Efficient IC Statistical Modeling and Extraction Using a Bayesian Inference Framework

by

Li Yu

B.S., Tsinghua University (2009)
M.S., Massachusetts Institute of Technology (2011)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2015

Author . . . .
Department of Electrical Engineering and Computer Science
Feb. 19, 2015

Certified by . . . .
Duane S. Boning
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Certified by . . . .
Dimitri A. Antoniadis
Professor of Electrical Engineering
Thesis Supervisor

Accepted by . . . .
Leslie A. Kolodziejski
Chairman, Department Committee on Graduate Theses

Efficient IC Statistical Modeling and Extraction Using a Bayesian Inference Framework

by

Li Yu

Submitted to the Department of Electrical Engineering and Computer Science
on Feb. 19, 2015, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

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 an ultra-compact transistor model, called the MIT virtual source (MVS) model, and a Bayesian extraction method. Based on statistical formulations extended from the MVS model, we propose algorithms for three applications that greatly reduce time and cost required for measurement of on-chip test structures and characterization of library cells.

We start with a novel DC and transient parameter extraction methodology for the MVS model and achieve a quantitative match with industry standard models for output characteristics of MOS transistor devices. We develop a physically based statistical MVS model extension and a corresponding statistical extraction technique based on the backward propagation of variance (BPV). The resulting statistical MVS model is validated using Monte Carlo simulations, and the statistical distributions of several figures of merit for logic and memory cells are compared with those of a 40-nm CMOS industrial design kit.

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 a mixture of on-chip test structures. Towards this goal, we propose a technique named *physical subspace projection* to transfer a mixture of measurements into a unique probability space spanned by MVS parameters. We search over MVS parameter combinations to find those with the maximum probability by extending the expectation-maximization (EM) algorithm and iteratively solve the maximum a posteriori (MAP) estimation problem. Finally, we develop a process shift calibration technique to estimate circuit performance by combining SPICE simulation and very few new measurements.

We further develop a parameter extraction algorithm to accurately extract all current-voltage $(I-V)$ parameters given limited and incomplete $I-V$ measurements, applicable to early technology evaluation and statistical parameter extraction.

An important step in this method is the use of MAP estimation where past measurements of transistors from various technologies are used to learn a prior distribution and its uncertainty matrix for the parameters of the target technology. We then utilize Bayesian inference to facilitate extraction and posterior estimates for the target technologies using a very small set of additional measurements.

Finally, we develop a novel flow to enable computationally efficient statistical characterization of delay and slew in standard cell libraries. We first propose a novel ultra-compact, analytical model for gate timing characterization. Next, instead of exploiting the sparsity of the regression coefficients of the process space with a reduced process sample size, we exploit correlations between different cell variables (design and input conditions) by a Bayesian learning algorithm to estimate the parameters of the aforementioned timing model using past library characterizations along with a very small set of additional simulations.

Thesis Supervisor: Duane S. Boning
Title: Professor of Electrical Engineering and Computer Science

Thesis Supervisor: Dimitri A. Antoniadis
Title: Professor of Electrical Engineering

Acknowledgments

I can only thank MIT for giving me the opportunity of attending such a prestigious institution, and meeting with so many fantastic and impressive people.

First and foremost, I would like to thank my research advisors, Professor Duane Boning and Professor Dimitri Antoniadis, for giving me the opportunity to work with two outstanding experts in the world that can be found nowhere else. I am really grateful to Duane, who is a great teacher for me. Without your unwavering support and clear guidance, my stay at MIT would not have been as exciting and fruitful. Your insight and dedication to excellence have always been inspirations to me throughout my Ph.D. research. I would like to express my gratitude to Dimitri, who sets the standard for how to be a great researcher. I really enjoyed our research discussions and I learned from you to always try to seek insight and intuition behind a problem.

I also owe my gratitude to my thesis committee members, Professor Luca Daniel and Professor Ibrahim (Abe) Elfadel, for reading my thesis and providing support throughout my Ph.D. work. I have greatly benefited from their expertise and suggestions each time I have interacted with them. Abe, thanks so much for all the collaboration and the shining suggestions for my work.

I would like to thank many of the research staff members at PDF Solutions. The internship opportunity extended to me by Dr. Christopher Hess, Dr. Sharad Saxena and Dr. Nigel Drego turned into an ongoing collaboration which eventually resulted in the major part of my thesis work. Christopher and Nigel, you were both very helpful and I thank you for your support and guidance. Sharad, thanks for many helpful discussions and suggestion. I also thank Dr. Larg Weiland, Mr. Daniel Firu, and Mr. Mehul Jain of PDF Solutions for their technical support.

I very much appreciate the collaboration and friendship from all my group mates at the Statistical Metrology group at MIT, present and past: Wei Fan, John Haeseon Lee, Joy Johnson, Hyun Ho Boo, Albert Chang and Karthik Balakrishnan, thank you for all your noble help. John, thanks for all the collaboration during class and all the good times. Wei, John, Albert and Joy, 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. There are many other friends who must be acknowledged for their friendship over the past four years, including Maokai Lin, Yuanyuan Cui, Tao Tong, Rong Yuan, Hongkai Dai, Zheng Zhang as well as Yu Xin. In addition, I would like to thank Debb Hodges-Pabon for making everyone's MTL experience so much better.

Last but not least, my family deserves all the credit for this accomplishment, as they have supported me every step of the way with complete selflessness. Thank you for your unconditional love. Lina and Yong, you are the best mom and dad one can ever wish for. You've taught me so much in every aspect of my life, academic or non-academic. Thank you for always being there with me to share the cheers during my success and comfort me during my down times. It is your encouragement that makes me go this far. To my wife Xiao - 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 really love you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In 1965, Gordon Moore observed that the number of transistors on a single chip doubles every 18 to 24 months [1], an observation now known as Moore's Law. This doubling of transistor density has served as the driving force for an astonishing increase in the functionality and computational capability of electronic devices from then to the present. Minimum transistor dimensions scale by a factor of 0.7 from generation to generation, which has enabled integration of more transistors with less power dissipation. In recent years, however, several bottlenecks have appeared as we have continued to scale down to sub $28nm$ technologies. One of the key issues related to deeply scaled semiconductor manufacturing is the *yield*, which is defined as the proportion of manufactured circuits that are functional and meet their performance requirement [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 of circuits is now due to process variations, which can be defined as the deviations of the manufactured circuit compared to the design [3].

With smaller transistors and increased transistor density, the effect of process and manufacturing variability is more significant and meeting performance and yield specifications is increasingly challenging. For example, Fig. 1-1 shows the general trend in the ratio between corresponding 3σ variation and mean value for some key

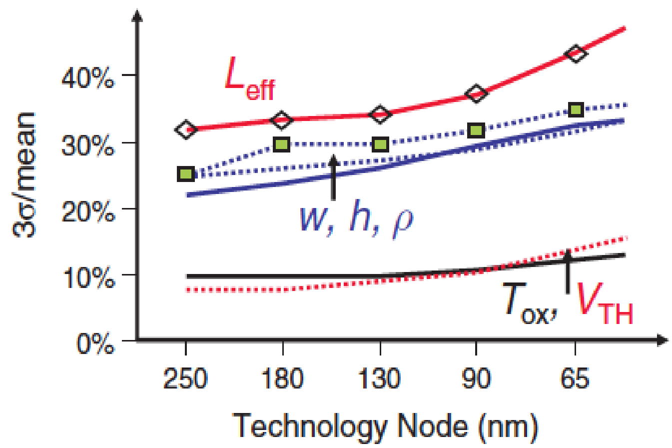


Figure 1-1: Increase in process variability for effective channel length, interconnect width and height, oxide thickness, and threshold voltage in conventional scaled MOS technology [2].

technology device and wire parameters from 250nm to 45nm. Over the time period 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 ρ also undergo a fairly major increase. Other parameters such as the threshold voltage V_{th} and oxide thickness T_{ox} increase at a lower rate.

Increasing process variations introduce significant uncertainty for both circuit performance and leakage power. It has been shown in that even for the 180nm technology, process variation can lead to 1.3X variation in frequency and 20X variation in leakage power, as illustrated in Fig. 1-2 [4]. In future technology generations, such impact will become even larger because the technology is approaching a fundamental randomness regime in the behavior of silicon structures, such as that due to random dopant fluctuations [5]. 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, variability needs to be characterized empirically for a specific semiconductor process in order to obtain a quantitative understanding of variability mechanisms. Such "statistical metrology" methods include measurement techniques for characterization of variability, and statistical modeling

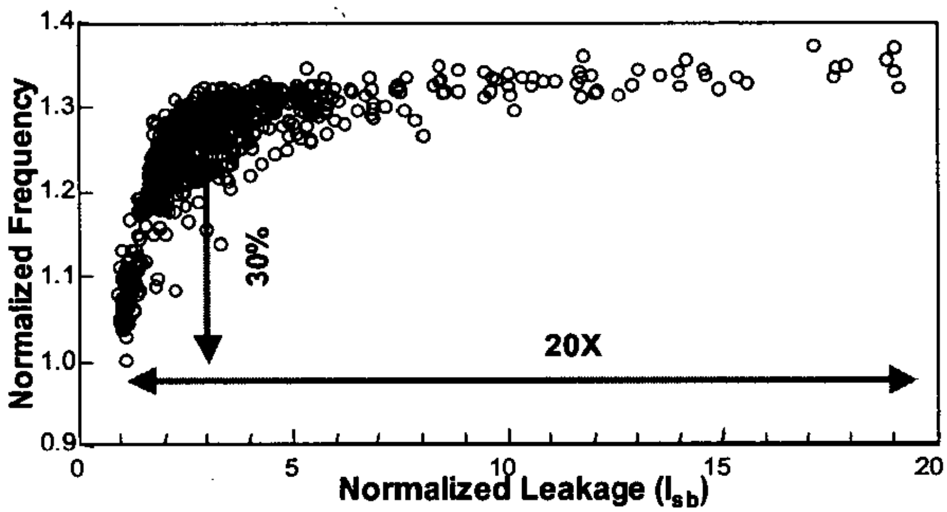


Figure 1-2: Leakage and frequency variations [4].

and extraction methods for properly interpreting measurement results.

1.1 Test Structures for Variation Characterization

In order to improve our understanding of process variation and ultimately reduce that 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. According to the objective of measurements, test structures fall into two classes; (a) test structures for process control, and (b) test structures for modeling [2].

Test structures for process control are used for monitoring and controling the fabrication line. They are typically small devices or circuits placed in the scribe line on all wafers and therefore are capable of modeling the history of the line. Monitors often consist of simple test structures that allow the measurements of current-voltage $(I - V)$ characteristics of MOSFETs [6, 7, 8], of the resistivity of wires and vias [9], and of interconnect capacitance [10].

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. Recent designs enable many individual transistor characteristics to be studied efficiently without the use of dedicated per-device probe pads [11, 12, 13, 14, 15, 16, 17, 18, 19]. Ref. [20] characterizes optical proximity behavior by measuring L_{gate} variability. Refs. [13, 21] present large transistor arrays dedicated to measure threshold voltage variation of each individual transistor, through measurements of gate-to-source voltage variation and leakage current. Ref. [19] describes a method to measure the contact resistance of individual contacts.

Another important category of commonly deployed test structure is ring oscillator (RO) structures that consist of an odd number of inverting stages [22, 23, 24, 25]. Compared to the single transistor test structures, ring oscillators reflect circuit operation under high-speed conditions as in an actual digital system, and thus are more strongly related to the performance of actual products [24]. Frequency and leakage measurements can be gathered from RO test structures, in which the frequency measurement can be easily measured with a low-cost frequency counter. A benefit of such RO based test structures is that they can be made to be sensitive to either device and interconnect variability.

1.2 Statistical Analysis of Measurement Data

After obtaining measurement results from the test structures described in Section 1.1, the next step is to apply statistical analysis techniques to interpret these measurement data. Statistical circuit analysis requires both a characterization of variability in device behavior, and the ability to translate device variability to circuit performance variability.

1.2.1 MOSFET Device Models and Extraction

Interconnect and device models are the critical interface between the underlying technology and integrated circuit design. Component models compute the current through every linear and non-linear interconnect and device component, and the derivatives of that current with respect to the terminal voltages of the device [2]. *Compact models* include key equations that describe the current and charge of a device as a function of its terminal voltages, to enable circuit simulation. MOSFET device models have a long and rich history spanning over 30 years. To meet accuracy requirements on device models, existing BSIM [26], PSP [27] and PTM [28] models are being constantly augmented to account for the emerging physical phenomena at the nanometer regime. An example of the evolution of parametric complexity of MOSFET models in industrial design kits is illustrated by the BSIM industry models. For the $0.5\mu m$ technology of the early nineties, this model had 99 parameters, 7% of which were physical [26]. Here physical quantities directly describe the physical attributes of the system [29]. In the deep sub-micron era, the BSIM4 generation of this model has 355 parameters (at the $65nm$ node), 2.5% of which are physical. The PSP compact model has similar parametric complexity [27].

The increasing number of parameters and complexity of equations of compact transistor models drive the need to accurately determine all the many dozens of model parameters in order to reproduce the behavior of a specific observed device, which is referred to as the *parameter extraction* problem [30]. The most widely used parameter extraction methods are based on the deterministic minimization of a nonlinear least square error function between model output and measurement data:

$$\mathcal{E}(\mathbf{P}) = \arg\min_{\mathbf{P}} \|\mathbf{F} - f(\mathbf{V}, \mathbf{P})\|^2 \tag{1.1}$$

where $\mathbf{F}$ is a vector of observed currents, $\mathbf{V}$ is a matrix of applied voltages at which $\mathbf{F}$ are observed, $\mathbf{P}$ is a vector of device model parameters to be extracted, and the function $f()$ represents the device model equations.

However, the minimization of Equ. (1.1) suffers from a uniqueness solution problem. The situation becomes even worse in the presence of one or more of the following conditions: (1) the dimension of the $\mathbf{P}$ vector is large; (2) the empirical or semi-empirical models unable to limit the parameter values to a specific domain; (3) the function f is substantially non-linear; and (4) there is physically collinearity among the parameters **P**. With many potential local minimum that might trap traditional minimization schedmes, it is often desirable to find the *global* minimum, in order to provide good input-output fidelity (accuracy) or confidence on the correctness of the model parameters (system identification and physical modeling). While there exist a number of optimization methods which are more likely to find the global minimum (e.g., Genetic Algorithms, or GA), nearly all optimization methods are iterative, and defining an appropriate starting point and parameter bounds is of crucial importance, and often rely on a deep understanding of the model to guide the minimization process.

1.2.2 Statistical Device Characterization

While a single device model can be characterized from observed $I-V$ measurements, a *statistical* device model is needed to express manufacturing variability [31]. Process variations usually manifest themselves as parameter fluctuations in nanoscale transistors physical dimensions or material/electronic properties, such as in the channel length, threshold voltage, and transistor parasitics [32]. The previous section noted that several problems can occur with nominal device characterization, such as local minima and unrealistic parameter values. These problems become more serious for statistical device characterization, since statistical extraction procedures rely on an accurate extraction of nominal parameters.

After characterizing appropriate test structures described in Section 1.1, 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*. Key statistical conclusions drawn from test structures measurements include magnitude of variation and average or nominal value of parameters. 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).

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 [33, 34]. 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 $\{F_i\}$. With variations σ_{F_i} $(i = 1, 2, ..., m)$ of electrical performance parameters (e.g., I_{dsat}, I_{off}, etc.) measured under different geometry and bias conditions, the BPV method calculates σ_{p_j} through [33]:

$$\sigma_{F_i}^2 = \sum_{j=1}^{n} (\frac{\partial F_i}{\partial p_j})^2 \sigma_{p_j}^2 \tag{1.2}$$

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. For example, variation in threshold voltage $\mathbf{V_{th}}$ includes both effects of random dopant fluctuation (RDF) and line-edge roughness (LER). However, the BPV method assumes all parameters $\{p_j\}$ to be *uncorrelated.*

Therefore, in BPV and many other 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) that generate an orthogonal basis set of model parameters. As a physical decoupling example, the aforementioned two variation sources for V_{th} fluctuation could be separated into two items:

$$\Delta V_{th} = \Delta V_{th0} + \Delta V_{th}(L_{eff}) \tag{1.3}$$

where V_{th0} represents threshold voltage of the long channel device which is only related to RDF, and the later item corresponds to drain-induced barrier lowering (DIBL) effect which is only related to LER. However, to account for and separate these correlation between model parameters and reach accuracy requirements on device models for deeply scaled devices, the parametric complexity of the underlying device model inevitably increases.

PCA is a commonly used statistical technique that transforms correlated measurements into a set of low-dimensional, uncorrelated factors [35]. Given N samples from a set of correlated random device variables $\mathbf{P}$ or correlated electrical measurements $\mathbf{F}$, PCA seeks a linear transformation of these variables into a new set of random variables $\mathbf{X}$ which are orthogonal. The procedure starts by forming the correlation matrix $\mathbf{M}$ amongst the output or measured samples. An eigenvalue decomposition of the correlation matrix $\mathbf{M}$ is then performed and combinations of eigenvalues and corresponding eigenvectors are obtained [36]. If we use the top eigenvalue/vector combinations, we can explain most of the overall observed variation among the output samples with just a few uncorrelated variables. In addition to orthogonalization, a major benefit of PCA is that the complexity of subsequent extraction and analysis techniques is considerably reduced by reducing the number of variables.

1.2.3 Statistical Circuit Modeling

After statistical device characterization, the next step to address the parametric yield problem is to effciently model circuit performance of circuit blocks (e.g., a standard library cell, 64-bit full adder, etc.) under process variation. There are generally two broad categories for such models: *behavior modeling* and $performance\ modeling$.

Model order reduction (MOR) is a systematic approach to create behavior models with reduced computational complexity. It takes high-order algebraic differential equations (e.g., modified-nodal-analysis equations for circuit simulation) as the input and creates a simplified (i.e., low-order) dynamic system to approximate the original input-output behavior. Then the extracted behavior models are utilized in a hierarchical simulation flow to reduce the simulation cost. For example, reduced-order interconnect models can be used to speed up the gate-interconnect cosimulation [37].

Different from the early stage MOR approaches that focus on a fixed dynamic system, parameterized model order reduction (PMOR) techniques can generate compact models reflecting the impact induced by design or process variations [38, 39, 37, 40, 41, 42]. Several PMOR techniques have been developed for both linear and nonlinear circuits [37, 41, 40, 42], and most of them are based on moment matching techniques which assume that the closed forms of the parameterized state-space models are given, or that the parameter statistical distributions are known.

Performance modeling is a mathematical approach that approximates the performance of interest (e.g., delay, leakage power, etc.) as a function of the parameters of interest (e.g., V_{th}, T_{OX}, L_{eff}, W, etc.) or uncorrelated variables after PCA [43]. The form of the function can be empirical (e.g., response surface modeling (RSM), usually with some restricted polynomial form), or physical (e.g, physically derived compact models or expressions, with some fitting coefficients and parameters, often where the functional form is non-linear). For RSM, a linear least square error function similar to (1.1) can be employed to optimize the results:

$$\mathcal{E}(\mathbf{P}) = \arg\min_{\mathbf{P}} \|\mathbf{F} - \alpha \cdot \mathbf{P}\|^2 \tag{1.4}$$

where $\mathbf{F}$ is a vector of observed performances, α is a vector of the unknown model coefficients, $\mathbf{P}$ is a vector of basis functions (e.g., linear or quadratic polynomials of principal components).

While the simplest performance modeling is based on linear approximation, it offers the least accurate predictions for capturing large-scale process variations. To achieve better accuracy, a quadratic approximation can be used, which, however, significantly increases the modeling cost. For example, even if we select the top 10 ranked random variables after PCA, a quadratic RSM model will need to fit 100 quadratic coefficients, which requires a large number of training samples and has a high fitting cost. When the measurement data set is not large enough to support the variable space, parameter estimates by RSM become non-unique; this phenomenon is known as over-fitting [44]. However, it is can be difficult or expensive to collect sufficient on-chip measurements to support full RSM approaches. Instead, we are often limited to a very small number of measurements, as post-Silicon characterization suffers from area and test time limitations.

One common strategy for preventing over-fitting in performance modeling is by adding regularization terms to error functions in an effort to reduce the number of significant or retained model parameters. An example of such a strategy is least-angle regression (LAR) which adds L_1-norm (the summation of the absolute values of all elements in the parameter coefficient vector) regularization $\|\alpha\|_1 \leq \lambda$ to (1.4) [45, 46]. One major benefit of regularizing with the L_1-norm is that it results in sample complexity logarithmic in the number of variables to be extracted. On the other hand, an L_2 regularization results in sample complexity that is linear in the number of features. A similar method, sparse regression, adds an L_0-norm (the total number of non-zeros in the parameter coefficient vector in the vector) constraint $\|\alpha\|_0 \leq \lambda$ instead of the L_1-norm term to find a unique solution α.

The connection between L_1-norm regularization and L_0-norm regularization is that by decreasing λ, we can impose a strong constraint for sparsity and achieve a sparse solution making use of fewer model coefficients or parameters. However, the selection of λ is of crucial importance for such optimizations and requires considerable experience.

Recent work has employed Bayesian inference and maximum posterior estimation (MAP) to address the over-fitting problem, where sparse model coefficients and correlated performance variability are exploited [47, 48]. According to Bayes' theory [49], the joint posterior distribution $pdf(\mathbf{F}, \boldsymbol{\mu}_{\mathbf{X}})$ is given by the product of the prior $pdf(\boldsymbol{\mu}_{\mathbf{X}})$ and the likelihood function $pdf(\mathbf{F}|\boldsymbol{\mu}_{\mathbf{X}})$, giving us the *posterior distribution*:

$$pdf(\mathbf{F}, \boldsymbol{\mu}_{\mathbf{X}}) = pdf(\boldsymbol{\mu}_{\mathbf{X}}) \cdot pdf(\mathbf{F}|\boldsymbol{\mu}_{\mathbf{X}}) \tag{1.5}$$

Bayesian model fusion (BMF) is an efficient algorithm exploiting "similarity" between model coefficients in different stages (e.g., post-layout performance model and schematic-level performance model). By "fusing" the schematic-level performance model with the post-layout performance model through a common template, the computational cost for post-layout performance modeling can be substantially reduced [50]. The limitation of such a method is that it requires the reuse of data collected from the same or very similar system, and modeling is especially challenging to combine measurements from a mixture of different systems. Virtual probe (VP) is another algorithm recently proposed, to accurately predict spatial variation across a wafer having a few test structures at pre-selected locations, by exploiting the underlying sparsity in the spatial frequency domain [51].

Bayesian inference and estimation approaches are attractive in several respects. First, they offer the possibility of using (or re-using) prior knowledge and experience. Second, they are well-suited to deal with cases where only a limited number of new observations are available. Finally, they support combination of multiple types of observations in a common framework. A core contribution of this thesis is to explore the use of a Bayesian framework for IC statistical extraction and modeling, taking advantage of these features.

1.3 Thesis Organization

In this thesis, we propose accurate and efficient statistical techniques to solve the following problem in DFM: given the measurements of one or several functions (e.g., transistor $I-V$ measurements, RO frequency measurements, etc.), we need to find the value of model parameters (such as parameters sensitive to process variation), to predict system performance, and eventually to improve product yield, with the complication that there are many possible mappings from the function space to the parameter space.

The overall structure of the thesis is shown in Fig. 1-3. We begin by presenting key elements of an ultra-compact MIT virtual source (MVS) model as well as its physical intuition, parameter extraction and statistical extensions in Chapter 2. Subsequent chapters leverage the MVS model, and focus on solving the aforementioned industrial practical problem by proposing a series of algorithms which facilitates semiconductor manufacturing yield control.

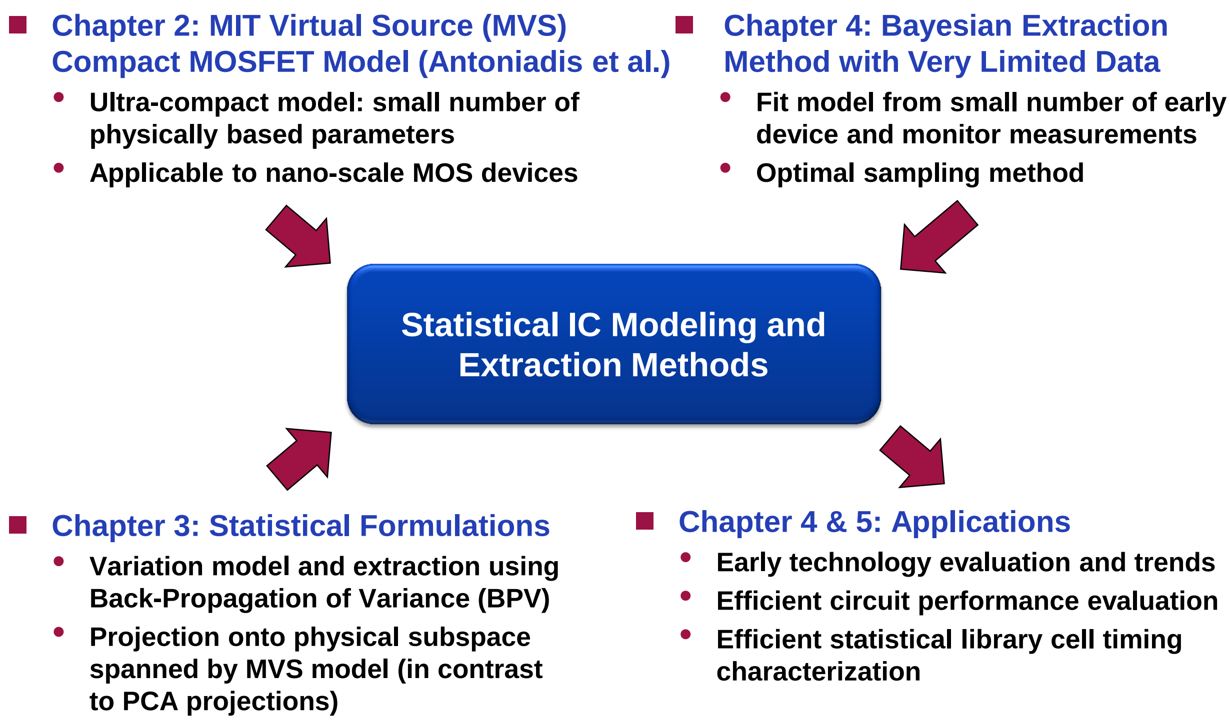


Figure 1-3: Overview of the thesis organization.

In Chapter 3, we propose an efficient method to build statistically valid prediction models of circuit performance based on a small number of mixture measurements (e.g., transistor $I-V$ measurements, Ring Oscillator (RO) frequency measurements). Different from the traditional approach which combines principal component analysis (PCA) and response surface modeling (RSM) to approximate the circuit performance, we propose *physical subspace projection* to project different groups of on-chip measurements to a unique subspace likelihood map spanned by a set of MVS model parameters. Then we introduce a Bayesian formalism to estimate the performance parameters by using maximum a posteriori (MAP) estimation defined over all of the group measurement distributions and utilizing the subspace variable prior distribution. Compared with the traditional PCA and RSM method, the proposed method preserves the physical correlation between MVS model variables and measurements such that the number of required measurements is greatly reduced for model establishment. The last step, process shift calibration, is an algorithm proposed to calibrate and minimize the difference between SPICE model predictions and measurements, while retaining statistical model information.

Another important problem in the IC process development cycle is to extract physically meaningful model parameters; these 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 is that measurements are collected from a limited number of early prototype devices rather than from a full suite of designed test structures. In Chapter 4, 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. The use of maximum a posteriori estimation allows two major improvements compared with traditional method. First, we learn a prior distribution from past measurements of transistors from various technologies. Second, we assign different weights for different measurements using the uncertainty matrix of the parameters of the target technology. That is to say, we can use our knowledge about the confidence in different measurements to judiciously update our model. We further extend the proposed extraction approach to enable us to characterize the statistical variations of MOSFETs, with the significant benefit that some constraints required by the backward propagation of variance (BPV) method are relaxed. Moreover, we propose an optimal $I-V$ measurement selection algorithm by finding those measurements which minimize the average uncertainty in our Bayesian framework, and we explore the the lower bound for the number of $I-V$ measurements required to fit a transistor compact model.

In Chapter 5, we extend the Bayesian framework to the standard cell level, and propose a novel flow to enable computationally efficient statistical characterization of delay and slew in standard cell libraries. The distinguishing feature of the proposed method is the usage of a limited combination of output capacitance, input slew rate and supply voltage parameters for the extraction of statistical timing metrics of an individual logic gate. The efficiency of the proposed flow stems from the introduction of a novel, ultra-compact, nonlinear, analytical timing model, having only four universal regression parameters. This novel model facilitates the use of maximum a posteriori belief propagation to learn the prior parameter distribution for the parameters of the target technology from past characterizations of library cells belonging to various other technologies, including older ones. The framework then utilizes the Bayesian inference approach developed in Chapters 3 and 4 to extract the new timing model parameters, with the benefit of only a small set of additional timing measurements required from the target technology.

Chapter 6 concludes the thesis. We summarize the key contributions of the thesis stemming from the use of Bayesian inference and learning technologies, combined with MVS and other ultra-compact models, toward the solution of IC design for manufacturability challenges. Areas for future research are suggested, including extension of MVS mode as a predictive statistical compact model and variation prediction in a process development cycle using Bayesian inference.

Chapter 2

MIT Virtual Source Model: Physical Intuition, Parameter Extraction and Statistical Modeling

As technology scales down to the nanometer range, accurate modeling and simulation of digital (and increasingly mixed signal) circuits are becoming more important, which requires understanding and management of increasing critical variations in transistor and other components and their impact on circuit performance. Existing BSIM [26], PSP [27] and PTM [28] models are being constantly augmented to account for the emerging physical phenomena at the nanometer regime. As a result, the management of transistor parameters in CAD environments or during library cell characterization for timing and power becomes a more complex undertaking. The issue becomes even more acute in the context of statistical simulations using Monte Carlo or other approaches if we have to account for parameter variations of a large number of parameters. One way to address these issues for digital design in the nanometer regime is to make use of ultra-compact transistor models specifically developed for short- and ultra-short-channel devices.

Key features of a desired ultra-compact transistor model include a simple parameter extraction procedure through measurements from simple on-chip test structures for monitoring the fabrication line, rather than requirings large number of measurements from a rich variety of complicated test structures for device modeling; excellent accuracy of current-voltage ($I - V$) and capacitance-voltage ($C - V$) characteristics in both the device operation domain as well as enabling excellent digital timing and power analysis in the circuit operation domain; and finally, the capability of mapping the variability characterization in device behavior onto a limited number of underlying model parameters, which in turn enables the efficient prediction of variations in circuit performance. The MIT virtual source (MVS) model is one such ultra-compact model [52, 53, 54, 55, 56].

The widely adopted threshold-voltage-based compact models (such as BSIM [26] and PTM [28]) and the surface-potential-based compact models (such as PSP [27]), include as many as several hundred parameters related to the manufacturing process, the geometry of the device, and to achieve smoothing or transitions between different equation regimes. On the other hand, the MVS model restricts itself to a simple physical description for channel minority carrier charges at the virtual source by substituting the quasi-ballistic carrier transport concept for the concept of drift-diffusion with velocity-saturation. In doing so, it achieves excellent accuracy for the $I - V$ and $C - V$ characteristics of the device throughout the domain of operation required for digital timing and power analysis. The number of parameters needed is considerably fewer (19 for DC and 23 in total) than in conventional models. It is worth noting that the ultra-compact model developed in [57] is based on the alpha-power model of [58]. Therefore Ref. [57] is purely empirical and aims at maximizing the timing accuracy of an inverter. The MVS model is physics-based and achieves higher timing accuracy than [57] with a similar number of parameters.

Furthermore, continued scaling of CMOS technology has introduced increased variations of process and design parameters, which profoundly affect all aspects of circuit performance [59]. While statistical modeling addresses the need for high product yield and performance, it inevitably increases the cost of computation. This problem is further exacerbated as future digital design becomes larger and more complex. Therefore, the simplicity of MVS models is a substantial help in effective statistical design flows. Previous compact transistor models consist of a large number of parameters and complex equations which do capture many (if not all) of the physical short-channel effects, but significantly slow down the simulation speed [32]. A distinct benefit of the statistical extension of the MVS model is that it directly addresses both the complexity and simulation problems of statistical circuit analysis for nanoscale CMOS devices [60]. Indeed, it provides a simple, physics-based description of carrier transport in modern short-channel MOSFETs, along with the capability of mapping the variability characterization in device behavior onto a limited number of underlying model parameters.

