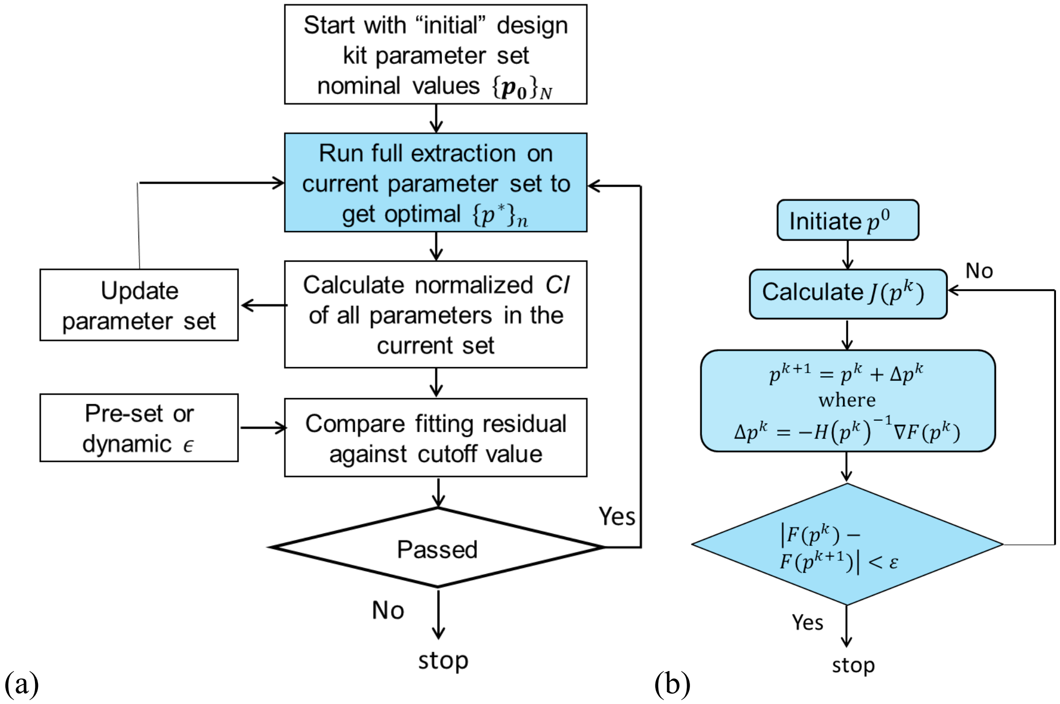

Start with "initial" design kit parameter set nominal values $\{\boldsymbol{p_0}\}_N$
Run full extraction on current parameter set to get optimal $\{p^*\}_n$
Update parameter set
Calculate normalized *CI* of all parameters in the current set
Pre-set or dynamic $\epsilon$
Compare fitting residual against cutoff value
Passed
Yes
No
stop
(a)
Initiate $p^0$
Calculate $J(p^k)$
No
$p^{k+1} = p^k + \Delta p^k$ where $\Delta p^k = -H(p^k)^{-1}\nabla F(p^k)$
$|F(p^k) - F(p^{k+1})| < \varepsilon$
Yes
stop
(b)

**Figure 3-2 (a) Stepwise compact model parameter selection algorithm. (b) Compact model parameter extraction with Newton's method.**

The key here is the definition of the extraction quality criterion. Ideally, it should represent how reliably the extracted parameter estimates its true value. However, in practice, there is no "true value" of model parameters from real transistors on silicon. Instead, we must define the criterion with metrics that can be calculated or observed from the extraction result itself. In this work, we choose the normalized confidence interval as defined in Equation 3-7. From the perspective of statistical circuit simulation where the models are utilized, we would like the statistical distribution as well as the correlation structure assigned for each of the model parameters to be as simple as possible. For example, a normally distributed parameter with a reasonable variance value is preferred over a multimodal distribution, which often indicates that there are two or more distinct transistor behaviors in the dataset[1].

### 3.2.2.2 Linear Spatial Backward Propagation of Variance

After getting the reduced transistor model parameter set, with supported measurement dataset, we can further improve the estimation of the model parameter statistical distribution estimation by performing our previously proposed method of *linear spatial backward propagation of variance* [3-12].

A statistical extraction method, namely Backward Propagation of Variance (BPV), has been proposed for iteratively solving the statistics of process parameters from the statistics of electrical performance measurements [3-8]. With the BPV approach, we can formulate statistical models as a set of independent, normally distributed process parameters, expressed as $\{p_j\}$. These parameters control the variations seen in device electrical performance $\{I_i\}$. With variations $\sigma_{I_i}$ $(i = 1,2 \dots m)$ of electrical performance parameters (e.g., $I_{dsat}$, $I_{off}$, etc.) measured under different geometry and bias conditions, the BPV method calculates $\sigma_{P_j}$ $(j = 1,2 \dots n)$ through [3-9]:

$$\sigma_{I_i}^2 = \sum_{j=1}^{n} s_{ij}^2 \sigma_{p_j}^2 \qquad (3\text{-}8)$$

[1] Multimodal distributions can also be represented via a "mixture of Gaussians" mechanism, discussed in Section 3.3.2.1.

The respective sensitivities are then defined around the nominal values as

$$\mathrm{s}_{ij} = \left.\frac{\partial I_i(\boldsymbol{p})}{\partial p_j}\right|_{\boldsymbol{p}=\overline{\boldsymbol{p}}} \qquad (3\text{-}9)$$

For a complete set of device performance electrical test measurements (large *m*) and carefully selected variation-aware compact model parameters (small *n*), the equations are solved using multivariate least squares fit or other linear regression methods. Electrical measurements of devices that strongly affect target circuit performances are selected as key quantities in $\{I_i\}$. Knowledge of circuit applications and device operations should guide this selection. Conversely, bias conditions or device geometries that are far removed from typical circuit applications are less likely to be chosen. Moreover, selection of $\{I_i\}$ is also required to make $\{p_j\}$ observable, and this requires that the sensitivity matrix $\boldsymbol{S} = \{s_{ij}\}_{m\times n}$, which is at the core of the BPV linear system, is well conditioned.

The BPV method is highly extendable to incorporate hierarchical variation models [3-3] which accurately describe the variability structure in the electrical measurement data. The superimposition property of linear systems makes linear BPV applicable to spatial variability characterization in compact models.

➢ *Hierarchical Variability Modeling*

Statistical process variations include both deterministic and random components. Certain types of deterministic variations are hierarchical in nature, while random variations can be modeled as white noise and added to the baseline. For simplicity, and without loss of generality, the total variation can be expressed as

$$\Delta p = \Delta p_{w2w} + \Delta p_{aw} + \Delta p_{awr} + \Delta p_{ad} + \Delta p_{adr} + \Delta p_{layout} + \varepsilon_{device} \qquad (3\text{-}10)$$

In this work, the measurement dataset does not have wafer-to-wafer random variations $\Delta p_{w2w}$, across-wafer systematic variations $\Delta p_{aw}$, or across-wafer random variations $\Delta p_{awr}$. The layout-dependent variations $\Delta p_{layout}$ can be ignored due to the regularized Device-Under-Test (DUT) design, and the residual $\varepsilon_{device}$ can be absorbed into the across-die random variations $\Delta p_{adr} \sim N(0, \sigma_{adr}^2)$. Across-die or within-die variation refers to the fluctuation of device properties on the same die/chip. The systematic across-die variation of a device with location $(x_d, y_d)$ on the die/chip are usually parabolic due to stepper-induced variations [3-13], as shown below.

$$\Delta p_{ad}(x_d, y_d) = a_d + b_d x + c_d y + d_d xy + e_d x^2 + f_d y^2 \tag{3-11}$$

Assuming a hierarchical variability structure, we apply an extension to the linear BPV statistical compact model characterization method [3-12]. The proposed method will obtain hierarchical spatial patterns of assigned compact model parameters (the optimal model parameter set selected by the algorithm described in Section 3.2.2.1) directly from test chip electrical measurement data. This propagation method applies linear regression to the coefficients of the hierarchical spatial variability model, i.e. $a_d \sim f_d$. The compact model sensitivity analysis is done at the nominal value, and the spatial coefficients of the hierarchical variability model in measured I-V data are linearly propagated to selected compact model parameters.

For a better application of our method, the hierarchical variability model is modified as (for chip-level data of transistor I-V test arrays):

$$\Delta p_j = \Delta p_{ad,j} + \Delta p_{awd,j} = \{\boldsymbol{xy}\}_{\boldsymbol{vec}}^T \cdot \boldsymbol{\lambda_j} + \epsilon$$

$$= \lambda_{0,i} + \lambda_{1,i} x + \lambda_{2,i} y + \lambda_{3,i} xy + \lambda_{4,i} x^2 + \lambda_{5,i} y^2 + \epsilon \tag{3-12}$$

The SBPV linear regression equation is:

$$\Delta \boldsymbol{I} = \boldsymbol{S} \Delta \boldsymbol{p} = \boldsymbol{S_{spatial}} \boldsymbol{\lambda}$$

$$\Delta \boldsymbol{I} = [\Delta \boldsymbol{I_1}(\boldsymbol{x}, \boldsymbol{y}), \dots, \Delta \boldsymbol{I_m}(\boldsymbol{x}, \boldsymbol{y})]^T \tag{3-13}$$

$$\Delta \boldsymbol{p} = [\Delta \boldsymbol{p_1}(\boldsymbol{x}, \boldsymbol{y}), \dots, \Delta \boldsymbol{p_n}(\boldsymbol{x}, \boldsymbol{y})]^T$$

The spatial coefficients and modified sensitivity matrix are

$$\boldsymbol{\lambda} = \left\{\lambda_{k,j}\right\}_{(\boldsymbol{n}\times\boldsymbol{6})\times\boldsymbol{1}} = (\boldsymbol{\lambda}_{\boldsymbol{1}}^{\boldsymbol{T}}, \dots, \boldsymbol{\lambda}_{\boldsymbol{n}}^{\boldsymbol{T}})^{\boldsymbol{T}}$$

$$\boldsymbol{S}_{\boldsymbol{spatial}} = \left\{s_{ij}^{spatial}\right\}_{m\times(n\times6)} = \left[\left.\frac{\partial I_i(\boldsymbol{p})}{\partial p_j}\right|_{\boldsymbol{p}=\boldsymbol{p}_{\boldsymbol{d}}^{\boldsymbol{0}}} \cdot \{\boldsymbol{xy}\}_{\boldsymbol{vec}}^{\boldsymbol{T}}\right]_{\substack{i=1,2,\dots,m;\\ j=1,2,\dots,n}} \tag{3-14}$$

Although process variation is correlated with parameters in a device model, it is rare that the sources of process variation are directly represented in the model parameters. Typically, device parameters include multiple sources of variation, and are therefore statistically correlated because of this common dependency. However, the conventional simple BPV method assumes all parameters $\{p_j\}$ to be uncorrelated. Therefore, in BPV and many other model parameter extraction approaches, it is necessary to transfer correlated parameters into a set of uncorrelated variables.

This can be achieved either by physically decoupling, with each parameter corresponding to a single physical effect, or by numerical methods such as Principal Component Analysis (PCA)[2]. As a physical decoupling example, the $V_{th}$ fluctuation could be separated into two items: $\Delta V_{th} = \Delta V_{th0} + \Delta V_{th}(L_{eff})$, where $V_{th0}$ represents the threshold voltage of the long channel device which is only related to random dopant fluctuation (RDF), and the later item corresponds to the drain-induced barrier lowering (DIBL) effect which is only related to line-edge roughness (LER). To account for and separate these correlations among model parameters and reach accuracy requirements on device models for deeply scaled devices, the parametric complexity of the underlying device model inevitably increases.

In this work, however, with the stepwise parameter selection algorithm performed first, the correlation structure among all the compact model parameters within the final optimal parameter set is well examined (and could be preserved with GMM, see Section 3.3.2.1).

[2] PCA generates an orthogonal basis set of parameters and its mathematical details will be discussed in Section 3.3.2

# 3.3 Statistical Circuit Verification

## 3.3.1 Statistical Parametric Yield Estimation

To ensure IC design robustness in the presence of variability, process variability must be properly modeled, and the model must be integrated into the design flow. Statistical circuit verification involves mainly two steps: worst-case analysis and yield estimation.

The circuit design is tested first through simulations on worst-case corner models [3-14] to see if it is robust under the influence of variability. If it fails, either the circuit topology or the design parameters should be changed. When the design passes the corner model test, it will be further evaluated by Monte Carlo simulations to estimate the parametric yield. If the yield is below the desired value, the design will also have to be rolled back and changed, and if the yield is satisfactory, the design is ready for tape-out.

Statistical circuit yield estimation starts with the assumption that the parameters (either process or compact model parameters) characterizing variability can be described by random variables, of which the distribution has a certain probability density function (PDF), for example, the commonly used Gaussian PDF. The parametric yield is associated with certain circuit performance metric(s). Based on the value of performance metric, one can judge whether the circuit succeeds or fails to meet specifications. Calculating the yield is equivalent to calculating the failure probability. Typically, the yield is high and the failure probability is low.

### 3.3.1.1 Customized Monte Carlo Simulation

The golden standard to estimate failure probability is the Monte Carlo method [3-15]. In Monte Carlo, a random sequence of length $R$ on transistor compact model variability parameter (denoted as a random vector) $\{\boldsymbol{X_r}\}_{r=1}^{R}$ is generated per its probability density function $pdf\ (\boldsymbol{X})$; then, the performance metric $f\ (\boldsymbol{X})$ is evaluated for all $\boldsymbol{X}$'s. The failure probability $\mathbb{P}_f$ is estimated as:

$$\widehat{\mathbb{P}}_f = \frac{1}{R}\sum_{r=1}^{R} \mathbb{I}(f(\boldsymbol{X}_r) \in \mathcal{F}) \qquad (3\text{-}15)$$

Here $\mathcal{F}$ represents the failure region and $\mathbb{I}\ (f(\boldsymbol{X}_r) \in \mathcal{F})$ is the indicator function of the failure event. Since these summands (indicator functions) are identically independent Bernoulli random variable with probability $\mathbb{P}_f$, $\widehat{\mathbb{P}}_f$ is an unbiased estimate of the true failure probability. When the sample number $R$ is large enough, the central limit theorem [3-15] shows that $\widehat{\mathbb{P}}_f$ is normally distributed with $\mathcal{N}\left(\mathbb{P}_f,\ \mathbb{P}_f(1-\mathbb{P}_f)/R\right)$.

To achieve a certain accuracy of the estimate, when the target failure probability is relatively low, the required sample size could be very high; for example, SRAM cell failure rate could be as low as $10^{-6}$ [3-16] in industrial standard production chips. This calls for a more efficient parametric yield MC simulation with a both simpler and quicker SPICE simulation on each run (calculating $f(\boldsymbol{X}_r)$) and more sensitivity of the failure region (evaluating $f(\boldsymbol{X}_r) \in \mathcal{F}$). Therefore, in this work, with full control starting from statistical device characterization, we have characterized our compact variability model specifically for a certain circuit performance metric (increased sensitivity) with a reduced number of model parameters (decreased simulation cost per run). These models can be incorporated into standard MC simulation flow with Python-based in-house customized scripts.

MC simulation based methods, when efficiently done, can naturally capture non-Gaussian circuit performance distributions through the non-linear SPICE simulation mapping from well-characterized compact variability models. We will show the results of customized MC simulation results in both Chapter 4 and Chapter 5.

### 3.3.2 Statistical Circuit Modeling

Another way to address the parametric yield problem is to efficiently model key performance metrics of circuit blocks (e.g., critical path delay of a digital circuit block, etc.) directly utilizing on-chip test structure measurement data, often referred to as statistical circuit *performance modeling*. This methodology in general requires more measurement data support (shown later in this section with details) than basic MC simulations with a compact variability model, and therefore, circuit verification results using our proposed methods described in this section will only be presented in Chapter 5.

Without loss of generality, we consider the problem of estimating a single performance metric, $f$ (we imply the assumption that $f \sim \mathcal{N}\ (\mu_f, \sigma_f)$, of which can be relaxed to a "mixture of Gaussians", later). Due to the constraints on testing costs, measurements of $f$ may not be directly available. Instead, groups of measurement data of other electrical performance (typically transistor I-V) are provided, denoted by $\boldsymbol{I} = \{I_1, I_2 \dots I_m\}$. For an example, consider the problem of post-silicon validation of a small digital system. The performance metric $f$ might be critical path delay and $I_i$ would be measurement results from on-chip monitoring transistor arrays [3-17]. The problem definition here is to predict the distribution of $f$ given $\boldsymbol{I}$ and consequently, predict the parametric yield.

One conventional approach to this problem is to apply principal component analysis (PCA) to $\boldsymbol{I}$ and select its top features $\boldsymbol{X}$. The rest of the problem is then converted into a response surface modeling (RSM) problem [3-19].

➢ *Principal Component Analysis*

PCA is a commonly used statistical technique that transforms correlated measurements into a set of low-dimensional, uncorrelated factors. Given $M$ samples from a set of correlated electrical measurements $\boldsymbol{I}$, PCA seeks a linear transformation of these variables into a new set of random variables $\boldsymbol{X}$ which are orthogonal. The procedure starts by forming the correlation matrix amongst the measured samples. An eigenvalue decomposition of the correlation matrix is then performed and combinations of eigenvalues and corresponding eigenvectors are obtained [3-18]. If we use the top eigenvalue/eigenvector combinations, we can explain most of the overall observed variation among the measured samples with just a few uncorrelated variables.

- *Response Surface Modeling*

For RSM, a linear least square error function can be employed to optimize the results $\boldsymbol{\alpha}^* = \text{argmin}_{\boldsymbol{\alpha}} \|\boldsymbol{f} - \boldsymbol{\alpha} \cdot b\ (\boldsymbol{X})\|^2$, where $\boldsymbol{f}$ is the vector of performances, $\boldsymbol{\alpha}$ is a vector of the unknown response surface model coefficients, $b\ (\boldsymbol{X})$ is a vector of orthogonal basis functions (e.g., linear or quadratic polynomials of principal components $\boldsymbol{X}$). When the measurement data set is not large enough to support the variable space, model coefficients estimates by RSM can become non-unique, which is known as overfitting [3-20]. One common strategy for preventing overfitting is by adding regularization terms to error functions to reduce the number of significant or retained model parameters. An example of such a strategy is least-angle regression (LARS) which adds the $L_1$-norm (the summation of the absolute values of all elements in the parameter coefficient vector) regularization $\|\alpha\|_1 \leq \lambda$ to the error function [3-21]. One major benefit of regularizing with the $L_1$-norm is that by decreasing $\lambda$, we can impose a strong constraint for scarcity and achieve a sparse solution making use of fewer model coefficients. However, the selection of $\lambda$ is of crucial importance for such optimizations and requires considerable domain knowledge.

However, with our pre-selected optimal compact model parameter set described in section 3.2.2.1, we naturally obtain a feature subspace of parameters with a good correlation structure and process variability interpretability. Here, $\boldsymbol{X}$ is re-used as the feature vector of compact model parameters (instead of principal components). The performance metric could then be approximated as:

$$f(\Delta\boldsymbol{X}) = \sum_{k=1}^{n} \alpha_k \cdot b_k(\Delta\boldsymbol{X}) \qquad (3\text{-}16)$$

Here $\{b_k(\Delta\boldsymbol{X});\ k = 1,2 \ldots n\}$ contains the basis functions (in this work, for simplicity, linear expansion is selected), and $\{\alpha_k;\ k = 1,2 \ldots n\}$ are the performance model coefficients, which are determined by solving a linear system with just a few sampling points of direct measurement data from test structures[3].

[3] With our measurement data in Chapter 5, the dataset is split into “training” (for fitting these coefficients) and “testing” (for statistical circuit performance verification).

#### 3.3.2.1 Enhanced Method with Gaussian Mixture Model

We propose Gaussian Mixture Models (GMM) to represent the distributions of compact model parameters, or equivalently the feature subspace for performance modeling. The GMM is a weighted sum of Gaussian distributions, and can represent non-Gaussian distributions generated by compact model parameter extraction properly. Moreover, GMMs can represent any correlation easily since each component is Gaussian; thus, we can easily compute correlation caused by dependencies intrinsic to some of compact model parameters [3-22].

The Expectation-Maximization (EM) algorithm has been widely used for building GMMs for digital statistical static timing analysis (SSTA) [3-23]; it is thus adopted here for fitting a GMM to the statistical compact model parameters. We study the following probabilistic model:

$$\mathbb{P}(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{j=1}^{K} \phi_j \mathbb{P}\left(\boldsymbol{X}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) \tag{3-17}$$

$$\mathbb{P}\left(\boldsymbol{X}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) = \frac{1}{(2\pi)^{n/2}\left|\boldsymbol{\Sigma}_j\right|^{1/2}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{X}-\boldsymbol{\mu}_j\right)^T \boldsymbol{\Sigma}_j^{-1}\left(\boldsymbol{X}-\boldsymbol{\mu}_j\right)\right\}$$

Here the mixture weight $\phi_j \geq 0$ and $\sum_{j=1}^{K} \phi_j = 1$ and $n$ is the dimensionality of the vector $\boldsymbol{X}$. The parameter vector $\boldsymbol{\theta}$ consists of $\phi_j$, mean vectors $\{\boldsymbol{\mu}_j\}$, and the covariance matrices $\{\boldsymbol{\Sigma}_j\}$. Given the number of mixtures $K$and $N$ independent, identically distributed (*i.i.d.*) samples $\{\boldsymbol{X}^{(t)}\}_1^N$, we obtain the following log-likelihood:

$$\ell(\boldsymbol{\theta}) = log \prod_{t=1}^{N} \mathbb{P}\left(\boldsymbol{X}^{(t)}|\boldsymbol{\theta}\right) = \sum_{t=1}^{N} log\mathbb{P}\left(\boldsymbol{X}^{(t)}|\boldsymbol{\theta}\right) \tag{3-18}$$

The EM algorithm updates the parameter vector as shown in Table 3-1.

When we get the statistics out of the feature vector X, we can then calculate the mean and standard deviation of the performance metric through Equation 3-16, with linear superposition of different mixtures. Furthermore, with GMM, we can utilize our customized MC simulation platform to further improve the accuracy of the parametric circuit yield estimation.

| **Algorithm** | **EM algorithm to solve GMM parameter vector** |
|---|---|
| *Initialization* | Choose an initial setting for the parameter vector $\boldsymbol{\theta}^{\boldsymbol{new}}$ (uniform mixture weight, global mean/variance for all) |
| *Repeat* | $\boldsymbol{\theta}^{\boldsymbol{old}} = \boldsymbol{\theta}^{\boldsymbol{new}}$ |
| *[E-step]* | Evaluate posterior probabilities $\boldsymbol{h}_{\boldsymbol{j}}^{\boldsymbol{old}}$: |
|  | $\boldsymbol{h}_{\boldsymbol{j}}^{\boldsymbol{old}}(\boldsymbol{t}) = \frac{\boldsymbol{\phi}_{\boldsymbol{j}}^{\boldsymbol{old}}\mathbb{P}\left(\boldsymbol{X}^{(\boldsymbol{t})} \mid \boldsymbol{\mu}_{\boldsymbol{j}}^{\boldsymbol{old}}, \boldsymbol{\Sigma}_{\boldsymbol{j}}^{\boldsymbol{old}}\right)}{\sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{K}} \boldsymbol{\phi}_{\boldsymbol{i}}^{\boldsymbol{old}}\mathbb{P}\left(\boldsymbol{X}^{(\boldsymbol{t})} \mid \boldsymbol{\mu}_{\boldsymbol{i}}^{\boldsymbol{old}}, \boldsymbol{\Sigma}_{\boldsymbol{i}}^{\boldsymbol{old}}\right)}$ |
| *[M-step]* | Evaluate $\boldsymbol{\theta}^{\boldsymbol{new}}$ given by: |
|  | $\boldsymbol{\phi}_{\boldsymbol{j}}^{\boldsymbol{new}} = \frac{\boldsymbol{1}}{\boldsymbol{N}}\sum_{\boldsymbol{t}=\boldsymbol{1}}^{\boldsymbol{N}} \boldsymbol{h}_{\boldsymbol{j}}^{\boldsymbol{old}}(\boldsymbol{t})$ |
|  | $\boldsymbol{\mu}_{\boldsymbol{j}}^{\boldsymbol{new}} = \frac{\sum_{\boldsymbol{t}=\boldsymbol{1}}^{\boldsymbol{N}} \boldsymbol{h}_{\boldsymbol{j}}^{\boldsymbol{old}}(\boldsymbol{t})\boldsymbol{X}^{(\boldsymbol{t})}}{\sum_{\boldsymbol{t}=\boldsymbol{1}}^{\boldsymbol{N}} \boldsymbol{h}_{\boldsymbol{j}}^{\boldsymbol{old}}(\boldsymbol{t})}$ |
|  | $\boldsymbol{\Sigma}_{\boldsymbol{j}}^{\boldsymbol{new}} = \frac{\sum_{\boldsymbol{t}=\boldsymbol{1}}^{\boldsymbol{N}} \boldsymbol{h}_{\boldsymbol{j}}^{\boldsymbol{old}}(\boldsymbol{t})\left[\boldsymbol{X}^{(\boldsymbol{t})} - \boldsymbol{\mu}_{\boldsymbol{j}}^{\boldsymbol{new}}\right]\left[\boldsymbol{X}^{(\boldsymbol{t})} - \boldsymbol{\mu}_{\boldsymbol{j}}^{\boldsymbol{new}}\right]^{\boldsymbol{T}}}{\sum_{\boldsymbol{t}=\boldsymbol{1}}^{\boldsymbol{N}} \boldsymbol{h}_{\boldsymbol{j}}^{\boldsymbol{old}}(\boldsymbol{t})\boldsymbol{X}^{(\boldsymbol{t})}}$ |
| *Until* | $\left\lvert \boldsymbol{\theta}^{\boldsymbol{old}} - \boldsymbol{\theta}^{\boldsymbol{new}} \right\rvert < \boldsymbol{\epsilon}$ |

**Table 3-1 EM algorithm to solve GMM maximum likelihood estimation**

## 3.4 Summary

Statistical compact models properly characterized for process variability are important for the design of high-yield integrated circuits. Although it is important to understand the underlying physical mechanisms that cause variations in device performances, the goal of compact statistical modeling is to accurately represent variations of simulated circuit characteristics. This chapter includes a brief review of statistical compact modeling methods, and then proposes methodology based on stepwise parameter selection and BPV for fast statistical compact model characterization.

Statistical circuit modeling methods are reviewed briefly, followed by our proposed customized MC simulation platform developed with the compatibility of industrial standard design flow, as well as the extensibility of incorporating GMM models into the statistical sub-model of transistor compact models. The demonstration of all proposed methodology on post-silicon data will be presented in the following chapters.

## References

[3-1] C.H. Lin, M. V. Dunga, D.D. Lu, A. M. Niknejad and C. Hu, "*Performance-Aware Corner Model for Design for Manufacturing," IEEE Trans Electron Devices, vol. 56, no. 4, pp. 595-600, Apr*. 2010.

[3-2] P. Asenov, D. New, D. Reid, C. Millar, S. Roy and A. Asenov, *"Evaluating the accuracy of SRAM margin simulation through large scale Monte-Carlo simulations with accurate compact models,"* IC Design & Technology (ICICDT), 2013 International Conference on, Pavia, 2013, pp. 29-32.

[3-3] K. Qian, *"Variability Modeling and Statistical Parameter Extraction for CMOS Devices"*, Ph.D. dissertation, EECS, UC Berkeley, 2015.

[3-4] M. Chan, K. Hui, C. Hu, and P. Ko, "*A robust and physical BSIM3 non-quasi-static transient and AC small-signal model for circuit simulation,"* IEEE Transactions on Electron Devices, vol. 45, no. 4, pp. 834 {841, Apr. 1998.

[3-5] G. Gildenblat, X. Li, W. Wu, H. Wang, A. Jha, R. van Langevelde, G. Smit, A. Scholten, and D. Klaassen, *"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.

[3-6] B. Calhoun, Y. Cao, X. Li, K. Mai, L. Pileggi, R. Rutenbar, and K. Shepard, *"Digital circuit design challenges and opportunities in the era of nanoscale CMOS,"* Proceedings of the IEEE, vol. 96, no. 2, pp. 343-365, Feb. 2008.

[3-7] J. Watts, "Modeling Circuit Variability," Workshop on Compact Variability Modeling (CVM), San Jose, CA, Nov. 2008

[3-8] X. Li, C. McAndrew, X. Wu, S. Chaudhry, J. Victory, and G. Gildenblat*," Statistical modeling with the PSP MOSFET model,"* IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 29, no. 4, pp.599-606, April 2010.

[3-9] C. McAndrew, *"Statistical modeling for circuit simulation,"* in International Symposium on Quality Electronic Design (ISQED), Mar. 2003, pp. 357-362.

[3-10] S. Boyd, L. Vandenberghe, "*Convex Optimization*", Cambridge University Press, 2004.

[3-11] R. M. O'Brien, "*A caution regarding rules of thumb for variance inflation factors,*" *Qual. Quant.*, vol. 41, no. 5, pp. 673–690, Mar. 2007

[3-12] Y. Qiao, M.S. report, *"Linear Propagation Methodology in Compact Variability Modeling"* UC Berkeley, 2011

[3-13] M. Orshansky, L. Milor, and C. Hu, *"Characterization of Spatial Intrafield Gate CD Variability, Its Impact on Circuit Performance, and Spatial Mask-Level Correction," IEEE Trans. Semicond. Manuf.*, vol. 17, no. 1, pp. 2–11, Feb. 2004.

[3-14] N. Arora, *"MOSFET modeling for VLSI simulation: theory and practice,"* reprint by World Scientific, 2007.

[3-15] A. C. Davison, *"Statistical models*," Cambridge University Press, 2008.

[3-16] S. O. Toh, *"Nanoscale SRAM Variability and Optimization"*. PhD thesis, University of California, Berkeley, 2011.

[3-17] S. Saxena, C. Hess, and etc, *"Variation in transistor performance and leakage in nanometer-scale technologies,"* IEEE Transactions on Electron Devices, vol. 55, no. 1, pp. 131-144, 2008.

[3-18] I. T. Jolliffe, Principal Component Analysis. Springer Series in Statistics, 2002.

[3-19] X. Li, J. Le, L. Pileggi, and A. Strojwas, *"Projection-based performance modeling for inter/intra-die variations,"* in International Conference on Computer-Aided Design (ICCAD), Nov. 2005, pp. 721-727.

[3-20] C. M. Bishop, Pattern Recognition and Machine Learning. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[3-21] X. Li, *"Finding deterministic solution from underdetermined equation: Largescale performance modeling by least angle regression,"* in Design Automation Conference, 2009, pp. 364-369.

[3-22] S. Takahashi, Yuki Yoshida and S. Tsukiyama, *"A Gaussian mixture model for statistical timing analysis,"* 2009 46th ACM/IEEE Design Automation Conference, San Francisco, CA, 2009, pp. 110-115.

[3-23] S. Tsukiyama, M. Fukui, *"A Statistical Maximum Algorithm for Gaussian Mixture Models Considering the Cumulative Distribution Function Curve"*, IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol. E94.A, pp. 2528-2536, 2011

# Chapter 4

# Variability-Aware Compact Modeling and Statistical Circuit Validation on SRAM Test Array

## 4.1 Introduction

Variability modeling at the compact transistor model level can enable statistically optimized designs in view of limitations imposed by the fabrication technology. In this chapter, we propose a variability-aware compact model characterization methodology based on stepwise parameter selection. Transistor I-V measurements are obtained from bit transistor accessible SRAM test array fabricated using a collaborating foundry's 28nm FDSOI technology. Our in-house customized Monte Carlo simulation bench can incorporate these statistical compact models; simulation results on SRAM writability performance are very close to measurements in distribution estimation. Our proposed statistical compact model parameter extraction methodology also has the potential to predict non-Gaussian behavior in statistical circuit performances through mixtures of Gaussian distributions.

## 4.2 SRAM Test Array Design and Measurement

Static memory (SRAM) is a critical component of VLSI systems today. SRAM can provide the fastest random access time to stored data, and is used for lower-level caches and registers [4-1]. To increase the size of the cache on a chip, it is desirable and economical to fit as many cells into an SRAM array as possible. However, as the memory cell areas is scaled down with each new technology node, the read and write margins are degraded due to increasing variability in transistor characteristics. It is therefore desirable to minimize the operating voltage $V_{DD}$ of an SRAM cell to minimize power consumption. But a small mismatch in $V_{th}$ values can significantly reduce cell stability, setting a lower limit for the minimum operating voltage $V_{DD\,min}$ of the cell [4-2][4-3].

A widely-used SRAM cell design is the six-transistor (6T) cell, consisting of one pair of NMOS pull-down transistors, one pair of PMOS pull-up transistors, and one pair of NMOS pass-gate transistors. Since the transistors in the SRAM cell are packed very close to one another to maximize storage density, the transistor pairs inside a 6T cell naturally form mismatch pairs that are ideal for studying random variability [4-4].

### 4.2.1 6T SRAM Macro

In a normal 6T SRAM cell, the only accessible nodes are the Word Line (WL), Bit Lines (BL, BLB), voltage supply (VDD), and ground (GND). This is acceptable if we just want to perform basic SRAM operations such as read, write, and hold. However, to gauge SRAM cell stability, it is necessary to access the internal storage node to sweep the voltage on the node [4-5]. To this end, padded out SRAM cells were previously employed such that every node of a transistor inside the cells can be directly accessed [4-6]. Not only does this permit the butterfly plot to be generated, but the standard transistor level characterization test such as current vs. voltage can also be performed, allowing a direct correlation between transistor performance parameters and SRAM cell metrics to be observed. However, this adds too much of area overhead onto the state-of-the-art test chip design; furthermore, the final SRAM array designed does not mimic what is commonly used in real-life applications, which allows unforeseen discrepancy between experimental results versus actual performances.