In this chapter we review the MIT virtual source model, including its physical intuition, and then introduce a consistent DC and AC parameter extraction flow, and derivation of a statistical MVS model. The rest of this chapter is organized as follows. Section 2.1 reviews the physical intuition of the MVS model in both its static and dynamic versions. Section 2.2 describes the transient and DC parameter extraction methodology flow of the MVS model. This flow aims at providing a consistent, highly-calibrated parameter set based on both $I-V$ and $C-V$ curve measurements. Section 2.3 presents timing analysis over a large set of standard library cells from an industrial design kit, demonstratrating the use of the MVS model for timing verification of MVS model in digital circuits. Section 2.4 shows transient and power simulations using a $40nm$ MVS model integrated in a vendor CAD environment. The simulation results for a 1001-inverter chain and a 32-bit ripple adder under a V_{dd} sweep demonstrate that our calibrated MVS model is suitable for use in a digital design environment. Sections 2.5 and 2.6 introduce the physical derivation of the statistical MVS model, and reviews of a second-order statistical extraction technique called the backward propagation of Variance (BPV) [61]. Although the extraction is performed for the nominal V_{dd}, the resulting statistical model is valid over a whole range of V_{dd}'s, thus enabling the efficient analysis of power-delay tradeoffs in the presence of parameter validations. And finally, Section 2.7 presents several statistical validation examples.

2.1 Review of the virtual source charge-based compact model

The MIT virtual source model [55] consists of a core static or DC model with 19 parameters [52], and an extended dynamic model with 4 additional parameters [54], both are summarized below.

2.1.1 Static MVS Model

In the MVS model, the drain current of a MOSFET normalized by width (I_D/W) can be described using the following general equation:

$$I_D/W = Q_{ix_o} v_{x_o} F_s \tag{2.1}$$

valid for both the saturation and non-saturation regions.

The virtual source velocity, denoted as v_{x_o}, refers to the velocity of carriers located in the MOSFET channel at the top of the energy barrier near the source (virtual source). The core concept in the MVS model is that in short-channel devices, v_{x_o} does not depend on V_{ds} except for drain-induced barrier lowering (DIBL) effects. This is to be contrasted with a drift-diffusion transport model where the velocity is directly proportional to low electrical field E and is saturated when E is larger than the critical electrical field. The MVS model also uses the fact that although the ballistic velocity increases with V_{gs}, the virtual source velocity v_{x_o} is almost constant at high V_{gs} [62].

The MVS inversion charge density Q_{ix_o} can be approximated by the empirical function [52][63]:

$$Q_{ix_o} = C_{inv} n\phi_t \ln(1 + \exp \frac{V'_{GS} - (V_T - \alpha\phi_t F_f)}{n\phi_t}) \tag{2.2}$$

where C_{inv} is the effective gate-to-channel capacitance per unit area in strong inversion, ϕ_t is the thermal voltage ($k_B T/q$), V'_{GS} represents the internal gate-source voltage, and n is the subthreshold coefficient. F_f denotes a Fermi function that allows a smooth 0 to 1 transition, and α is introduced to adjust V_T shift for which $3.5\phi_t$ is a good approximation.

The function F_s in (2.1) serves to account for the continuous transition from non-saturation to saturation, and is given by

$$F_s = \frac{V_{ds'}/V_{dsat}}{(1 + (V_{ds'}/V_{dsat})^{\beta})^{1/\beta}} \tag{2.3}$$

where $V_{ds'}$ accounts for the intrinsic drain-to source voltage after deducting the resistive voltage drop for both the source R_s and drain R_d resistances using $V_{ds'} = V_{ds} - I_D(R_s + R_d)$. β is an empirical parameter for the transition from the low-field non-saturation region to high-field saturation region with a typical value of about 1.8 [52].

Other key static MVS model parameters include drain-induced-barrier-lowering (DIBL) coefficient δ, subthreshold swing factor n_0, low-field carrier "apparent mobility" μ, effective body factor γ, and carrier effective mass m_c with a fixed value of $0.2m_e$. Among all 19 DC parameters, seven key parameters $\{\delta, n_0, Rs_0, Rd_0, v_{x0}, mu, V_{T0}\}$ need to be extracted to achieve calibration with experimental $I - V$ data.

2.1.2 Dynamic MVS Model

The transient behavior of the MVS model is described in [54], where the intrinsic channel charge is partitioned into that of the source and drain terminals:

$$\begin{cases} Q_S = \displaystyle\int_0^{L_g} (1 - x/L_g) Q'_i(x)]dx \\ Q_D = \displaystyle\int_0^{L_g} \frac{x}{L_g} Q'_i(x)dx \end{cases} \tag{2.4}$$

The channel charge areal density, $Q'_i(x)$, is calculated according to [54], using a non-saturation (NVsat), saturation (Vsat), or quasi-ballistic (QB) model. Since the

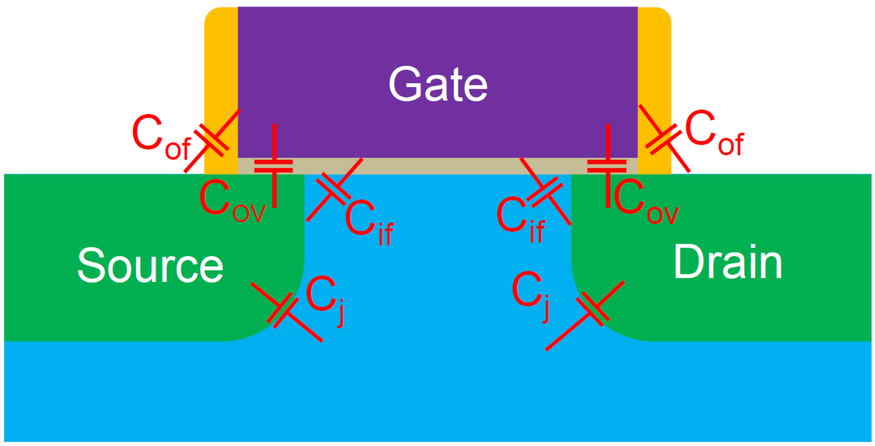


Figure 2-1: Schematic of a short-channel n-MOSFET showing the MVS model parasitic capacitances.

nominal transistor gate length in this work is $40nm$, we have used the quasi-ballistic version of the channel charge model.

Under the assumption that the source and drain are symmetric, we have four capacitances to model parasitic effects, as shown in Fig. 2-1. C_{ov} is the overlap capacitance, C_{of} is the outer-fringing capacitance, C_{if} is the inner-fringing capacitance and C_j is the junction capacitance. In this work, C_{ov} and C_{of} are considered voltage independent, while C_{if} and C_j are considered voltage dependent.

2.2 Parameter Extraction and Model Calibration

The MVS model described in (2.1) through (2.4) is calibrated using data for transistors with different sizes. To extract all of the parameters of the MVS model, a full set of $I-V$ and $C-V$ measurement data is needed. The parameter extraction flow is described in Section 2.2.1, followed by validation through analysis of $I-V$ curves in Section 2.2.2.

2.2.1 Parameter Extraction Flow

While previous work has presented the key parameters extracted for the MVS model [52, 54], a clear optimization flow to extract full $I-V$ and $C-V$ parameters has been missing. Fig. 2-2 shows a parameter extraction flow for the MVS model which proceeds as follows.

First, the effective gate-to-channel capacitance C_{inv} is extracted by subtracting the C_{gs} curves of two long channel devices at a point where the short-channel parasitic capacitances are negligible. This step needs to be done before other DC parameter extraction since C_{inv} affects the distribution of charges Q_{ix_o}. Once C_{inv} is properly extracted, the $I-V$ curve calibration is achieved under sequential flows in the sub-threshold and above-threshold regions. First, V_{th0} is adjusted to achieve consistency with respect to Q_{ix_o}. Then the sub-threshold parameter set (S, δ) and full region parameter set (v_{xo}, μ, etc.) are optimized sequentially using non-linear least-squares error minimization. The coupling between the two is achieved through iteration, until final convergence. Good convergence is achieved for transistors with various sizes in our tests. The back bias coefficient γ is extracted from the $I-V_{bs}$ measurement as the last step in DC parameters extraction procedure. We then extract the parasitic capacitances (C_{if} and C_{of}) by fitting the $C_{gg}-V_g$ curve, as shown in Fig. 2-3(a). The equations employed to extract parasitics in the MVS model are shown in (2.5).

Table 2.1 lists the key parameters of the MVS model and the parasitic capacitances obtained from the parameter extraction methodology illustrated in Fig. 2-2, for a typical $40nm$ technology. Note that the MVS model calculates the drain current normalized by width, so that the extracted parameters are applicable for all rectangular devices having the same channel length. Good consistency and accuracy are achieved in devices with various widths using the one parameter set extracted by the aforementioned parameter extraction flow. This is illustrated in Fig. 2-3(b). Since the $40nm$ channel is close to the ballistic transport regime, a quasi-ballistic (QB) model is employed to calculate the channel charge areal density [54]. We note, however, that for this technology, the non-saturation (NVsat) model achieves a similar accuracy.

$$\begin{cases} C_{gg}(V_g=0) = 2W(C_{if}+C_{of}+C_{ov}) \\ C_{gg}(V_g=V_{dd}) = W[(C_{inv}(L-L_{ov})+2C_{of}+2C_{ov})] \end{cases} \tag{2.5}$$

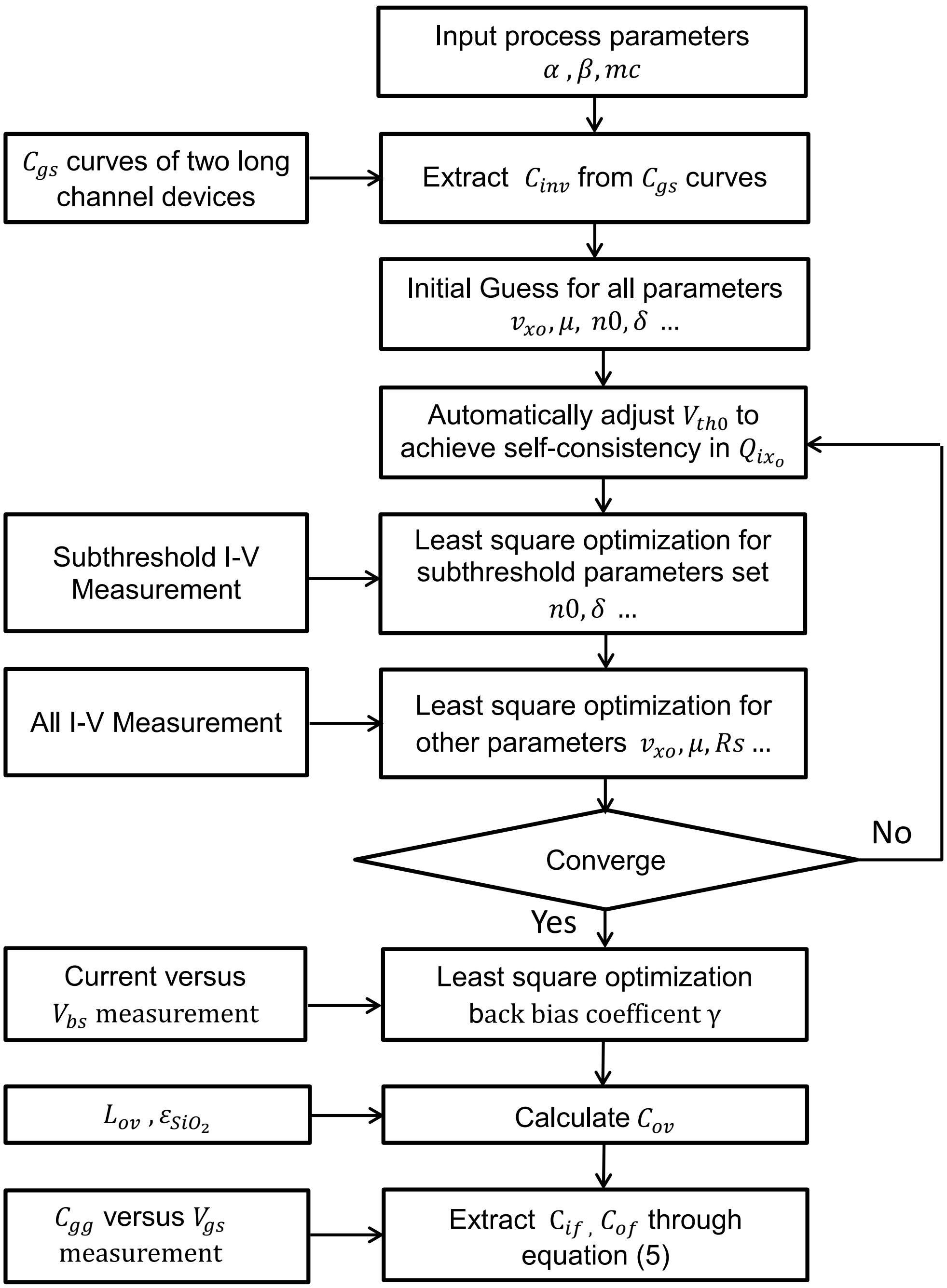


Figure 2-2: An optimization flow for $I - V$ parameter extraction in the MVS model.

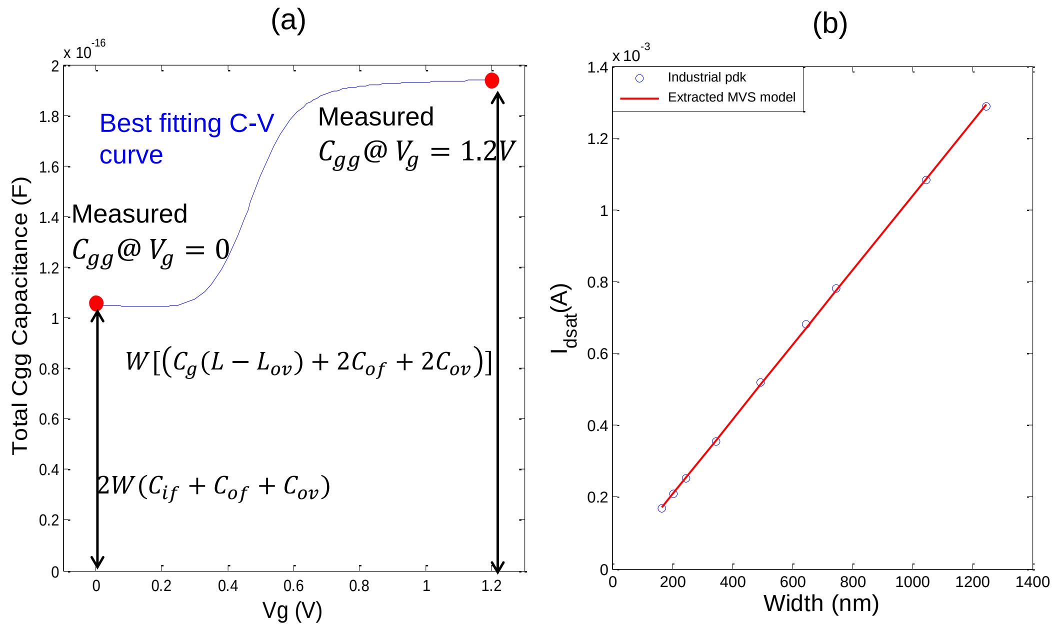


Figure 2-3: (a) Gate capacitance versus V_{gs} at $V_{ds} = 0$V with the equation used to extract parasitic capacitances. (b) I_{dsat} versus effective channel width.

Table 2.1: Key parameters for the MVS model fitted to a 40nm industrial design kit. Channel widths are 300nm and 600nm for NFET and PFET, respectively.

Parameters	NMOS	PMOS	Description
$L_g(nm)$	40	40	Channel length drawn
$L_{ov}(nm)$	8	8	Total overlap channel length on both side
$C_{inv}(\mu F/cm^2)$	1.40	1.35	Effective gate-to-channel capacitance per unit area
n_0	1.47	1.55	Subthreshold swing
$\delta(mV/V)$	93	159	Drain-induced barrier lowering
$v_{xo}(cm/s)$	1.39e7	0.855e6	Virtual source velocity
$\mu(cm^2/V \cdot s)$	248	146	Low-field mobility
$R_s(ohm - \mu m)$	60	80	Series resistance per side
γ	0.34	0.39	Body effect coefficient
m_c	$0.2m_e$	$0.2m_e$	Carrier effective mass

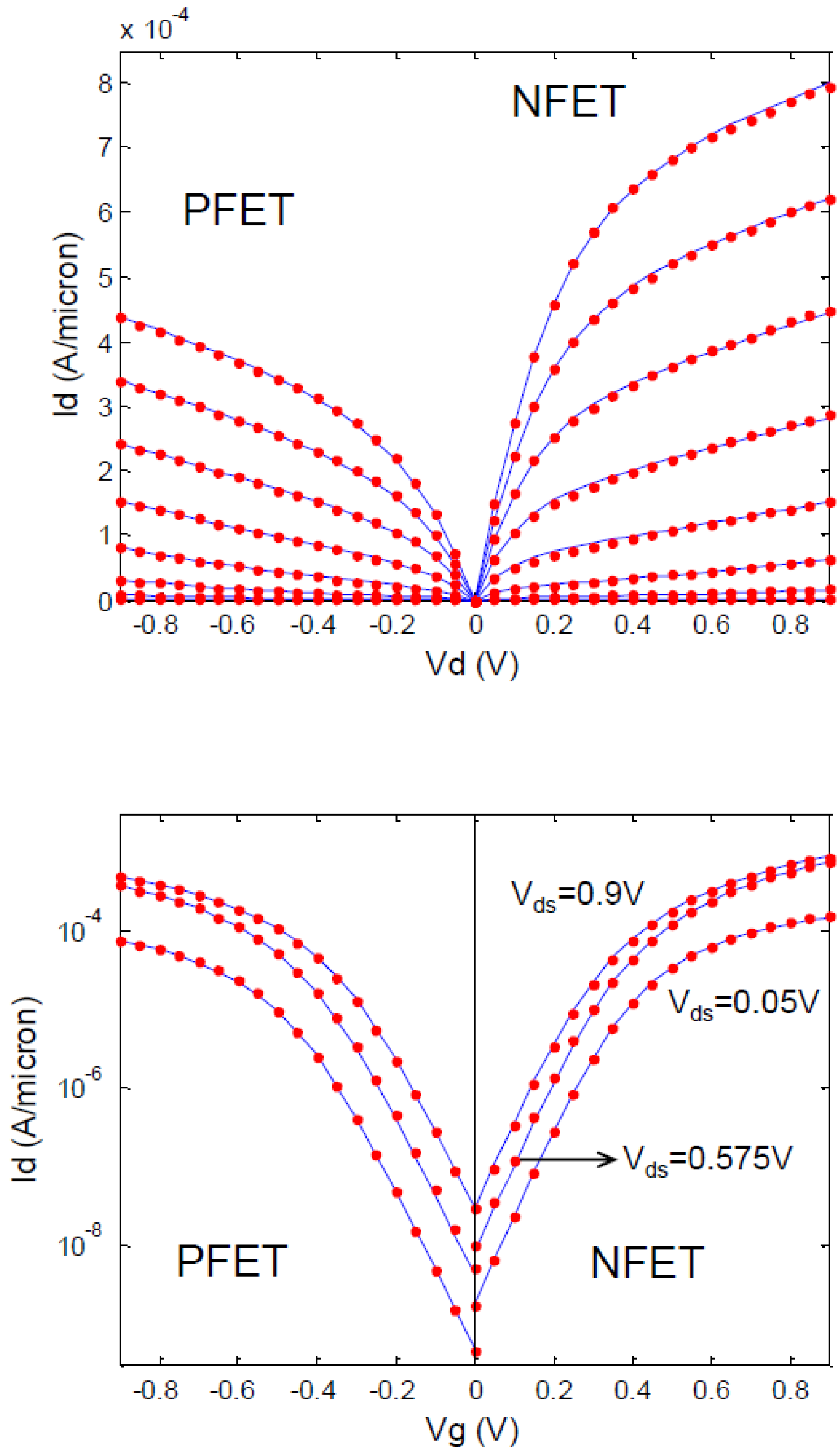


Figure 2-4: MVS model fitting with data from a 40nm BSIM4 industrial design kit. The channel width is 300nm and 600nm for NFET and PFET, respectively.

2.2.2 $I-V$ Curve Analysis

Once all of the parameters are extracted, the MVS model is validated by comparing its $I-V$ curve with that of a BSIM4 model from a $40nm$ bulk industrial design kit, as shown in Fig. 2-4. We see good agreement in both sub-threshold and above threshold regions for both NFETs and PFETs. The accuracy of the MVS model fitting is comparable to other popular industrial models [27][28] and much better than other ultra-compact models with similar complexity as the MVS model [57]. Previous work has demonstrated that the MVS model has indeed good DC agreement with real measurement data fabricated in various technologies ($32nm$, $45nm$, $65nm$) and processes (poly-SiON gate, metal-gate high-k) from various foundries (IBM, Intel) [52, 54]. However, systematic validation of the MVS model for timing verification of digital circuits has been missing so far in the literature; the results of this thesis presented in Section 2.3 are the first such validation [64]. Once the device $I-V$ and $C-V$ curves are calibrated, we can proceed to timing or power comparison for standard library cells and other large-scale digital circuits. This is illustrated in the next two sections.

2.3 Standard Cell Characterization Using Calibrated MVS Model

To validate the accuracy of the MVS model as calibrated in the previous section, we implement it using Verilog-A under the Cadence Virtuoso Design Environment. We then use it to characterize the SPICE-level circuits of a set of generic library standard cells from an industrial design kit in $40nm$ bulk CMOS technology.

The first circuit we consider is an inverter undergoing trapezoidal input transitions. This basic example is used to illustrate several important features of our calibrated MVS model. It is well known that the charging and discharging activities during input gate transitions require precise balancing of both static and dynamic behavior of the NFET and PFET transistors. The output voltage waveforms using the MVS model in comparison with the industry-standard BSIM4 model are depicted in Fig. 2-

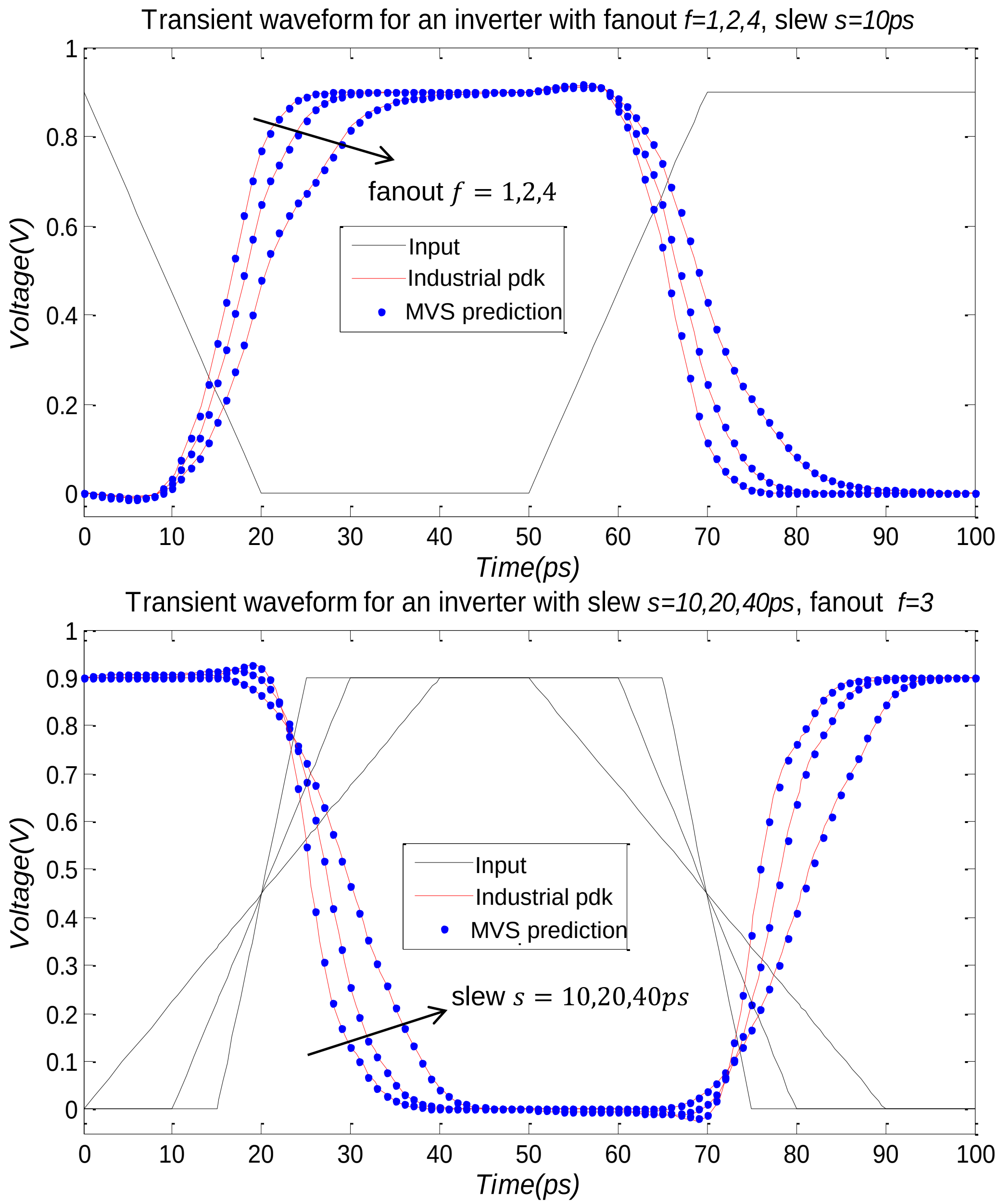


Figure 2-5: Transient response waveform for an inverter chain with various input slews and fanouts.

5. The conducted tests are similar to those used in static timing cell characterization as they have sweeps with respect to both output loads and input slews. In the first sweep test, the input slew is fixed at $10ps$ and the load (fanout) is 1, 2 and 4, while in the second test, the load (fanout) is fixed at 3 and the input slew rate is $10ps$, $20ps$, and $40ps$. The average delay error between the MVS model and the "golden" or baseline BSIM4 model is 0.88% and the 10% − 90% rising/falling time error is 0.92%/1.11%. This is a good indication of the accuracy of the transient calibration of the extracted MVS model.

Table 2.2: Delay (in ps) comparison between MVS and BSIM4 for various gates with input slew of $10ps$ and a fanout of 3. D_{l-h} represents low-to-high propagation delay and D_{h-l} represents high-to-low propagation delay.

(ps)	VS		BSIM 4		Error	
	D_{l-h}	D_{h-l}	D_{l-h}	D_{h-l}	E_{l-h}	E_{h-l}
INV	5.75	5.59	5.71	5.62	0.7%	0.5%
$NAND2_1$	6.3	9.86	6.3	10.2	0.1%	2.6%
$NAND2_2$	6.97	11.5	6.97	11.3	0.1%	2.3%
$NAND3_1$	7.4	16.2	7.4	16.5	0.1%	1.9%
$NAND3_2$	7.25	16.2	7.23	15.9	2.7%	2%
$NAND3_3$	7	15.1	6.97	14.8	0.4%	2.5%
$NOR2_1$	9.92	6.4	10.1	6.4	1.6%	0.1%
$NOR2_2$	8.84	6.12	8.93	6.08	0.9%	0.6%
$NOR3_1$	15.3	6.96	15.58	6.96	1.9%	0.1%
$NOR3_2$	14.8	6.81	15.01	6.8	1.4%	0.1%
$NOR3_3$	13.0	6.49	13.07	6.44	0.4%	0.7%

In Tables 2.2 and 2.3 we summarize the computed delays and rise/fall times for 2- and 3-input symmetrical NAND/NOR gates, with all variables defined in Fig. 2-6. The input slew is $10ps$ and the output load is a fanout of 3. The average and the maximum relative error, MVS vs. BSIM4, for all gates under test are 1.5% and 2.6%, respectively. This compares favorably with [57] where their ultra-compact model achieves only 6% (delay) and 11% (slew) accuracy on a similar set of tests.

Table 2.3: Output slew comparison between MVS and BSIM4 for various gates with input slew of $10ps$ and a fanout of 3.

(ps)	VS		BSIM 4		Error	
	rise	fall	rise	fall	E_r	E_f
INV	9.13	8.84	9.03	8.67	1.1%	2.0%
$NAND2_1$	10.3	17.7	10.5	18.1	1.7%	2.2%
$NAND2_2$	11.4	18	11.2	17.6	1.6%	2.2%
$NAND3_1$	12.3	27.8	12.2	27.1	0.9%	2.5%
$NAND3_2$	12.1	27.8	11.9	27.1	1.6%	2.5%
$NAND3_3$	11.8	27.8	11.6	27.1	1.9%	2.5%
$NOR2_1$	16.0	9.94	15.7	9.82	1.0%	1.2%
$NOR2_2$	16	9.37	15.9	9.25	0.7%	1.3%
$NOR3_1$	25.0	11.22	24.9	11.2	0.7%	0.2%
$NOR3_2$	25.0	10.51	24.8	10.38	0.8%	1.2%
$NOR3_3$	24.9	10.07	24.6	9.9	1.3%	1.7%

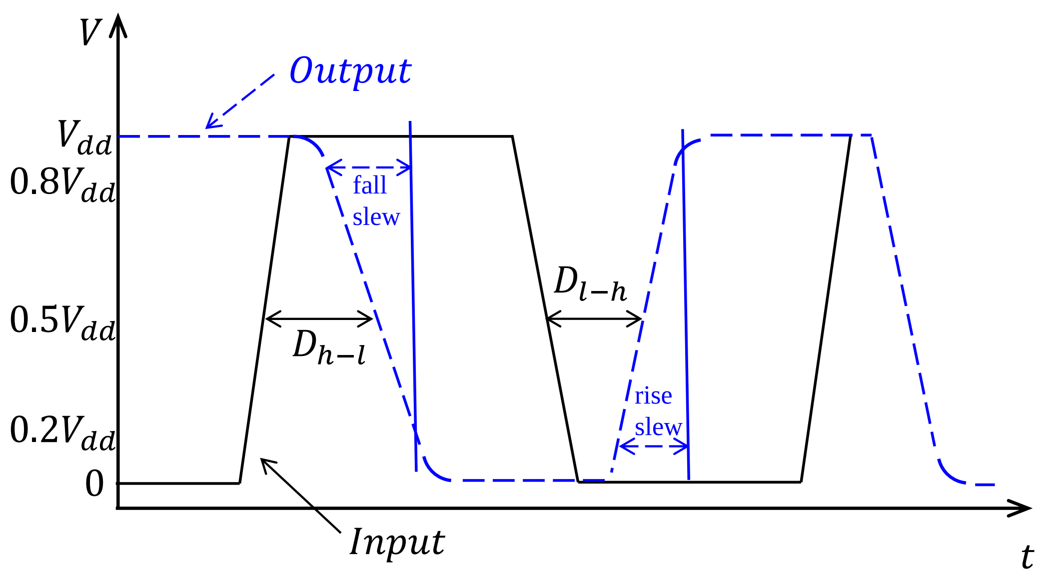


Figure 2-6: Delay and slew between input and output signals. Delay is measured at the 50% V_{dd} points; slew is measured between the 20% and 80% V_{dd} points.

2.4 MVS Model Validation for Digital Circuits

To further verify the calibrated MVS model, a 32-bit ripple-carry adder is designed in the targeted technology ($40nm$ bulk CMOS) and the transient waveform of the critical path compared using MVS and BSIM4 models. The simulation environment and the SPICE convergence setting using both models are exactly the same; this is important to demonstrate that no major work is required to adapt the circuit simulation setting to the presence of the new transistor model.

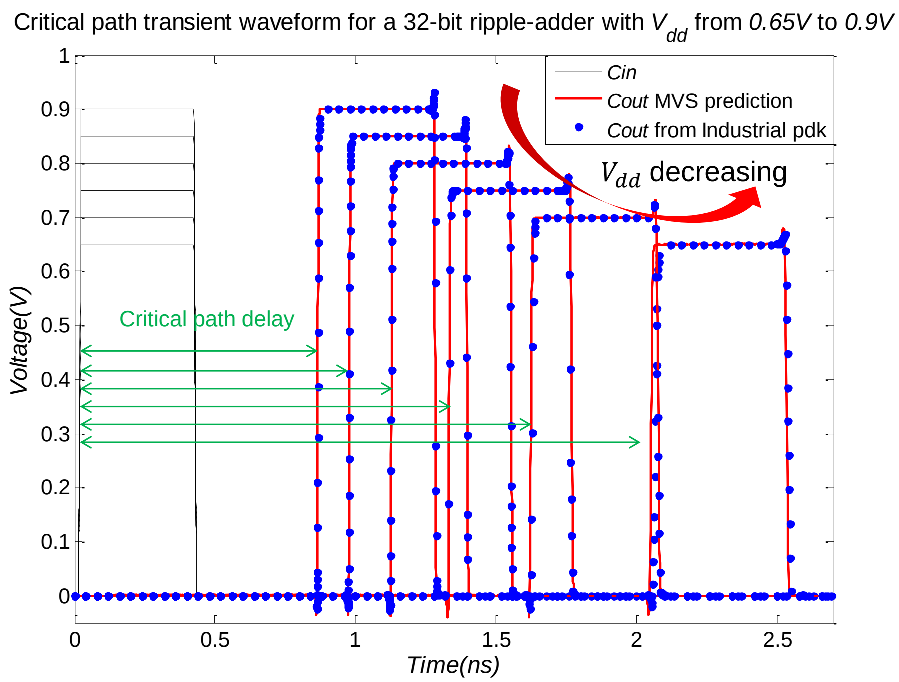


Figure 2-7: Critical path transient waveform for a 32-bit ripple-adder with V_{dd} from $0.65V$ to $0.9V$, in $0.05V$ increments.