As shown in Figure 4-1, a direct bit transistor accessible (DBTA) SRAM is identical to a typical functional SRAM except that it has direct bit-line access and independent controls of array, word-line, and bit-line power supplies. In DBTA test mode, the bit lines of the addressed bit are multiplexed to the respective pads through multiple levels of analog multiplexers. The word-line of the addressed bit is selected and the bit is initialized. When testing pass-gate transistor Q1, an overdrive voltage VH is set on array and one of the bit lines (BLB) to maintain the source node SNT close to 0V. The gate voltage on the word-line and drain voltage on the other bit-line (BLT) are individually swept while the drain current on BLT is measured at each step [4-7].

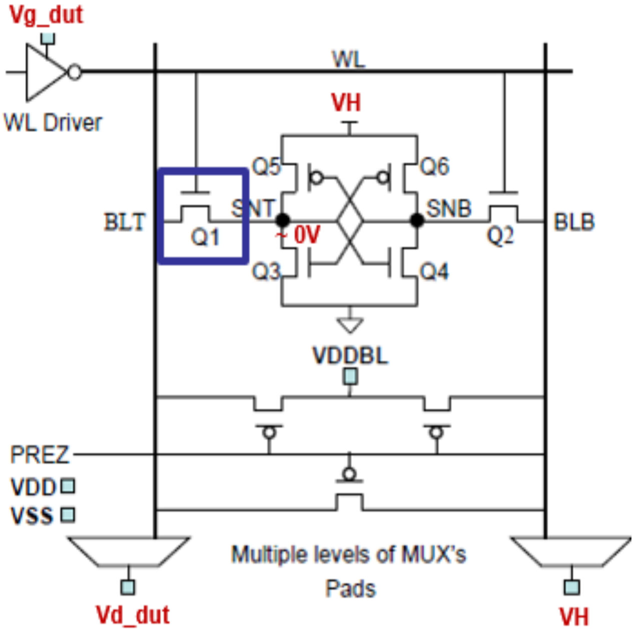


**Figure 4-1 Direct bit transistor access cell pass-gate transistor I-V measurement scheme. [4-7]**

### 4.2.2 Test Circuitry

Our collaborating design team has designed a bit-transistor accessible SRAM test array [2-8], as shown in Figure 4-2, using foundry's pre-production 28*nm* FDSOI technology [4-9]. This design enables fast high-volume transistor I-V measurements as well as cell DC characteristics measurements. The same design mask set was used to manufacture both bulk and FDSOI wafers, which enables direct comparison between bulk and FDSOI transistor gate stacks [4-10].

Figure 4-2(a) shows the transistor layout for a given cell. Note that the pull-down devices are larger than the pass gates (to ensure read stability) and the pull-up devices are smaller than the pass gates (to ensure write-ability). Figure 4-2(b) shows the metal layers that connect to each cell. VDD runs vertically along a column, and GND is supplied from different wires for the left and right sides and is shared with the rows above and below. Figure 4-2(c) shows a small 8 by 8 array with 64 total cells. All cells on the same row share the same word-line (WL) wire and all cells on the same column share the same bit-line (BLL and BLR) wires. The full array implementation is formed by combining these sub-arrays to form an 8*kB* memory, the effective size of which is 128 rows by 512 columns.

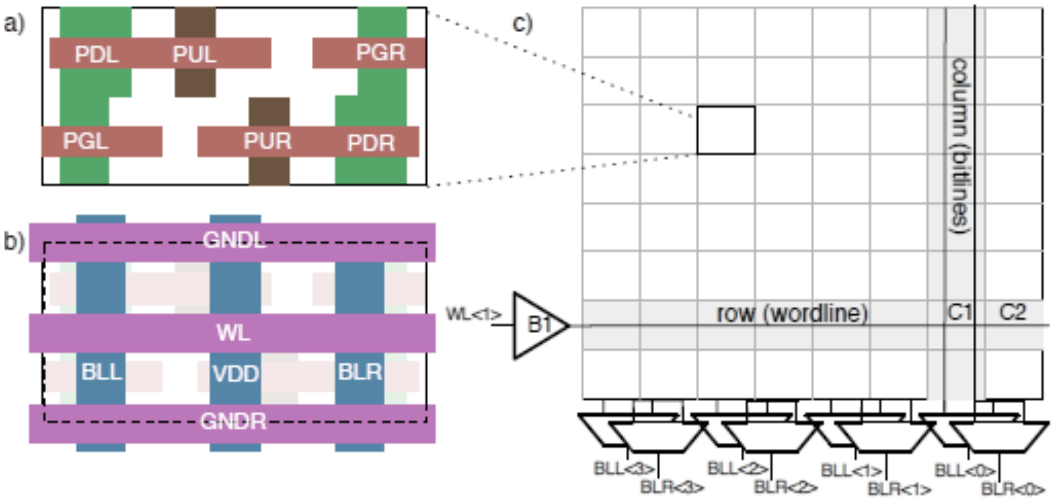


**Figure 4-2 (a) Layout of transistors in a 6T SRAM cell; (b) Layout of metal layers for a 6T SRAM cell (c) Generic overview of SRAM array design [4-10]**

Separated supplies, combined with a bit line multiplexer, are central to static *I-V* measurements. When assigning different bias voltages to a node inside the 6T cell, this allows for *I-V* measurement of any one of the 6 transistors inside the cell. The measurement setup was shown in Figure 4-1.

The digital control signal that is fed to the test array is controlled by off-chip FPGA on an in-house designed test PCB. Once a set of binary sequences has been loaded into the SRAM control BIST and bit line multiplexer, logic gates are set such that the desired operation is performed. For example, if the binary sequence is meant to enable a $I_D$-$V_G$ sweep of a PG transistor on the left half of the SRAM cell, the circuitry will use this binary sequence to connect the source and gate terminals of the left PG transistor to the correct analog signals on the I/O pads, while setting the cell to storage ‘0’.

### 4.2.3 Measurement Setup

The fabricated test chips were sent back in a form of a 12-inch wafer. Several dies were then packaged into sockets fit for our test PCB, making it compatible with all electrical measurement equipment.

Due to the limited number of I/O pads, attaching all necessary digital signals onto the main PCB directly is not practical. Also, with the addition of the weight of all analog triaxial connectors, it can result in the test PCB being flexed too much. To alleviate this problem, a small breakout board is used instead to route the digital signals from the main test PCB to a FPGA control board for easier digital signal control and programming. All the sensitive analog signals passing to the SMUs are connected through coaxial and triaxial cables to mitigate leakage and excessive voltage drop. Digital signals are jumped using header pins and ribbon cables with alternating ground between the adjacent wires to help shield the signals from cross talk and ambient electric noise.

The test configuration is shown in Figure 4-3. A laptop computer is used to control and remotely program all characterization equipment through GPIB connections.

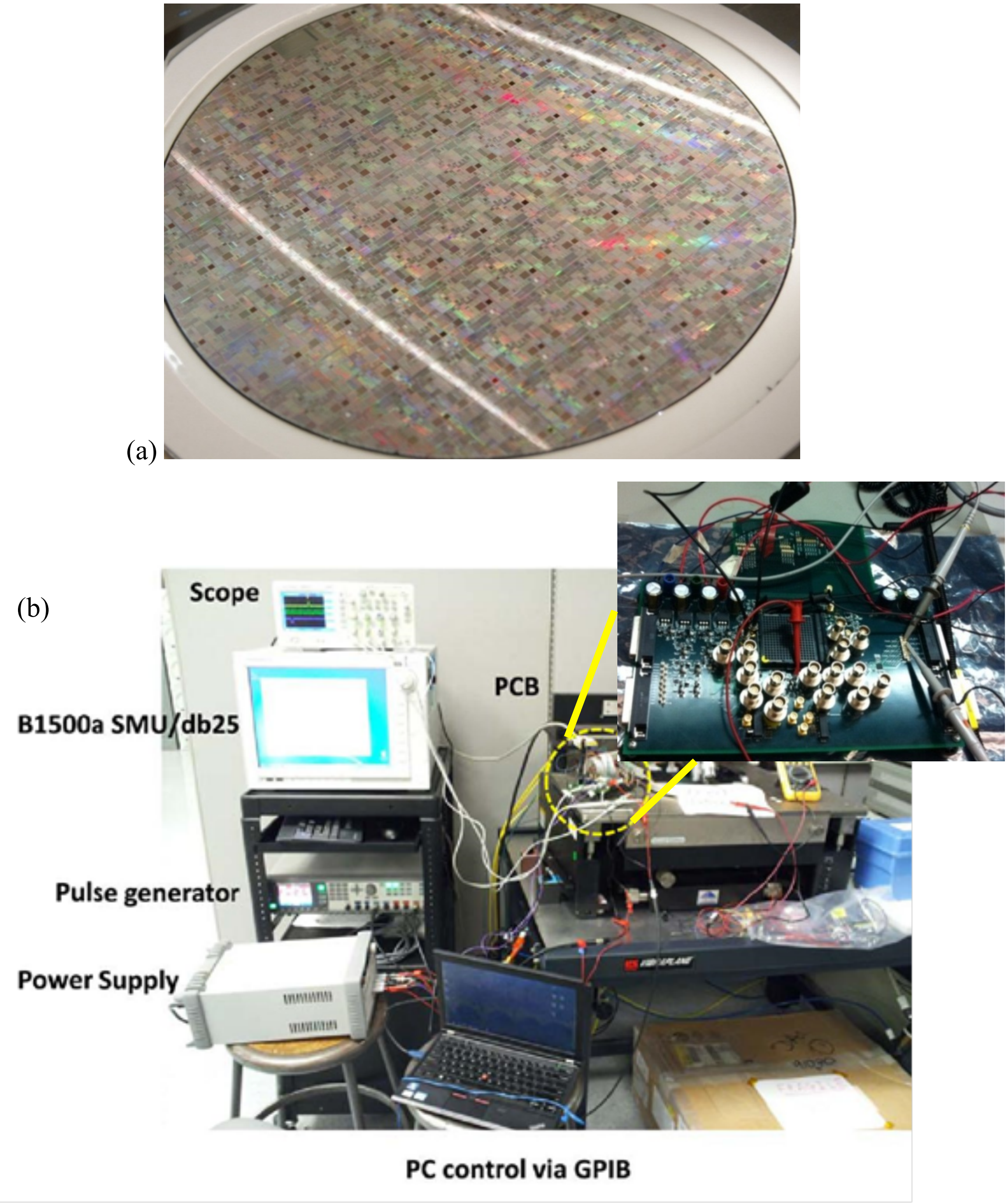


**Figure 4-3 (a) Wafer photo containing multiple dies; (b) Overall test setup consisting of semiconductor parametric analyzer (Agilent B1500A), arbitrary waveform generator, DC power supply, and a laptop computer; enlarged photo of test PCB and die socket.**

We have collected I-V measurement data of the pass-gate transistor within each SRAM cell across 5 rows and 512 columns of the test structure array on chip1 (G10) as well as data across 55 rows and 512 columns on chip2 (G7). These I-V curves are then used for compact variability model characterization. In this work, bit-line write trip voltage (BWTV) [2-11] is measured as a circuit performance example instead of other static metrics because it can be easily measured by sweeping the bit-line voltages of the SRAM cell while monitoring the currents flowing through the bit-lines. BWTV data are collected on chip1 (G10) only.

## 4.3 SRAM Transistor PSP Model Statistical Characterization

### 4.3.1 Review of PSP Model Parameters

The PSP model is an advanced surface potential based compact SPICE model, intended for digital, analog and RF design, which has been jointly developed by Philips Research and The Pennsylvania State University [4-12]. It includes all relevant physical effects including mobility reduction, velocity saturation, DIBL, gate current, and STI stress to model Nano-scale CMOS technologies. Wrapper models can be easily developed upon this core model at the foundry's site for adaptation to specific manufacturing process.

The PSP model has two sets of model parameters: the global-level parameter set, which describes entire space of device geometries, and the local-level parameter set, which models transistors with specific device dimensions [4-13]. In this experiment, we only extract parameters of the pass-gate transistors within the bit cell, focusing on the derived local-level parameters. Per the recommended local parameter extraction procedure in the PSP manual [4-13] and the *I-V* data available in the experiment, 18 parameters are chosen as pre-candidates for our experiment in parameter extraction. The parameter names and their physical meanings are listed in Table 4-1.

| **Param.** | **Description** | **Param.** | **Description** |
|---|---|---|---|
| *vfbo* | Geometry-independent flat-band voltage | *toxo* | Gate oxide thickness |
| *nsubo* | Geometry-independent substrate doping | *npo* | Geometry-independent gate polysilicon doping |
| *dphibo* | Geometry-independent offset of $\phi_B$ | *rsw1* | Source/drain series resistance |
| *cto* | Geometry-independent part of interface states factor CT | *thesato* | Geometry-independent velocity saturation |
| *cfl* | Length dependence of CT | *alpl* | Length dependence of CLM pre-factor ALP |
| *uo* | Zero-field mobility at TR | *alp1l1* | Length dependence of CLM enhancement factor above threshold |
| *xmueo* | Geometry-independent mobility reduction coefficient | *novo* | Effective doping of overlap region |
| *themuo* | Mobility reduction exponent | *vpo* | CLM logarithmic dependence |
| *xcoro* | Geometry-independent non-universality | *cso* | Geometry-independent Coulomb scattering |

**Table 4-1 Candidates of PSP model parameters for extraction [4-15]**

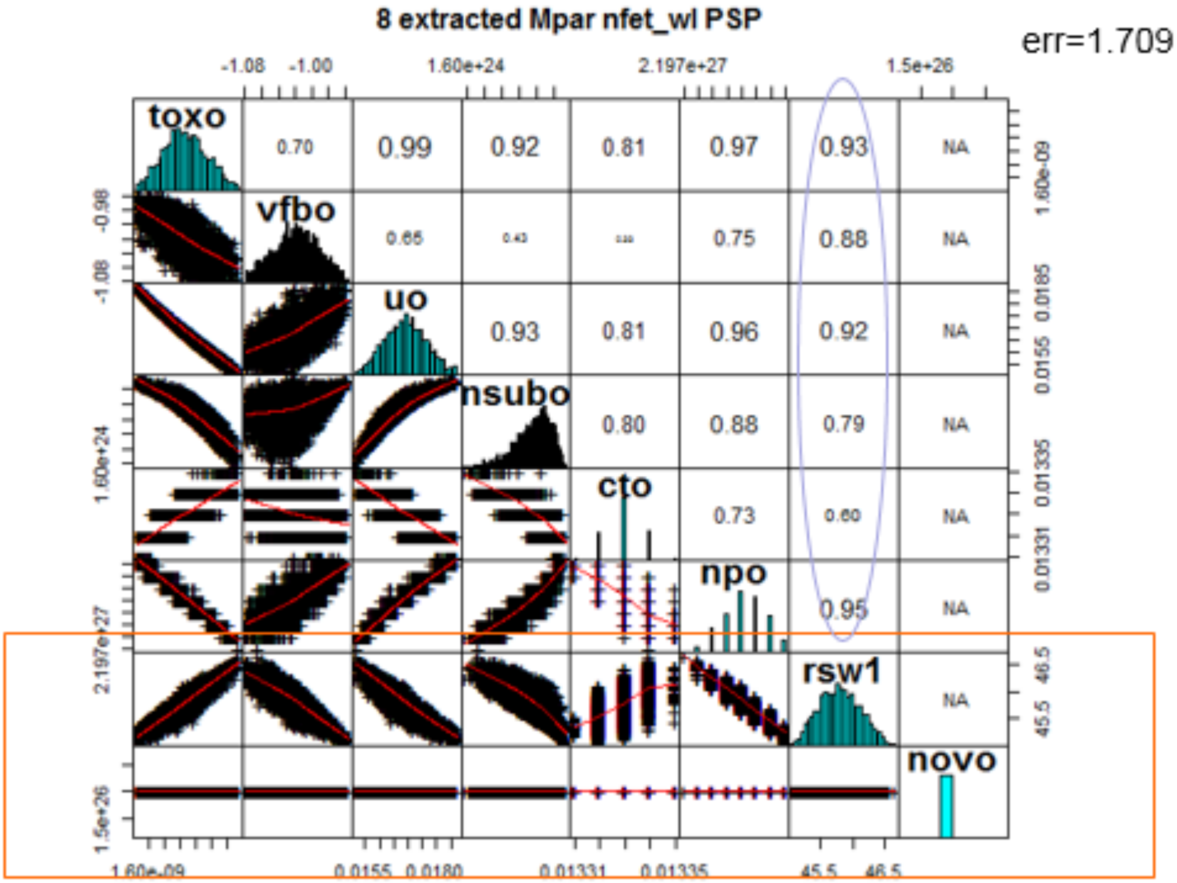
### 4.3.2 Statistical Model Characterization Results[4]

As discussed in [4-14], properly selecting model parameters for statistical characterization is essential for compact variability modeling. We have proposed and implemented a stepwise parameter selection procedure to obtain an optimal set of model parameters for statistical model parameter extraction. Based on the test structure design, we have chosen the industry standard PSP 103.1 model with nominal model parameter values from the collaborating foundry's default model. The dataset presented here is obtained from chip1 (G10), with pass-gate NMOS transistor full I-V curves among cells across 5 rows and 512 columns.