The test circuit includes 0.9k transistors in total belonging to various library cell types (INV, NAND, NOR and XOR). We select the worst-case delay for a 32-bit add operation. This requires setting input A at 100...00 and input B at 111...11. The input carry on signal of the very first bit C_{in0} has a $0 \rightarrow 1$ transition and then a $1 \rightarrow 0$ transition, and the output carry on signal of the very last bit will reflect the critical path delay. To show the robustness of the MVS model for low-power design, the supply voltage V_{dd} is swept from $0.6V - 0.9V$. The transient signal C_{in0} and C_{out32} at different V_{dd} from both MVS and BSIM4 model are shown in Fig. 2-7, which demonstrates that the output signals of the two models have excellent matching. The average delay mismatch under all V_{dd} conditions is about 0.3%.

The second digital circuit we consider is a 1001-stage inverter chain designed in the same technology. This test circuit includes 2k transistors in total and the delays under different V_{dd} from $0.6V$ to $0.9V$ are compared between the MVS and BSIM4 models. The average delay mismatch under all V_{dd} is about 0.25%. In both digital circuit cases, the delay mismatch for MVS vs. BSIM4 model is smaller than the mismatch observed in library cell delays. This is because the rise/fall mismatches tend to cancel each other.

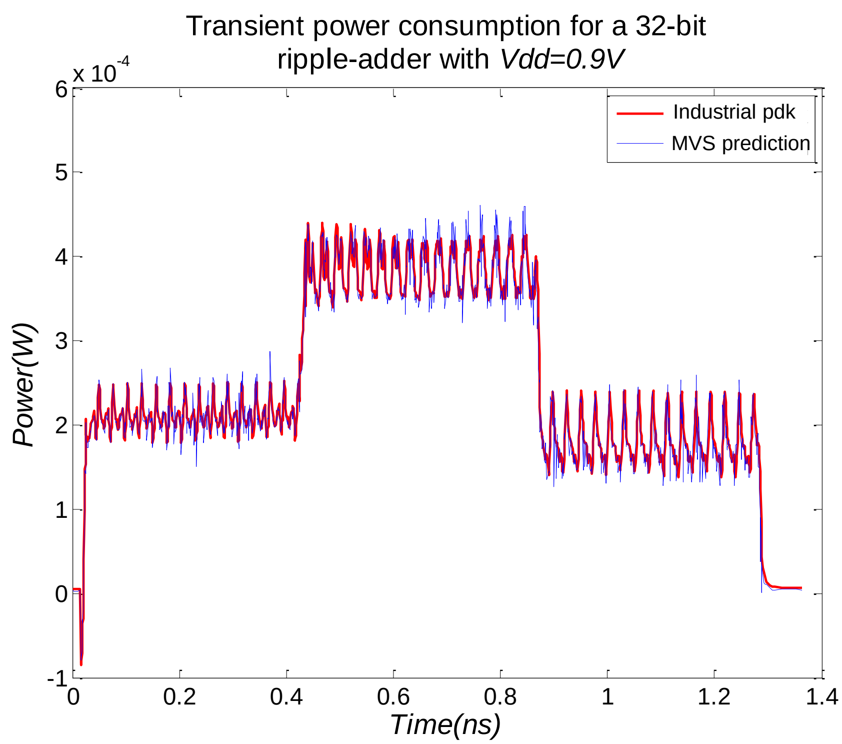


Figure 2-8: Transient power consumption for a 32-bit ripple-adder with V_{dd}=0.9V.

Power consumptions of the critical path transitions in the aforementioned test cases are also compared. Transient power consumed by the 32-bit ripple-adder under $V_{dd} = 0.9V$ is shown in Fig. 2-8, which demonstrates good agreement between the MVS and BSIM4 models. The average power consumption mismatch under all V_{dd}'s is 1.3% for the 32-bit adder and 1.8% for the 1001-stage inverter chain. Finally, the power-delay curves for both cases under different V_{dd} are shown in Fig. 2-9.

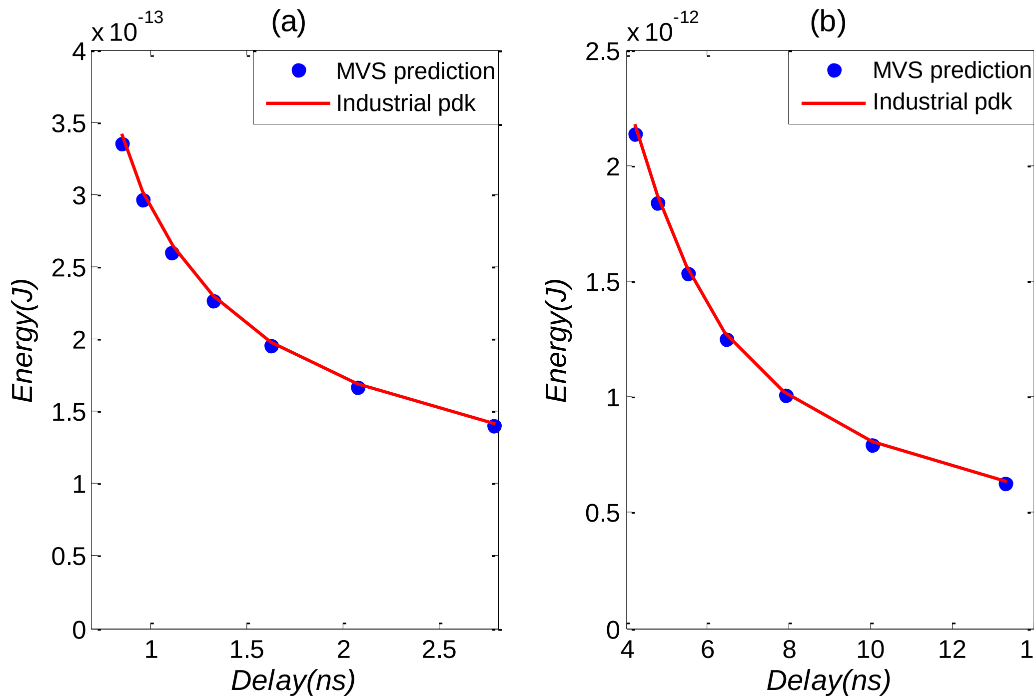


Figure 2-9: Energy delay curve for (a) a 32-bit ripple-adder and (b) 1001 stage inverter chain under V_{dd} from 0.65 to $0.9V$.

The runtime speedup of the MVS model is further compared with an open source BSIM SOI compact model implemented in Verilog-A [65], which has similar model complexity with BSIM4. The comparison is between MVS and BSIM SOI instead of BSIM4 because BSIM4 has been implemented through C code which has higher efficiency and comparison between Verilog-A and C code is not appropriate. The transient simulation runtime comparison in the two aforementioned test circuits are shown in Table 2.4. An average speed up of 7.6× is achieved which is in line with the order of magnitude reduction in the number of MVS parameters. The simulation environment, the SPICE convergence setting and maximum iteration setting for both models are exactly the same to achieve a fair comparison. Also, in both models, the transition time and the delay time of the library cells are tuned to be similar to ensure a comparable computing effort in both cases. Since the simplicity of device models is key to a statistical design flow [32], the 7.6× speed up achieved by the MVS model points to its potential for variation-aware statistical analysis with reduced computational cost.

Table 2.4: Transient simulation speed comparison between MVS model and BSIMSOI model [65].

Model	MVS run time	BSIMSOI run time	MVS speed up
1001-stage inverter chain	$621s$	$3470s$	$5.6\times$
32-bit ripple-adder	$111s$	$1060s$	$9.6\times$

In summary, the above results demonstrate that the MVS model, fully integrated in a vendor CAD environment equipped with a $40nm$ industrial design kit, is capable of dealing with SPICE-level timing and power analysis tasks at an industrial degree of accuracy, while having an order of magnitude fewer parameters than the BSIM4 industry standard.

2.5 Parameter Variations in MVS Model

To support statistical circuit simulation, the measured $I-V$ and $C-V$ statistics need to be converted into variations of a complete set of independent MVS model parameters. For modern MOSFETs, the primary sources of within-die variations include random dopant fluctuation (RDF), line-edge roughness (LER) and oxide thickness fluctuation (OTF), as well as local fluctuations of mechanical stress [66]. To maintain the simplicity of the statistical MVS model, we relate most of its parameters directly to standard device measurements rather than to manufacturing process parameters. The MVS model parameters used for statistical modeling are listed in Table 2.5.

Table 2.5: MVS model parameters list

Source	Model Parameter	Description
LER	$L_{eff}(nm)$	Effective channel length
LER	$W_{eff}(nm)$	Effective channel width
RDF	V_{T0}	Zero-bias threshold voltage
OTF	$C_{inv}(\mu F/cm^2)$	Effective gate-to-channel capacitance per unit area
Stress	$\mu(cm^2/V\cdot s)$	Carrier mobility
Stress	$v_{xo}(cm/s)$	Virtual source velocity

In the MVS model, the threshold voltage is modeled as

$$V_T = V_{T0} - \delta(L_{eff})V_{DS} \tag{2.6}$$

where $\delta(L_{eff})$ is the L_{eff}-dependent DIBL coefficient [52]. The threshold voltage variation in Table 2.5 is determined by the variations in implantation energy and dose as well as fluctuations in substrate doping. These effects are modeled through variation in V_{T0}, while length-dependent threshold voltage variation is captured through variation in $\delta(L_{eff})$. Note that V_{T0} has a weak dependency on L_{eff} over the range considered here thus this effect is negligible. A special feature of the MVS model is that v_{xo} is independent of the bias voltages. Previous work has shown that the relative change in virtual source velocity is related to the change in mobility [67]. According to [68], v_{xo} also has a dependency on $\delta(L_{eff})$. Therefore variation on L_{eff} also has an impact on v_{xo}. In the MVS model, both effects are described using an approximation for the sensitivity of v_{xo} with respect to μ and $\delta(L_{eff})$, as shown in the following expression:

$$\frac{\Delta v_{xo}}{v_{xo}} = [\alpha + (1-B)(1-\alpha+\gamma)]\frac{\Delta\mu}{\mu} + \frac{\partial v_{xo}}{v_{xo}\partial\delta(L_{eff})}\Delta\delta(L_{eff}) \tag{2.7}$$

Here $\alpha \approx 0.5$ and $\gamma \approx 0.45$ are both fitting indices to a power law and B is the ballistic efficiency given by the expression

$$B = \lambda/(\lambda + 2l) \tag{2.8}$$

where λ is the mean free path and l is the critical length for backscattering to the source at nominal L_{eff}. An approximate value for $\frac{\partial v_{xo}}{v_{xo}\partial\delta(L_{eff})}$ in the targeted technology is 2.

2.6 Statistical Extraction Method

A well-characterized nominal MVS model is the foundation of variability analysis. The nominal values of important effects, such as DIBL, mobility and virtual source velocity are critical for determining the model sensitivity to parameter variations. The basis for mismatch modeling was proposed by Pelgrom, *et al* [69]. For local variation, the fluctuations in the observed variation of parameters have a uniform area dependency

$$\frac{\sigma_p^2}{p^2} \propto \frac{1}{LW} \tag{2.9}$$

where the subscript p represents a process parameter such as effective channel length and width, channel dopant concentration, mobility, and effective gate-to-channel capacitance per unit area. For local mismatch, we have $\sigma_L = \sigma_{L_{eff}}$ and $\sigma_W = \sigma_{W_{eff}}$, and a complete equation considering the geometric dependence of each parameter is

$$\begin{bmatrix} \sigma_{V_{T0}} \\ \sigma_L \\ \sigma_W \\ \sigma_\mu \\ \sigma_{C_{inv}} \end{bmatrix} = [\alpha_1\ \alpha_2\ \alpha_3\ \alpha_4\ \alpha_5] \begin{bmatrix} \frac{1}{\sqrt{WL}} \\ \sqrt{\frac{L}{W}} \\ \sqrt{\frac{W}{L}} \\ \frac{1}{\sqrt{WL}} \\ \frac{1}{\sqrt{WL}} \end{bmatrix} \tag{2.10}$$

The ultimate goal of this statistical modeling is to extract a group of α_{1-5} that is appropriate for all transistor geometries and that match the statistical circuit performance. The mismatch variances of p_j cannot be characterized directly from measurement or device simulations. Instead, variations σ_{F_i} $(i = 1, 2, ..., m)$ of electrical performance parameters (e.g., I_{dsat}, I_{off}, etc.) are measured under different geometry and bias conditions and the σ_{p_j} are calculated from backward propagation of variance (BPV) [33] according to the formula

$$\sigma_{F_i}^2 = \sum_{j=1}^{n} (\frac{\partial F_i}{\partial p_j})^2 \sigma_{p_j}^2 \tag{2.11}$$

$$\begin{bmatrix} \begin{pmatrix} \sigma^2_{I_{dsat}} & -(\frac{\partial I_{dsat}}{\partial C_{inv}})^2 \sigma^2_{C_{inv}} \\ \sigma^2_{log_{10} I_{off}} & -(\frac{\partial log_{10} I_{off}}{\partial C_{inv}})^2 \sigma^2_{C_{inv}} \\ \sigma^2_{C_{gg@V_g}} & -(\frac{\partial C_{gg@V_g}}{\partial C_{inv}})^2 \sigma^2_{C_{inv}} \end{pmatrix}_1 \\ . \\ . \\ . \\ \begin{pmatrix} \sigma^2_{I_{dsat}} & -(\frac{\partial I_{dsat}}{\partial C_{inv}})^2 \sigma^2_{C_{inv}} \\ \sigma^2_{log_{10} I_{off}} & -(\frac{\partial log_{10} I_{off}}{\partial C_{inv}})^2 \sigma^2_{C_{inv}} \\ \sigma^2_{C_{gg@V_{dd}}} & -(\frac{\partial C_{gg@V_{dd}}}{\partial C_{inv}})^2 \sigma^2_{C_{inv}} \end{pmatrix}_m \end{bmatrix} =$$

$$\begin{bmatrix} \begin{pmatrix} (\frac{\partial I_{dsat}}{\partial V_{T0}})^2 \frac{1}{WL} & (\frac{\partial I_{dsat}}{\partial L})^2 \frac{L}{W} & (\frac{\partial I_{dsat}}{\partial W})^2 \frac{W}{L} & (\frac{\partial I_{dsat}}{\partial \mu})^2 \frac{1}{WL} \\ (\frac{\partial log_{10} I_{off}}{\partial V_{T0}})^2 \frac{1}{WL} & (\frac{\partial log_{10} I_{off}}{\partial L})^2 \frac{L}{W} & (\frac{\partial log_{10} I_{off}}{\partial W})^2 \frac{W}{L} & (\frac{\partial log_{10} I_{off}}{\partial \mu})^2 \frac{1}{WL} \\ 0 & (\frac{\partial C_{gg@V_{dd}}}{\partial})^2 \frac{L}{W} & (\frac{\partial C_{gg@V_{dd}}}{\partial W})^2 \frac{W}{L} & (\frac{\partial C_{gg@V_{dd}}}{\partial \mu L})^2 \frac{1}{WL} \end{pmatrix}_1 \\ . \\ . \\ . \\ \begin{pmatrix} (\frac{\partial I_{dsat}}{\partial V_{T0}})^2 \frac{1}{WL} & (\frac{\partial I_{dsat}}{\partial L})^2 \frac{L}{W} & (\frac{\partial I_{dsat}}{\partial W})^2 \frac{W}{L} & (\frac{\partial I_{dsat}}{\partial \mu})^2 \frac{1}{WL} \\ (\frac{\partial log_{10} I_{off}}{\partial V_{T0}})^2 \frac{1}{WL} & (\frac{\partial log_{10} I_{off}}{\partial L})^2 \frac{L}{W} & (\frac{\partial log_{10} I_{off}}{\partial W})^2 \frac{W}{L} & (\frac{\partial log_{10} I_{off}}{\partial \mu})^2 \frac{1}{WL} \\ 0 & (\frac{\partial C_{gg@V_g}}{\partial L})^2 \frac{L}{W} & (\frac{\partial C_{gg@V_g}}{\partial W})^2 \frac{W}{L} & (\frac{\partial C_{gg@V_g}}{\partial \mu})^2 \frac{1}{WL} \end{pmatrix}_m \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} \tag{2.12}$$

Equ. (2.11) assumes p_j and p_k for any $j \neq k$ are independent. And Equ.(2.11) furthur assumes Gaussian distributions for both groups of $\{F_i\}$ and $\{p_j\}$. This assumption requires a careful selection of both $\{F_i\}$ and $\{p_j\}$. In our work, and unlike the statistical modeling approach of [33], measurements at some conditions, such as I_{off} or I_d at the transition region between linear and saturation, do not strictly follow a Gaussian distribution. It follows that such conditions do not result in suitable F_i parameters, or require restricted conditions or transformations to be used. In this work, the F_i are selected to be I_{dsat}, $log_{10} I_{off}$ and $C_{gg}@V_{dd}$.

The accuracy of Equation (2.11) hinges on the validity of approximating the electrical performance parameters as linear functions of the process parameters. We have found that such linear approximation is sufficiently accurate to extract σ_{p_j}.

A system of linear equations is set up after stacking a group of equations with different transistor sizes, as is shown in (2.12). The sensitivity matrix in (2.12) is calculated from SPICE simulation using the MVS model. To ensure the independence of p_j's as required by (2.11), the virtual source velocity is not considered as a separate variation parameter in Equation (2.12), since its effect has been captured in the variation of L_{eff} and μ. Also, silicon dioxide films are created with a thermal oxidation process which historically has been extremely tightly controlled [2] with the σ variation of C_{inv} being less than 0.5% in our case. Because the BPV process tends to overestimate variation in tightly controlled process parameters, we directly measure C_{inv} through the oxide thickness, as suggested in [34].

Since the primary intrinsic mismatch corresponding to gate length and width variation is due to line edge roughness (LER), which is caused by etching and sub-wavelength photo-lithographic process, it is reasonable to assume the same roughness for both length and width. Therefore an empirical relationship $\alpha_2 = \alpha_3$ ($\sigma_L/\sigma_W = L/W$) is assumed to further reduce the unknown parameters in (2.12). A good match to data is achieved ($\alpha_2/\alpha_3 = 0.95 - 0.99$ under different geometries) in a $40nm$ CMOS technology.

Parameter coefficients α_{1-4} are solved separately using individual transistor data without using (2.10), or solved together using transistors with different geometries through a least square fit. The solution given by solving the stacked equations with different geometries provides a more consistent and scalable result across these geometries while the solution given by using individual transistor data is more accurate for each geometry. Therefore a trade-off between accuracy and consistency is made according to the difference between the two solutions. Less than 10% difference between the two methods is observed and the solution with different geometries is employed in this work, as shown in Fig. 2-10.

2.7 Statistical Verification

To validate the accuracy of the MVS statistical model as well as the statistical extraction method, we implement it using Verilog-A under the Cadence Virtuoso Design Environment and run camparisons against BSIM simulations. The method described

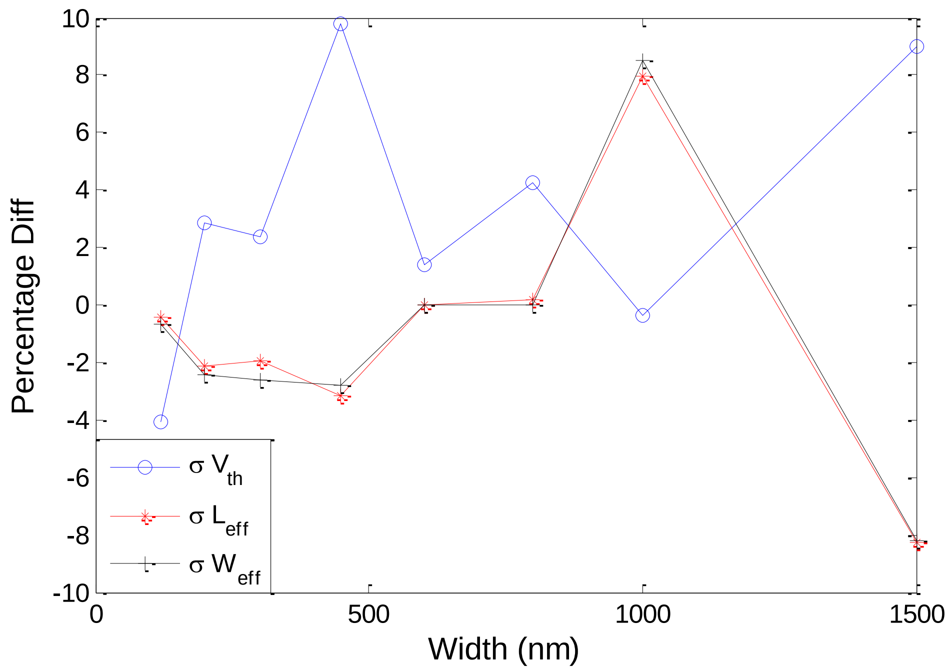


Figure 2-10: Relative difference in $\sigma_{V_{T0}}$, $\sigma_{L_{eff}}$ and $\sigma_{W_{eff}}$ between solving (2.12) individually and together.

in Section 2.6 is applied to characterize the SPICE-level benchmark circuit statistics of a 40nm bulk CMOS technology. Although the BPV method is applicable to measurement data, here we have employed data generated using a BSIM based industrial design kit to validate the proposed MVS statistical model. The benchmark circuits include both digital (standard cell library and D flip-flop) and analog circuits (SRAM). Monte-Carlo simulations are run by randomly generating samples of each process parameter based on the independent Gaussian distributions extracted from Section 2.6 Various Monte Carlo simulations are performed, including several geometries of MOSFETs and different electrical tests ($I-V$ and $C-V$). The sample sizes are more than 1000 to characterize the statistical variation and correlation for F_i. The extracted parameter statistics α_{1-5} are listed in Table 2.6.

2.7.1 Validation of Device Variability

The percentage standard deviation of σ/μ for I_{dsat} and the underlying process parameter contributions are shown in Fig. 2-11. Compared with previous results in

Table 2.6: Extracted standard deviation coefficient using the BPV method.

	NMOS	PMOS
α_1 $(V \cdot nm)$	2.3	2.86
α_2 (nm)	3.71	3.66
α_3 (nm)	3.71	3.66
α_4 $(nm \cdot cm^2/V \cdot s)$	944	781
α_5 $(nm \cdot \mu F^2/cm^2)$	0.29	0.81

a similar technology [70], we observe a similar extracted $\sigma_{V_{T0}}/\mu_{V_{T0}}$ and $\sigma_{L_{eff}}/\mu_{L_{eff}}$ but smaller σ_μ/μ_μ in the MVS model. The latter result is due to the fact that in the context of the MVS model, mobility and virtual source velocity have meanings that differ with those of [70]. Table 2.7 shows Monte Carlo simulation results for both I_{dsat} and $log_{10}I_{off}$ for various transistor sizes, and a comparison between the MVS and an industrial statistical BSIM model. The simulated variation shows good matching between the MVS and BSIM models, thus confirming the accuracy of our statistical MVS model and the correctness of the BPV extraction procedure.

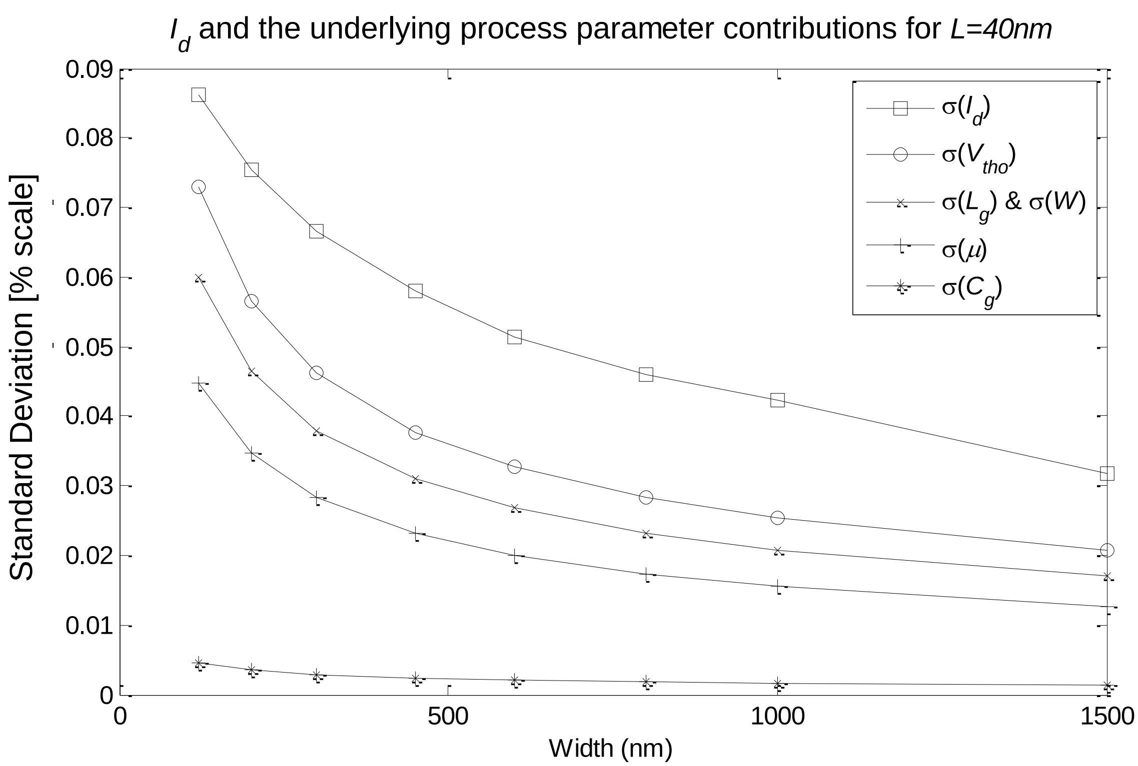


Figure 2-11: Standard deviation of I_{dsat} and the underlying process parameter contributions for $L = 40nm$.

Table 2.7: Standard deviation of the MVS Monte-Carlo simulation compared with industrial model.

Device (W/L nm)		NMOS		PMOS		
	F_i	BSIM σ	MVS σ	BSIM σ	MVS σ	Unit
Wide	I_{dsat}	33.1	32.7	21.6	21.7	uA
(1500/40)	$log_{10}I_{off}$	0.13	0.13	0.15	0.15	
Medium	I_{dsat}	20.2	19.9	14.8	14.8	uA
(600/40)	$log_{10}I_{off}$	0.17	0.17	0.24	0.23	
Short	I_{dsat}	8.7	8.8	6.95	6.86	uA
(120/40)	$log_{10}I_{off}$	0.33	0.33	0.49	0.47	

I_{dsat} and $log_{10}I_{off}$ bivariate scatter plots for BSIM model and 1σ, 2σ and 3σ confidence ellipses for both MVS and BSIM model are shown in Fig. 2-12. Note that in the statistical MVS model, the generated variation parameters L_{eff}, V_{T0}, and μ are non-correlated. This behavior confirms that the I_{dsat} and $log_{10}I_{off}$ variations are fully decoupled during the statistical extraction procedure.

2.7.2 Statistical Validation Using Benchmark Circuits

We have performed statistical experiments on both the BSIM model and the MVS model using a set of benchmark circuits, including standard library logic cells (INV, NAND2, DFF, etc.) and an SRAM cell.

Our first standard cell is a fanout-of-3 static INV gate having the geometry: $1\times$, $2\times$ and $4\times$. For each of BSIM and MVS, 2500 Monte Carlo simulations are run to generate delay probability density functions as shown in Fig. 2-13. The V_{dd} in all cases is $0.9V$, which is the standard supply voltage for this particular technology. Delay variations generated from both models follow a Gaussian distribution. Excellent matching is achieved across a wide range of transistor sizes, which confirms that the geometric dependencies of the MVS variation are well characterized. It is important to note that our statistical extraction procedure remains valid regardless of the specific functional dependence of the variations on device geometry.

Not only does the MVS statistical model enable the characterization of the impact of variability in L_{eff}, W_{leff}, V_{T0}, μ, v_{x0} and C_{inv} on timing, but also it may be used to

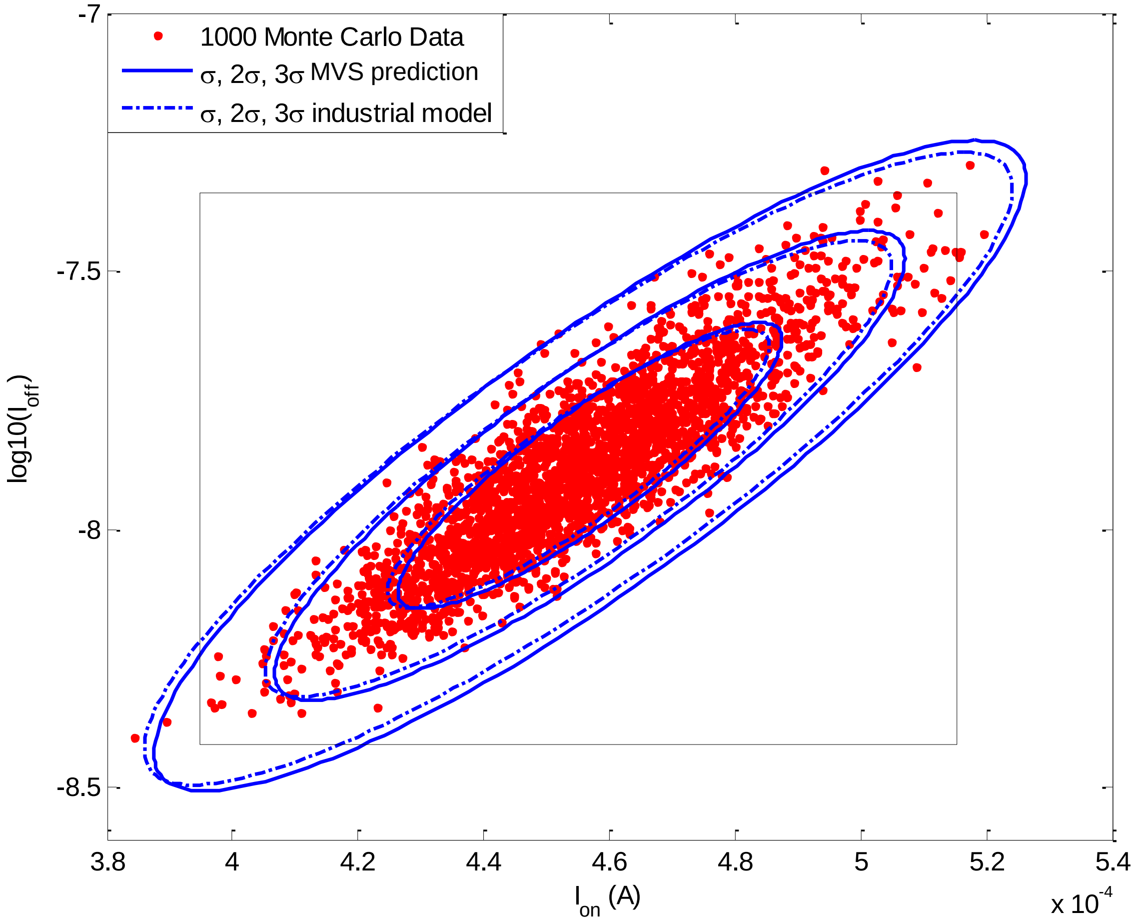


Figure 2-12: Comparison of 1000 Monte Carlo simulation results for medium sized device ($W/L = 600nm/40nm$) between MVS and BSIM statistical model. 1σ, 2σ and 3σ confidence ellipses for both model are also shown. The solid box represents $\pm 3\sigma$ limits for each variable from the BSIM model.

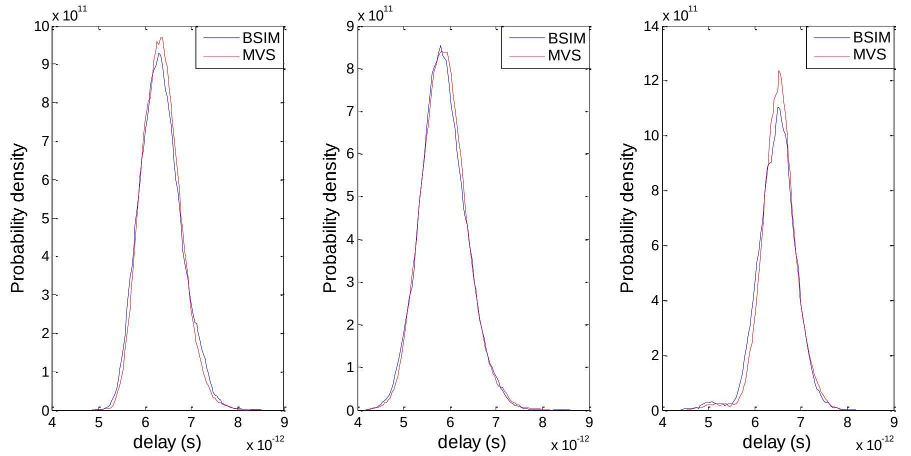


Figure 2-13: Delay probability density comparison between BSIM and MVS models for an INV gate (fanout of 3) with different sizes.

predict the distribution of frequency, leakage power, and parametric yield, as shown in Fig. 2-14. The leakage-frequency scatter plots, as well as mean and standard deviations predicted by the BSIM and MVS models, are almost identical. In both cases, the total spread of leakage is as much as 37×. The impact of within-die variation on frequency variation is 45% and 50% of the mean frequency for BSIM and MVS models, respectively.

Our second standard cell is a fanout-of-3 static NAND2 gate operating under a V_{dd} of $0.9V$, $0.7V$ and $0.55V$. Although power consumption decreases with supply voltage, local variations increase significantly, and as a result parametric yield is decreased. Even worse, the probability density of the delay becomes highly non-Gaussian at low supply voltage, and as a result, the application of statistical static timing analysis (SSTA) becomes more difficult [71]. Although all variation parameters in the MVS model are assumed to be independent Gaussian variables, the non-Gaussian property of the delay distribution is correctly captured, as is shown in Fig. 2-15. The quantile-quantile plot for delay variation starts to deviate from a linear relationship when $V_{dd} = 0.7V$, and the non-linearity becomes pronounced at $V_{dd} = 0.55V$. In both cases, the MVS prediction shows a good match with the BSIM model at the 3σ scale. Unlike the PSP model [33] where variances of extra electrical performance parameters

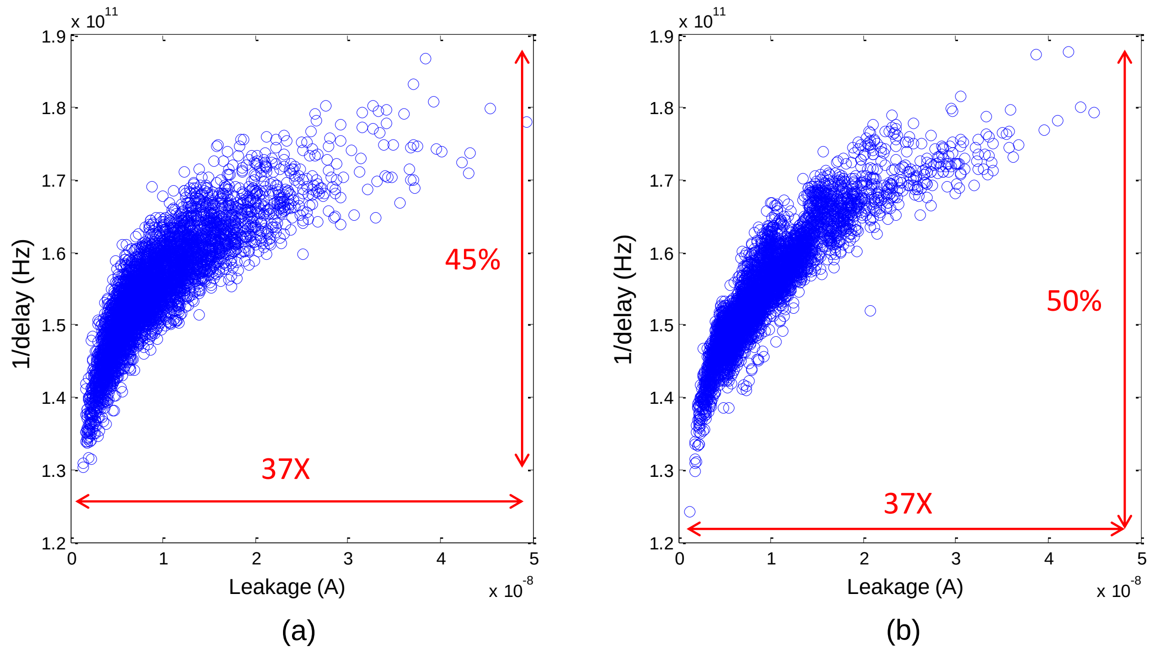


Figure 2-14: Scatter plot generated by 5000 Monte Carlo samples showing the distribution of the total circuit leakage versus frequency (1/delay) for an INV gate (fanout of 3) in (a) BSIM model, and (b) MVS model.

have to be added to match the variance at different V_{gs}, no extra statistical fitting is needed in the MVS model to adjust timing distributions, in cases where dynamic voltage scaling is used.

After verifying the approach on combinational logic cells, we now extend it to perform setup and hold time analysis on a D flip-flop. The schematic of the benchmark master-slave register is shown in Fig. 2-16 (a). Fig. 2-16 (b) shows a typical timing path for setup/hold analysis. Considering statistical variations, the hold and setup constraints are:

$$t_1 - t_2 \quad > \quad T_{hold} \tag{2.13}$$

$$t_1 - t_2 \quad < \quad T_{clk} - T_{setup} \tag{2.14}$$

where T_{clk} is the clock period for the design. The PDF's for setup/hold time for the registers simulated from MVS model and BSIM models are shown in Fig. 2-16(c). One important note is that the characterization of the setup/hold time requires about

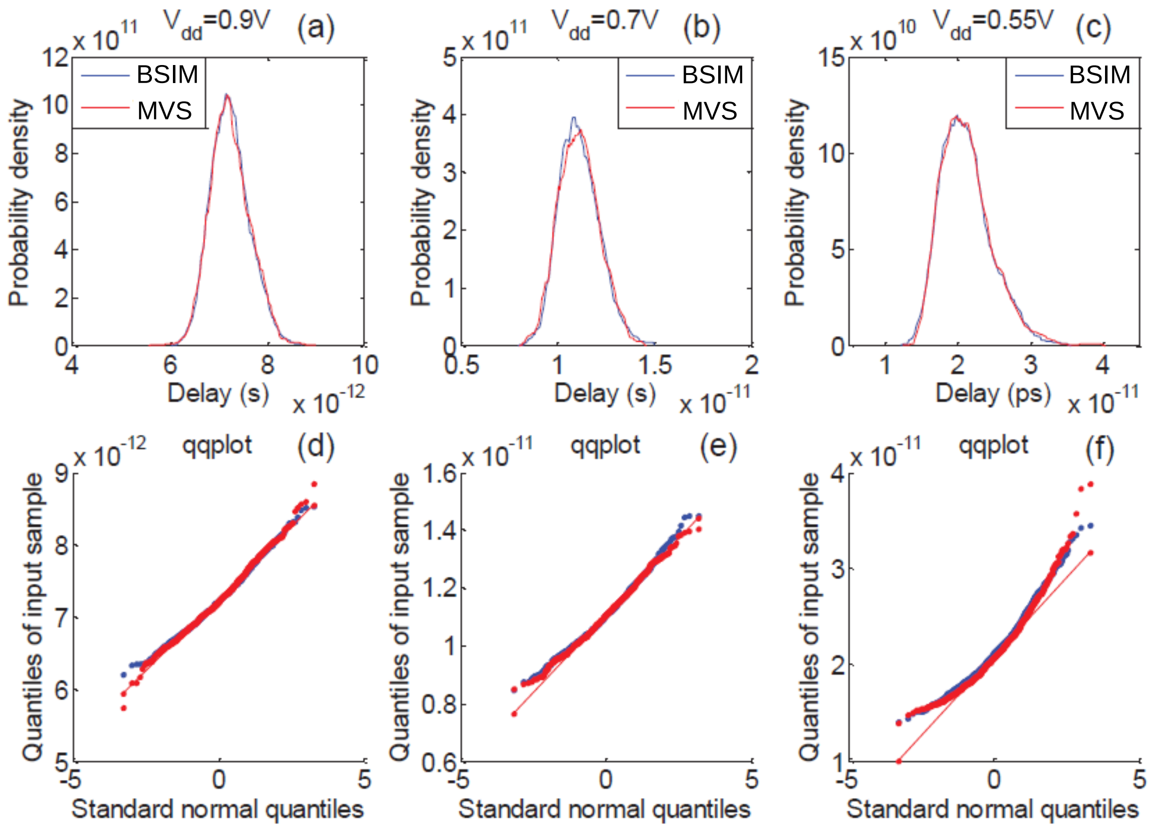


Figure 2-15: Delay probability density comparison between BSIM and MVS models for a NAND2 gate (fanout of 3) with a supply voltage of (a) $0.9V$, (b) $0.7V$ and (c) $0.55V$. The quantile-quantile plots for delay variation under each supply voltage in (d) $0.9V$, (e) $0.7V$ and (f) $0.55V$ show a strongly nonlinear pattern in low power application.

20 times more SPICE simulations than those of a combinational cell having the same number of transistors. This is because the setup/hold time can only be measured indirectly by varying clock to input signal delay. The ultra compact MVS model plays a more important role in this case, where tens of thousands of SPICE simulations are required.

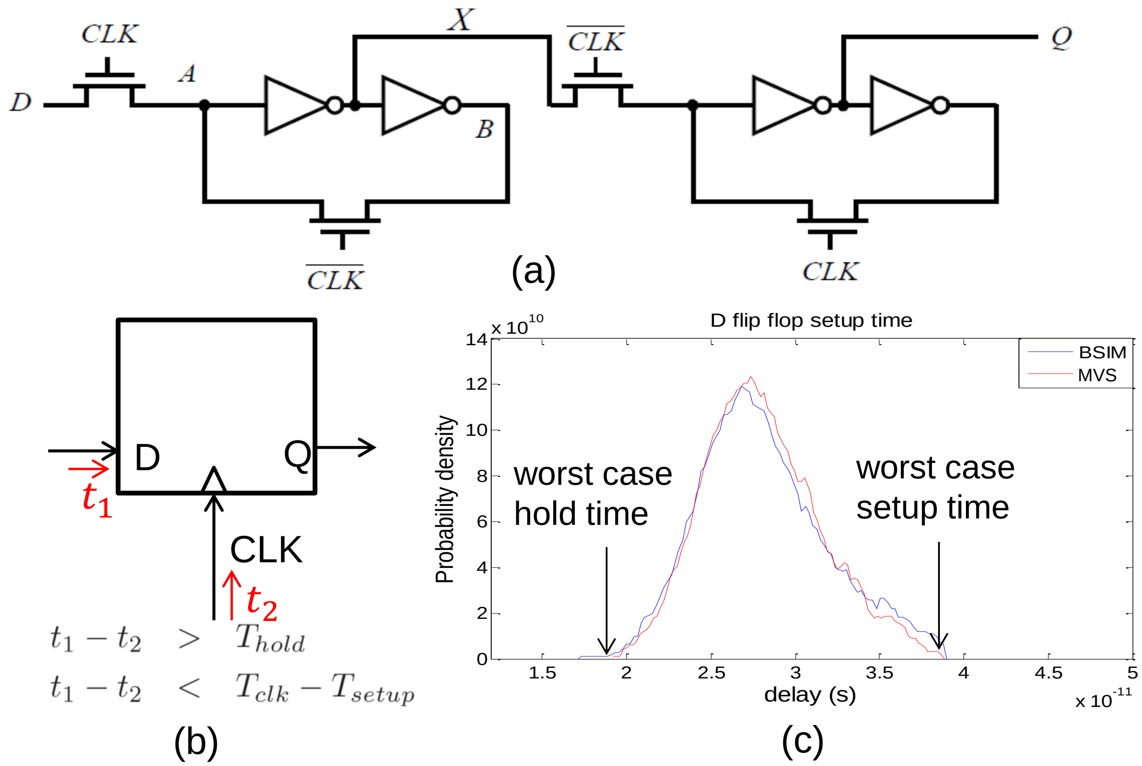


Figure 2-16: (a) Master-slave register based on NMOS-only pass transistors, P/N sizes are $600nm/40nm$ and $300nm/40nm$, respectively; (b) typical timing path for setup/hold analysis; and (c) probability density of the setup time in circuit (a) with 250 Monte Carlo runs.

The last circuit in our validation is a 6T SRAM cell, which is known to be highly sensitive to within-die variations, as shown in Fig. 2-17. The N/P sizes are $150nm/40nm$. Both the MVS and BSIM models are employed to simulate the variability in SRAM READ and HOLD static noise margin (SNM). The characteristic butterfly patterns generated with the statistical MVS model are shown in Fig. 2-17(a) and (d), for READ and HOLD, respectively. The SNM comparisons between the two models for READ and HOLD are shown in Fig. 2-17 (b) and (e). Even with this highly sensitive analog circuit, the ultra compact statistical MVS model provides an excellent match to the "golden" BSIM model. In Fig. 2-17(f), the quantile-quantile plot for SRAM HOLD SNR using both models shows a slightly non-Gaussian distribution.

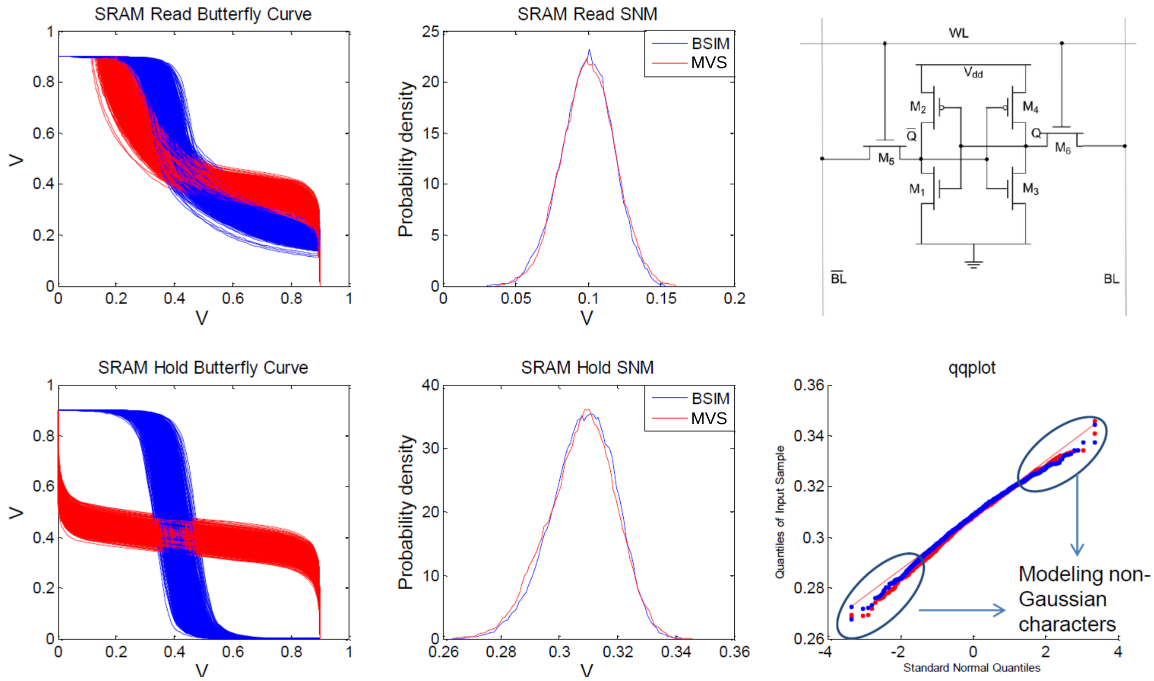


Figure 2-17: 2500 Monte Carlo simulation for a 6T SRAM cell; (a) butterfly pattern from MVS model in static READ mode; (b) probability density for SRAM READ static noise margin (SNR); (c) schematic of the 6-T SRAM; (d) butterfly pattern from MVS model in static HOLD mode; (e) probability density for SRAM HOLD SNR; and (f) quantile-quantile plot for SRAM HOLD SNR.

Finally, the runtime speedup of the MVS model (Verilog-A) with respect to BSIM4 (C code) is shown in Table 2.8. We notice a 4.2× speedup and 8.7× reduction in memory usage. These favorable results can be further improved using an optimized C code implementation of the MVS model in line with the optimized C code used for BSIM4.

Table 2.8: Speed and memory comparison for Monte Carlo simulation between MVS (in Verilig-A code) and BSIM4 model (in C code).

			MVS		BSIM 4	
Cell	Sim.	Sample	Runtime	Memory	Runtime	Memory
NAND2	Tran	2000	$225s$	$14.9M$	$855s$	$126M$
DFF	Tran	250	$3.86ks$	$23.2M$	$13.5ks$	$157M$
SRAM	AC	2000	$405s$	$17M$	$2.15ks$	$187M$

Chapter 3

Physical Subspace Projection: An Efficient Statistical Framework for Performance Estimation from on-Chip Test Structures

3.1 Introduction

With the success of semiconductor scaling, predicted by Moore's law, and the vastly increased complexity of nanometer scale processes and the billion-device circuits they allow, there is a need for comprehensive and efficient approaches for high-yielding designs in the state of the art VLSI technology [2, 72, 59, 32]. 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. These prediction models can then be used in many circuit applications such as parametric yield prediction and robust circuit design [73].

The measurements used to predict system performance are generally taken from different configurations of on-chip test structures which are used for monitoring and controlling the fabrication line. These structures are often small circuits placed in the scribe line on all wafers and therefore capable of modeling the history of the line. Test circuits include simple arrays of transistor structures that allow the measurements of $I-V$ characteristics of MOSFETs [11, 74], and ring-oscillator based on-chip digital circuits which convert an analog signal to more robust digital (frequency) measurements [75]. Compared with test structures for device modeling which are typically composed of a rich variety of structures (e.g., across channel width, length, and other geometric or configuration combinations), these on-chip test structures are often designed in a way that enhances their sensitivities to key physical parameters under process variation (and to decrease their sensitivity to other parameters, when possible). Although it is possible to identify certain parameter variations through measurements using only a single type of test structure, a challenging issue here is how to combine measurements from a mixture of test structures (e.g., different device *and* circuit test structures), and make good predictions on a target circuit performance, as shown in Figure 3-1.

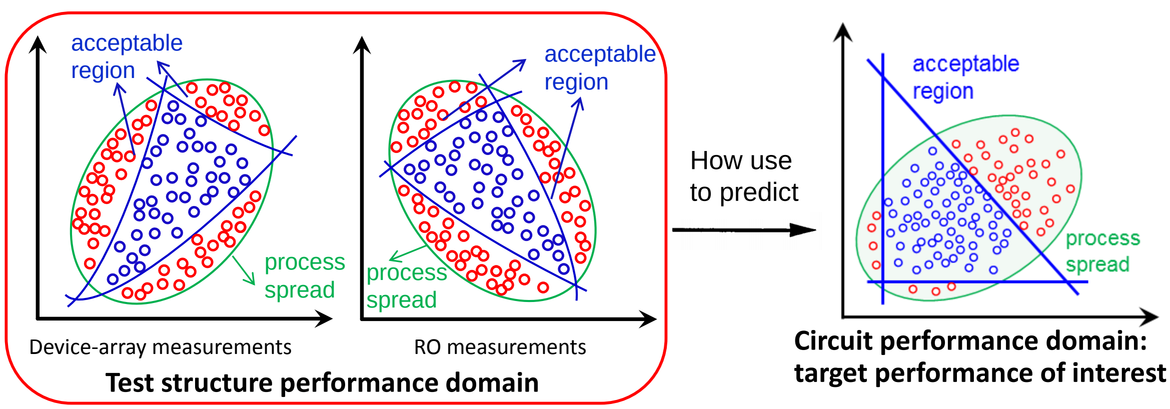


Figure 3-1: Performance estimation problem from a mixture of on-chip test structures.

For standard device models that have large numbers of empirical or semi-empirical parameters, measurements from a single test structure nevertheless end up representing or being sensitive to variations in multiple parameters, and the measurements are therefore statistically correlated. Because of this fact, it is preferable to transform the correlated measurements into a set of low-dimensional, uncorrelated factors. Principal component analysis (PCA) is a commonly used statistical technique that performs this task [35]. Given N samples from a set of correlated random device variables **P** or correlated electrical measurements **E**, PCA seeks a linear transformation of these variables into a new set of random variables **X** which are orthogonal. In other words, PCA identifies non-redundant combinations of the measurements, and therefore achieves dimensionality reduction using only unlabeled measurements.

In conventional approaches, the next step is to perform response surface modeling (RSM) and approximates the circuit performance (e.g., delay, power, etc.) as an analytical (typically linear or quadratic) function of the orthogonal variables **X** [43]. However, even if we select top ranked variables **X** after PCA as basis functions of the performance modeling, the required training sample size is typically quite large. When the measurements data set is not large enough to support the variable space, over-fitting problem described in Section 1.2.3 will appear. Unfortunately, it is can be difficult or expensive to collect sufficient on-chip measurements to support full RSM approaches. In stead, we are often limited to a very small number of measurements as post-Silicon characterization suffers from two major issues: (1) a limited number of replicated devices under test (DUTs) per die due to limited area and pads for on-chip monitor circuits; and (2) a limited number of training dies are measured due to test time limitations.

To address this issue, feature selection algorithms has been introduced to eliminate variables irrelevant to the targeted circuit performance. Least-angle regression (LAR) adds L_1-norm regularization terms to error functions and results in sample complexity logarithmic in the number of features [45, 46]. On the other hand, an L_2 regularization results in sample complexity that is linear in the number of features.

This chapter proposes an efficient method to build statistically valid prediction models of circuit performance based on a small number of mixture measurements. The key idea is to exploit two types of physical correlation. The first is the one that exists between different groups of performance measurements. We exploit this correlation by a technique named physical subspace projection that maps different groups of on-chip measurements onto an unique likelihood subspace spanned by a set of physical variables of the MIT virtual source (MVS) transistor model. As a ultra-compact transistor model with 23 parameters, the MVS model has a benefit in that most of its

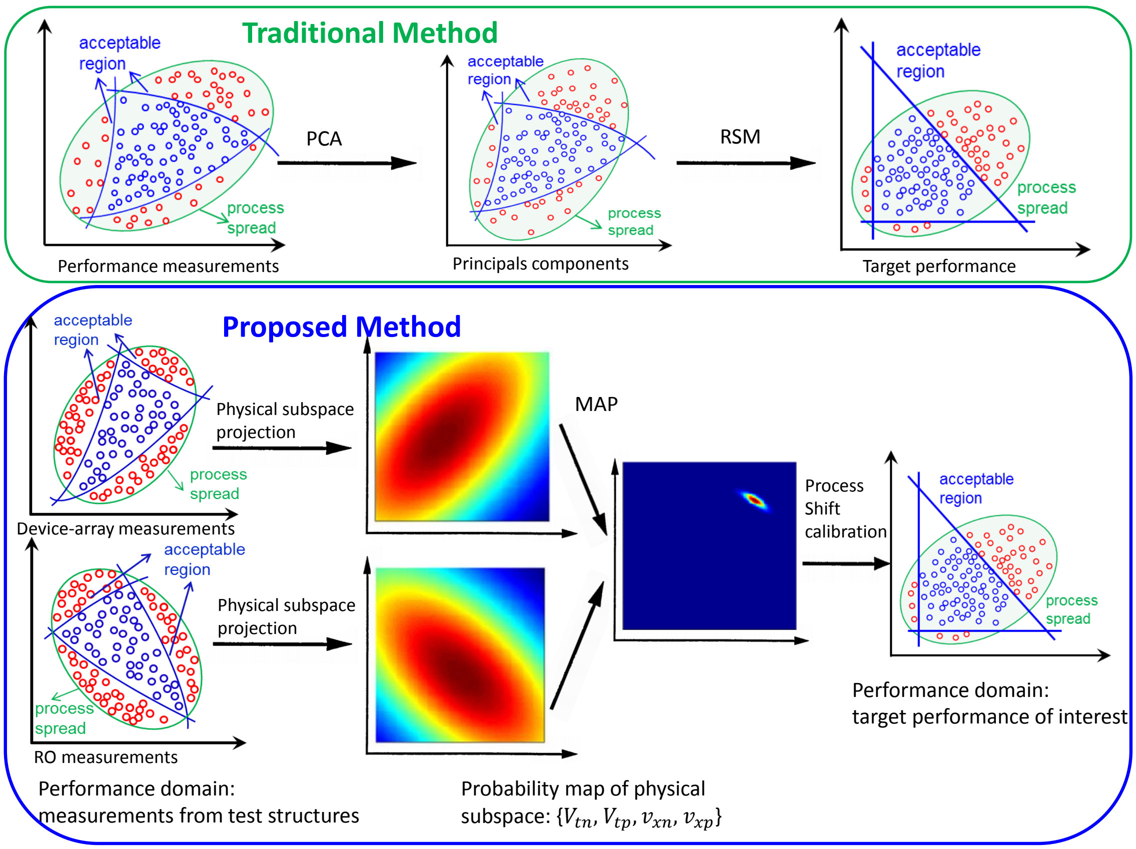


Figure 3-2: Proposed method: physical subspace projection, maximum a posteriori estimation and process shift calibration v.s. traditional method: PCA and RSM

parameter are physical. A prior distribution is defined over these subspace variables and a Bayesian formalism is introduced to estimate the performance parameters. This is achieved using maximum a posteriori (MAP) estimation defined over all the group measurement distributions and the subspace variable prior. A modification of an expectation-maximization (EM) algorithm is employed to iteratively solve the MAP estimation problem. The second physical correlation exists between SPICE simulation and measurements, and a technique named process shift calibration is introduced to estimate circuit performance offset between SPICE simulation and measurements. Compared with the traditional PCA and RSM method, the proposed method reserves the physical link between MVS model variables and measurements, such that the required number of measurements is greatly reduced for model establishment, as illustrated conceptually in Figure 3-2.

3.2 Background and Problem Definition

Without loss of generality, we consider the problem of estimating a single performance of interest, denoted by g. Assume that g follows a Gaussian distribution $g \sim \mathcal{N}(\mu_{\mathbf{g}}, \sigma_g)$:

$$pdf(g) = \frac{1}{\sqrt{2\pi} \cdot \sigma_g} \cdot exp[-\frac{(g - \mu_g)^2}{2 \cdot \sigma_g^2}] \tag{3.1}$$

where μ_g and σ_g are, respectively, the mean and standard deviation of the performance distribution. However, due to constraints on testing costs, measurements of g may not be directly available. Instead, groups of measurement data of other performance (usually electriccal) parameters are provided that we denote by $\mathbf{F} = \{F_1, F_2, ..., F_m\}$.

As an example, consider the problem of post-Silicon validation of a digital system. In this application, the performance metric g might be critical path delays or leakage power across a die and F_i would be measurement results from on-chip monitoring arrays (e.g., threshold voltages, I_{dsat} for transistors, frequencies for ring oscillators (ROs) [74, 76]). Here the variability of g is mainly caused by process and operating parameter variations such as V_{th} and V_{dd}. Our task therefore is to predict the distribution of g given $\mathbf{F}$ and, consequently, predict the parametric yield.

To formalize the above description, we define a measurement *group* to be a performance measured under a certain circuit topology and configuration. We assume that there are m such groups. To each group i ($i \in [1, m]$), we associate a random variable F_i to model the variability of the measurement under a certain circuit configuration. Therefore the aforementioned $\mathbf{F}$ could be represented by $\{F_1, F_2, ..., F_m\}$. We also assume that each F_i follows a Gaussian distribution $F_i \sim \mathcal{N}(\mu_{F_i}, \sigma_{F_i})$:

$$pdf(F_i) = \frac{1}{\sqrt{2\pi} \cdot \sigma_{F_i}} \cdot exp[-\frac{(F_i - \mu_{F_i})^2}{2 \cdot \sigma_{F_i}^2}] \tag{3.2}$$

For each group F_i, we obtain a set of independent observations $\{F_i\} = \{F_i^{(1)}, F_i^{(2)}, ..., F_i^{(N_i)}\}$, where N_i is the sample size of the i-th group. The problem we aim to address is to estimate μ_g and σ_g given the observations $\{F_1, F_2, ..., F_m\}$ with the constraint that N_i are very small. For simplicity, we consider the case where $N_1 = N_2 = ... = N_m = N$.

This problem cannot be addressed by the conventional moment estimation techniques because it is hard to assign a weight to each group and because the relationships between g and the F_i's are unclear [77]. One possible approach is to apply principal component analysis (PCA) to $\mathbf{F}$ and select its top features $\mathbf{X}$. The rest of the problem is then converted into a performance modeling problem. The performance function could then be approximated as:

$$g(\Delta\mathbf{X}) = \sum_{k=1}^{M} \alpha_{gk} \cdot b_k(\Delta\mathbf{X}) \tag{3.3}$$

where $\{b_k(\Delta\mathbf{X}); k = 1, 2, ..., M\}$ contains the basis functions (e.g, linear, quadratic, etc.), and $\{\alpha_{gk}; (k = 1, 2, ..., M)\}$ are the model coefficients. The unknown model coefficients α_{gk} are usually determined by solving a linear system with N sampling points:

$$\mathbf{G} = \mathbf{B} \cdot \alpha_g \tag{3.4}$$

where

$$\alpha_g = [\alpha_{g1} \quad \alpha_{g2} \quad ... \quad \alpha_{gM}]^T \tag{3.5}$$

$$\mathbf{G} = [G^{(1)} \quad G^{(2)} \quad ... \quad G^{(N)}]^T \tag{3.6}$$

$$G^{(i)} = g(\Delta \mathbf{X}^{(i)})$$

$$\mathbf{B} = \begin{bmatrix} b_1(\Delta \mathbf{X}^{(1)}) & b_2(\Delta \mathbf{X}^{(1)}) & \cdots & b_M(\Delta \mathbf{X}^{(1)}) \\ b_1(\Delta \mathbf{X}^{(2)}) & b_2(\Delta \mathbf{X}^{(2)}) & \cdots & b_M(\Delta \mathbf{X}^{(2)}) \\ \vdots & \vdots & & \vdots \\ b_1(\Delta \mathbf{X}^{(N)}) & b_2(\Delta \mathbf{X}^{(N)}) & \cdots & b_M(\Delta \mathbf{X}^{(N)}) \end{bmatrix} \tag{3.7}$$

However, the relationship between $\mathbf{X}$ and $\mathbf{G}$ is unknown and we have no prior information on $\{\alpha_{gk}; (k = 1, 2, ..., M)\}$. Under the constraint of very small N, strong over-fitting would appear and the prediction would be unreliable. Although least-angle regression (LAR) or sparse regression could add a regularization term on $\{\alpha_{gk}; (k = 1, 2, ..., M)\}$, an appreciable number of samples is still required. This is the main motivation for the development of a new performance estimation method via physical subspace projection and maximum a posteriori (MAP) estimation. In contrast to PCA, we project $\mathbf{F}$ onto a physical variable subspace $\mathbf{X}$, and Bayesian inference is used to learn a prior distribution on the $\mathbf{X}$ parameters using measurement data from all the groups. The estimates of μ_g and σ_g are also obtained via a projection onto the $\mathbf{X}$ subspace, and the projection operation itself is facilitated using the MVS MOSFET model [52].

3.3 Physical Subspace Projection

As discussed in Section 2.2, the MVS model is an ultra compact, charge-based MOSFET model that provides a simple, physics-based description of carrier transport in modern short-channel MOSFETs [52, 54, 55]. It substitutes the quasi-ballistic carrier transport concept for the concept of drift-diffusion with velocity-saturation. In doing so, it achieves excellent accuracy for the I-V and C-V characteristics of the device throughout the various domains of circuit operation. The number of parameters needed is considerably fewer (19 for DC and 23 in total) than in conventional models [64], making it attractive for our goal of modeling with very few measurement points.

Another feature of the MVS model is that most of its parameters are physical and can be related strongly to well-chosen measurement points. In Chapter 2, we described a statistical extension of MVS with the capability of mapping the variability characterization in device behavior onto a limited number of underlying model parameters, which in turn enables the efficient prediction of variations in circuit performance [60].

3.3.1 Definition of Physical Subspace

We define *physical subspace* as a variable space spanned by model parameters in the MVS model (e.g., V_{tn}, V_{tp}, etc.). Notice that model parameters are different from measured parameters. For example, V_t is commonly measured through the so-called "constant current method" where threshold voltage is the gate bias corresponding to an arbitrary value of drain current, for instance $0.1\mu A$ [78]. Such measured V_t relates to factors such as transistor geometries and configuration of devices under test (DUTs). Hence its absolute value does not have unique physical meaning. An MVS model parameter, in contrast, is a physical parameter with fixed value shared by all transistors with different geometries.