Simulation-based sensitivity analysis of pass-gate NMOS threshold voltage on the nominal model card is applied to select a "starting" subset of the standard PSP statistical model parameters (~20). The normalized CIs, described in Chapter 3, of these parameters are treated as indicators for parameter selection. This set can capture the variations in device performance within the low drain voltage operation region, where our measurements are collected.

The stepwise parameter selection analysis starts with an "initial" set of 8 PSP model parameters. Model parameters with the "worst" extraction quality are removed iteratively. This procedure stops when excluding more parameters will severely deteriorate the fitting quality, as represented by the relative residual fitting error. As shown in Figure 4-4, the optimal four-parameter-set, {*vfbo*, *uo*, *toxo*, *nsubo*}, has clear physical property representations in the model equation, reasonable model fitting error (1.833) and a far simpler statistical correlation structure. The parameters that are excluded during the procedure either have extremely high correlations with all the other parameters (*rsw1*, *npo*) or have unreliable extraction values of hitting the preset boundary values (*cto*).

[4] The contents of this and the following sections are generated from the previously published article [4-17].



**Stepwise parameter selection results with pass-gate I-V data from chip1 on PSP model (to be continued)**

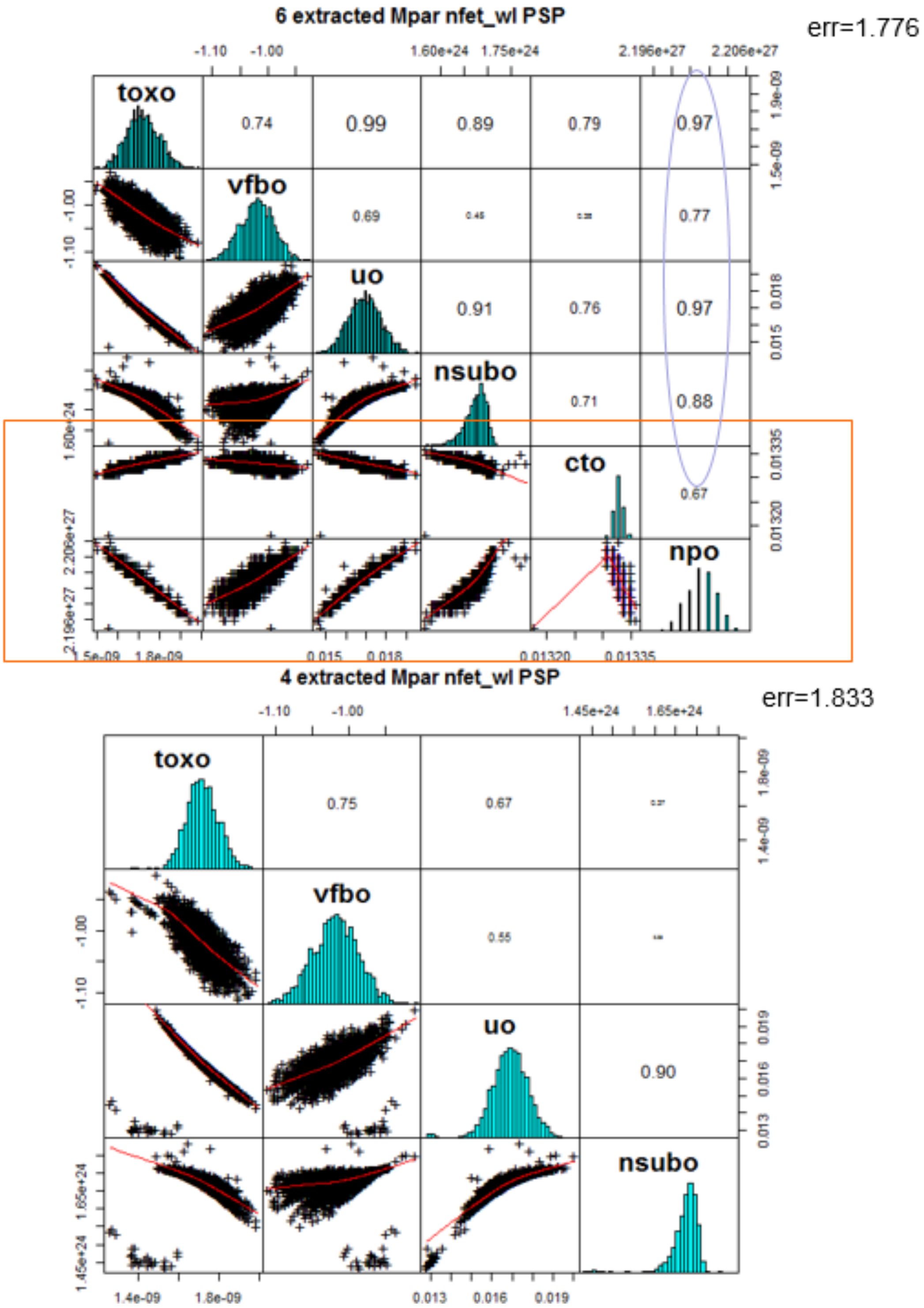


**Figure 4-4 Stepwise parameter selection results with pass-gate I-V data from chip1 on PSP model**

## 4.4 Statistical Circuit Simulation Results on SRAM writability

In this work, we have developed a customized MC simulation bench to conduct statistical circuit performance simulations using our variability-aware compact models. Python-based in-house customized scripts work as a wrapper driving standard HSPICE MC Simulations on circuits. The scripts can accommodate pre-defined compact model statistical parameters without hacking into the standard MC simulation flow.

Circuit performance measurements are obtained simultaneously with device characterization data to reduce systematic error. The data presented here are collected from chip1 (G10), fabricated in 28*nm* FDSOI technology, with SRAM cells across 5 rows and 512 columns.

Figure 4-5 shows the comparison results of measurements vs. simulations. QQ-plots are added for a complete comparison and clear visualization of the distribution. The "full standard" MC simulations are based on the compact model embedded in the default design kit. The results show that customized MC simulations using the far simpler statistical compact models, represented only by parameters that were selected from Figure 4-4, match the full MC simulation results in distribution estimation. Both present very similar histograms and standard deviation numbers compared to the measurement data, though missing the non-Gaussian tail. Furthermore, as shown in Figure 4-5, for this SRAM dataset, the optimal statistical model parameter set for simulation is the same as for extraction, which is {*vfbo*, *uo*, *toxo*, *nsubo*}, based on the nominal PSP model. Redundant compact model statistical parameters slow down the simulation without improving the overall accuracy in circuit performance distribution estimation. This is a strong validation of our proposed model characterization methodology.

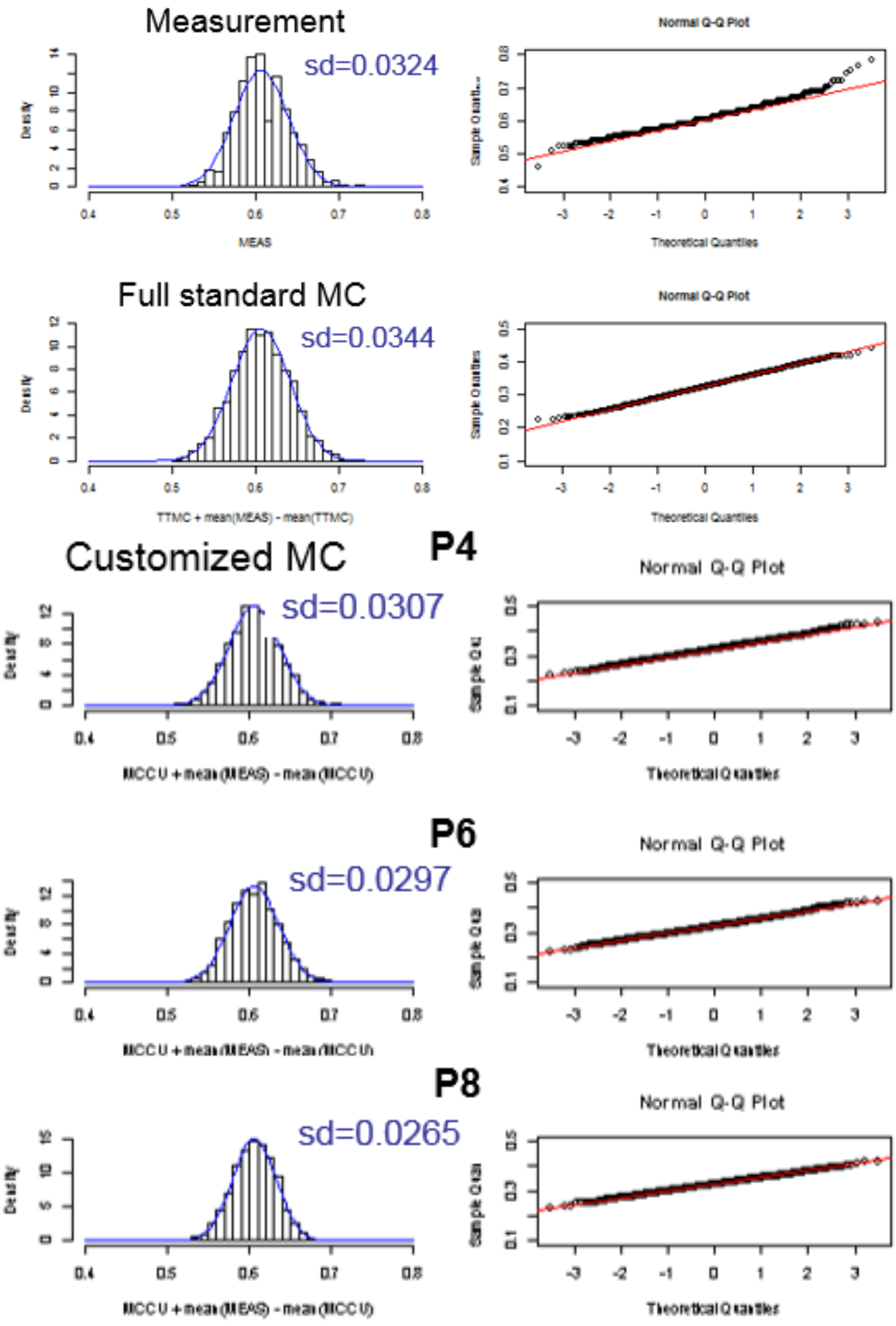


**Figure 4-5 SRAM DC write noise margin BWTV metric**
**Comparison of measurements vs. full standard Monte Carlo simulations vs. customized Monte Carlo simulations using differ reduced set of extracted model parameters.**

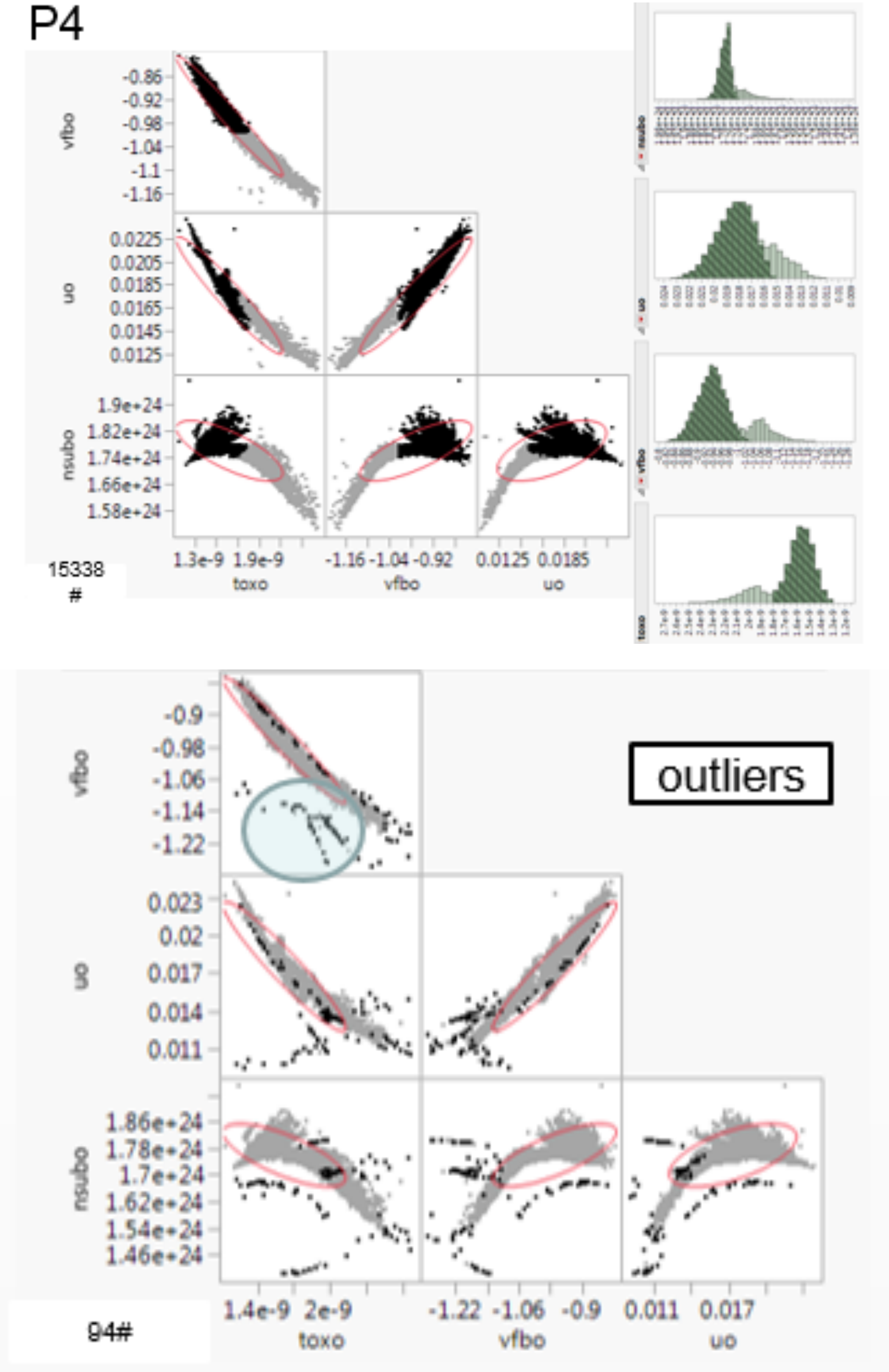


**Figure 4-6 Statistical model parameter extraction results from chip2 data (55 rows and 512 columns) on PSP model, extracted parameter set of {*vfbo*, *uo*, *toxo*, *nsubo*} shows multimodal distribution as well as outliers in the data set.**

## 4.5 Summary

We have proposed a statistical compact model extraction methodology, which has been implemented on a PSP model with 28nm SRAM cell bit transistors data. We built a CAD flow to utilize the models for statistical circuit performance estimation; simulation results on SRAM writability performances are very close to measurements in statistical distributions. In the future, we will exploit the statistical structure of the extracted parameters to capture the nonlinear correlations and the non-Gaussian distributions through mixture of Gaussian distributions, as in Figure 4-6. The goal is to demonstrate that our simplified model can capture significant non-normality in the measured data. Such non-normality is often evident at the tails of the performance distributions, and capturing that is necessary for the statistical modeling of inherently high-yielding IC designs.

## Reference

[4-1] K. Zhang, Embedded Memories for Nano-Scale VLSIs, Integrated Circuits and Systems Series, Springer, 2009.

[4-2] Z. Guo; A. Carlson; L.-T. Pang; K.T. Duong; T.-J. King; B. Nikolic, *"Large-Scale SRAM Variability Characterization in 45 nm CMOS,"* Solid-State Circuits, IEEE Journal of, vol.44, no.11, pp.3174-3192, Nov. 2009.

[4-3] A. E. Carlson, *"Device and circuit techniques for reducing variation in nanoscale SRAM,"* Ph.D. dissertation, UC Berkeley, May 2008.

[4-4] S.V. Kumar; C.H. Kim; S.S. Sapatnekar, *"Impact of NBTI on SRAM read stability and design for reliability,"* Quality Electronic Design, 2006. ISQED '06. 7th International Symposium on, pp.6 -18, 27-29 March 2006.

[4-5] K. Takeuchi; T. Nagumo; K. Takeda; S. Asayama; S. Yokogawa; K. Imai; Y. Hayashi, *"Direct observation of RTN-induced SRAM failure by accelerated testing and its application to product reliability assessment,"* VLSI Technology (VLSIT), 2010 Symposium on , pp.189,190, 15-17 June 2010.