Although parameters measured from different groups have large differences in their absolute values, they are strongly correlated. This assertion is not only valid for the same parameter measured from different configurations (e.g., V_t measurement for transistors with different geometries), but is often also valid for different parameters measured from different configurations (e.g., I_{dsat} for a transistor and frequency for a ring oscillator (RO)). Fig. 3-3 shows different groups of on-chip monitoring measurements. All parameters in the red box refer to parameters that are directly measurable. They are governed by a hidden model parameter, namely, V_{tn} (other hidden parameters and their link to measured parameters are not shown in this figure).

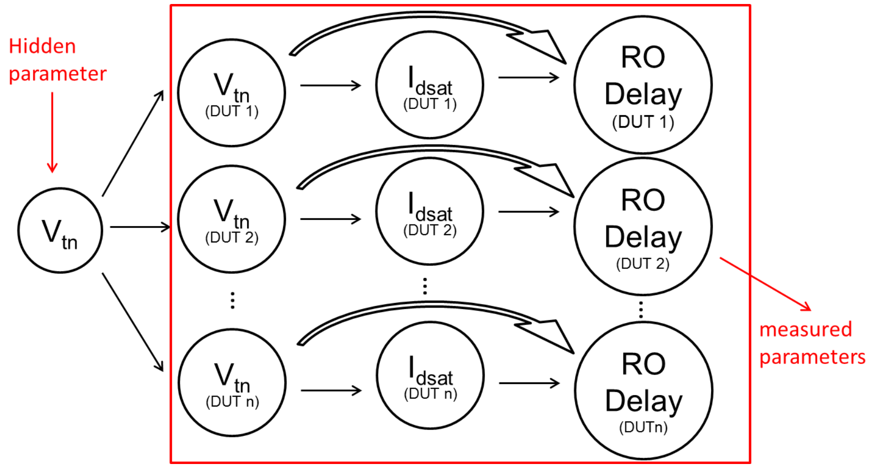


Figure 3-3: A graphical model linking hidden (internal) model parameters and correlated measured parameters for parameter correlations.

3.3.2 Physical Subspace Selection

The selection of physical subspace $\mathbf{X}$ is a key step in physical subspace projection. Here we propose a least-angle regression (LAR) method to solve this feature selection problem based on simulation, prior to using measurement results. A set of MVS model parameters $\mathbf{Y} = \{Y_1, Y_2, ..., Y_s\}$ are preselected as candidate subspace variables. Then Monte-Carlo simulations are run to compute target performance g by randomly generating samples of each MVS model parameter. $\mathbf{X}$ is initially set to be $\{\oslash\}$. Next, LAR finds the vector Y_{si} that is most correlated with g. Once Y_{si} is identified, Y_{si} is removed from $\mathbf{Y}$ and added to $\mathbf{X}$. The model coefficient α is determined by solving the linear equation $\mathbf{G} = \Delta\mathbf{X} \cdot \alpha$ and the residual of the approximation is calculated by:

$$Res = \mathbf{G} - \Delta\mathbf{X} \cdot \alpha \tag{3.8}$$

Then a new vector Y_{new} which has the largest correlation with the residual Res is found. The whole process is repeated until Res is smaller than a given threshold. For performance (e.g., delay, power, etc.) of a typical digital system, we found that $\mathbf{X} = \{V_{tn}, V_{tp}\}$ would be sufficient for a Res criterion of $0.02\mathbf{G}$ and $\mathbf{X} = \{V_{tn}, V_{tp}, v_{xon}, v_{xop}\}$ would be sufficient for a Res criterion of $0.01\mathbf{G}$, where v_{xon} and v_{xop} are the virtual source velocity for NMOS and PMOS, respectively. This is consistent with intuition that V_{tn} and V_{tp} are dominant variation because random dopant fluctuation and channel length variability are highly important physical sources of performance variation. For simplicity and visualization purposes, we select $\mathbf{X} = \{V_{tn}, V_{tp}\}$ for the rest of this chapter.

3.3.3 Physical Subspace Projection

The purpose of *physical subspace projection* is to transfer measurement data from different groups into a unique physical subspace $\mathbf{X}$. This is a one-to-many function that cannot be resolved using deterministic methods. However, given coefficients α, we can calculate the *pdf* on $\mathbf{X}$, and maximize the joint likelihood of each sample using maximum a posteriori (MAP) estimation.

In line with our previous assumptions, the subspace $\mathbf{X}$ satisfies a multivariate Gaussian distribution $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\theta})$:

$$pdf(\mathbf{X}) = \frac{1}{\sqrt{(2\pi)^k \, |\boldsymbol{\theta}|}} \cdot exp[-\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})^{\mathbf{T}} \boldsymbol{\theta}^{-1} (\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})] \tag{3.9}$$

where $\boldsymbol{\mu}_{\mathbf{X}}$ is the mean vector of $\mathbf{X}$, $\boldsymbol{\theta}$ is the covariance matrix of $\mathbf{X}$ that captures the intra-die variation and correlation of MVS parameters, and k is the dimension of $\mathbf{X}$.

We also assume that the "uncertainty" of $\boldsymbol{\mu}_{\mathbf{X}}$ follows a conjugate Gaussian prior distribution $\boldsymbol{\mu}_{\mathbf{X}} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, where $\boldsymbol{\Sigma}_0$ is introduced to capture the covariance of MVS parameters under inter-die variation.

$$pdf(\boldsymbol{\mu}_{\mathbf{X}}) = \frac{1}{\sqrt{(2\pi)^k \, |\boldsymbol{\Sigma}_{\mathbf{0}}|}} \cdot exp[-\frac{1}{2}(\boldsymbol{\mu}_{\mathbf{X}} - \boldsymbol{\mu}_0)^{\mathbf{T}} \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\mu}_{\mathbf{X}} - \boldsymbol{\mu}_0)] \tag{3.10}$$

Here, μ_{F_i} and σ_{F_i} are calculated by:

$$\mu_{F_i} = \mu(F_i(\Delta\mathbf{X})) = \sum_{k=1}^{M} \alpha_{F_i k} \cdot \mu(b_k(\Delta\mathbf{X})) = \mu_{F_i}(\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\theta}) \tag{3.11}$$

$$\sigma_{F_i}^2 = \sum_{j=1}^{M}\sum_{k=1}^{M}\alpha_{F_ij}\alpha_{F_ik}\cdot\sigma(b_j(\Delta\mathbf{X}),b_k(\Delta\mathbf{X})) - (\sum_{k=1}^{M}\alpha_{F_ik}\cdot\mu(b_k(\Delta\mathbf{X})))^2$$
$$= \sigma_{F_i}^2(\boldsymbol{\mu}_{\mathbf{X}},\boldsymbol{\theta}) \tag{3.12}$$

where $\mu(b_k(\Delta\mathbf{X}))$ and $\sigma(b_j(\Delta\mathbf{X}),b_k(\Delta\mathbf{X}))$ are the mean and covariance of the basis function, respectively.

Therefore the probability of observing data point $F_i^{(n_i)}$ in the ith group associated with the subspace distribution is

$$pdf(F_i^{(n_i)}|\boldsymbol{\mu}_{\mathbf{X}},\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma_{F_i}(\boldsymbol{\mu}_{\mathbf{X}},\boldsymbol{\theta})}exp[-\frac{(F_i^{(n_i)}-\mu_{F_i}(\boldsymbol{\mu}_{\mathbf{X}},\boldsymbol{\theta}))^2}{2\cdot\sigma_{F_i}(\boldsymbol{\mu}_{\mathbf{X}},\boldsymbol{\theta})^2}] \tag{3.13}$$

which is the complete form of physical subspace projection.

3.4 Maximum A Posteriori Estimation

Our proposed physical subspace projection method is facilitated by Bayesian inference approaches which efficiently exploit the correlation between different groups of measurements to improve the accuracy of the estimator.

3.4.1 Initial setting

Before we start, a proper physical variable subspace $\mathbf{X}$ is selected and prior knowledge is learned by fitting α_g and α_{F_i}, as described in Section 3.3.2. Initial guesses of parameters $\boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\theta}$ are also be selected; $\boldsymbol{\mu}_0$ is the nominal value for the subspace variables and $\boldsymbol{\Sigma}_0$ is the covariance matrix of MVS subspace variables under inter-die variation. The initial value of $\boldsymbol{\theta}$ equals the covariance of MVS subspace variables under only intra-die variation.

3.4.2 Learning a Prior Distribution

The first step is to project each sample in different measurement groups $\{\{F_i^{(n_i)}; n_i = 1,2,...,N_i\}; i = 1,2...,m\}$ to the selected subspace $\mathbf{X}$ and obtain the probability of observing $\mathbf{F_i}$ given $\boldsymbol{\mu}_{\mathbf{X}}$ and $\boldsymbol{\theta}$. Then we further combine the probability with the prior distribution $pdf(\boldsymbol{\mu}_{\mathbf{X}})$ in (3.10) to accurately estimate $\boldsymbol{\mu}_{\mathbf{X}}$ and $\boldsymbol{\theta}$, and through those parameters obtain μ_g and σ_g.

Assuming that the sampling process for different measurement groups are i.i.d., we can write the likelihood function $pdf(\mathbf{F}|\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\theta})$ as:

$$pdf(\mathbf{F}|\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\theta}) = \prod_{i=1}^{m} pdf(F_i|\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\theta}) \tag{3.14}$$

Similarly, assuming that the sampling process for samples in the same measurement groups are i.i.d., the likelihood function $pdf(F_i|\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\theta})$ is written as:

$$pdf(F_i|\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\theta}) = \prod_{n_i=1}^{N_i} pdf(F_i^{(n_i)}|\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\theta}) \tag{3.15}$$

According to Bayes' theory, the joint distribution $pdf(\mathbf{F}, \boldsymbol{\mu}_{\mathbf{X}}|\boldsymbol{\theta})$ is given by the product of the prior $pdf(\boldsymbol{\mu}_{\mathbf{X}})$ and the likelihood function $pdf(\mathbf{F}|\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\theta})$, giving us the *posterior distribution*:

$$pdf(\mathbf{F}, \boldsymbol{\mu}_{\mathbf{X}}|\boldsymbol{\theta}) = pdf(\boldsymbol{\mu}_{\mathbf{X}}|\boldsymbol{\theta}) \cdot pdf(\mathbf{F}|\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\theta}) \tag{3.16}$$

Substituting (3.14) and (3.15) into (3.16) and noticing that $pdf(\boldsymbol{\mu}_{\mathbf{X}}|\boldsymbol{\theta}) = pdf(\boldsymbol{\mu}_{\mathbf{X}})$ gives:

$$\begin{aligned} pdf(\mathbf{F}, \boldsymbol{\mu}_{\mathbf{X}}|\boldsymbol{\theta}) &= pdf(\boldsymbol{\mu}_{\mathbf{X}}) \cdot \textstyle\prod_{i=1}^{m} \prod_{n_i=1}^{N_i} pdf(F_i^{(n_i)}|\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\theta}) \\ &= pdf(\boldsymbol{\mu}_{\mathbf{X}}) \cdot pdf(F_1^{(n_1)}|\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\theta})) \cdot \ldots \cdot pdf(F_m^{(N_m)}|\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\theta}) \end{aligned} \tag{3.17}$$

This demonstrates the sequential nature of Bayesian learning in which the current posterior distribution forms the prior when a new data point is observed. Fig. 3-4 shows the results of Bayesian learning on $\boldsymbol{\mu}_{\mathbf{X}}$ as the portfolio of the measurement groups expands. The first column of this figure corresponds to the situation before any data points are observed, and shows a plot of the prior distribution $\boldsymbol{\mu}_{\mathbf{X}} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$; again, we are using $\mathbf{X} = \{V_{tn}, V_{tp}\}$ as our simplified basis function to enable visualization of the method. The first row shows the likelihood function $pdf(F_i|\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\theta})$

for different individual measurements, taken alone. The second row shows posterior distribution $pdf(\boldsymbol{\mu}_{\mathbf{X}}, \mathbf{F}|\boldsymbol{\theta})$ obtained by multiplying its likelihood function from the top row by the prior. As this process continues, the posterior distribution becomes much sharper and in the limit of an infinite number of data points, the posterior distribution would become a delta function centered on the true parameter values.

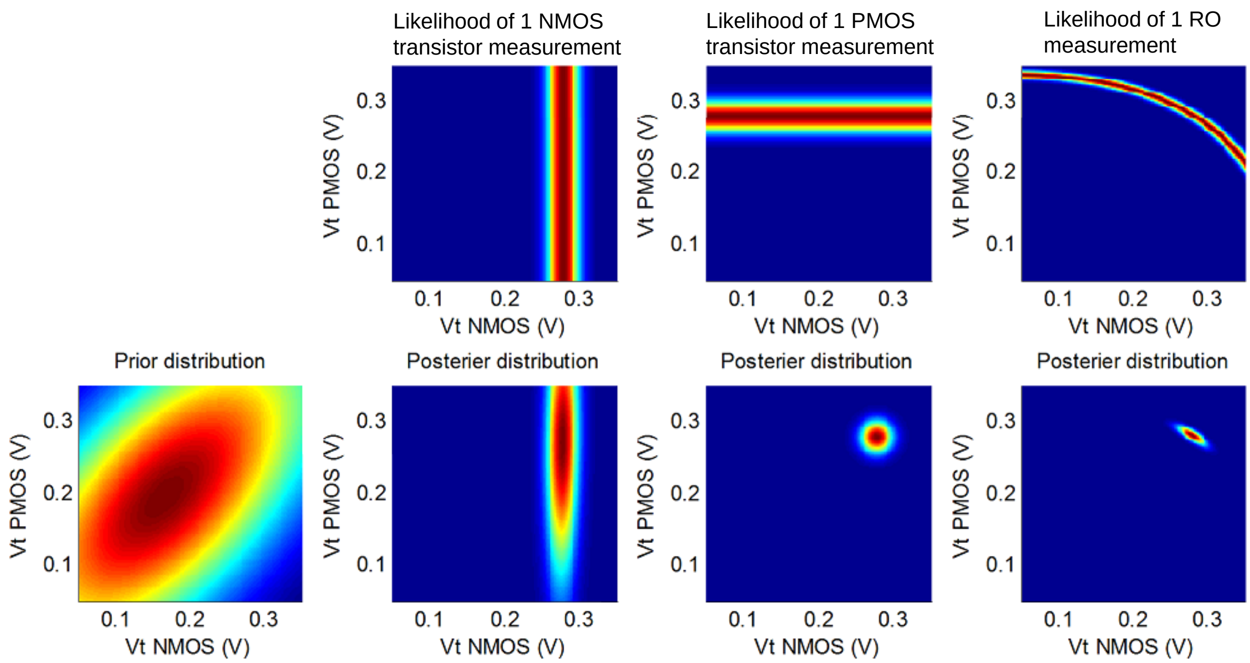


Figure 3-4: Illustration of sequential Bayesian learning of $\boldsymbol{\mu}_{\mathbf{X}}$ from prior and on-chip monitor circuits.

3.4.3 Maximum A Posteriori Estimation

Our final goal is to find an optimal estimation of $\boldsymbol{\mu}_{\mathbf{X}}$ which maximizes the log likelihood of posterior distribution $\ln pdf(\boldsymbol{\mu}_{\mathbf{X}}, \mathbf{F}|\boldsymbol{\theta})$. However, a key step is still missing, which is to determine the hidden variable $\boldsymbol{\theta}$ which maximizes the log likelihood function

$$\mathbf{ln} pdf(\mathbf{F}|\boldsymbol{\theta}) = \mathbf{ln} \int_{\mathbf{X}} pdf(\boldsymbol{\mu}_{\mathbf{X}}, \mathbf{F}|\boldsymbol{\theta}) d\boldsymbol{\mu}_{\mathbf{X}} \tag{3.18}$$

The difficulty arises from the presence of integration that appears inside the logarithm in (3.18), so that the logarithm function no longer acts directly on the Gaussian. If we set the derivatives of the log likelihood to zero, we will no longer obtain a closed form solution. The idea presented in this chapter follows the expectation maximization (EM) algorithm [79].

For any normalized distribution $q(\boldsymbol{\mu}_{\mathbf{X}})$, we have

$$\begin{aligned}
&\mathbf{ln}pdf(\mathbf{F}|\boldsymbol{\theta}) = 1 \cdot \mathbf{ln}pdf(\mathbf{F}|\boldsymbol{\theta}) = \int_{\mathbf{X}} q(\boldsymbol{\mu}_{\mathbf{X}})d\boldsymbol{\mu}_{\mathbf{X}} \cdot \mathbf{ln}pdf(\mathbf{F}|\boldsymbol{\theta}) \\
&= \int_{\mathbf{X}} q(\boldsymbol{\mu}_{\mathbf{X}})\mathbf{ln}pdf(\mathbf{F}|\boldsymbol{\theta})d\boldsymbol{\mu}_{\mathbf{X}} = \int_{\mathbf{X}} q(\boldsymbol{\mu}_{\mathbf{X}})\mathbf{ln}\tfrac{pdf(\boldsymbol{\mu}_{\mathbf{X}},\mathbf{F}|\boldsymbol{\theta})}{pdf(\boldsymbol{\mu}_{\mathbf{X}}|\mathbf{F},\boldsymbol{\theta})}d\boldsymbol{\mu}_{\mathbf{X}} \\
&= \int_{\mathbf{X}} q(\boldsymbol{\mu}_{\mathbf{X}})\left(\mathbf{ln}pdf(\boldsymbol{\mu}_{\mathbf{X}},\mathbf{F}|\boldsymbol{\theta}) - \mathbf{ln}q(\boldsymbol{\mu}_{\mathbf{X}}) - \mathbf{ln}\tfrac{pdf(\boldsymbol{\mu}_{\mathbf{X}}|\mathbf{F},\boldsymbol{\theta})}{q(\boldsymbol{\mu}_{\mathbf{X}})}\right)d\boldsymbol{\mu}_{\mathbf{X}}
\end{aligned} \tag{3.19}$$

Here the second item $-\int_{\mathbf{X}} q(\boldsymbol{\mu}_{\mathbf{X}})\mathbf{ln}q(\boldsymbol{\mu}_{\mathbf{X}})d\boldsymbol{\mu}_{\mathbf{X}}$ is always a constant. The third item $\int_{\mathbf{X}} q(\boldsymbol{\mu}_{\mathbf{X}})\mathbf{ln}\frac{pdf(\boldsymbol{\mu}_{\mathbf{X}}|\mathbf{F},\boldsymbol{\theta})}{q(\boldsymbol{\mu}_{\mathbf{X}})}d\boldsymbol{\mu}_{\mathbf{X}}$ is the Kullback-Leibler divergence between $pdf(\boldsymbol{\mu}_{\mathbf{X}}|\mathbf{F},\boldsymbol{\theta})$ and $q(\boldsymbol{\mu}_{\mathbf{X}})$ which is ≥ 0, with equality if and only if $q(\boldsymbol{\mu}_{\mathbf{X}}) = pdf(\boldsymbol{\mu}_{\mathbf{X}}|\mathbf{F},\boldsymbol{\theta})$.

Algorithm 1 Algorithm to solve maximum a posteriori estimation

Require: a joint distribution $pdf(\boldsymbol{\mu}_{\mathbf{X}},\mathbf{F}|\boldsymbol{\theta})$ over observed variables $\mathbf{F}$ and latent variables $\boldsymbol{\mu}_{\mathbf{X}}$, governed by parameters $\boldsymbol{\theta}$, convergence requirement ϵ.

Ensure: $\boldsymbol{\theta}$ which maximizes the likelihood function $\mathbf{ln}pdf(\mathbf{F}|\boldsymbol{\theta})$ and $\boldsymbol{\mu}_{\mathbf{X}}$ which maximizes the likelihood function $pdf(\boldsymbol{\mu}_{\mathbf{X}},\mathbf{F}|\boldsymbol{\theta})$

```
1: Choose an initial setting for the parameters θ^new;
2: repeat
3:     θ^old = θ^new;
4:     Evaluate pdf(μ_X|F, θ^old) = pdf(μ_X, F|θ^old) / pdf(F|θ^old);
5:     Evaluate θ^new given by
6:         θ^new = arg max_θ Q(θ, θ^old);
7:     Q(θ, θ^old) = ∫_X pdf(μ_X|F, θ^old) ln pdf(μ_X, F|θ) dX;
8: until |θ^old − θ^new| < ε
9: μ_X = arg max_{μ_X} ln pdf(μ_X, F|θ^new);
```

This suggests an iterative algorithm, as summarized in Algorithm 1. Given an initial value of $\boldsymbol{\theta}^{old}$, the first step is to maximize likelihood $\mathbf{ln}pdf(\mathbf{F}|\boldsymbol{\theta}^{old})$ with respect to $q(\boldsymbol{\mu}_{\mathbf{X}})$, which gives $q(\boldsymbol{\mu}_{\mathbf{X}}) = pdf(\boldsymbol{\mu}_{\mathbf{X}}|\mathbf{F},\boldsymbol{\theta}^{old})$. The second step is to fix the distribution $q(\boldsymbol{\mu}_{\mathbf{X}})$ and maximize $\mathbf{ln}pdf(\mathbf{F}|\boldsymbol{\theta}^{old})$ with respect to $\boldsymbol{\theta}^{old}$. The whole process is repeated until convergence and estimations of $\boldsymbol{\theta}$ and $\boldsymbol{\mu}_{\mathbf{X}}$ are obtained.

3.5 Process Shift Calibration and Circuit Performance Calculation

Once the physical subspace $\mathbf{X}$ and its corresponding basis functions $\mathbf{B} = \{b_k(\Delta\mathbf{X}); k = 1, 2, ..., M\}$ are obtained, we are able to determine model coefficients α_g and α_{F_i} by solving linear equations $\mathbf{G} = \mathbf{B} \cdot \alpha_g$ and $F_i = \mathbf{B} \cdot \alpha_{F_i}$, respectively. This allow us to build a one-to-many function from $\mathbf{X}$ to target performance g and measurements $\mathbf{F}$. In order to reuse prior information, to this point we assume that the coefficients α_g and α_{F_i} of post-layout simulations are identical with the α_g and α_{F_i} of measurement results. While this assumption usually holds in many practical applications, it is sometimes the case that there is mismatch between the nominal performance values of post-layout simulations and the measurements with a typical shift of 15% or less. The shifts in the corresponding performance distributions are due to modeling and extraction inaccuracy. Since only a very small sample size is needed to correct $\mathbf{F}_{nom}$ and g_{nom}, we are able to calibrate $\mathbf{F}$ and g with a simple correcting mean shift.

To further illustrate the calibration, Fig. 3-5 shows an example of RO stage delay measurements versus V_{tn} and V_{tp} extracted from same-die test arrays and compared with modeling prediction. A prediction of nominal performance without process shift calibration is also shown in the figure. Note that measurements in Fig. 3-5 are sampled from dies on various wafers and lots, and only a few dies (< 10) with both measurements from on-chip test structures and measurements from target performance are needed at the same time, to calibrate the nominal shift.

Finally, after building the link between physical subspace and performance, we are able to generate the performance maps for different systems. Fig. 3-6 shows the INV and NAND RO stage delay versus V_{tn} and V_{tp} after process shift calibration. A high similarity is observed between the two maps; this suggests that having the measurement result for one digital system would give us confidence in predicting other digital system performances.

Combining α_g and α_{F_i} with $\boldsymbol{\theta}$ and $\boldsymbol{\mu}_{\mathbf{X}}$ obtained after MAP, we estimate the mean

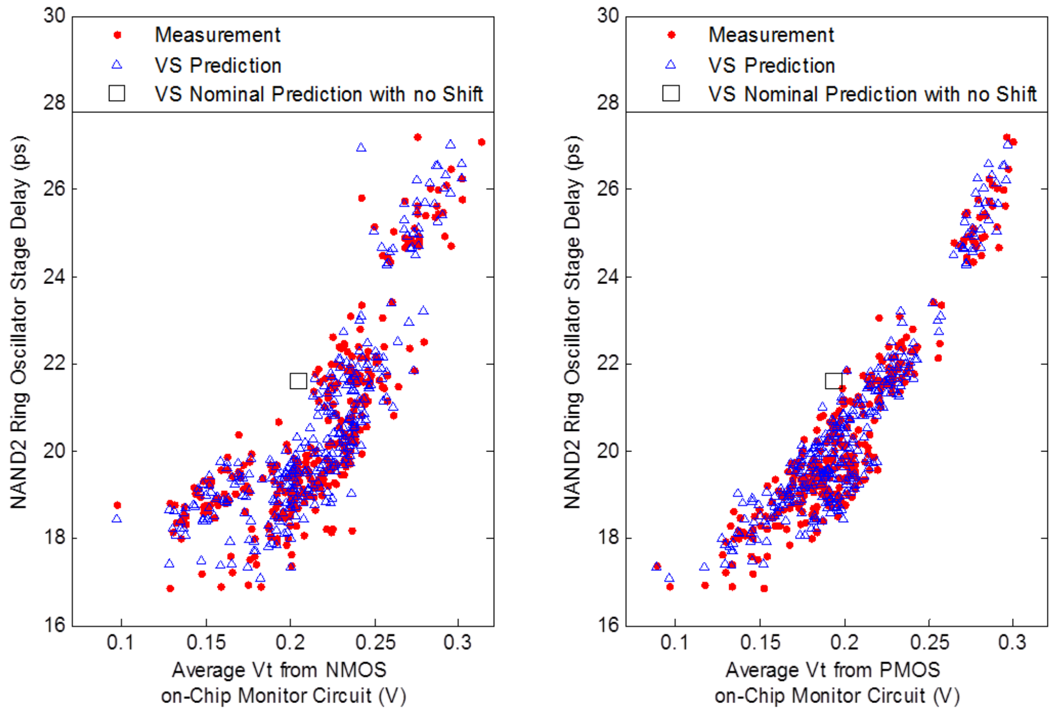


Figure 3-5: A comparison of measured and MVS model predicted ring oscillator (RO) stage delay versus (a) NMOS V_t, and (b) PMOS V_t. Nominal post-layout simulation without any shift and variation is marked as square.

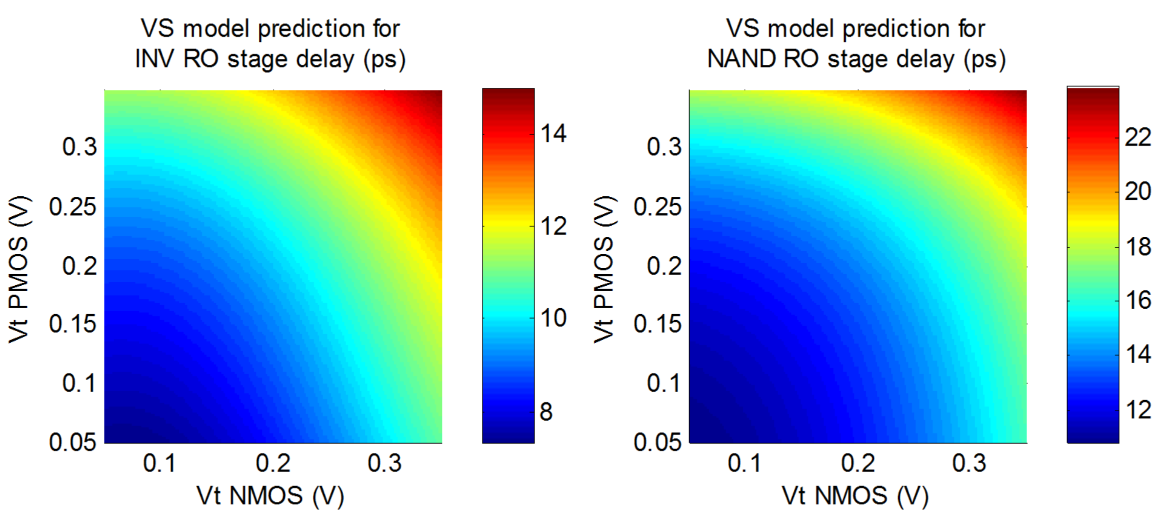


Figure 3-6: Sensitivity analysis on (a) INV, and (b) NAND2 ring oscillator (RO) stage delay using MVS model fit using proposed approach.

and standard deviation of target performance μ_g and σ_g:

$$\mu_g = \mu(g(\Delta\mathbf{X})) = \sum_{k=1}^{M} \alpha_{gk} \cdot \mu(b_k(\Delta\mathbf{X})) \tag{3.20}$$

$$\sigma_g^2 = \sigma^2(g(\Delta\mathbf{X})) = \sum_{i=1}^{M}\sum_{j=1}^{M} \alpha_{gi}\alpha_{gj} \cdot \sigma(b_i(\Delta\mathbf{X}), b_j(\Delta\mathbf{X})) - (\sum_{k=1}^{M} \alpha_{gk} \cdot \mu(b_k(\Delta\mathbf{X})))^2 \tag{3.21}$$

3.6 Summary

A summary of our proposed physical subspace projection method and maximum a posteriori estimation is shown in Fig. 3-7.

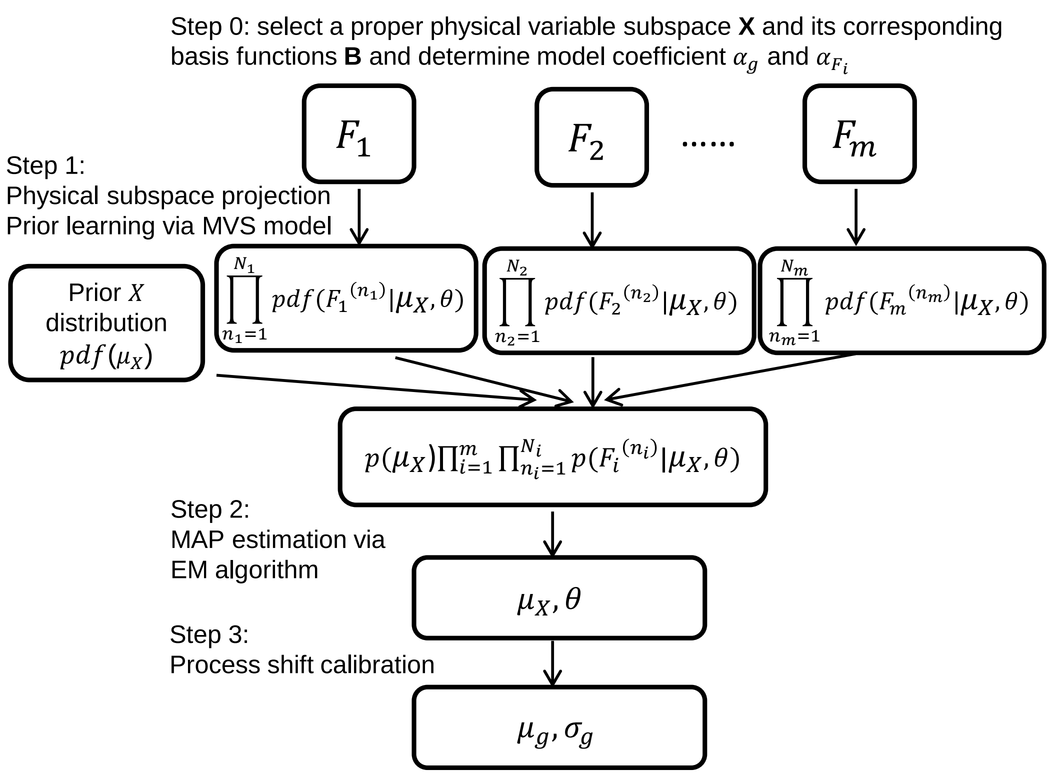


Figure 3-7: Proposed method employing Bayesian inference and maximum a posteriori estimation.

Before we start (step 0), a proper physical variable subspace $\mathbf{X}$ and its corresponding basis functions $\mathbf{B} = \{b_k(\Delta\mathbf{X}); k = 1, 2, ..., M\}$ are selected, and prior knowledge is learned by fitting α_g and α_{F_i}. The first step (step 1) is to perform physical subspace projection which maps different groups of on-chip measurements onto an unique likelihood subspace spanned by the set of physical variables of the MVS model. Then a prior distribution is defined over these subspace variables and a Bayesian formalism is introduced to estimate the performance parameters. The next step (step 2 in Fig. 3-7) is using maximum a posteriori (MAP) estimation defined over all the group measurement distributions and the subspace variable prior, which is achieved using an expectation-maximization (EM) algorithm. Finally, step 3 ultilizes process shift calibration to estimate circuit performance by combining SPICE simulation and very few new measurements.

Although both intra-die variation ($\boldsymbol{\theta}$) and inter-die variation ($\boldsymbol{\Sigma}_0$) are introduced with initial values from the design kit, it is worth noting that they are dealt with separately in this work. For intra-die variation ($\boldsymbol{\theta}$), different dies have different parameter values and these need to be updated through the EM algorithm. However, inter-die variation ($\boldsymbol{\Sigma}_0$) only serves as prior information and different dies share the same value. While the whole process may be repeated for many dies, we can also update $\boldsymbol{\Sigma}_0$ if necessary. Once we have updated $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}_0$, together with the extracted mean vector, we can estimate the parametric yield of the product.

3.7 Validation

In this section, we demonstrate the accuracy and efficacy of our proposed physical subspace projection and MAP algorithm using measurement results. We consider on-chip measurement results collected from 3186 dies in 27 wafers in a 28-nm bulk CMOS process. Each chip contains different test structures, including transistor-arrays and ring oscillator (RO)-arrays, which are often used as monitor circuits due to their simplicity and small area overhead. Configurations of transistor and RO test structures are summarized in Tables 3.1 and 3.2, respectively.

In group #1, $I-V$ curves of single NMOS and PMOS devices, which correspond to the driving strength of INV RO, are characterized. Similarly, $I-V$ curves of

Table 3.1: A summary of transistor-array test structures.

Measurement group #	1		2		3	
DUT	NMOS	PMOS	NMOS	PMOS	NMOS	PMOS
Connection	single	single	stacked	parallel	parallel	stacked
Measurements	I_{dsat},I_{off},...					
Replicas	4	4	4	4	4	4

Table 3.2: A summary of RO-array test structures.

Measurement group #	4	5	6
DUT	INV	NAND	NOR
Circuit topology	RO	RO	RO
Measurements	$frequency$		
Stages	97	97	97
Replicas	4	4	4

stacked NMOS and parallel PMOS transistors are characterized in group #2; and correspond to the driving strength of NAND RO, while $I - V$ curves of parallel NMOS and stacked PMOS transistors are characterized in group #3, corresponding to the driving strength of NOR RO. For validation purposes, we use one group of RO frequency as the performance of interest to mimic the operation of a digital system, while the rest of the measurement groups are used to train the model.

3.7.1 Comparison with a Naive Approach

As a first example for demonstration, we compare our proposed method with a naive approach where measurements from a single group are used to construct a response surface method (RSM) model. Fig. 3-8 shows relative error on group #6 frequency predictions as a function of N_i replicas on the same die. All 3186 dies are used for training and the number of samples per die varies. Our proposed method is able to achieve higher accuracy using multiple-group measurements, compared with RSM using single group measurements. We observe consistently 2x sample accuracy improvement over sample mean, which represents a 32.5x sample size reduction.

Table 3.3 shows cross-group validation errors in RO frequency predictions using the proposed physical subspace projection and MAP method, with different mix-

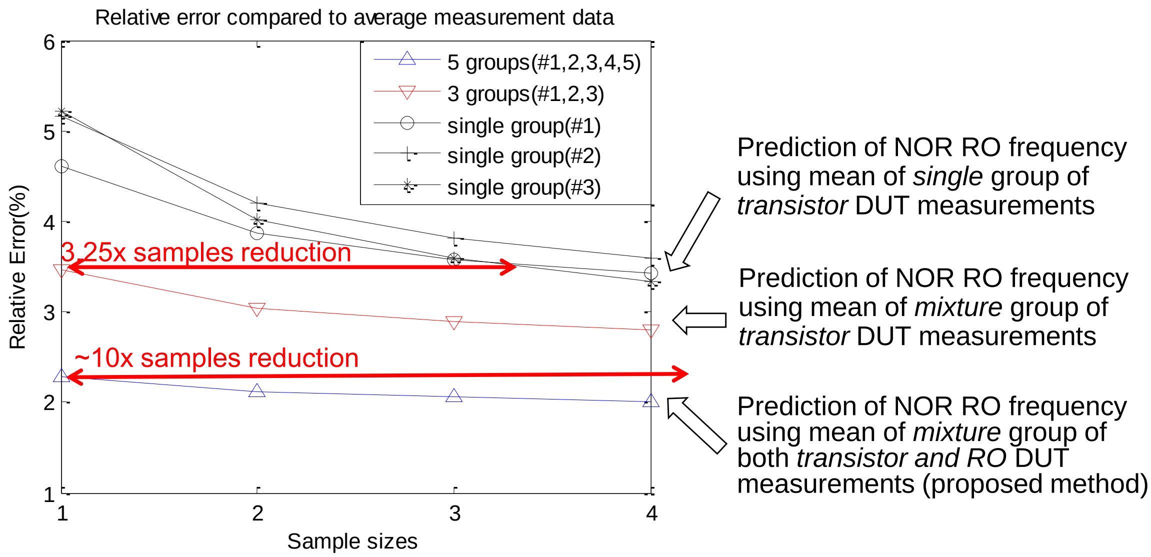


Figure 3-8: Relative prediction error for group #6 versus replicate samples per die. A mixture of measurement groups is compared.

Table 3.3: Relative prediction error for cross-group validation.

#	Measurement group	Prediction group	%error
1	1,2,3	4	3.22
2	1,2,3	5	2.99
3	1,2,3	6	2.70
4	5,6	4	2.40
5	4,6	5	2.19
6	5,6	4	3.54
7	1,2,3,4	5	2.26
8	1,2,3,4	6	2.17
9	1,2,3,5	5	2.26
10	1,2,3,5	6	2.06
11	1,2,3,6	4	2.32
12	1,2,3,6	5	2.15
13	1,2,3,4,5	6	2.10
14	1,2,3,4,6	5	1.98
15	1,2,3,5,6	4	2.01

tures of device- and RO-array measurements. As we compare cross-group prediction errors, we observe consistently smaller prediction errors as the number of different measurement groups grows.

Fig. 3-9 shows a prediction of INV RO frequency ultilizing the proposed method, generated using a mixture of training groups (but not including INV RO test structures) from on on-chip monitor data, compared with measurement of INV RO frequency. A high similarity is observed between the two wafer maps; having measurement results for on-chip test structures give us high confidence in predicting other digital system performances.

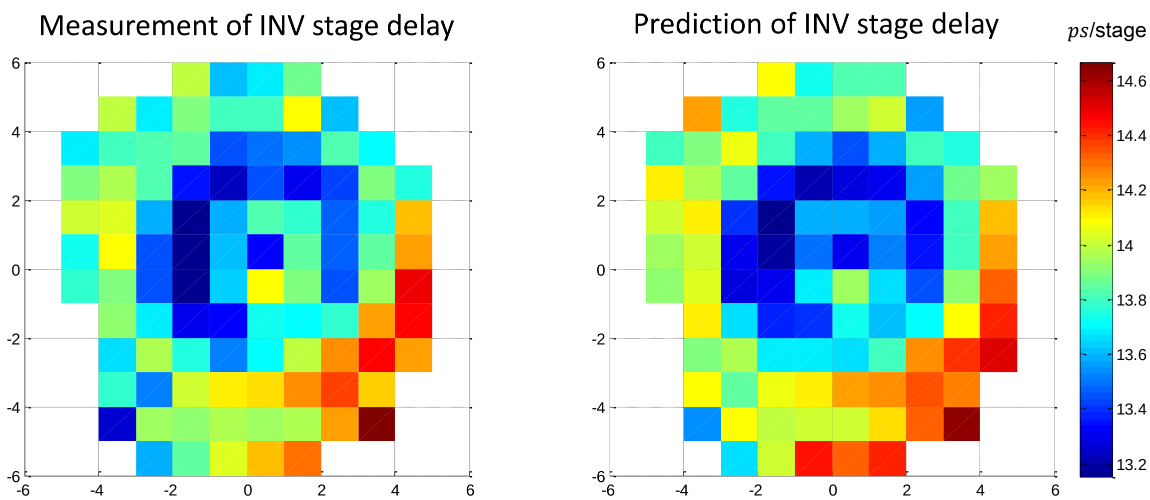


Figure 3-9: Wafer map comparison between measurement and proposed method prediction.

3.7.2 Comparison with PCA and RSM Approach

As a second example for demonstration, we compare our proposed method with the traditional PCA and RSM approach, where in both cases a mixture of measurement groups are used for model training. Fig. 3-10 shows how the average performance prediction error varies with the number of training dies for three different techniques: PCA and least-squares regression (LSR), PCA and least-angle regression (LAR), and proposed physical subspace projection (PSP) and MAP method. All replicas on each training dies are used, but the number of training dies varies in this case to mimic the situation where measurements for the target system are difficult to obtain. Both "PCA+LAR" and the proposed method require fewer training samples than "PCA+LSR", because they do not solve the unknown model coefficients from an over-determined equation. Meanwhile, our proposed "PSP+MAP" method is able to achieve substantially higher accuracy compared with "PCA+LSR" and "PCA+LAR", with 70x and 150x sample reduction, respectively. Here the measurements from device-array test structures and RO-array test structures are sampled from 27 wafers. Each time we select measurements from one or several wafers to train the model, and use measurements from the rest of the wafers to do validation. The whole process is repeated and the prediction error is averaged.

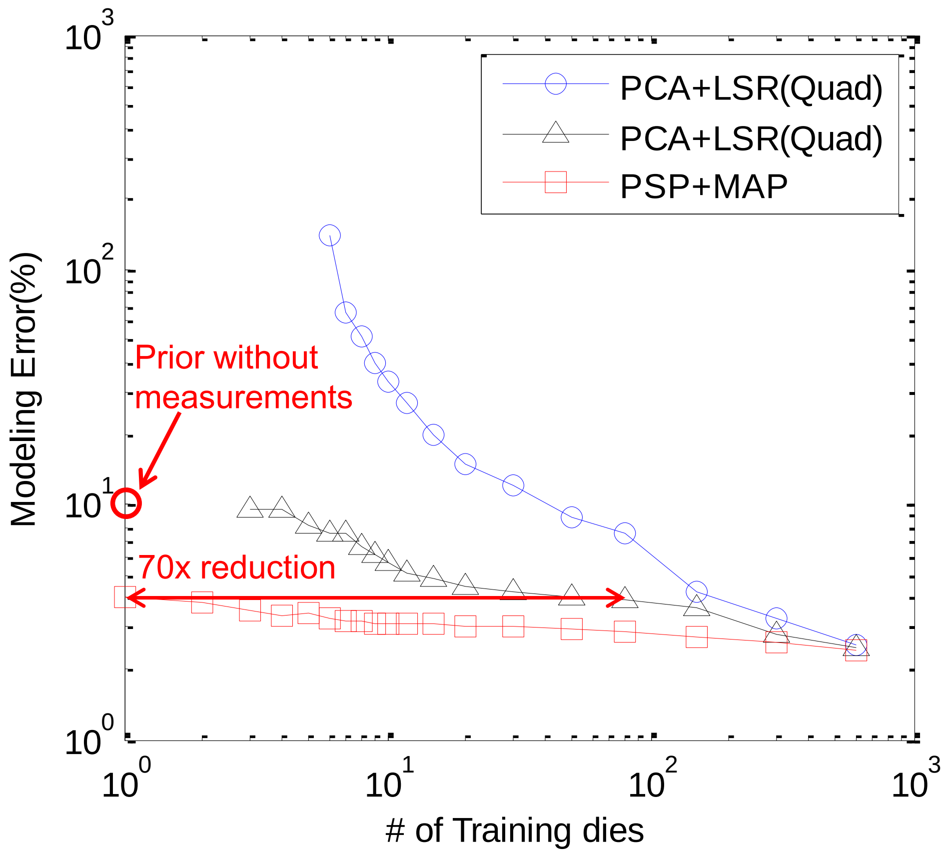


Figure 3-10: Relative prediction error for group #6 versus number of training dies. Various algorithms are compared. Candidate variables for "PCA+LSR" and "PCA+LAR" include both linear and quadratic items for variables after PCA process. Prior without any measurements for "PSP+MAP" is also labeled with an average modeling error of 9.5%

Fig. 3-11 shows how the average performance prediction error varies with the number of training dies for different starting prior errors: Similar prediction errors are observed for priors with 7.5% and 26.4% modeling error. However, a slightly larger error is observed for the latter case, as the final model parameters deviate more from the nominal (assumed prior) value.

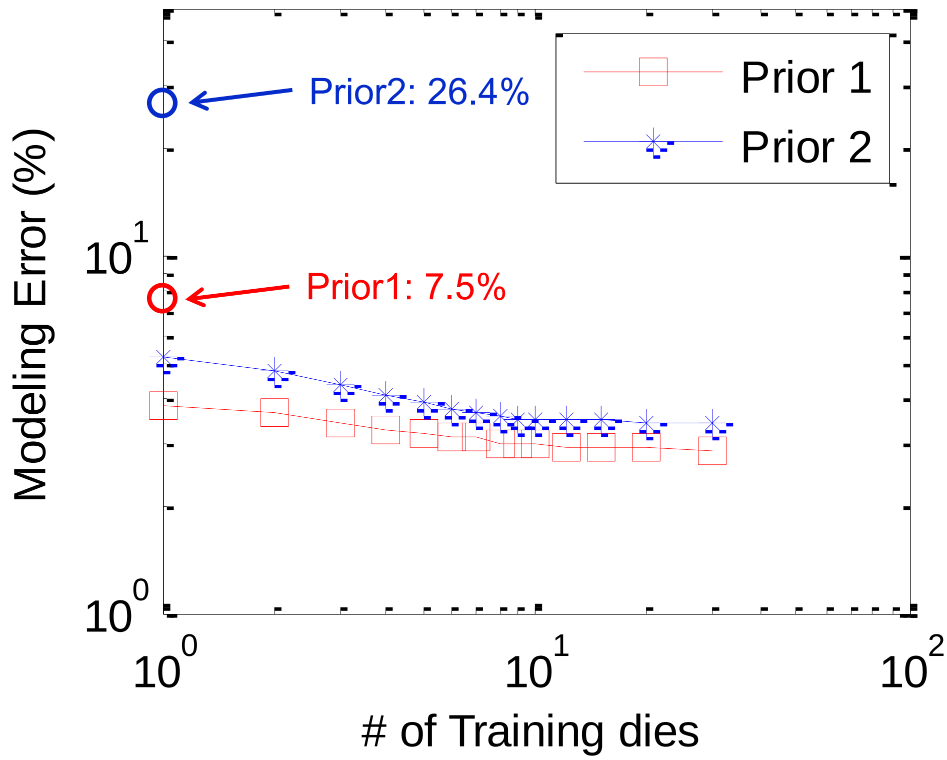


Figure 3-11: Relative prediction error for group #6 versus number of training dies. Different starting prior error cases are compared.

Chapter 4

Compact Model Parameter Extraction Using Incomplete New Measurements and a Bayesian Framework

4.1 Introduction

Continued scaling of CMOS technology has introduced new physical mechanisms for short-channel devices that significantly increase the number of parameters and the complexity of equations of compact transistor models. To be effectively used in circuit design simulations, all the many dozens of model parameters need to be carefully extracted from multiple test structures (e.g., $I-V$ structures, $C-V$ structures, ring oscillators, etc.) so that the model can accurately reproduce the transistor electrical characteristics. Usually the set of model parameters is divided into subsets of local and global parameters, where the local parameters apply to a single device dimension while global parameters apply to all relevant device geometries [80]. Therefore experimental data for devices with different geometries and replicas are needed to find the global set of parameters. The most widely used parameter extraction methods are based on the deterministic minimization of an error function between model output and measurement data, as noted in Chapter 1. Algorithms used to solve the optimization problem are either gradient-based (e.g., Levenberg-Marquardt) or gradient-free (e.g., Genetic Algorithm - GA). GA mimics the natural selection and evolution process and is more likely to find the global minimum [81]. All of these minimization methods are iterative, and defining an appropriate starting point and parameter bounds is of crucial importance and requires considerable experience. Furthermore, for a given set of measurements, there may be multiple minimizing solutions (sets of parameters that reproduce input-output data), and selecting the one most compatible with the physics of the device is a difficult task. The situation becomes even worse in the presence of significant measurement "noise" which introduces unavoidable errors in the extracted model parameter, thus compromising even further their physical significance.

Traditional silicon characterization and extraction flows suffer from 1) large area overhead due to the complexity of different test structures and transistor geometries; and 2) long testing time due to a very limited number of I/O ports through which all measurement data for the test structures have to be collected. This problem is further exacerbated in statistical parameter extraction as required for statistical IC analysis and optimization, e.g., statistical static timing analysis and post-silicon tuning. In nanoscale technologies, IC testing has contributed to a significant portion of the total manufacturing cost, to the point that it is now almost impossible to proceed with all $I-V$ measurements for every on-chip monitoring device on each die in a wafer. As discussed in Chapter 2, existing statistical parameter extraction methods such as the backward propagation of variance (BPV) [60, 82] are advantageous only when the number of measurements is larger than the number of model parameters. They also typically impose the stringent constraint that extracted parameters must be statistically uncorrelated. Such limitations mean, among several similar situations, that the correlated variations in sub-threshold swing (SS) and threshold voltage (V_{th0}) cannot be extracted at the same time.

In this chapter, we exploit recent advances in statistics and semiconductor metrology to develop a novel and unified MOSFET parameter extraction method for low-cost silicon testing and characterization. While the virtual probe described in [51] and in [83] focuses on reducing the number of measured dies needed to characterize spatial variation, our work focuses on reducing testing cost per die. Our new method is general, allows for missing $I-V$ measurements in the data set, removes the independence restriction on the model parameters and can be used to conduct both deterministic and statistical model parameter extraction. While our theory and algorithms are independent of the underlying transistor model (BSIM, PSP, EKV, MVS, etc.), we mainly use the MIT virtual source (MVS) model to illustrate the applicability of our work to deeply-scaled devices where the main mode of charge transport is quasi-ballistic. The intrinsic simplicity of the MVS model combined with the Bayesian inference [44] framework enables the statistical extraction of an entire parameter set using only six noisy $I-V$ measurements. A key step in this new method is maximum a posteriori (MAP) estimation where past $I-V$ measurements of older transistor technologies are used and learned to obtain a prior distribution on the parameter set along with its uncertainty matrix.

4.2 MVS Model and Parameters Revisited

As presented in Chapter 2, the MIT virtual source (MVS) model is an ultra compact, charge-based MOSFET model that provides a simple, physics-based description of carrier transport in modern short-channel MOSFET [52, 55, 54]. It utilizes a quasi-ballistic carrier transport concept rather than drift-diffusion with velocity saturation. In doing so, it achieves excellent accuracy for the $I-V$ and $C-V$ characteristics with continuity of current and its derivatives throughout all regions of operation. The MVS model has the advantage of using a limited number of input parameters, most of which have straightforward physical meanings and can be easily measured using traditional device characterization.

Fig. 4-1 shows the parameter extraction methodology for MVS described in detail in Chapter 1, with the key parameters extracted from $I-V$ measurements highlighted. Table 4.1 summarizes these key parameters along with their physical meaning.

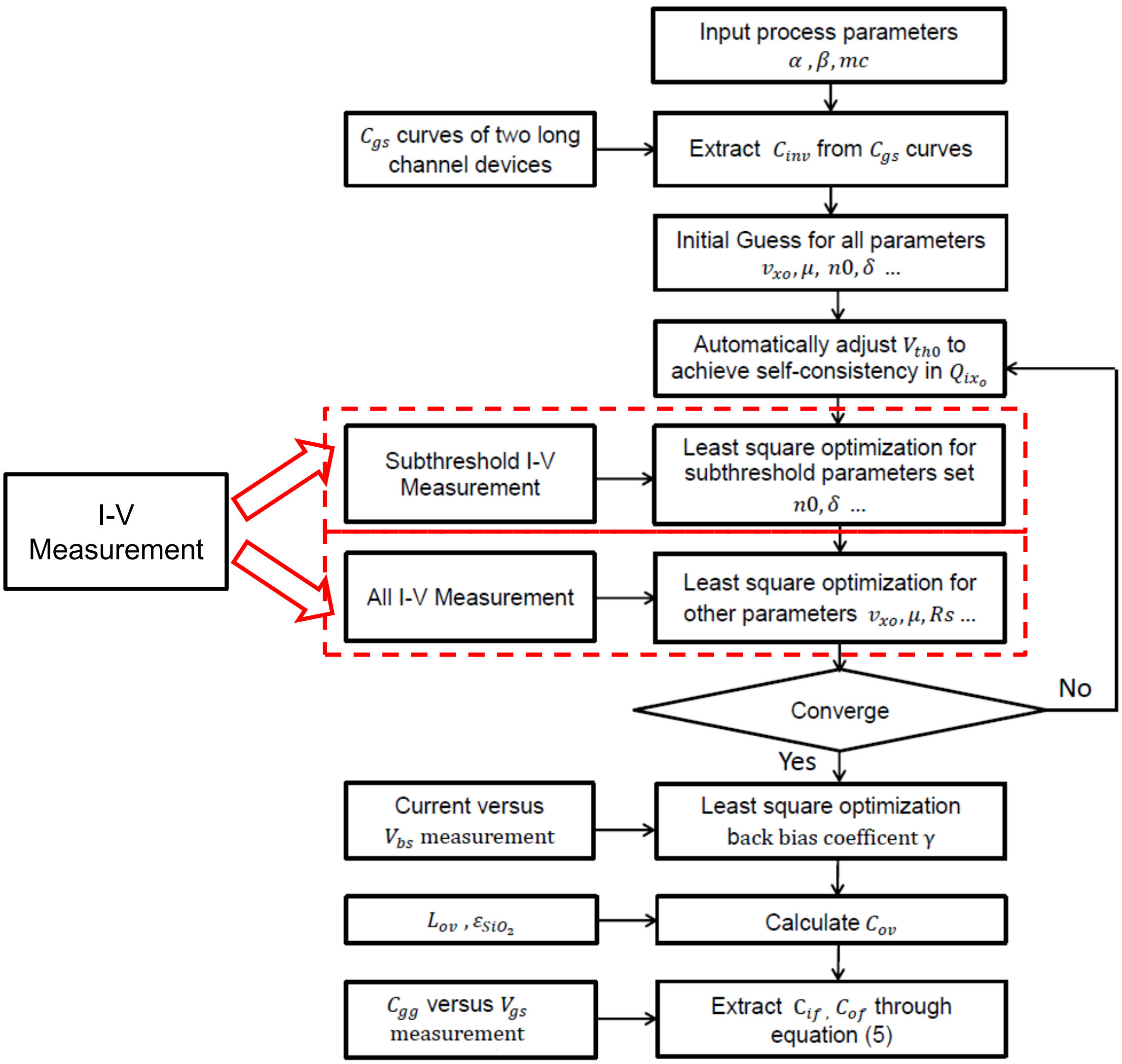


Figure 4-1: Optimization flow for I-V parameter extraction in the MVS model.

Table 4.1: Key parameters extracted with experimental data for the MVS model with physical meaning.

Parameters	Description
$V_{t0}(V)$	Strong inversion threshold voltage
n_0	Sub-threshold swing factor
$\delta(mV/V)$	Drain-induced barrier lowering
$v_{xo}(cm/s)$	Virtual source carrier velocity
$\mu(cm^2/V \cdot s)$	Low-field mobility
$Rs_0(ohm \cdot \mu m)$	Series resistance per side

The extracted parameters are divided into two groups: one for the sub-threshold region (V_{t0}, n_0 and δ) and one for the above-threshold region (v_{xo}, μ, Rs_0 and Rd_0). In traditional device characterization, each group is optimized separately using non-linear least-squares error minimization [64]. In this chapter, we propose an alternative approach to parameter extraction, using a Bayesian framework. The key additional idea is to use uncertainty about the *a priori* model to guide and improve parameter estimation given new data.

4.2.1 Problem Definition

To formalize the parameter extraction problem, we consider a measurement set of currents $\{F_1, ..., F_N\}$ with corresponding inputs $\mathbf{V} = \{\mathbf{V}^1, ..., \mathbf{V}^N\}$. We group the target variables $\{F_n\}$ into a vector that we denote by $\mathbf{F}$. Each input contains voltages from four terminals, $\mathbf{V^n} = \{V_g^n, V_d^n, V_s^n, V_b^n\}$. We define $\mathbf{P_{sub}} = \{V_{t0}, n_0, \delta\}$ as the sub-threshold parameters, and $\mathbf{P_{above}} = \{v_{xo}, \mu, Rs_0, Rd_0\}$ as the above-threshold region parameters. As discussed in detail in Chapter 2, we call the submanifolds of $\mathbf{P_{sub}}$ and $\mathbf{P_{above}}$ the *physical subspace* since each is a multidimensional surface parameterized with the physical parameters in the MVS model. The output of the MVS model would then be $f(\mathbf{V}, \mathbf{P_{sub}}, \mathbf{P_{above}})$. The problem we aim to address is to estimate $\mathbf{P_{sub}}$ and $\mathbf{P_{above}}$ given the observations $\{F_1, ..., F_N\}$, with the challenge that the size of the measurement set N is very small.

Based on the physics of transistor operation in the subthreshold regime, we assume that the measurement of subthreshold currents F_n follows a log-normal distribution:

$$\mathbf{ln}F_n \sim \mathcal{N}(\mathbf{ln}f(\mathbf{V}^n, \mathbf{P_{sub}}, \mathbf{P_{above}}), \beta_{\mathbf{ln}F_n}^{-1}) \tag{4.1}$$

where $\beta_{\mathbf{ln}F_n}$ is the precision (inverse variance) of $\mathbf{ln}F_n$.

Similarly, we assume that the above-threshold current F_n follows a Gaussian distribution:

$$F_n \sim \mathcal{N}(f(\mathbf{V}^n, \mathbf{P_{sub}}, \mathbf{P_{above}}), \beta_{F_n}^{-1}) \tag{4.2}$$

where β_{F_n} is the precision (inverse variance) for F_n.

The least-squares error function of sub-threshold and above-threshold regions are, respectively,

$$\mathcal{E}(\mathbf{P_{sub}}) = \frac{1}{2}\sum_{n=1}^{N}\{\mathbf{ln}(F_n) - \mathbf{ln}(f(\mathbf{V^n}, \mathbf{P_{sub}}, \mathbf{P_{above}}))\}^2 \tag{4.3}$$

$$\mathcal{E}(\mathbf{P_{above}}) = \frac{1}{2}\sum_{n=1}^{N}\{F_n - f(\mathbf{V^n}, \mathbf{P_{sub}}, \mathbf{P_{above}})\}^2 \tag{4.4}$$

4.3 Maximum A Posteriori Estimation

In this section, we present maximum a posteriori (MAP) estimation of $\mathbf{P_{above}}$ and $\mathbf{P_{sub}}$, where instead of minimizing the error function in (4.3) and (4.4), we will maximize the probability of observing $\mathbf{F}$.

4.3.1 Physical Subspace Projection

The purpose of *physical subspace projection* is to relate an observed measurement to an output of the physics-based model and use both to derive a probability distribution on the physical subspace [49]. In the transistor model extraction context, this means that we seek to relate current measurements at different voltage biases to the outputs in the sub-threshold and above-threshold physical subspace $\mathbf{P_{sub}}$ and $\mathbf{P_{above}}$, respectively. The *pdf*'s on $\mathbf{P_{sub}}$ or $\mathbf{P_{above}}$ can then be calculated and the parameter extraction problem solved using maximum a posteriori (MAP) estimation.

Without loss of generality, we describe the MAP estimation for sub-threshold parameters. Above-threshold parameters could be estimated in a similar manner. First, we assume that $\mathbf{P_{sub}}$ follows a multivariate Gaussian distribution $\mathbf{P_{sub}} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{P_{sub}}}, \mathbf{\Sigma}_{\mathbf{P_{sub}}})$:

$$\begin{aligned} pdf(\mathbf{P_{sub}}) = & \frac{1}{\sqrt{(2\pi)^{k_{sub}}\,|\Sigma_{P_{sub}}|}} \\ & \cdot exp[-\frac{1}{2}(\boldsymbol{\mu}_{P_{sub}} - P_{sub})^{\mathbf{T}}\mathbf{\Sigma}_{P_{sub}}^{-1}(\boldsymbol{\mu}_{P_{sub}} - P_{sub})] \end{aligned} \tag{4.5}$$

where $\boldsymbol{\mu}_{\mathbf{P_{sub}}}$ and $\boldsymbol{\Sigma}_{\mathbf{P_{sub}}}$ are the mean vector and covariance matrix of the sub-threshold region parameters, respectively, and where k_{sub} is the dimension of $\mathbf{P_{sub}}$. The covariance between $\mathbf{P_{sub}}$ and $\mathbf{P_{above}}$ is handled through the iteration process in Fig. 4-1. Next, we assume that the "uncertainty" of $\boldsymbol{\mu}_{P_{sub}}$ follows a conjugate Gaussian prior distribution $\boldsymbol{\mu}_{P_{sub}} \sim \mathcal{N}(\boldsymbol{\mu}_{s0}, \boldsymbol{\Sigma}_{s0})$.

$$\begin{aligned} pdf(\boldsymbol{\mu}_{P_{sub}}) = & \frac{1}{\sqrt{(2\pi)^k \left|\boldsymbol{\Sigma}_{\mathbf{s0}}\right|}} \\ & \cdot exp[-\frac{1}{2}(\boldsymbol{\mu}_{P_{sub}} - \boldsymbol{\mu}_{s0})^{\mathbf{T}} \boldsymbol{\Sigma}_{s0}^{-1} (\boldsymbol{\mu}_{P_{sub}} - \boldsymbol{\mu}_{s0})] \end{aligned} \tag{4.6}$$

where $\boldsymbol{\mu}_{\mathbf{s0}}$ and $\boldsymbol{\Sigma}_{\mathbf{s0}}$ are the mean vector and covariance matrix of $\boldsymbol{\mu}_{\mathbf{P_{sub}}}$, respectively. Given $\boldsymbol{\mu}_{\mathbf{P_{sub}}}$ and $\beta_{\mathbf{ln}F_n}$, we calculate the probability of observing each data point $\mathbf{ln}F_n$ associated with subspace distribution $pdf(\mathbf{P_{sub}})$ as

$$\begin{aligned} pdf(\mathbf{ln}F_n | \boldsymbol{\mu}_{P_{sub}}, \beta_{\mathbf{ln}F_n}) = & \sqrt{\frac{\beta_{\mathbf{ln}F_n}}{2\pi}} \\ & \cdot exp[-\frac{(\mathbf{ln}F_n - \mathbf{ln}f(\mathbf{V}^n, \boldsymbol{\mu}_{P_{sub}}, \boldsymbol{\mu}_{P_{above}}))^2}{2 \cdot \beta_{\mathbf{ln}F_n}^{-1}}] \end{aligned} \tag{4.7}$$

The above equation (4.7) is the complete form of the physical subspace projection for the sub-threshold region.

4.3.2 Learning Precision at Different Biases

The learning of precision $\beta_{\mathbf{ln}F_n}$ is a key step in physical subspace projection. In practice, many reasons may contribute to the uncertainties of measurements at each bias. These reasons include modeling errors due to the inability of MVS to capture certain physical effects or measurment errors due to inaccuracies in current measurements. While they depend on the details of the fabrication or measurement process, these uncertainties show a strong systematic trend at different biases.

In this work, the $I-V$ curves of transistors from past technologies or past transistor data from current technologies are used to learn the systematic MVS model uncertainty trend at different voltage biases. Such data may come from either test-site measurement or simulations using mature or early product design kits. The detailed learning process proceeds as follows. First, a group of historical transistors are selected depending on the fabrication process of the target transistor. For example, if we intend to fit a transistor fabricated in a low power process, appropriate historical transistors would also be transistors in a low power process. In cases (such as that considered in this chapter) where no detailed information about the target transistor is available, a mix of short-channel transistors in several fabrication processes and technology nodes (six in this chapter) are employed to improve our confidence in predicting $\beta_{\mathbf{ln}F_n}$ on an unknown transistor. This assumes that although a new process introduces different lithography, structures and materials, the basic transistor operations and trends remain. Therefore the MVS parameters do not drastically change, because most of them are based on underlying solid-state physics rather than on the fabrication process. After selection of a group of historical transistors, each selected transistor is fitted into the MVS model with a complete set of $I-V$ measurements using the deterministic non-linear least-squares (NLS) error function, (4.3) and (4.4).

Given experimental $I-V$ measurements of a transistor and simulation results generated from the MVS model, inverse uncertainty quantification estimates the discrepancy between the experiment and the mathematical model, as shown in (4.8). The standard deviation $\sqrt{\beta^{-1}_{\mathbf{ln}F_n}}$ is calculated by the average differences between measurements and MVS model predictions using the NLS extracted parameters:

$$\mathbf{ln}f(\mathbf{V}^n, \mathbf{P_{sub}}, \mathbf{P_{above}}) = \mathbf{ln}F_n + \delta + \epsilon \tag{4.8}$$

where δ denotes the additive discrepancy function, and ϵ denotes the experimental uncertainty (measurement noise). For example, Fig. 4-2 (a) shows errors resulting from the MVS model being an ultra-compact model that is unable to capture the gate tunneling effect for certain technologies. Fig. 4-2 (b) shows measurement errors due to the inaccuracies of current measurement in the sub-nA region.