[4-6] S.O. Toh; T.-J. King; B. Nikolic, *"Impact of random telegraph signaling noise on SRAM stability,"* VLSI Technology (VLSIT), 2011 Symposium on, pp.204-205, 14-16 June 2011.

[4-7] X. Deng , W. K. Loh , B. Pious , T. W. Houston , L. Liu , B. Khan and D. Corum "*Characterization of bit transistors in a functional SRAM*", IEEE Symp. VLSI Circuits Dig., pp.44 -45, 2008.

[4-8] B. Zimmer, *"Resilient Design Methodology for Energy-Efficient SRAM"*, M.S. report, EECS, UC Berkeley, 2012.

[4-9] N. Planes, O. Weber, V. Barral, S. Haendler, D. Noblet, D. Croain, M. Bocat, P. Sas-soulas, X. Federspiel, A. Cros, A. Bajolet, E. Richard, B. Dumont, P. Perreau, D. Pe-tit, D. Golanski, C. Fenouillet-Beranger, N. Guillot, M. Rak, V. Huard, S. Puget,X. Montagner, M. A. Jaud, O. Rozeau, O. Saxod, F. Wacquant, F. Monsieur, D. Barge,L. Pinzelli, M. Mellier, F. Boeuf, F. Arnaud, and M. Haond, *"28nm FDSOI technology platform for high-speed low-voltage digital applications,"* in VLSIT, pp. 133-134, 2012.

[4-10] B. Zimmer, O. Thomas, S.O. Toh, T. Vincent, K. Asanovic and B. Nikolic, *“Joint Impact of Random Variations and RTN on Dynamic Writability in 28nm Bulk and FDSOI SRAM”*, 2014 44th European Solid State Device Research Conference (ESSDERC), Venice, 2014, pp. 98-101.

[4-11] Z. Guo, A. Carlson, L. T. Pang, K. Duong, T.-J. K. Liu and B. Nikolic, *"Large-scale read/write margin measurement in 45nm CMOS SRAM arrays,"* VLSI Circuits, 2008 IEEE Symposium on, Honolulu, HI, 2008, pp. 42-43.

[4-12] W. Wu, X. Li, G. Gildenblat, G. Workman, S. Veeraraghavan, C. Mcandrew, R.Vanlangevelde, G. Smit, a Scholten, and D. Klaassen, *“PSP-SOI: An advanced surface potential based compact model of partially depleted SOI MOSFETs for circuit simulations,”* Solid. State. Electron., vol. 53, no. 1, pp. 18–29, Jan. 2009.

[4-13] PSP Group, “PSP Manual V103p1,” http://pspmodel.asu.edu/downloads/psp103p1_summary.pdf.

[4-14] X. Li, C.C. McAndrew, W. Wu, S. Chaudhry, J. Victory and G. Gildenblat, “*Statistical Modeling with the PSP MOSFET Model*,” IEEE Trans on CAD of Integrated Circuits and Systems, vol. 29, no. 4, pp.599-606, 2010.

[4-15] W. Zhao and 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.

[4-16] Y. Qiao, K. Qian, C.J. Spanos, *“Variability aware compact model characterization for statistical circuit design optimization”*, SPIE Advanced Lithography, 83270J-83270J-9, 2012.

[4-17] Y. Qiao, C. J. Spanos, *“Variability-aware compact modeling and statistical circuit validation on SRAM test array”*, SPIE Advanced Lithography, 97810D-97810D-9, 2016.

# Chapter 5

# Statistical Device Characterization and Circuit Modeling with Imaging ROIC Data

## 5.1 Introduction

This chapter proposes an efficient method to build statistically valid prediction models for circuit performances based on transistor test array measurement data. We exploit the improvement of statistical model parameter extraction procedure by a stepwise parameter selection algorithm with confidence intervals (CI) as quantitative *extraction quality* criteria. This projection maps on-chip electrical measurements onto a subspace spanned by a set of physical variables of compact transistor models. The key idea is to reduce the dimensionality (model parameters) of the statistical sub-model while keeping the physical correlation between the model prediction and the device electrical measurement.

In addition, an expectation-maximization (EM) algorithm is employed to iteratively solve the GMM estimation problem of the extracted model parameters. Our customized MC simulation platform therefore incorporates this advanced statistical compact variability model for better circuit performance modeling. Compared with the traditional PCA and RSM method, the proposed method reserves the physical link between compact model variables and measurements while still compatible with industrial standard IC design flow, as illustrated conceptually in Figure 5-1.

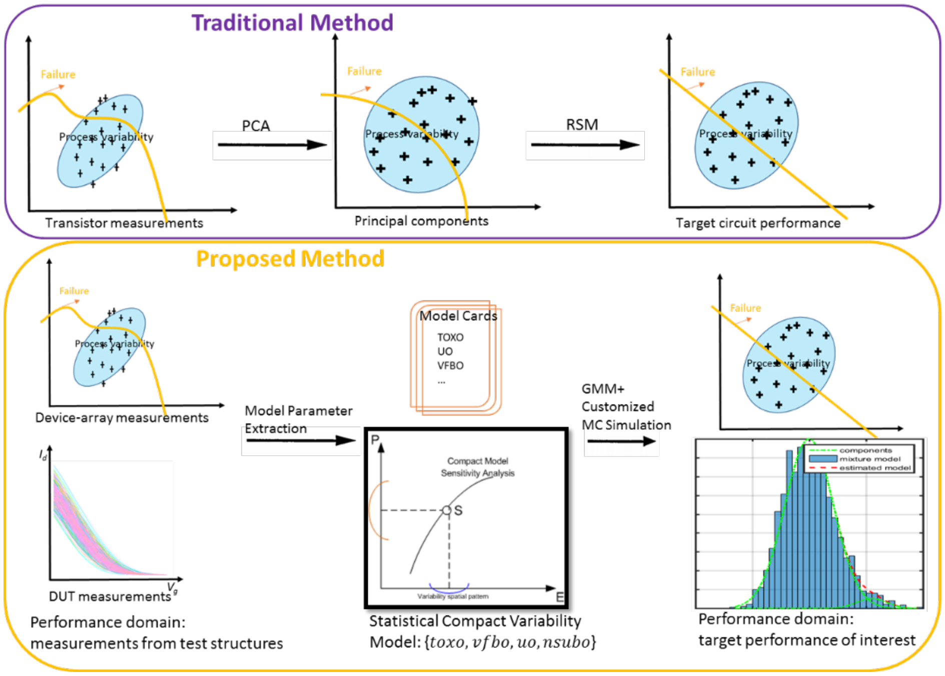


**Figure 5-1 Conceptual illustration of our proposed method**

## 5.2 Imaging ROIC Design and Measurement

In CMOS infrared (IR) imagers, a non-silicon photodetector, which generates current proportional to the amount of light it absorbs, is fabricated on top of a standard CMOS readout integrated circuit (ROIC). With IC technology advancing, this photocurrent is typically converted to a digital signal with an analog to digital converter (ADC) built at the pixel level, using a current to frequency converter and a digital counter [5-1]. However, with shrinking device sizes down to nanometer range, degradation of performance shows up in imaging sensor cell resolution improvements with in-pixel computation [5-2].

On the other hand, large-scale device characterization studies can help provide information about the performance variation within a CMOS technology node [5-3]. With prior work that requires direct addressable device arrays using separate test bench supplies for each device terminal [5-4], our collaborating design team builds on well-tested techniques in imaging ROIC design to investigate methods of variability characterization with fully digital interfaces [5-5]. The design team uses an array of test cells with in-pixel integrated ADCs to achieve high measurement throughput for characterizing performance variation in a $28nm$ bulk CMOS technology, to determine the feasibility of implementing precision ROICs [5-6].

### 5.2.1 DUT and Measurement Cell

The device under test (DUT) unit cell, shown in Figure 5-2, generates a test current fed into the adjacent measurement cell. Within the DUT cell, digital scan chain signals ensure that only the drain current of the desired DUTs contributes to the test current. The scan input (S_in) and scan clock (S_clk) are shared between all unit cells, but the scan chain output (S_out) only propagates if both row and column enable signals (array-level controls) are high. In the characterization testing, only the $|V_{GS}|$ of selected devices is set to the input value, while other terminals are fixed at preset voltages.

Within the measurement cell, $I_{test}$ from the DUT cell is first subtracted from a bias current $I_{bias}$ (due to design topology) before being integrated onto a reference capacitor. This is to implement the current-to-frequency converter, where the capacitor is reset whenever the ramp voltage, generated by the test current capacitive integration, exceeds a preset threshold within the RST (reset) logic block. These pulses are counted through a 16-bit counter for digitalizing the current-dependent frequency. Measurement cells will be calibrated using the chip-level bias current reference to mitigate the effect of variations in the integration capacitance and the preset ramp voltage threshold.

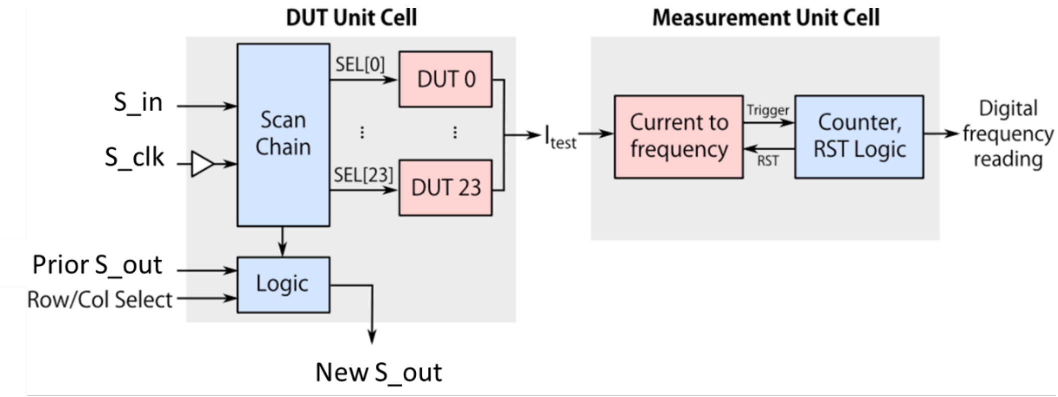


**Figure 5-2 General structure of unit cell pair, containing one device under test (DUT) unit cell (left) and one measurement unit cell (right) [5-6].**

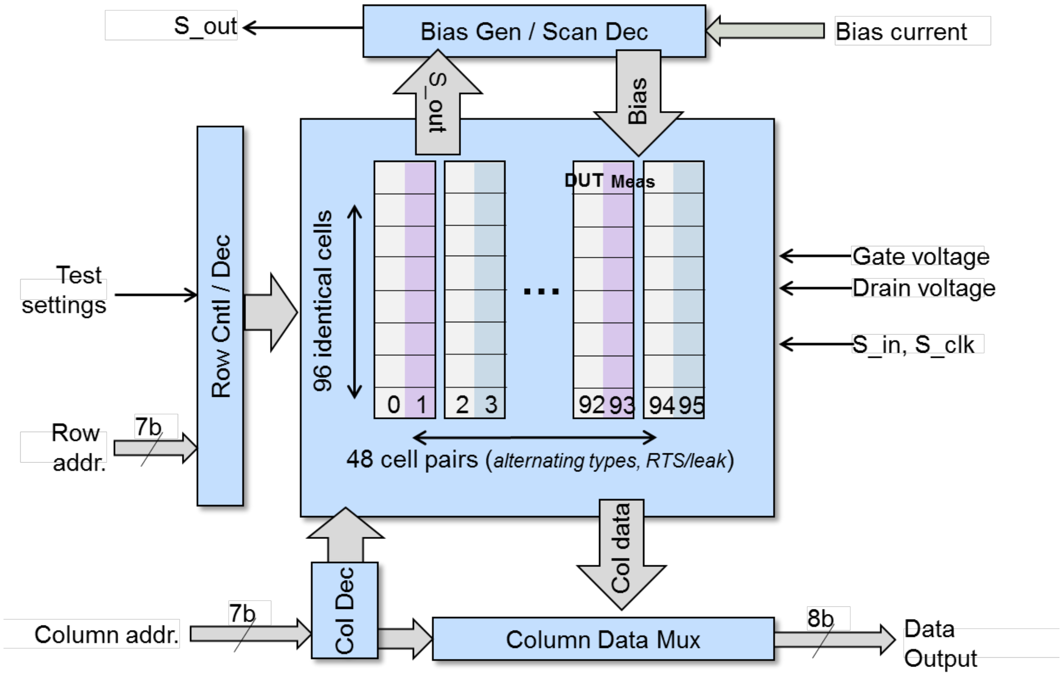


**Figure 5-3 Characterization chip system architecture [5-6]**

### 5.2.2 Test Chip Architecture

This characterization chip mirrors the design of a digital focal plane array (FPA) ROIC [5-7]. Figure 5-3 shows that it consists of a central array of 96x48 DUT and measurement cell pairs, with peripheral circuitry to address each cell pair. Digital frequency/counter readings from each cell pair are collected through the chip-level output using the row/column decoders. A simple current mirror generates reference bias currents for measurement calibration. Adopting an architecture that is similar to a FPA ROIC allows the digital interface circuitry as well as the test chip measurement processing platform be reproduced from previous work [5-8].

A die photo of the test chip fabricated in a 28*nm* bulk CMOS process is shown in Figure 5-4. The total die area is 3.24$mm^2$, with approximate core area of 1.82 $mm^2$ [5-9].

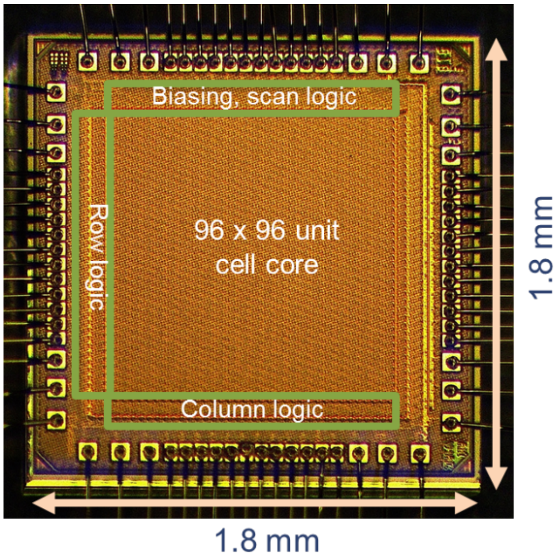


**Figure 5-4 Die photo of the fabricated chip in 28nm bulk process [5-6]**

### 5.2.3 Measurement Procedure and Calibration

To obtain statistically meaningful results, this dataset of characterization test chip measurement includes over 6,000 functional DUTs for each type of device. NMOS and PMOS devices with multiple (high/regular/low) threshold voltage ($V_{th}$) flavors and dimensions of interest (80*nm*/30*nm*, 80*nm*/86*nm*) to digital designs are included. Within each cell, the serial scan chain signal is translated into 24 parallel select signals that control the 24 devices within the DUT cell. Thus, each transistor is identified by a scan channel number, as well as the row and column addresses of the unit cell.

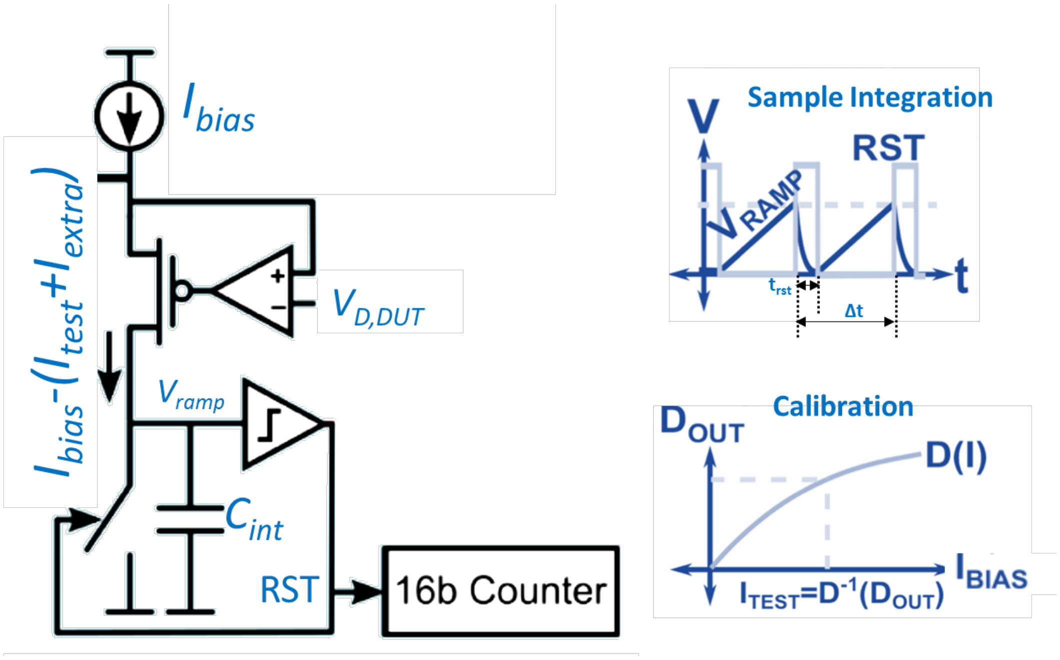


**Figure 5-5 Example of calibration and measurement [5-6]**

The illustration of the typical integration period and possible measurement parasitics are show in Figure 5-5 and Equation 5-1:

$$D(I) \sim f_{meas} = \Delta t^{-1} = \left[t_{rst} + \frac{C_{int}\Delta V_{ramp}}{I_{bias} - (I_{test} + I_{extra})}\right]^{-1}$$

$$= D(I_{bias} - I_{test} | t_{rst}, C_{int}\Delta V_{ramp}, I_{extra}) \quad (5\text{-}1)$$

$$I_{test} = D^{-1}(f_{meas}@I_{test} = 0 | I_{bias}) - D^{-1}\left(f_{meas}@I_{test} = I_{D,DUT} | I_{bias}\right)$$

Here, $t_{rst}$ is the reset (RST) pulse duration, $I_{test}$ is the desired device test current, $I_{extra}$ is the unwanted parasitic current in the test array (gate leakage, additional leakage from non-DUT transistors, etc.), $I_{bias}$ is the measurement cell bias current, $\Delta V_{ramp}$is the voltage swing determined by the measurement cell delay path trip point, and $C_{int}$ is the total integration capacitance at the ramp output node.

To account for the nonlinearity in the formula [5-11], each unit cell pair must be calibrated individually by measuring a full $I_{bias}$ vs. frequency counter digital output *D* curve over the desired operating range, which can be used as a lookup table to find the corresponding test current given a measured counter output. Here, the test current is sampled differentially before and after the desired test device is selected by the scan chain in order to eliminate possible parasitic background current $I_{extra}$. Meanwhile, this calibration procedure will generate the indirect measurement of $t_{rst}$ during the function $D^{-1}$ parameter fitting, which we will use later in our circuit performance modeling study.

## 5.3 Transistor Model Statistical Characterization

### 5.3.1 Dataset Description and PSP Model Parameter Review

The measured test chip (labeled: chip3) contains 96 cell pairs per column and 48 cell pairs per row. There are two types of unit cell pairs: a normal *I-V* characterization unit, in which the $V_{GS}$ of each DUT can be controlled independently to sweep across desired measurement voltage range ($V_{DS}$ is set at

$50mV$ for better $V_{th}$ and noise characterization for other collaborating work), and a leakage current characterization unit, in which $V_{DS}$ modulation is used to control drain to source leakage current for DUTs that already have $V_{GS}$=$0V$. Therefore, there are only 96*18=1728 cell pairs measured for *I-V* characterization data (out of the 24 *I-V* cell pairs per row, 6 of them are not included due to ill-functioned behavior caused by their periphery array location).

Table 5-1 summarizes the quantity of various PMOS device types included in the chip. Here, we pick PMOS I-V data, which is more prone to error as it is not limited by the predefined $I_{bais}$. Each I-V DUT cell contains 24 PFETs, with six unique device types. The high $V_{TH}$ *(HVT)*, low $V_{TH}$ *(LVT)*, and regular $V_{TH}$ *(RVT)* devices are used to validate leakage current measurement trends as well as explore variability as a function of threshold voltage. Only two device dimensions are considered: minimum-sized transistors (80*nm*/30*nm*) for digital gates and longer devices (80*nm*/86*nm*) for analog applications. It allows more devices of each type to be incorporated within each DUT cell by including only a few dimensions, which will create a larger statistical dataset and improve the likelihood of capturing corner cases of interest.

| **PMOS DUT Quantity** | **DUT $V_{TH}$ flavors** | ***HVT*** | ***RVT*** | ***LVT*** |
|---|---|---|---|---|
| **DUT dimensions (W/L)** | 1728*24= 41472 | 1728*8= 13824 | 1728*8= 13824 | 1728*8= 13824 |
| ***80nm/30nm*** | 1728*12= 20736 | 1728*4= 6912 | 1728*4= 6912 | 1728*4= 6912 |
| ***80nm/86nm*** | 1728*12= 20736 | 1728*4= 6912 | 1728*4= 6912 | 1728*4= 6912 |

**Table 5-1 Summary of DUT types included in the measurement dataset.**

The foundry-wrapped PSP models applied within the design kit of this characterization chip have its own set of statistical model parameters. We adopt a subset of the foundry default parameter set based on both the design kit and Table 4-1, shown in Table 5-2.

| **Param.** | **Description** | **Param.** | **Description** |
|---|---|---|---|
| *vfbo* | Geometry-independent flat-band voltage | *toxo* | Gate oxide thickness |
| *nsubo* | Geometry-independent substrate doping | *npo* | Geometry-independent gate polysilicon doping |
| *dphibo* | Geometry-independent offset of $\phi_B$ | *rsw1* | Source/drain series resistance |
| *cto* | Geometry-independent part of interface states factor CT | *thesato* | Geometry-independent velocity saturation |
| *cfl* | Length dependence of CT | *novo* | Effective doping of overlap region |
| *uo* | Zero-field mobility at TR | *xcoro* | Geometry-independent non-universality |
| *themuo* | Mobility reduction exponent | *cso* | Geometry-independent Coulomb scattering |

**Table 5-2 Candidates of PSP model parameters for statistical extraction (excerpt from design kit manual)**

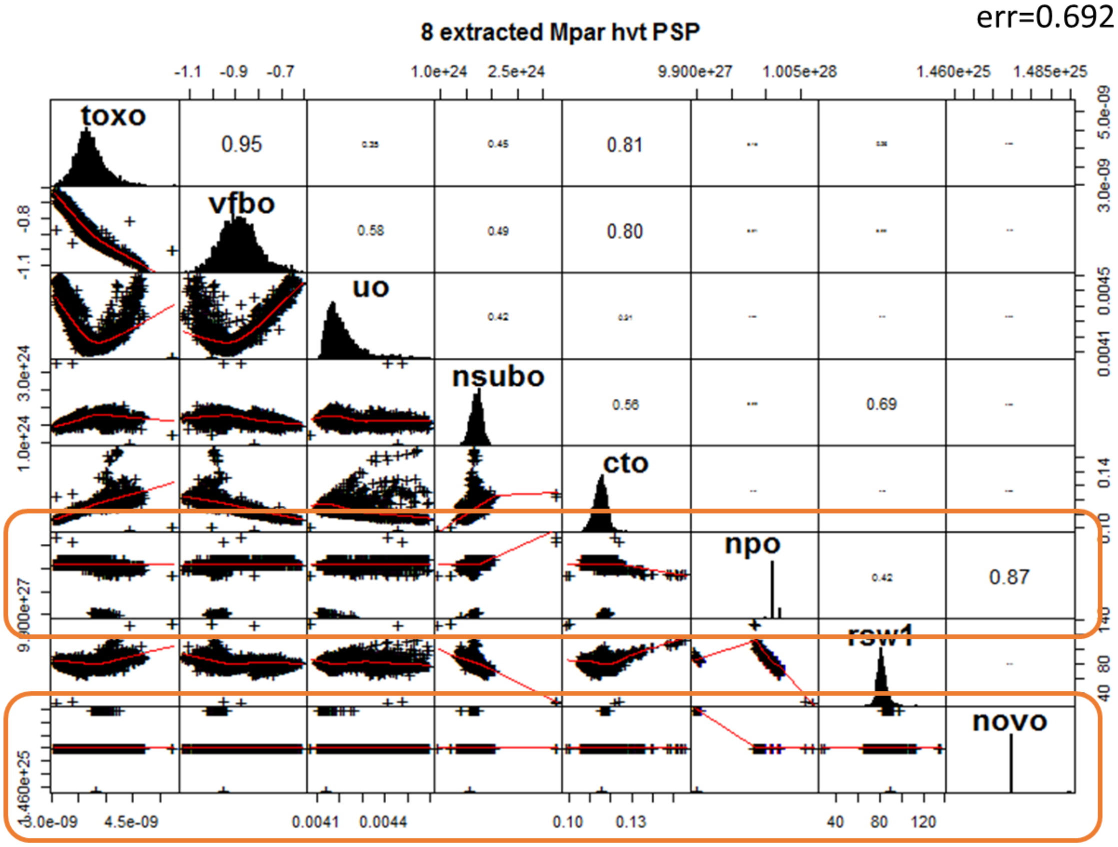
## 5.3.2 Stepwise Compact Model Parameter Selection Results

As discussed in Section 3.2.2, we have proposed and implemented a stepwise compact model parameter selection procedure to obtain an optimal set of model parameters for statistical compact model parameter extraction. Based on the chip design kit, we have chosen the industry standard PSP model wrapped with foundry's confidential driver.

SPICE simulation based sensitivity analysis of the HVT PMOS threshold voltage on the nominal model card (with nominal compact model parameter values from the foundry's default model) is applied to select a "starting" subset of the standard PSP statistical model parameters. The normalized confidence intervals (CIs), described in the Equation 3-7, of these parameters are treated as extraction quality criteria for parameter selection. This starting parameter set can capture the variations in device performance with low-$V_{DS}$ operations, where our measurements are collected. The dataset presented here is obtained from chip3, with minimum-sized HVT PMOS transistor full *I-V* curves among cells across 96 rows and 18 columns.

The stepwise compact model parameter selection algorithm starts with an "initial" set of 8 PSP model parameters. Model parameters with the "worst" *extraction quality* are removed iteratively. This procedure stops when excluding more parameters will severely deteriorate the *fitting quality*, as represented by the relative residual fitting error. As shown in Figure 5-6, the optimal four-parameter-set, {*vfbo*, *uo*, *toxo*, *cto*} has clear physical property representations in the model equation, a reasonable model fitting error (1.138) and a far simpler statistical correlation structure.

The parameters that are excluded during the procedure either have no variability representation (*novo*) or have unreliable extraction values of hitting the preset boundary values (*npo*). Furthermore, from the correlation plot of the final optimal parameter set, we can see a clear "grouping" effect, which indicates clusters (or groups) within the dataset. Just by preliminary data structure examination, we can speculate that the different scan channel number within each DUT cell can contribute to this dissection of dataset. GMM estimation in the later section will help verify this hypothesis by better visualization.



**Stepwise parameter selection results with HVT PMOS I-V data from chip3 on PSP model (continued on next page)**

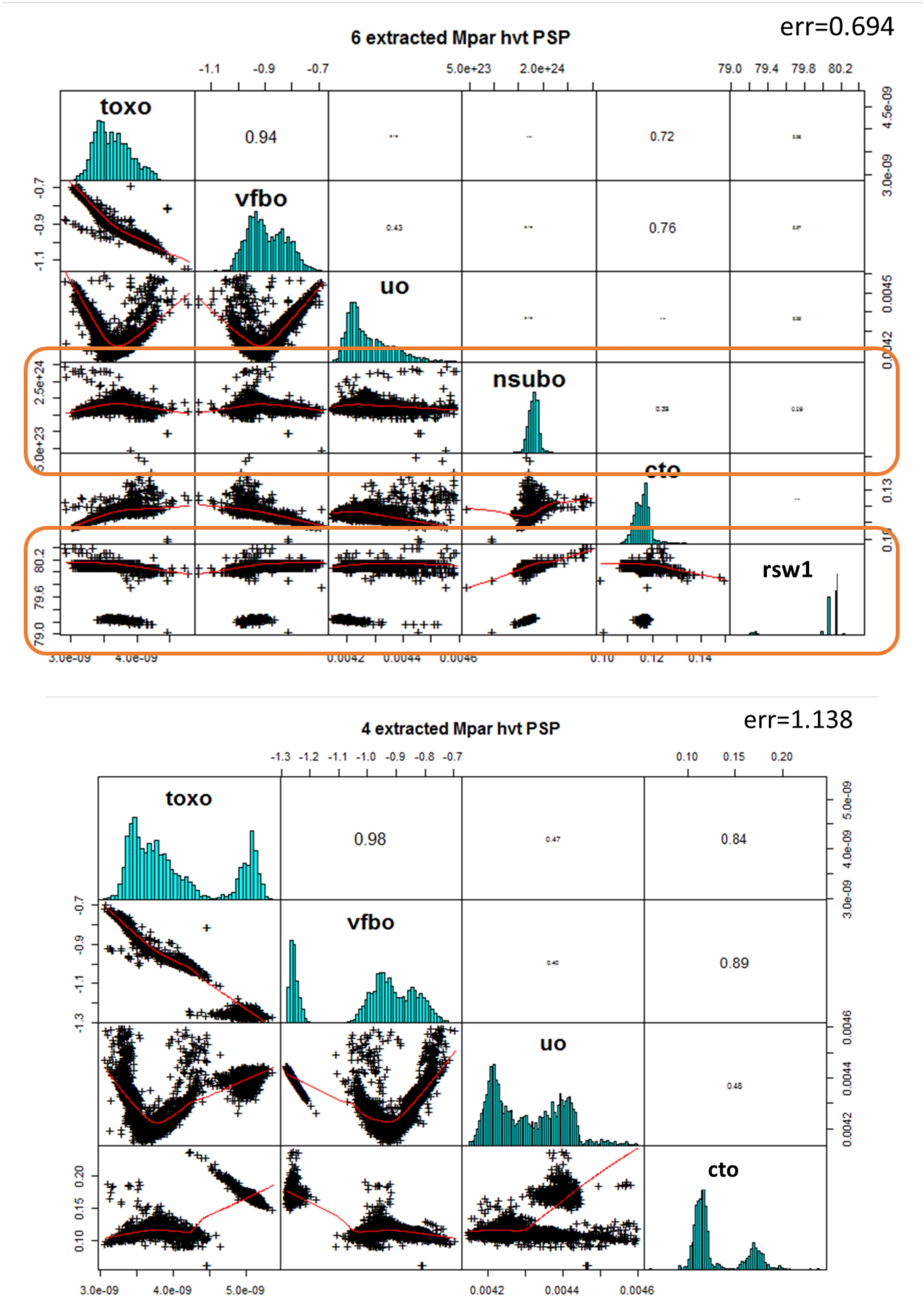


**Figure 5-6 Stepwise parameter selection results with HVT PMOS I-V data from chip3 on PSP model.**

### 5.3.3 Linear SBPV Device Characterization Results

The within-chip spatial pattern of the optimal four-parameter-set, {*vfbo*, *uo*, *toxo*, *cto*} of the HVT PMOS transistors are shown in Figure 5-7.