Fig. 4-3 shows the learned standard deviation ($\sqrt{\beta^{-1}_{\mathbf{ln}F_n}}$) at different voltage biases. Fig. 4-3 (a) is an extraction of average uncertainty from design kits when a clean data set with no measurement error is involved. A high uncertainty is observed on I_{dsat}

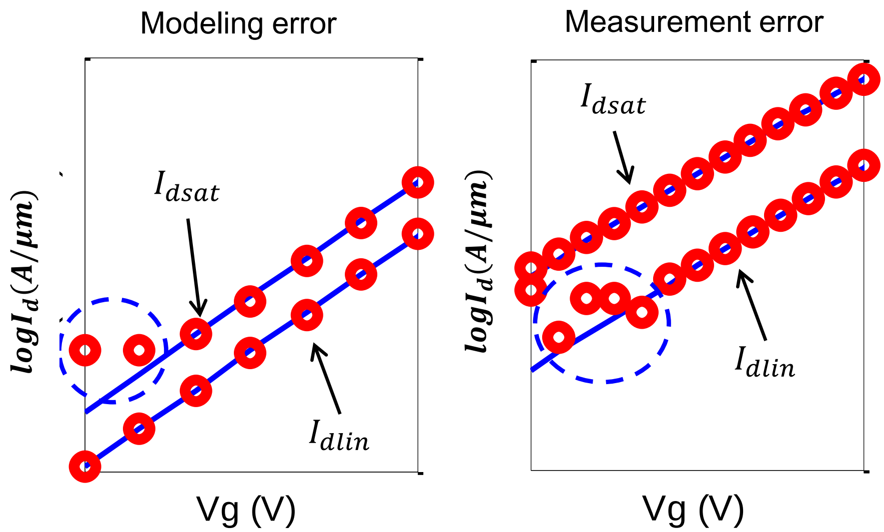


Figure 4-2: Sources of uncertainties (a) modeling error, and (b) measurement error.

with very low gate voltage where the gate tunneling effect appears. Fig. 4-3(b) is an extraction of average uncertainty from measurement data, where the uncertainty includes both modeling and measurement errors. A rapid increase of uncertainty is observed on $I_{d_{lin}}$ in the very low gate voltage region. This is due to the inaccuracies of current measurement in the sub-nA region. The increased uncertainty around $V_{gs} = 0.5V$ is due to the inaccuracy of the MVS in modeling the transition region.

Similarly, Fig. 4-4 shows the extraction of average uncertainty $\sqrt{\beta_{F_n}^{-1}}$ from measurement data for the above-threshold region. It will be used to extract $\mathbf{P_{above}}$.

4.3.3 Learning a Prior Distribution

After physical subspace projection and precision estimation, we are able to project very small numbers of samples in current measurements $\{F_1, ..., F_N\}$ to parameter subspace $\mathbf{P_{sub}}$ and obtain the conditional probability of observing $\mathbf{ln}F_n$ given $\boldsymbol{\mu}_{P_{sub}}$ and $\beta_{\mathbf{ln}F_n}$. We then combine this conditional probability with the prior distribution $pdf(\boldsymbol{\mu}_{P_{sub}})$ in (4.6) to accurately estimate $\boldsymbol{\mu}_{P_{sub}}$. Assuming each of our N current measurements is i.i.d., we can write the likelihood function $pdf(\mathbf{F}|\boldsymbol{\mu}_{P_{sub}}, \beta_{\mathbf{ln}F_n})$ as:

$$pdf(\mathbf{F}|\boldsymbol{\mu}_{P_{sub}}, \beta_{\mathbf{ln}F_n}) = \prod_{n=1}^{N} pdf(F_n|\boldsymbol{\mu}_{P_{sub}}, \beta_{\mathbf{ln}F_n}) \tag{4.9}$$

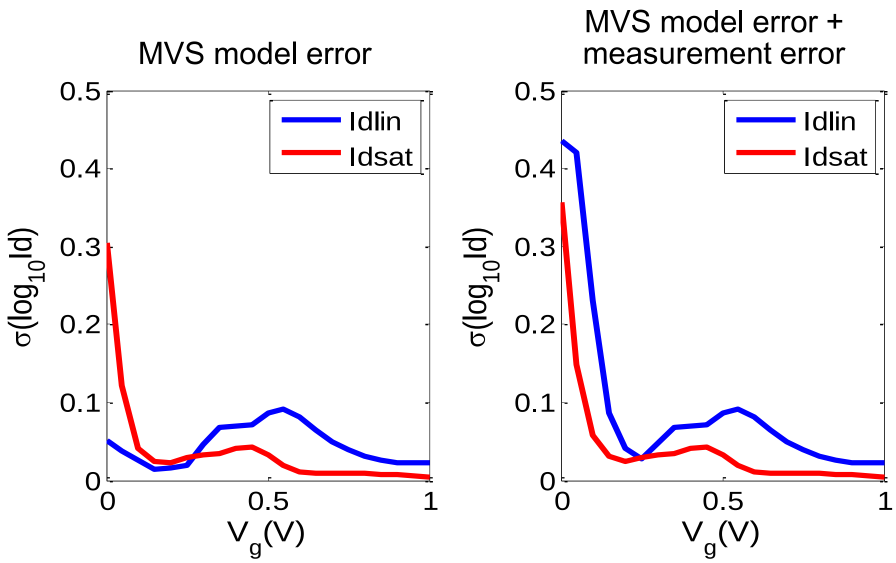


Figure 4-3: Extraction of average uncertainty $\sqrt{\beta_{\mathbf{ln}F_n}^{-1}}$ at different bias for 6 different technologies from (a) design kits, and (b) measurement results.

According to Bayes' theory, the conditional distribution $pdf(\boldsymbol{\mu}_{P_{sub}}|\mathbf{F})$ equals the product of the prior $pdf(\boldsymbol{\mu}_{P_{sub}})$ and the likelihood function $pdf(\mathbf{F}|\boldsymbol{\mu}_{P_{sub}})$ divided by total likelihood of observing $\mathbf{F}$, $pdf(\mathbf{F})$:

$$pdf(\boldsymbol{\mu}_{P_{sub}}|\mathbf{F}) = pdf(\boldsymbol{\mu}_{P_{sub}}) \cdot pdf(\mathbf{F}|\boldsymbol{\mu}_{P_{sub}})/pdf(\mathbf{F}) \tag{4.10}$$

The precision $\beta_{\mathbf{ln}F_n}$ is learned from historical transistor data and is therefore independent of the measurement set $\mathbf{F}$. Consequently,

$$pdf(\mathbf{F}|\boldsymbol{\mu}_{P_{sub}}, \beta_{\mathbf{ln}F_n}) = pdf(\mathbf{F}|\boldsymbol{\mu}_{P_{sub}}) \tag{4.11}$$

Substituting (4.9) and (4.11) into (4.10) yields:

$$pdf(\boldsymbol{\mu}_{P_{sub}}|\mathbf{F}) = pdf(\boldsymbol{\mu}_{P_{sub}}) \cdot \prod_{n=1}^{N} pdf(F_n|\boldsymbol{\mu}_{P_{sub}}, \beta_{\mathbf{ln}F_n}) / \prod_{n=1}^{N} pdf(F_n) \tag{4.12}$$

The above equation demonstrates the sequential nature of Bayesian learning in which the "old" posterior distribution becomes the "new" prior when a new data point

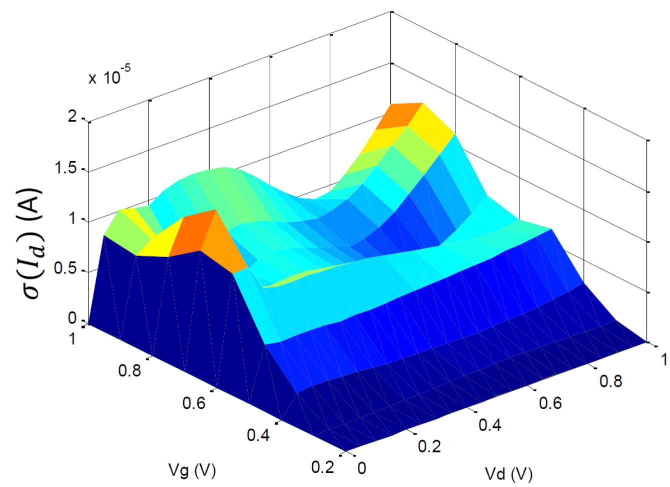


Figure 4-4: Extraction of average uncertainty $\sqrt{\beta_{F_n}^{-1}}$ at different biases for six different technologies from measurement results.

is added to the measurement set. Fig. 4-5 shows the results of Bayesian learning on $\boldsymbol{\mu}_{P_{sub}}$ as the portfolio of the measurement groups is expanded. For illustration, we only show a 2-D map of δ and SS (sub-threshold swing which is the physical expression of n_0), and the third sub-threshold parameter V_{t0} is fixed for better comparisons across all measurement updates. The bottom left figure corresponds to the situation before any data points are observed, and shows a plot of the prior distribution $\boldsymbol{\mu}_{P_{sub}} \sim \mathcal{N}(\boldsymbol{\mu}_{s0}, \boldsymbol{\Sigma}_{s0})$. Note that $\boldsymbol{\mu}_{s0}$ and $\boldsymbol{\Sigma}_{s0}$ are learned from historical transistor data in exactly the same way as $\beta_{\mathbf{ln}F_n}$.

A complete set of $I-V$ measurements for short-channel transistors with six different fabrication processes and technology nodes are fitted into the MVS model using the least-squares error functions, giving us the *a priori* $\boldsymbol{\mu}_{s0}$ and $\boldsymbol{\Sigma}_{s0}$ MVS model parameters. We then derive the mean and standard deviation of transistor currents at each bias voltage using $\boldsymbol{\mu}_{s0}$ and $\boldsymbol{\Sigma}_{s0}$. Fig. 4-6 shows the mean and standard deviation of the transistor $I-V$ curves using the extracted $\boldsymbol{\mu}_{s0}$ and $\boldsymbol{\Sigma}_{s0}$; these provide the prior distributions before collecting any new measurements for the target technology.

However, the uncertainty band of the *a priori* in Fig. 4-6 is still relatively wide. If more information from the target technology is available, e.g., repeated parameter extractions from on-chip monitor structures, the *a priori* $\boldsymbol{\mu}_{s0}$ and $\boldsymbol{\Sigma}_{s0}$ could provide

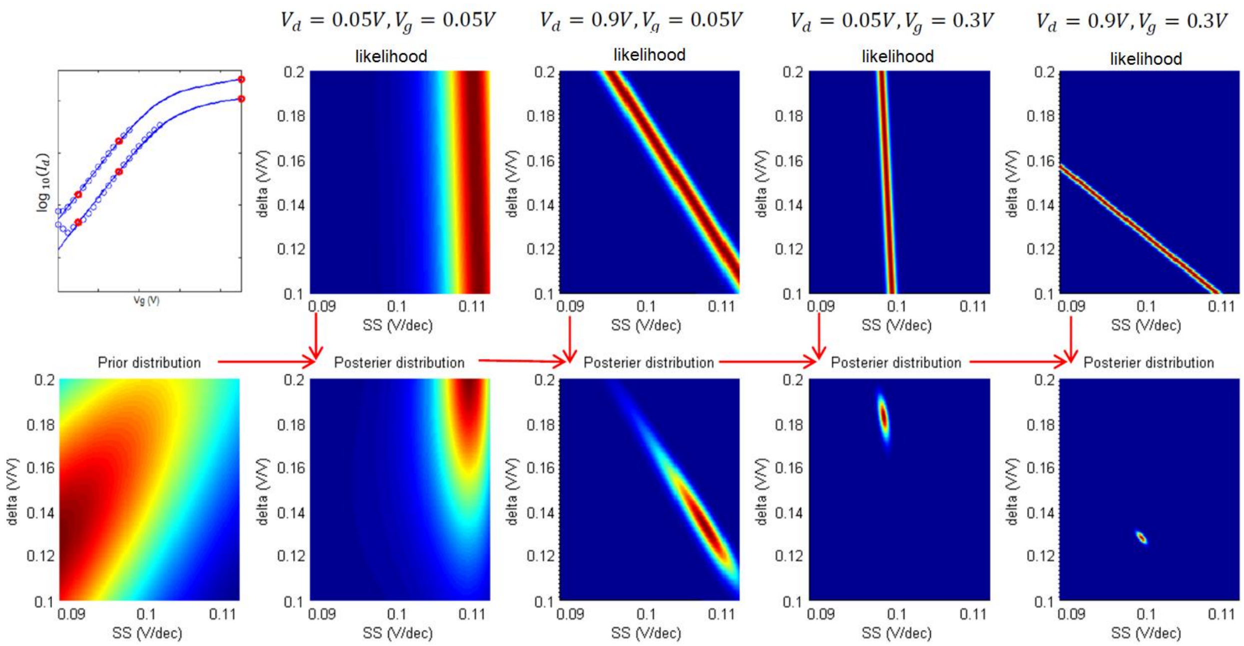


Figure 4-5: Illustration of sequential Bayesian learning of $\boldsymbol{\mu}_{P_{sub}}$ using priors and $I - V$ measurements. The two parameters shown are the sub-threshold swing factor SS and the drain-induced barrier lowering δ. The red color represents estimates with high likelihood while the blue color represents estimates with low likelihood. As more measurements are added, the MAP parameter estimates become more accurate. Note that the actual extraction is not done sequentially, as summarized in Section 4.3.4.

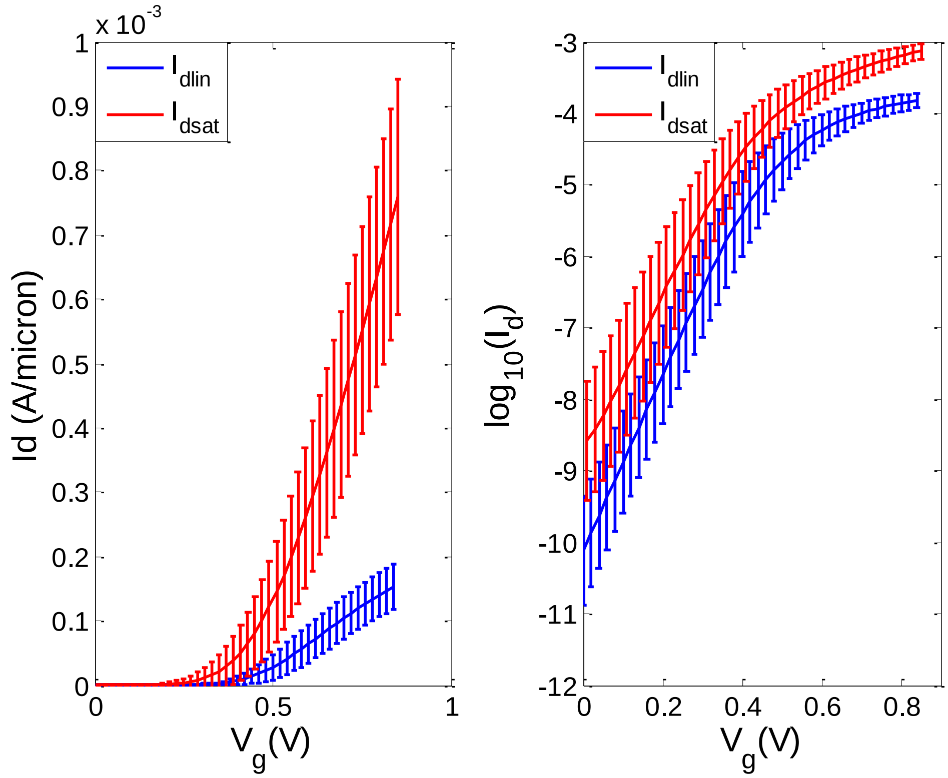


Figure 4-6: Mean and standard deviation of the transistor $I - V$ curve using $\boldsymbol{\mu}_{s0}$ and $\boldsymbol{\Sigma}_{s0}$ learned from historical transistor data when no measurements from target technology are available.

much tighter prior distributions before collecting measurements of a new transistor in the same technology. For example, Fig. 4-7 shows the mean and standard deviation of the transistor $I - V$ curves using the extracted $\boldsymbol{\mu}_{s0}$ and $\boldsymbol{\Sigma}_{s0}$ from repeated measurements of a single technology, with correspondingly tighter prior confidence.

The first row of Fig. 4-5 shows the resulting likelihood function $pdf(F_n|\boldsymbol{\mu}_{P_{sub}})$ for measurements at different biases alone. Different widths of red regions at each bias represents historical learning of $\beta_{\mathbf{ln}F_n}$. If two measurements have large discrepancy (e.g., SS from the first and third samples), the extraction results will be more strongly adjusted toward the measurement with the higher precision (narrower width). The second row shows the posterior distribution $pdf(\boldsymbol{\mu}_{P_{sub}}|\mathbf{F})$ that results from multiplying its likelihood function from the top row by the prior (bottom left). As this process continues, the posterior distribution becomes much sharper, and in the limit of an infinite number of data points, the posterior distribution would become a delta

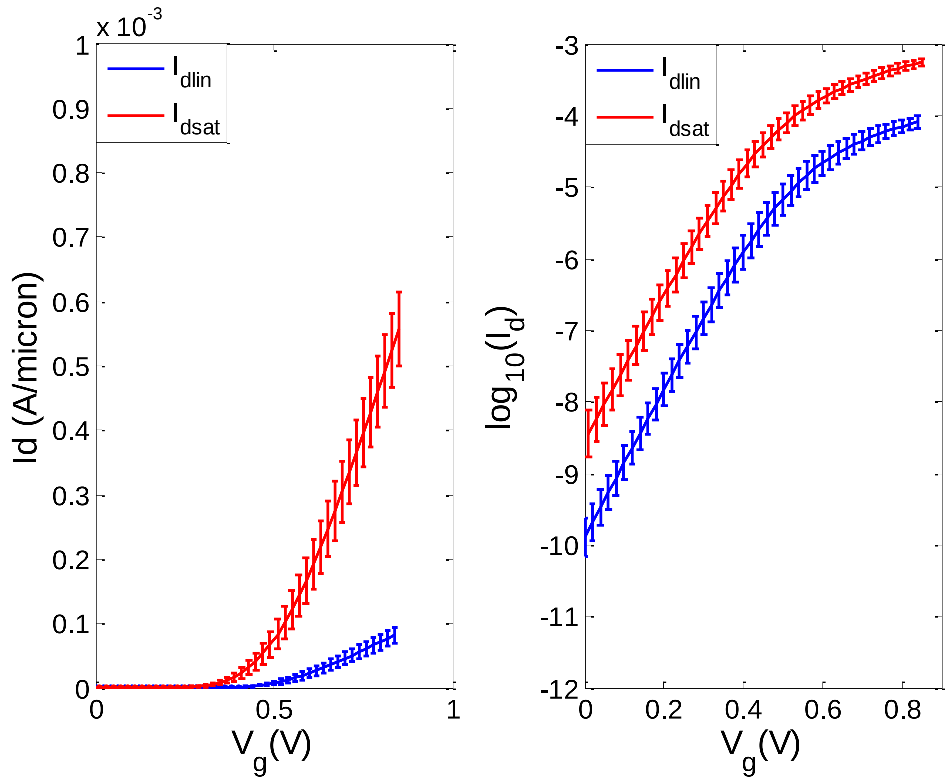


Figure 4-7: Mean and standard deviation of the transistor $I-V$ curve using $\boldsymbol{\mu}_{s0}$ and $\boldsymbol{\Sigma}_{s0}$ learned from a large number of historical transistor measurements from target technology, showing tighter confidence intervals compared to Fig. 4-6.

function centered on the true parameter values.

4.3.4 Maximum A Posteriori Estimation

Our final goal is to find optimal estimates of $\boldsymbol{\mu}_{P_{sub}}$ and $\boldsymbol{\mu}_{P_{above}}$ that maximize the log likelihood of the posterior distributions $\ln pdf(\boldsymbol{\mu}_{P_{sub}}|\mathbf{F})$ and $\ln pdf(\boldsymbol{\mu}_{P_{above}}|\mathbf{F})$, respectively. Substituting (4.6) and (4.7) into (4.12) and removing the constant items yields:

$$\begin{aligned}\mathcal{E}(\mathbf{P_{sub}}) = \frac{1}{2}(\boldsymbol{\mu}_{P_{sub}} - \boldsymbol{\mu}_{s0})^{\mathbf{T}}\boldsymbol{\Sigma}_{s0}^{-1}(\boldsymbol{\mu}_{P_{sub}} - \boldsymbol{\mu}_{s0}) \\ +\frac{1}{2}\sum_{n=1}^{N}\beta_{\mathbf{ln}F_n}\{\mathbf{ln}(F_n) - \mathbf{ln}(f(\mathbf{V_n}, \mathbf{P_{sub}}, \mathbf{P_{above}}))\}^2\end{aligned} \tag{4.13}$$

Similarly, we have

$$\begin{aligned}\mathcal{E}(\mathbf{P_{above}}) = &\frac{1}{2}(\boldsymbol{\mu}_{P_{above}} - \boldsymbol{\mu}_{f0})^{\mathbf{T}}\boldsymbol{\Sigma}_{f0}^{-1}(\boldsymbol{\mu}_{P_{above}} - \boldsymbol{\mu}_{f0}) \\ &+\frac{1}{2}\sum_{n=1}^{N}\beta_{F_n}\{F_n - f(\mathbf{V_n}, \mathbf{P_{sub}}, \mathbf{P_{above}})\}^2\end{aligned} \tag{4.14}$$

Equations (4.13) and (4.14) are the new error functions given by maximum a posteriori (MAP) estimation. Compared with (4.3) and (4.4), we note the following two advantages: 1) a bias-dependent precision allocating weights to sample measurement errors as contrasted with the uniform weights of NLS; 2) an appropriate prior distribution that has been learned from historical transistor data and which provides a parameter probability distribution before any measurement. In particular, the BPV restriction that the number of electrical measurements should be larger than the number of extracted parameters is removed.

4.4 Validation

In this section, two model parameter extraction examples in several cutting-edge CMOS technologies are used to demonstrate the efficiency of our method. To test and compare with the prior art, we have also implemented deterministic extraction using the non-linear least-squares error function and statistical extraction using backward propagation of variance (BPV).

4.4.1 Example I: Early Technology Evaluation

The first example is to use the MVS model for early technology evaluation. The difficulty of this problem is that measurements are collected from a limited number of early prototype devices rather than from a full suite of designed test structures. Therefore it is highly unlikely that the limited data will be sufficient to fit complex compact models such as BSIM. As for the ultra compact MVS model, separate measurements of $I_d - V_{ds}$ and $I_d - V_{gs}$ are needed in order to apply the traditional deterministic method of non-linear least squares error function.

Fig. 4-8 shows the MVS model fitting results using our new parameter extraction method for four technologies from $14nm$ through $45nm$. Notice that some technologies are not used for learning the prior information, and only six points from $I_d - V_{gs}$ are used to fit the entire model. The prediction results on the rest of the $I - V$ measurements match well throughout the operating region, including the $I_d - V_{ds}$ curves. Even for transistors with gate tunneling effects (e.g., Technology 3), the proposed method still extracts the correct trend in the sub-threshold region. This is because, according to the historical record, the MVS model suffers a large modeling error in this region and therefore less weight is assigned for the I_{off} measurement.

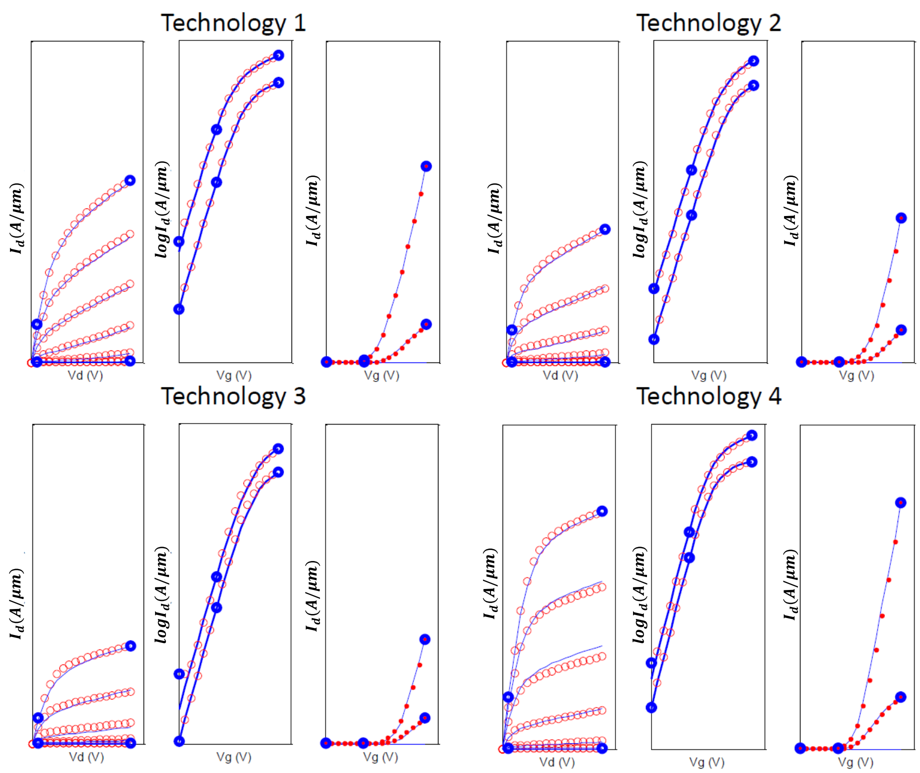


Figure 4-8: MVS model fitting results using MAP parameter extraction method for four technologies in $14nm$-$45nm$. Blue circles are fitted measurements using the MAP method and red circles are test measurements for validation.

Fig. 4-9 shows an average MVS model prediction error for Technology 3 using both our MAP method and the least-squares error function. A 3X sample size reduction is observed to achieve comparable error.

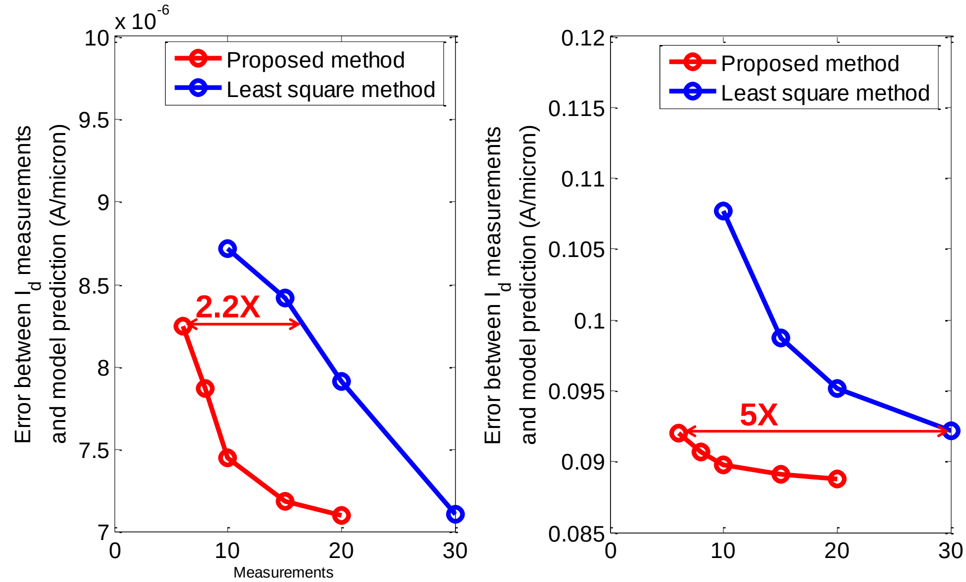


Figure 4-9: Average model prediction error for Technology 3 on I_d for above-threshold region and $\log_{10} I_d$ for the sub-threshold region, showing reduced error of proposed method compared to the traditional NLS method.

Fig. 4-10 shows errors for compared with baseline extraction MVS model parameters extracted using LSE method and proposed Bayesian extraction framework versus number of measurements. Here $I - V$ curves with $20mV$ V_{gs} intervals (around 100 measurements total) are used for baseline extraction. For extraction in all five parameters, better consistency is observed for the proposed Bayesian extraction framework compared with traditional LSE method. Due to high correlation between v_{x0} and μ, these sub-threshold subgroup parameters show higher percentage errors than the above-threshold parameters. However, the proposed Bayesian extraction shows lower error and more consistent extraction compared with the traditional LSE method.

4.4.2 Example II: Statistical Extraction for post-Silicon Validation

The second example is to use the MVS model to estimate post-Silicon circuit performance and to conduct statistical parameter extraction. In both cases, measurements are taken from different on-chip monitor circuits in which large numbers of transistors are involved. Unlike the first example where we focus on characterizing a target tran-

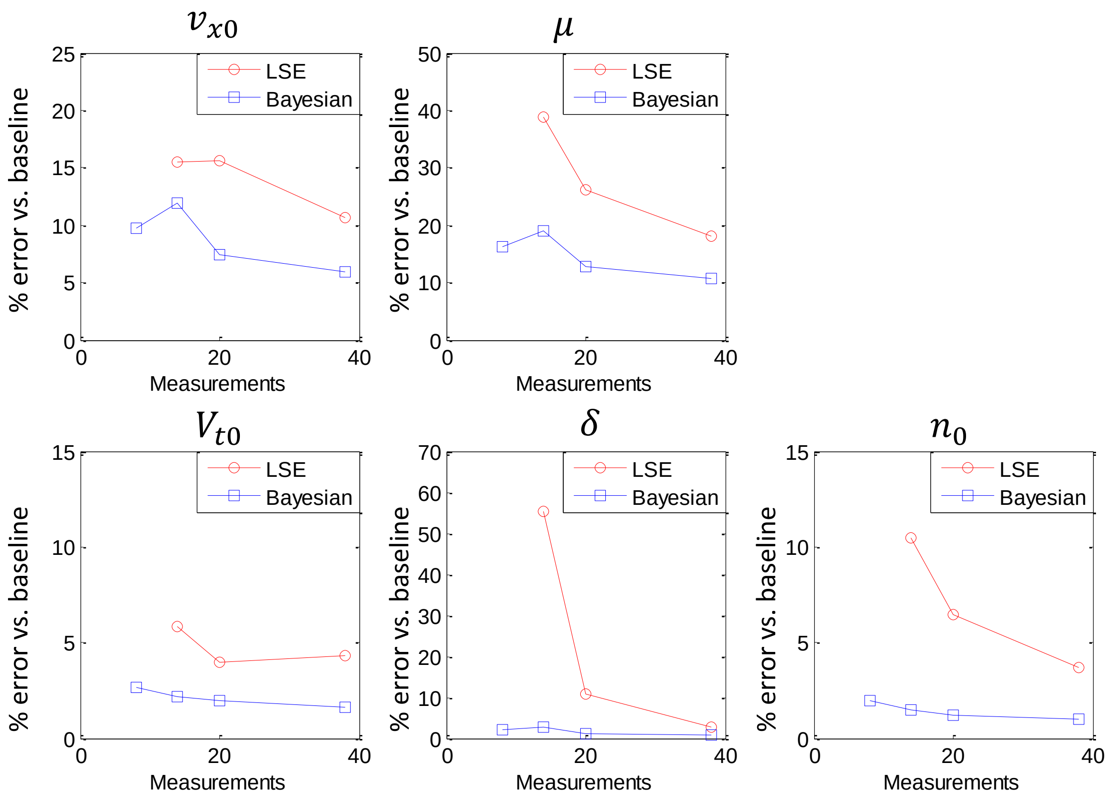


Figure 4-10: Parameter consistency (percentage error for parameters compared with baseline extraction) for extraction of the MVS model using LSE method and proposed Bayesian extraction framework versus number of measurements.

sistor, here we emphasize modeling statistical distributions of groups of transistors and seek to extract the distribution of their MVS parameters. The bottleneck here is that in order to obtain a statistically sufficient number of transistors, we traditionally suffer from the testing costs from the tens of thousands of transistors required. To extract key parameters (e.g., I_{dsat}, I_{dlin}, V_{tsat} and V_{tlin}) and save testing costs, we would prefer to require that only a small set of $I-V$ curve data is collected. In a standard BSIM model, these few $I-V$ measurements are not sufficient to extract *all* parameters. Some knowledge of, or restrictions on, many of the BSIM parameters is needed. On the other hand, Fig. 4-11 shows MVS model fitting for two transistors in a 28nm technology using our MAP parameter extraction method. Only six measurements are used to fit eight parameters in the MVS model for each device, and the prediction matches well with other test measurements.