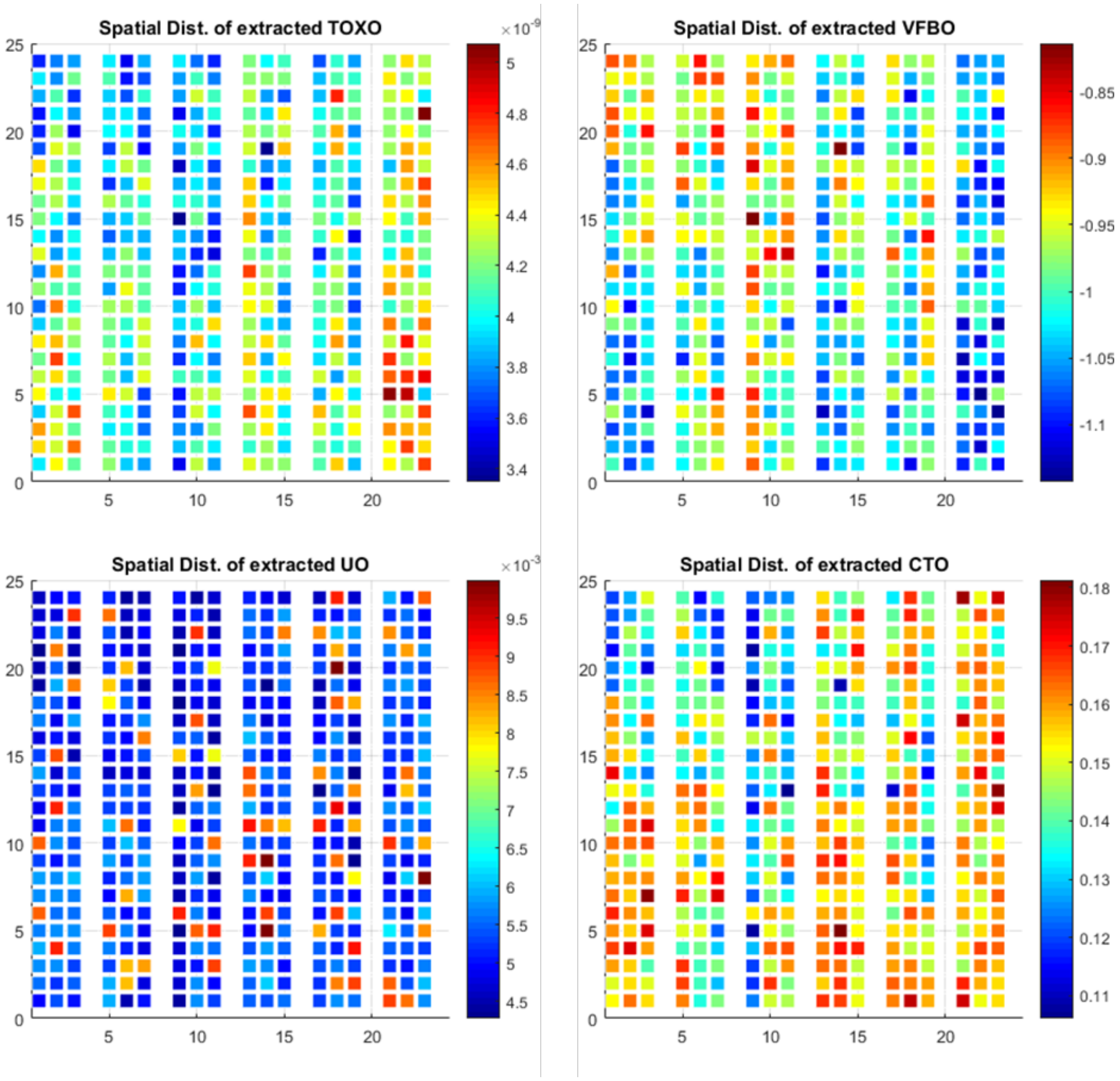
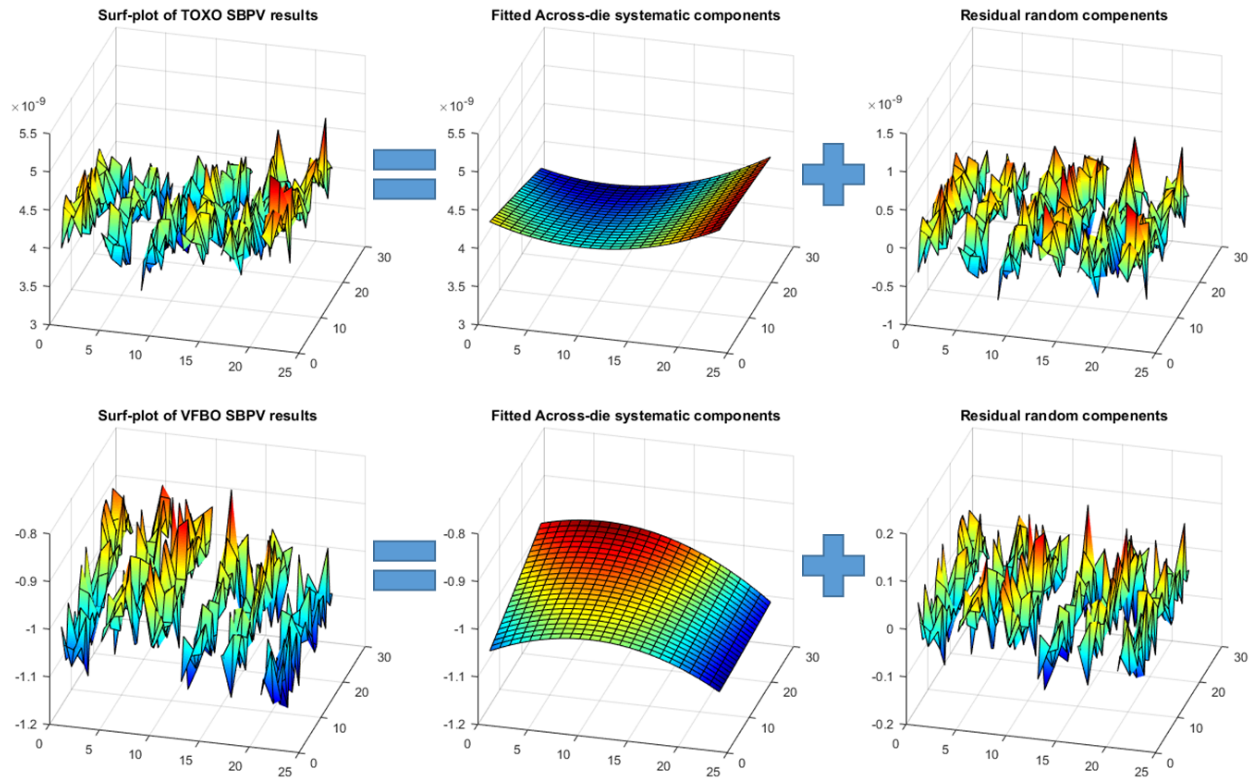


**Figure 5-7 Chip maps of extracted compact model parameters {*toxo, vfbo*, uo, cto};**

The mobility parameter *{uo}* does not show any significant across-chip pattern (but rather flat with a few outliers), which is in line with the fact that mobility variation in PSP compact model is mainly associated with the result of random dopant fluctuation and is largely dominated by the random components [5-10]. On the other hand, the parameters {*vfbo, cto*} both show a clear cross-chip pattern that varies along the columns of the DUT cell array, which will be captured by the SBPV method described in Chapter 3. We decompose the variability in the selected transistor *I-V* measurement points and apply our SBPV method to extract spatial patterns (parabolic surfaces along chip columns plus random components) of model parameters through linear propagation. The results are shown in Figure 5-8.



**Chip level variation decomposition for *{toxo, vfbo}* calculated from SBPV method (more parameters next)**

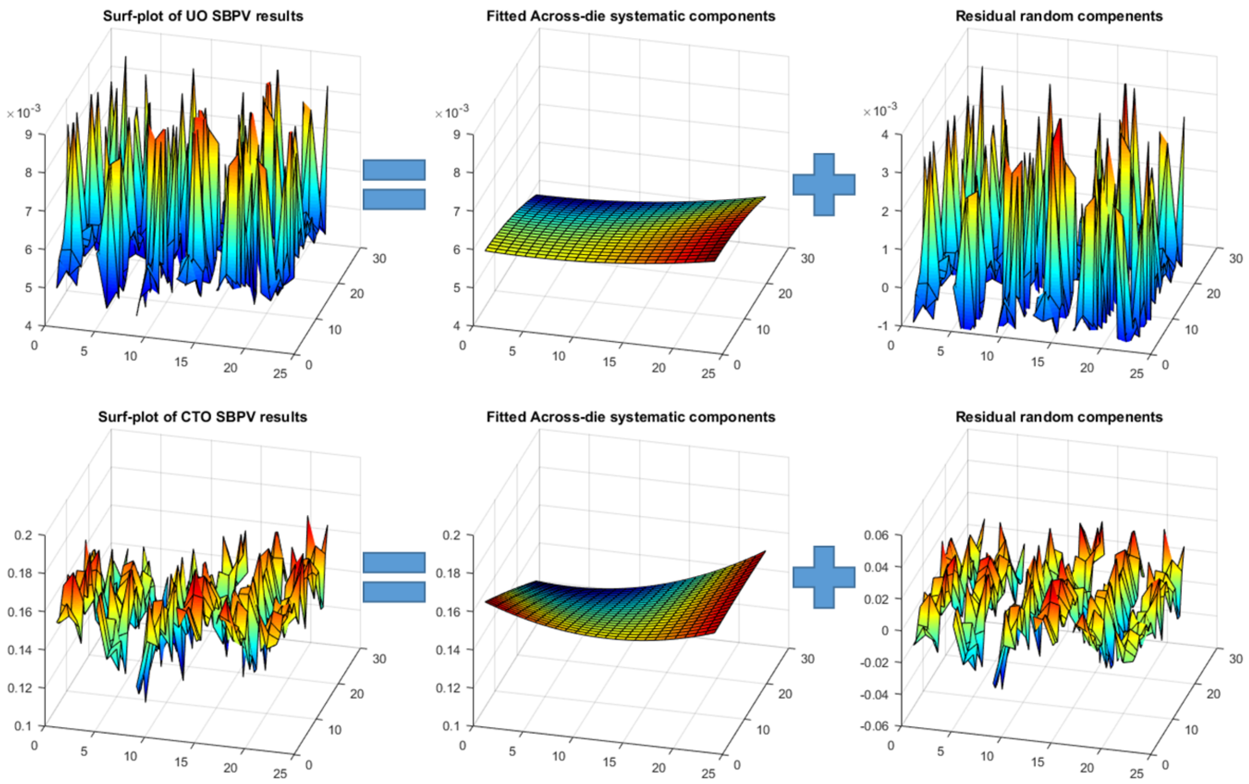


**Figure 5-8 Chip level variation decomposition for *{uo, cto}* calculated from SBPV method;**

## 5.4 Statistical Circuit Modeling and Verification Results on Critical Path Delays

### 5.4.1 Circuit Performance GMM Results

We have developed Python-based customized scripts to drive standard HSPICE Monte Carlo (MC) simulations incorporating our variability-aware compact models. The scripts can accommodate pre-defined compact model statistical parameters without hacking into the standard MC simulation flow.

However, with a complex correlation structure among extracted model parameters, Gaussian mixture models will be applied first to build mean/variance estimations of potential mixtures through the EM algorithm.

Our employment of GMM on the extracted model parameters falls within a general clustering problem (we do not have the "group label info"), which requires a systematic exploration of the number of mixture components. Here, with prior knowledge of four different scan channel number within the dataset, we limit the maximum mixture components to 4. The visualization of two parameters *{vfbo, cto}* is shown in Figure 5-9, with the final selected 2-component GMM details shown in Figure 5-10.

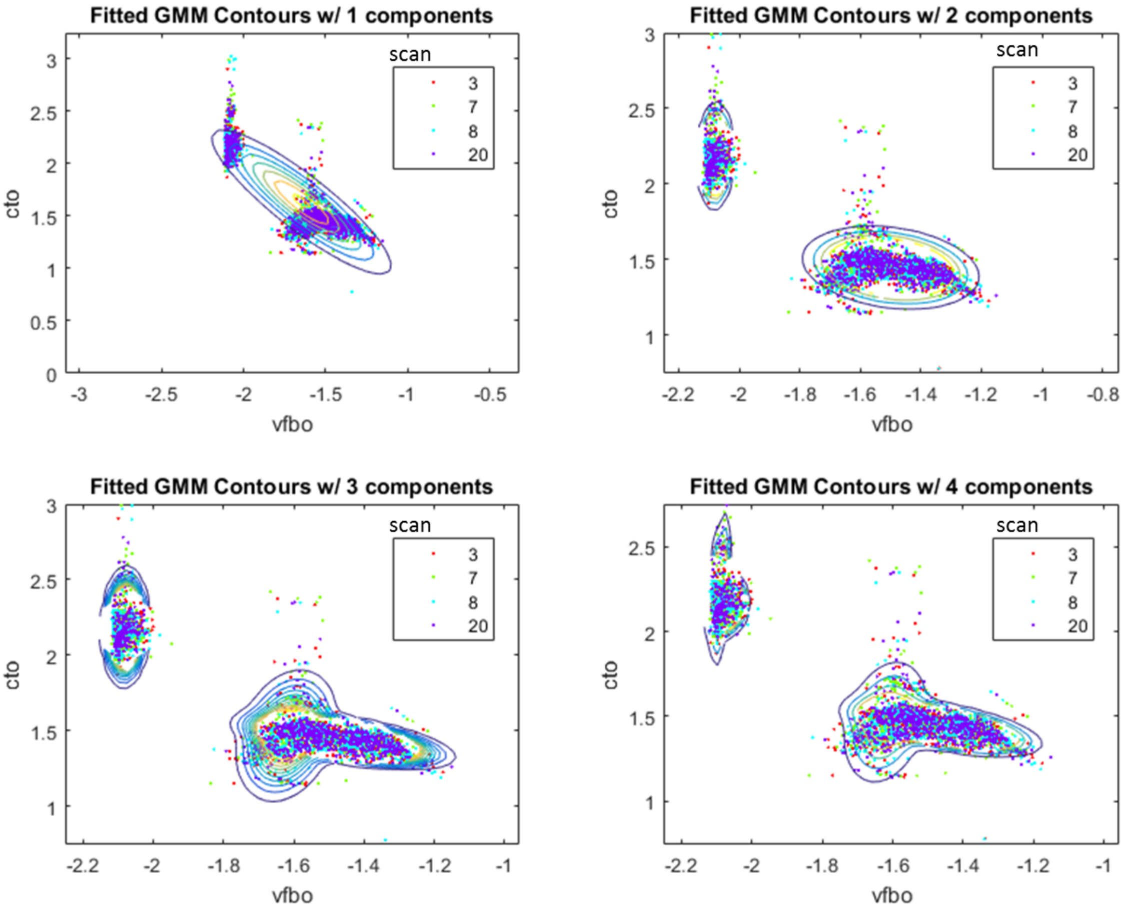


**Figure 5-9 Fitting GMM contours with different number of mixture components on extracted model parameter pair *{vfbo, cto}* for visualization;**

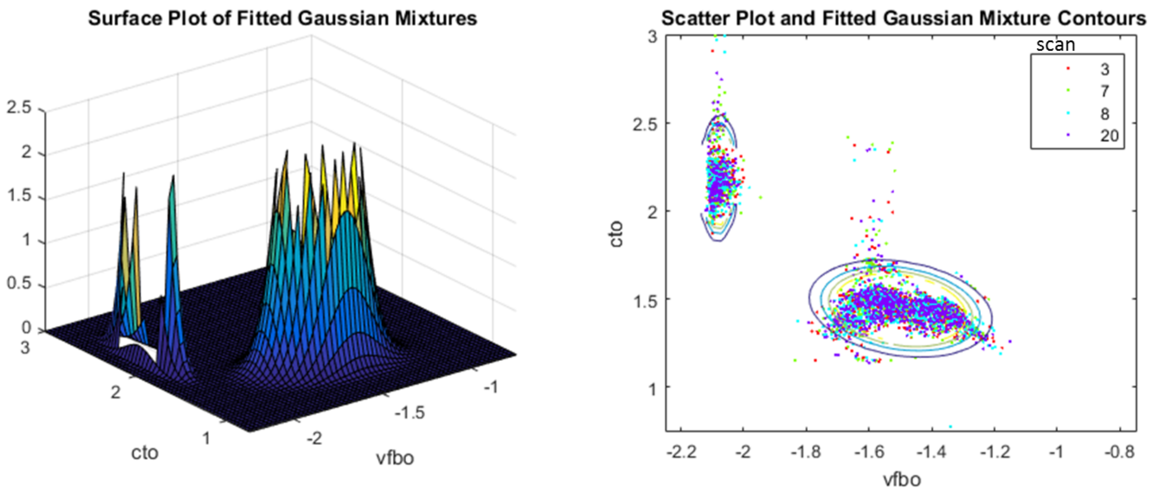


**Figure 5-10 Surface plot, scatter plot and fitted GMM contours with 2 mixture components on extracted model parameter pair *{vfbo, cto}* for visualization;**

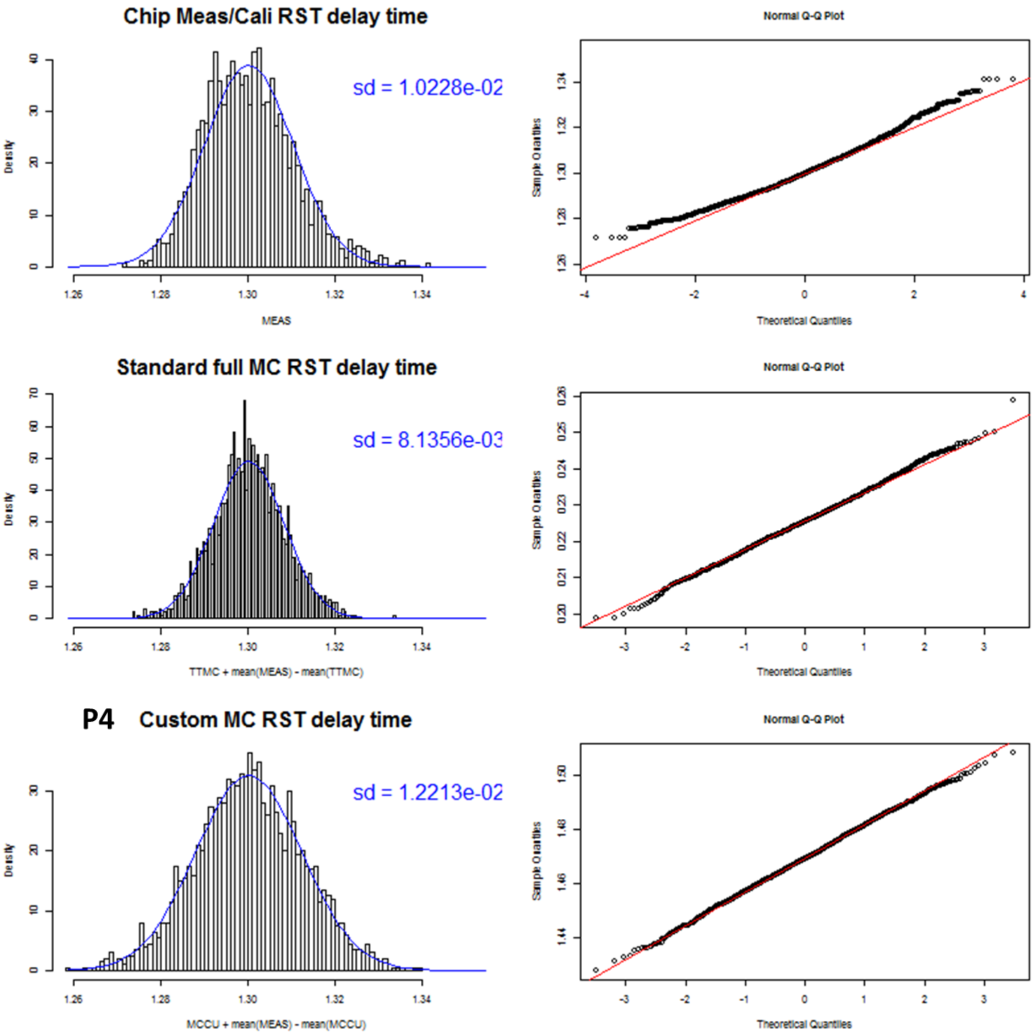
## 5.4.2 Customized MC Simulation Results

We pick the critical path delay of the RST block as our target performance metric, because measurements of this circuit performance metric can be obtained simultaneously with device characterization data (during the calibration phase) to reduce systematic error. The data presented here are collected from chip3, fabricated in 28*nm* bulk technology, with DUT cells across 96 rows and 18 columns.

Figure 5-11 shows the comparison results of measurements vs. simulations. QQ-plots are added for complete comparison and clear visualization of the distribution. The “full standard” Monte Carlo simulations are based on the full compact model embedded in the default design kit. The results show that customized MC simulations using the compact variability model with only parameters that were selected from Figure 5-8, match the full MC simulation results in distribution estimation.

The 4-parameter customized MC simulation results present very similar histograms and standard deviations compared to the measurement data, though missing the non-Gaussian tail. Furthermore, as shown in Figure 5-11, the optimal statistical model parameter set for simulation is the same as for extraction, which is {*toxo, vfbo, uo, cto*} based on the nominal PSP model. Redundant compact model statistical parameters (results with P6 and P8) slow down the simulation without improving the overall accuracy in circuit performance distribution estimation. This is a strong validation of our proposed model characterization methodology.

In order to capture the final missing components in the distribution estimation, GMMs are introduced to capture the non-Gaussian tail of the measurement data. As shown in Figure 5-12, standard full MC results tend to give a strict normal distribution (linear QQ plot) due to its intrinsic parameter sampling algorithm. Our GMM-incorporated customized MC simulation results, though have different mixture components, give a pretty good prediction of the overall statistical distribution of the target circuit performance.



**Comparison of measurements vs. full standard MC simulations vs. customized MC simulations using extracted model parameters**

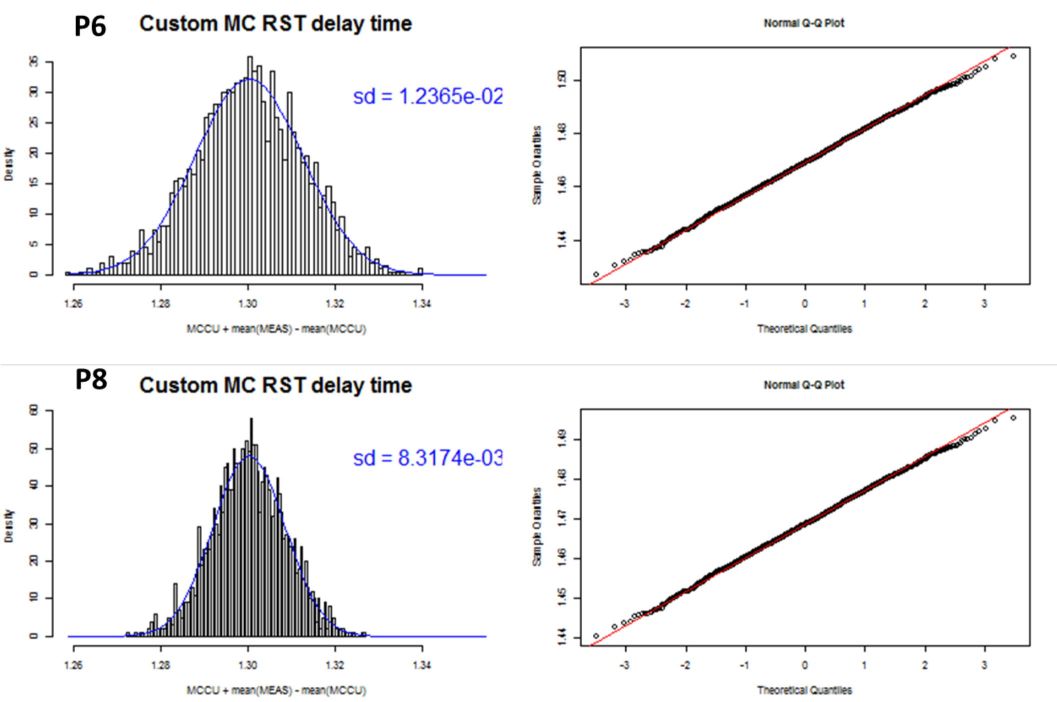


**Figure 5-11 RST critical path delay metric**
**Comparison of measurements vs. full standard MC simulations vs.**
**customized MC simulations using different reduced set of extracted model parameters.**

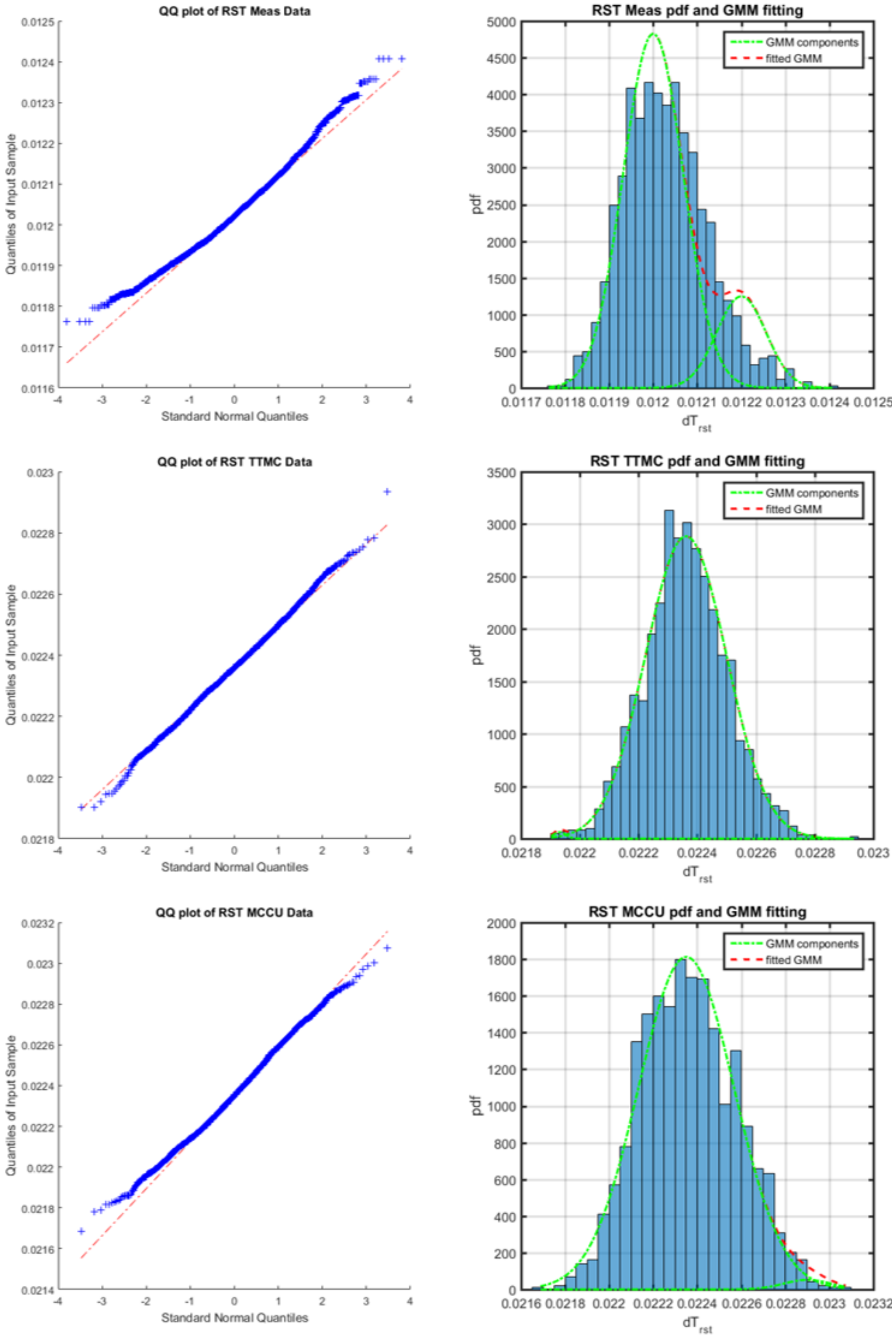


**Figure 5-12 Comparison of measurement, standard full MC and customized MC with GMM models.**

## 5.5 Summary

Our proposed statistical compact model characterization methodology has been implemented on a PSP model with 28nm imaging ROIC based transistor array test chip measurement data. We built a CAD flow with a customized MC simulation platform to utilize the SBPV models generated from the extracted compact model parameters for statistical circuit performance estimation. Simulation results on RST block critical path delay performances are very close to measurements in statistical distributions. Furthermore, we exploit the statistical structure of the extracted parameters to capture the nonlinear correlations and the non-Gaussian distributions through a mixture of Gaussian distributions [5-11]. We demonstrate that our simplified model, when incorporated into our customized MC simulations, can capture non-normality in the measured data.

## Reference

[5-1] M. Kelly et al., *"Design and testing of an all-digital readout integrated circuit for infrared focal plane arrays,"* SPIE Focal Plane Arrays for Space Telescopes II, vol. 5902, 2005.

[5-2] M. Kelly et al., *"Digital focal-plane arrays,"* MIT Lincoln Labs Tech Notes, 2014.

[5-3] S. Realov and K. Shepard, *"On-chip combined C-V / I-V characterization system in 45-nm CMOS technology,"* IEEE Journal of Solid-State Circuits, vol. 48, 2013.

[5-4] S. Saxena, *"Variation in transistor performance and leakage in nanometer-scale technologies,"* IEEE Transactions on Electron Devices, vol. 55, 2008.

[5-5] C. M. Mezzomo et al., *"Characterization and modeling of transistor variability in advanced CMOS technologies,"* IEEE Transactions on Electron Devices, vol. 58, 2011.

[5-6] A. Whitcombe, *"Variability Characterization of Imaging Readout Integrated Circuits in Deeply Scaled CMOS"*, M.S. report, UC Berkeley, 2016.

[5-7] M. G. Brown et al., *"Digital-pixel focal plane array development,"* Quantum Sensing and Nanophotonic Devices VII, 2010.

[5-8] P. Martin-Gonthier and P. Magnan, *"Novel readout circuit architecture for CMOS image sensors minimizing RTS noise,"* IEEE Electron Device Letters, vol. 32, 2011.

[5-9] A. Whitcombe, S. Taylor, M. Denham, V. Milovanović, and B. Nikolic, *"On-chip I-V variability and random telegraph noise*

*characterization in 28 nm CMOS,"* in European Solid-State Device Research Conference (ESSDERC), 2016, pp. 248-251.

[5-10] K. Qian *"Variability Modeling and Statistical Parameter Extraction for CMOS Devices"*, PhD Dissertation, UC Berkeley 2015.

[5-11] X. Yu et al., *"Accurate chip leakage prediction: Challenges and solutions,"* Symposium on VLSI Technology, 2012.

# Chapter 6

# Conclusion

## 6.1 Summary and Thesis contribution

Variability modeling and extraction in advanced process technologies is a key challenge to ensure robust circuit performance as well as high manufacturing yield. In this thesis, we present an efficient framework for device and circuit variability modeling and extraction by combining a variability-aware compact transistor model characterized using customized test structure designs, and customized Monte Carlson simulation method for utilizing these models.

We start with a test chip design and measurement of device array including different combinations of gate length/width dimensions and different layout proximities, which allow for the collection of data that can be used to analyze random and systematic variability. We take care of the limited I/O pads by designing selection circuitry with a decoder or multiplexer to access a device in the array. We examine the leakage path and minimize its impact on the measured device characteristics. The variation sources identified in this experiment will help build variability-aware compact models of transistors with physically meaningful parameters sensitive to variability observed in the manufacturing process.

A critical problem in design for manufacturability (DFM) is to build statistically valid prediction models of circuit performance based on a small number of measurements taken from on-chip test structures. Towards this goal, we propose a statistical compact modeling methodology based on stepwise parameter selection and linear spatial BPV for fast statistical compact model characterization. The methodology has been implemented on a PSP model with 28nm FDSOI SRAM cell bit transistor data, as well as 28nm bulk imaging ROIC transistor I-V data (which includes chip spatial info and linear SBPV part of the methodology is validated on this dataset).

Finally, the key contribution of this thesis work is that we built a CAD flow with a customized MC simulation setup to utilize these variability-aware compact transistor models for statistical circuit performance estimation. Simulation results on SRAM writability performances are very close to on-chip measurements in statistical distributions. In the imaging ROIC data, we further exploit the extracted compact model parameters through a mixture of Gaussian modeling, and demonstrate that significant non-normality in the measured data can be captured by our customized MC simulation platform using these simplified statistical compact transistor models.

## 6.2 Suggestions for Future Work

### 6.2.1 Process Variation Ahead

The semiconductor industry has seen a remarkable progression in integrated circuit technology over the years, which helped to usher in the personal computing era, mobile and cloud based computing, and the emerging market of Internet of Things (IoT). Underlying this success is the unwavering determination and resiliency to keep up with Moore's Law, despite all the technical challenges. Unfortunately, all exponential growth trends eventually come to an end, and Moore's Law is no exception. There are signs that the end of Moore's Law may not be far out. Market analysis shows that the cost of manufacturing a transistor is no longer decreasing, starting from the 28*nm* technology node as shown in [6-1]. Since the motivation behind Moore's Law is to reduce cost, the recent trend for transistor manufacturing costs is of major concern. Chip design costs have also risen rapidly with each new technology node. In fact, an apparent slowdown can be observed when looking at the volume production at a given technology node vs.

the year it is introduced. Starting from the 32*nm* technology node, the introduction of new generation of chips is taking longer than before, shown in [6-2].

### 6.2.2 Variation Prediction within a Process Development Cycle

For the statistical compact model characterization methodology described in Chapter 3, we made an important assumption that the nominal physical core part of the compact transistor model is fixed given a detailed technology. However, this assumption is only valid for a mature technology where key process steps are stable. In practical product development, winning in the marketplace requires system development teams to bring a better product to the market ahead of the competition and to continuously improve the yield of that product. In addition, to continue design success and make an impact on leading products, advanced circuit design exploration must begin in parallel with early silicon development.

For example, Intel has adopted a development cycle model named “Tick-Tock", where a new line of processors is released shortly after a shrinking of the process technology [6-3]. However, simulations with an early version of the design kit may have large differences with realistic or later manufacturing output of the technology. For each process development cycle, the specification for each technology is typically tighter than that which can be achieved early in its life cycle. By the time the designs enters the fab in volume, the technology would have been refined to the point that it can achieve tighter tolerances that typically at the beginning of the technology introduction.

It would be interesting and important to dynamically predict parameter variations for a later targeted release date, with only early stage process information for the targeted technology, and with historical information on how past technologies have evolved for a complete process development cycle. Extrapolations of the shrinking covariance matrices for model parameters over time as the process matures can be used to predict yield improvement trends or expectations. A methodology could be introduced using the Bayesian framework

[6-4][6-5], with sophisticated prior selection and belief propagation to learn the evolution of process tolerances over the lifetime of a manufacturing technology.

## References

[6-1] Broadcom, *Analyst day 2013*, Dec. 2013.

[6-2] C. G. Dieseldorff, C. Tseng, *"Technology Node Transitions Slowing Below 32nm,"* Industry Research & Statistics, SEMI.org, June 2014.

[6-3] "Intel tick-tock model." [Online]. Available: http://www.intel.com/content/www/us/en/silicon-innovations/intel-tick-tock-model-general.html

[6-4] C. Currin, T. Mitchell, M. Morris, and D. Ylvisaker, *"Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments,"* Journal of the American Statistical Association, vol. 86, no. 416, pp. 953–963, 1991.

[6-5] L. Yu, S. Saxena, C. Hess, I. Elfadel, D. Antoniadis, and D. Boning, *"Remembrance of transistors past: Compact model parameter extraction using bayesian inference and incomplete new measurements,"* in Design Automation and Conference (DAC), 2014.