This suggests an alternative method for extracting variations of MVS model pa-

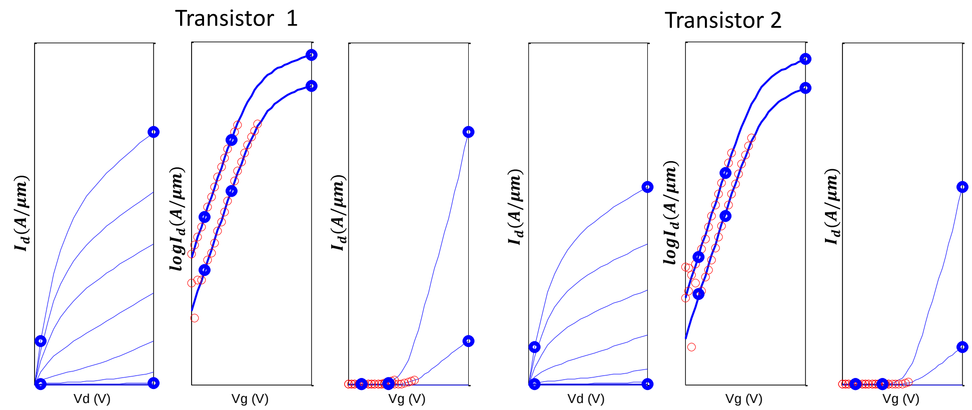


Figure 4-11: MVS model fitting results using the MAP parameter extraction method in a 28-nm technology from wafer measurements. The blue circles are fitting measurements used by the proposed method. The red circles are additional test measurements used for validation.

rameters from wafer level measurements. The method consists in simply extracting deterministic MVS parameters separately for many devices and calculating their covariance matrix. Compared with the traditional BPV method, this direct method has several advantages. First, the relationship between measurements at different biases is maintained. For example, the extraction of sub-threshold swing SS needs measurements at more than two biases in the sub-threshold region, and therefore it is hard to extract its variance through BPV. Second, traditional BPV assumes independent parameters, and correlated variables cannot be extracted with the same backward propagation. Third, BPV requires the number of electrical measurements to be larger that the number of extracted parameters, which in turn requires a significant number of measurements, especially for complex compact models such as BSIM and PSP. In contrast, this direct statistical extraction is more accurate and is less constrained. Fig. 4-12 shows I_d probability density simulated using variance extracted through our MAP method compared with variance extracted using BPV, together with measurements at two different biases. Both methods show an accurate distribution for I_{dsat}, but the MAP method shows a much better prediction for I_d current distribution in the transition region ($V_d = V_{dd}$, $V_g = V_{dd}/2$).

Fig. 4-13 shows statistically-extracted wafer maps of key MVS model parameters V_{t0}, δ, SS (extracted as n_0), v_{xo} and μ in a 28-nm technology. A strong correlation is

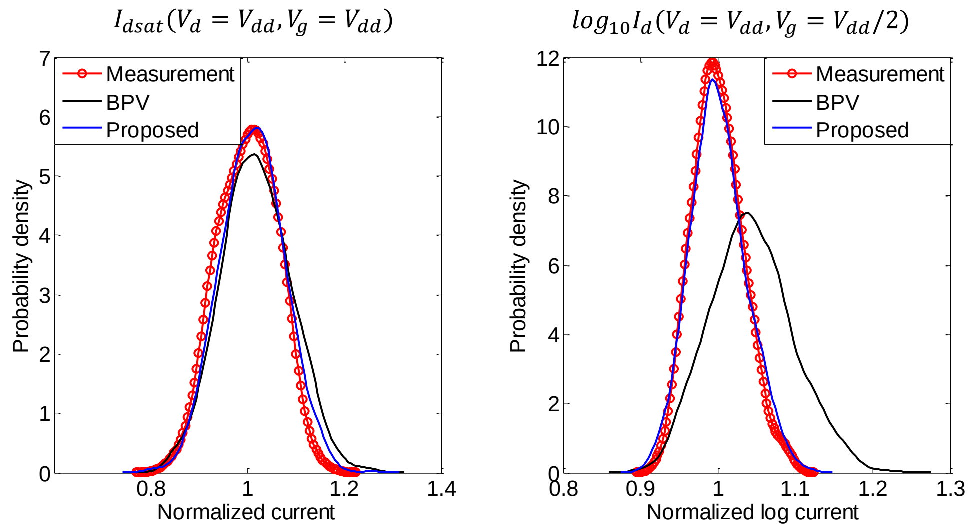
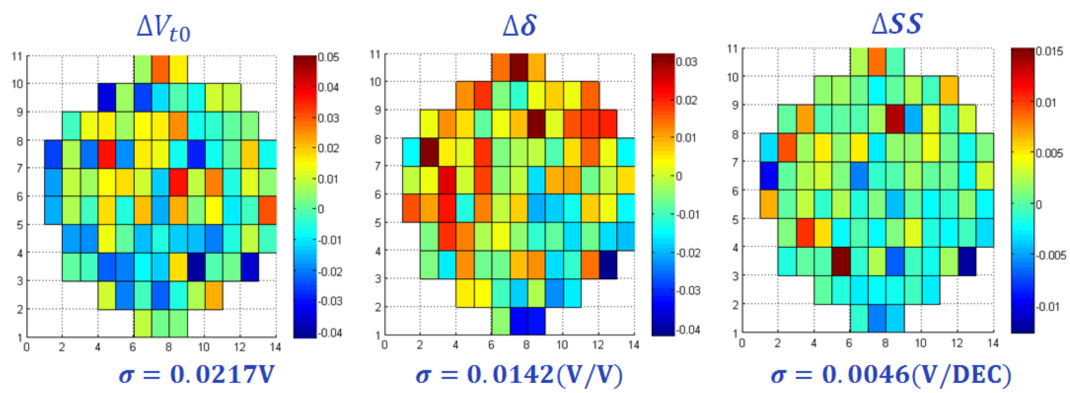
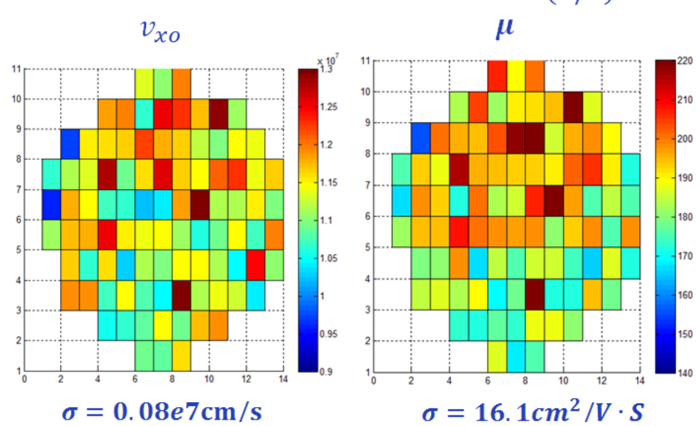


Figure 4-12: I_d probability density simulated using variance extracted through proposed method compared with variance extracted through BPV, together with measurements at two different biases.





V_{t0}	δ	SS	v_{xo}	μ
1	-0.104	0.363	0.339	0.328
-0.104	1	-0.001	-0.311	0.075
0.363	-0.001	1	0.616	0.529
0.339	-0.311	0.616	1	0.819
0.328	0.075	0.529	0.819	1

Figure 4-13: Statistically-extracted wafer maps of key VS model parameters $Vt0$, δ, SS (extracted as $n0$), v_{xo} and μ in a $28nm$ technology. Parameter correlations are shown in the bottom right table.

observed between v_{xo} and μ, which matches the theoretical prediction given by [67] that the relative change in virtual source velocity is proportional to the change in mobility. Fig. 4-14 shows a scatter plot of virtual source velocity and mobility from extraction of on-chip monitor circuits of one $28nm$ wafer. We observe that 83% of the relative change in mobility is converted to relative change in virtual source velocity, which is very close to 85% given by [67]. This shows that the proposed method is able to correctly extract and separate strongly correlated parameters even with very limited measurements.

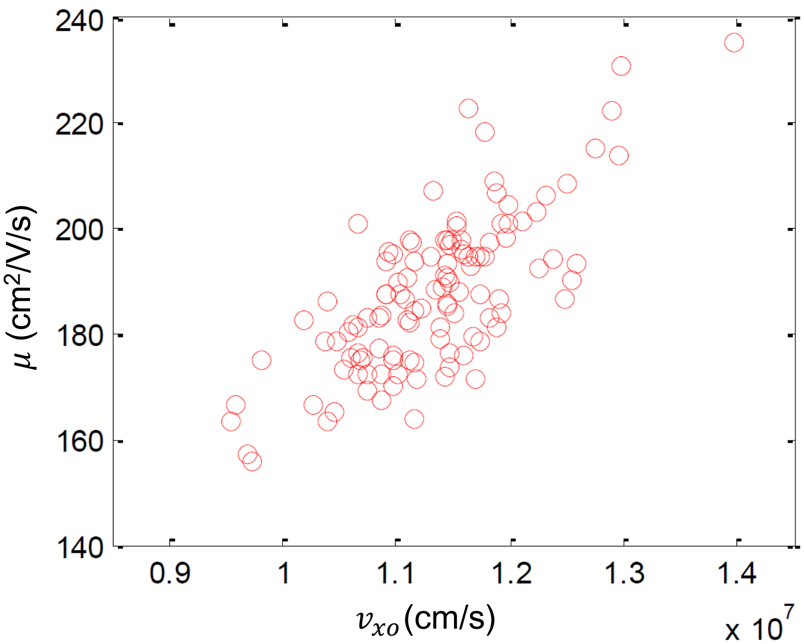


Figure 4-14: Scatter plot for statistically extracted virtual source velocity and mobility from on-chip monitor circuits of one wafer from a $28nm$ technology.

4.5 Optimal Sampling of Transistor Measurements

In Section 4.4, we demonstrated the efficiency of our proposed Bayesian method by fitting the full region of transistor operation of the MVS model using only six measurements. An interesting question then arises: what is the lower bound for the number of $I-V$ measurements needed to fit a compact model and how should they be selected? In this section, we will solve the problem quantitatively by constructing simultaneous confidence intervals using the historical transistor measurements. We also propose an efficient algorithm for selecting the optimal measurement biases by minimizing the average uncertainty.

This is essentially an optimal design of experiments problem. A non-optimal design (such as the fixed interval sampling method by sweeping the entire I_{dsat} and I_{dlin} curves) requires a greater number of measurements to estimate the parameters with the same precision as an optimal design. Therefore optimal experiments can reduce the costs of experimentation. Typically, such an optimal selection maximizes the expected information gain, which is measured by the change in entropy of the distribution ΔS [84, 85]. In general, it can be shown that the prior expectation of entropy can be minimized over designs in measurement space ξ by choosing a subset of ξ on which the prior entropy is maximized. For the optimal sampling of transistor measurements, this is equivalent to choosing a subset of voltage biases that minimize the integrated measurement uncertainties on the $I-V$ curves. There exist several possible ways for such optimal selection. Ref. [84] proposed an algorithm that minimizes the joint entropy of the output variables by measuring how effectively these selected output variables work together to reduce the joint uncertainty in the output. Ref. [86] proposed a sequential sampling algorithm such that at each stage of experimentation the experimenter is free to choose the random variable that he will observe from a given class of random variables. For the optimal sampling of transistor measurements, it would be hard to implement a sequential sampling algorithm because of the difficulties to hard code the algorithm and the priors into the probe station. Least Angle Regression (LAR) is another method to select the testing measurements that is only applicable to orthogonal testing points. Therefore it is difficult to solve the selection problem using a traditional greedy approximation, because the output of our candidate measurements are highly correlated [87, 88]. Our approach is similar to that proposed in [84], enabled by the availability in our case by large numbers of prior data allowing us to establish output precision and uncertainty.

The testing measurements are selected by two steps. First, a group of N candidate measurement biases are generated from different transistor operating regimes. K

measurement biases ($K << N$) are then selected according to their importance in the widths of output estimate confidence intervals. Next, simultaneous confidence intervals are constructed using the selected measurement biases.

4.5.1 Candidate Bias and Historical Correlation Generation

Let $\xi_i(i \in (1, N))$ be the bias conditions for eithor I_{dsat} or I_{dlin}. Every candidate has a V_{gs} biasing interval of $10mV$ with its nearest neighbors. (We could presumably choose other interval discretizations.) Variance $\sigma_I^2(\xi_i)$ at each bias candidate is computed using the data set described in Section 4.3.2 which covers a broad range of fabrication technologies along with measurement noises. The 95% confidence intervals using prior information are then defined by $[\mu_I(\xi_i) - 1.95\frac{\sigma_I(\xi_i)}{\sqrt{n}}, \mu_I(\xi_i) + 1.95\frac{\sigma_I(\xi_i)}{\sqrt{n}}]$. We define variance $\sigma_I^2(\xi_i|\xi_j)$ as the current variance of ξ_i given nearby measurements ξ_j. Similarly, variance $\sigma_I^2(\xi_i|\xi_j, \xi_k)$ is defined as the current variance of ξ_i given two nearby measurements ξ_j and ξ_k. With m measurements nearby, the variance $\sigma_I^2(\xi_i)$ is calculated by the minimum of unconditional variance $\sigma_I^2(\xi_i|\xi_j)$, conditional variance $\sigma_I^2(\xi_i|\xi_j)$ with the nearest measurement ξ_j and conditional variance $\sigma_I^2(\xi_i|\xi_j, \xi_k)$ with two nearest measurements ξ_j, ξ_k, as shown in (4.15).

$$\sigma_I^2(\xi_i|\xi^{1^{st}}, ..., \xi^{m^{th}}) = \sum_{i=1}^{N} min(\sigma_I^2(\xi_i), \sigma_I^2(\xi_i|\xi_j), \sigma_I^2(\xi_i|\xi_j, \xi_k)) \tag{4.15}$$

The nearest measurements for ξ_i are defined by measurements bias combinations which give the minimum conditional variance. For example, the nearest single measurement for ξ_i is defined by:

$$\xi_{nearest} = \arg\min_{\xi_j} \sigma_I^2(\xi_i|\xi_j) \tag{4.16}$$

4.5.2 Selecting Optimal Measurements

K measurement biases are selected from the N candidate biases based on their statistical importance, i.e., those biases with the jointly narrowest confidence intervals. The MATLAB pseudo code for selecting the optimal m measurement biases is provided in

Algorithm 2. The algorithm is iterative, selecting an additional single measurement each time, and given that the new additional measurement, to revisit prior selections for possible replacement. Finally, the optimal set of measurements is provided as a group.

Algorithm 2 Optimal measurement biases selection

Compute $\sigma_I^2(\xi_i)$, $\sigma_I^2(\xi_i|\xi_j)$ and $\sigma_I^2(\xi_i|\xi_j, \xi_k)$ for any combination of i, j and k;
$\xi^1 = \arg\min_{\xi_j}(\sum_{i=1}^{N} min(\sigma_I^2(\xi_i), \sigma_I^2(\xi_i|\xi_j)))$
% the optimal single measurement bias
for $m = 2, ..., K$ **do**
  % loop for searching optimal biases with m total measurements
  for $n = 1, ..., m-1$ **do**
    $\xi^{m,n^{th}} = \xi^{m-1,n^{th}}$;
  end for
  $\xi^{m,m^{th}} = \arg\min_{\xi_i} \sigma_I^2(\xi_i|\xi^{m,1^{st}}, ..., \xi^{m,m-1^{th}}, \xi_i)$;
  for $n = 1, ..., m$ **do**
    $\xi^{m,n^{th}} = \arg\min_{\xi_i} \sigma_I^2(\xi_i|\xi^{m,1^{st}}, ..., \xi^{m,n-1^{th}},$
    $\xi^{m,n+1^{th}}, ..., \xi^{m,m^{th}}, \xi_i)$;
  end for
end for

The basic idea of Algorithm 2 is as follows. The first measurement bias for the single measurement is selected such that the sum of conditional variance is minimized. Then, we add a second measurement bias by minimizing the sum of conditional variance given the single measurement and one of the candidates. Next, we revisit each of the earlier selected measurements and seek a best replacement for each based on minimizing the sum of the joint conditional variances given the new added measurement together with the rest of the current measurement set. The whole process is repeated until the required number of measurements is selected. A different set of points is selected for, say, six points than just adding an optimal sixth point to the optimal five points. This is because the candidate measurements are highly correlated and it is possible that the best five points given the sixth point may not be the same as the five optimal points.

4.5.3 Analysis for Optimal Measurements

Fig. 4-15 shows the optimal measurement selection from zero measurements to three measurements, as well as their 95% confidence bounds for both $V_{ds} = 0.05V$ and $V_{ds} = 0.85V$. It is clear that adding the first measurement squeezes the most uncertainty compared with the second and third measurement. The bias of the optimal single measurement is close to the measurement of threshold voltage, which matches our intuition that threshold voltage has the largest impact on transistor operation. The optimal two measurements are similar to measurements of threshold voltage and I_{on}, except that the optimal measurements are more close to the transition region. However, in all these cases the 95% confidence intervals are still relatively wide and it is difficult to extract a full set parameters even from optimal measurement biases with only one to three measurement points.

Fig. 4-16 shows the optimal measurements selection from four measurements to 12 measurements, as well as their 95% confidence intervals for both $V_{ds} = 0.05V$ and $V_{ds} = 0.85V$. The optimal four biases measure two points for the above-threshold region and two points for the sub-threshold region, which corresponds to measuring I_{dsat}, I_{dlin}, V_{tsat} and V_{tlin} in traditional transistor testing. However, the optimal measurements for $V_{ds} = 0.05V$ and $V_{ds} = 0.85V$ do not precisely align since the offset measurements tend to squeeze more uncertainty in the transition region. One interesting result of our optimal measurements is that biases in the deep sub-threshold region are not selected except for the last case, because of the exponential measurement error for $sub - nA$ current measurements. Instead, extrapolation using current measurement from $V_{gs} = 0.2V$ to $V_{gs} = 0.4V$ brings higher confidence than direct measurement. This explains why we prefer measuring V_{tsat} and V_{tlin} compared with measuring I_{off} directly.

In order to show the benefit of the proposed optimal sampling over a standard fixed interval sampling method, we plot average decade error for $log_{10}I_d$ for the Bayesian extraction method in Fig. 4-17(a). It shows that the Bayesian extraction method is further facilitated by the proposed optimal sampling approach by achieving sta-

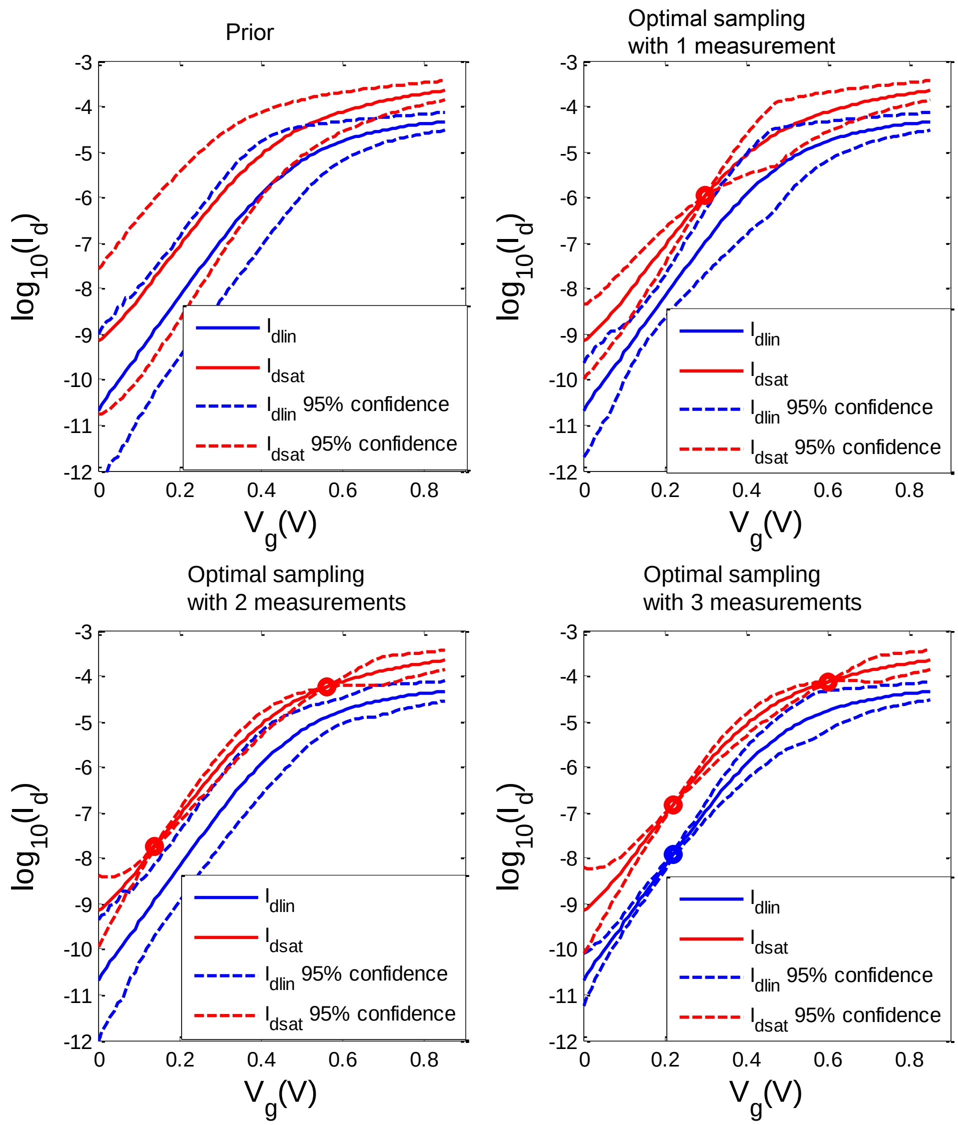


Figure 4-15: 95% confidence intervals for $V_{ds} = 0.05V$ and $V_{ds} = 0.85V$ with (a) no measurement, only prior, (b) optimal single measurement, (c) optimal two measurements, and (d) optimal three measurements. Measurement noise has been included.

ble $log_{10} I_d$ error with only six measurements. Since we cannot guarantee convergence in traditional LSE extraction, in Fig. 4-17(b) we show the percentage of non-convergent transistor extractions for the LSE method using the proposed optimal sampling method and the fixed interval sampling method. Fewer non-convergent extractions are observed for the proposed optimal sampling method.

Fig. 4-18 shows average uncertainty (quantified by $\sigma(log_{10}(Id))$) versus number of measurements. Although a proper prior has been applied using historical transistor

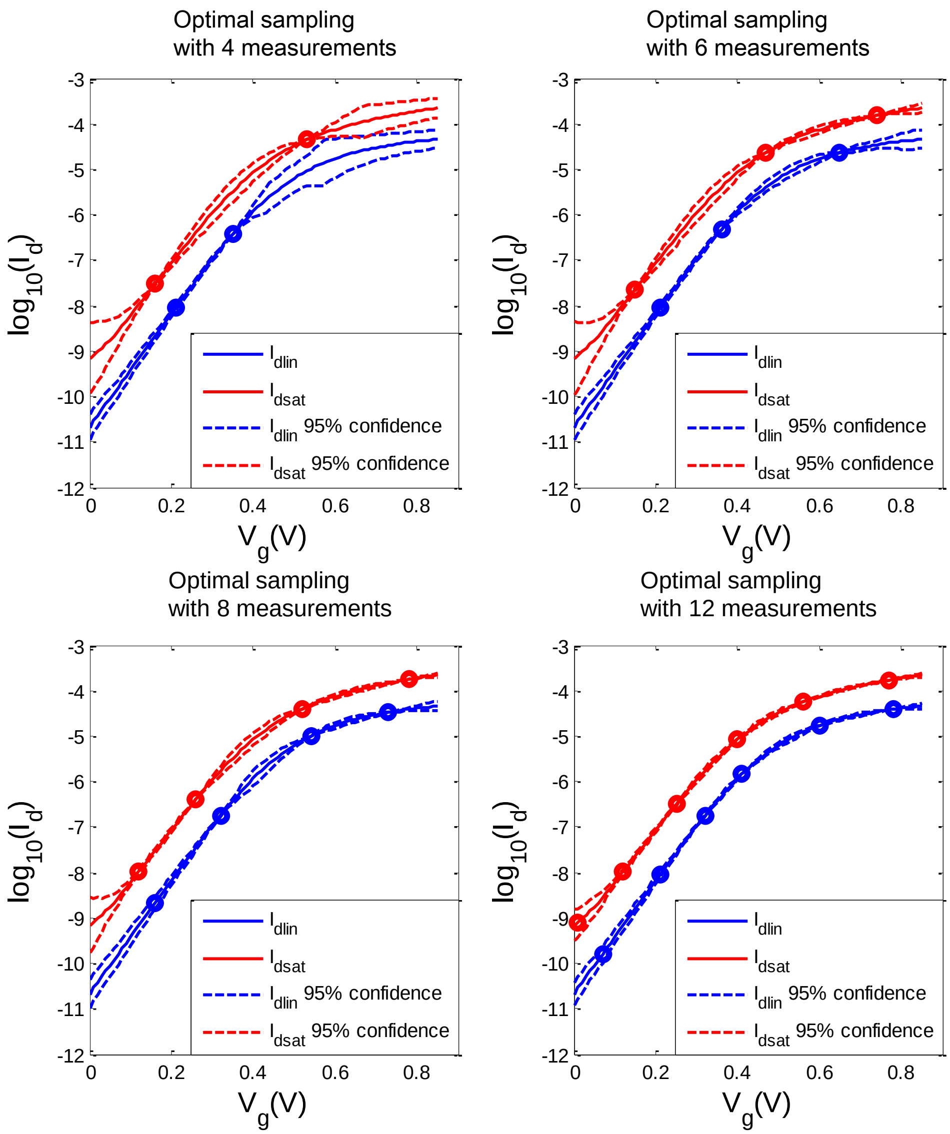


Figure 4-16: 95% confidence intervals for $V_{ds} = 0.05V$ and $V_{ds} = 0.85V$ with (a) optimal four measurements, (b) optimal six measurement, (c) optimal eight measurements, and (d)optimal 12 measurements. Measurement noise has been included.

measurements, the initial current uncertainty still reaches a relatively high value of nearly a half decade. However, the uncertainty drops quickly after a few measurements. With only six measurements, the uncertainty scales down to 10% of its initial value. When the average measurement uncertainty equals the average modeling error given a target compact model, we define the corresponding number of measurements as the optimal number of measurements. In this work, the theoretical prediction for the optimal number of measurements for the MVS model is eight, which matches well

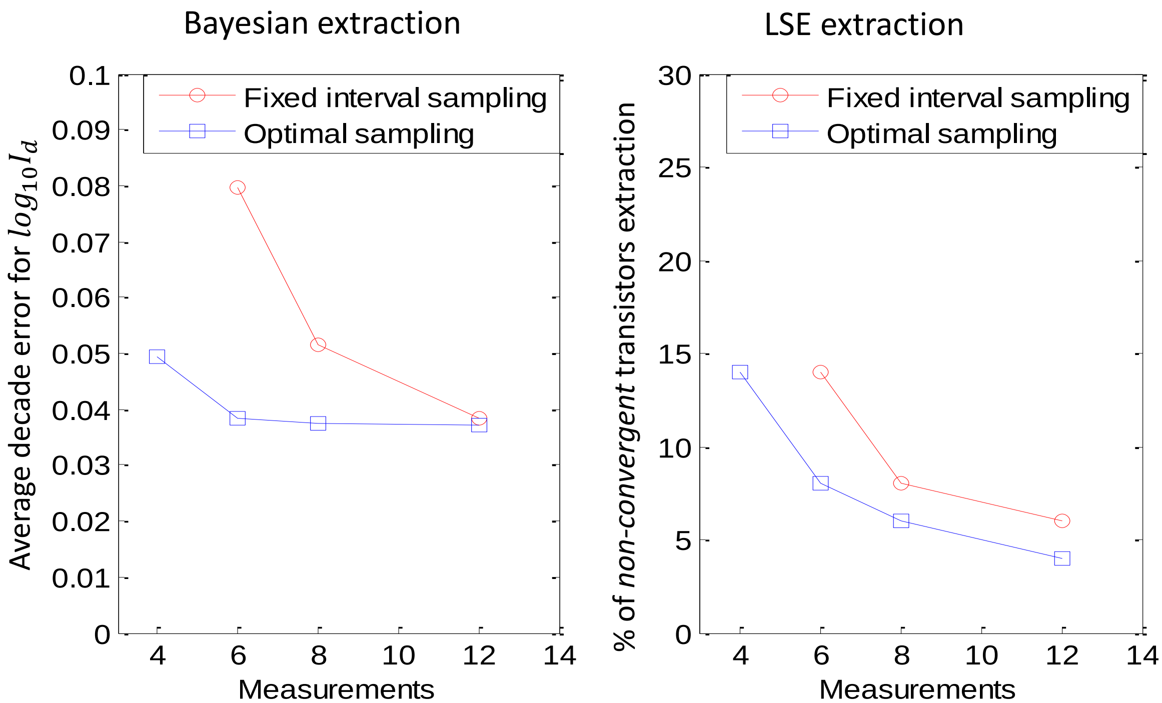


Figure 4-17: Comparison between fixed interval sampling and proposed optimal sampling: (a) average decade error for $log_{10}I_d$for Bayesian extraction method, and (b) percentage of non-convergent transistor extractions for LSE method.

with other parameter extraction results in Section 4.4.

An alternative optimal measurements solution using the MVS model is shown in Fig. 4-19. It is divided into two steps: (1) estimation of MVS model parameters $\{p_i\}$ (i=1,2,...,7) using current measurements $\{F_{meas}\}$ as described in Section 4.3, and (2) projection from parameter space to output space $\mathbf{F}$ and minimization of the total sensitivity "volume." The objective function is:

$$argmin_{F_{meas}} \sum_i w_i \cdot \mathcal{E}(p_i | \{F_{meas}\}) \tag{4.17}$$

where $w_i = \partial(\sum_{j=1}^{N} F_j)/\partial p_i$ is the normalization factor between parameters, and $\mathcal{E}(p_i | \{F_{meas}\})$ is the new error function described in (4.13) and (4.14).

The proposed optimal measurements solution in Section 4.5.2 is essentially a maximal precision estimator that uses the theory of optimal experimental design to specify inputs. Compared with the alternative approach in (4.17) where the physically based MVS model is employed, our optimal measurements solution in Section 4.5.2 only

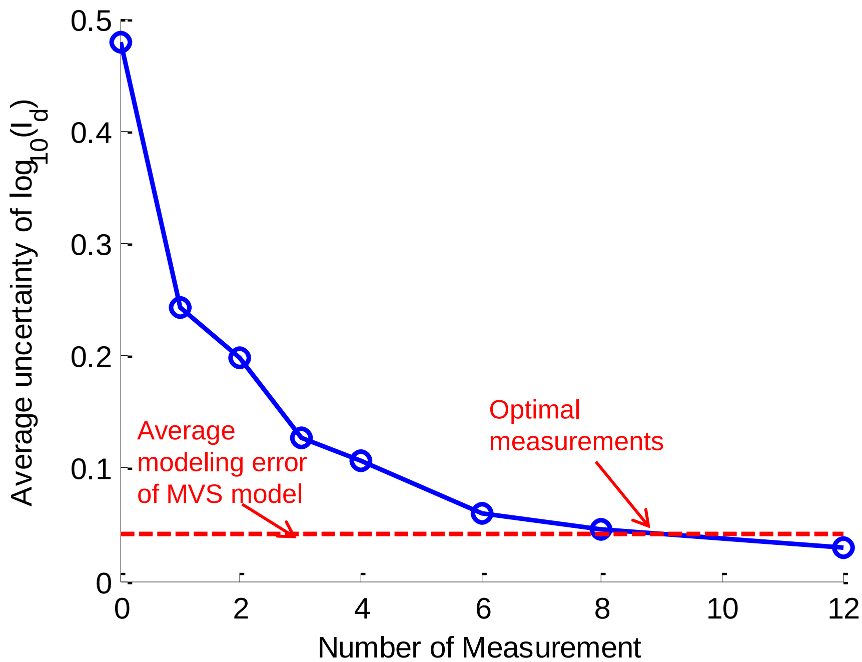


Figure 4-18: Average uncertainty (quantified by $\sigma(log_{10}(Id))$) versus number of measurements. The average modeling error of the MVS model is shown as the red dashed line.

relies on historical $I-V$ measurements by using a system identification approach. This allows us to separate measurement errors from modeling errors; the optimal measurement solution is driven only by minimizing output (selected measurement point) uncertainty, and is thus universal to all physical models (is not limited to the MVS model). However, the optimal number of measurements depends on the modeling error of the physical model being employed, as additional measurement points beyond that error limit does not improve accuracy. Future work could explore alternative optimal sampling approaches such as in (4.17), or by a combination of measurement *and* model parameter uncertainty minimization.

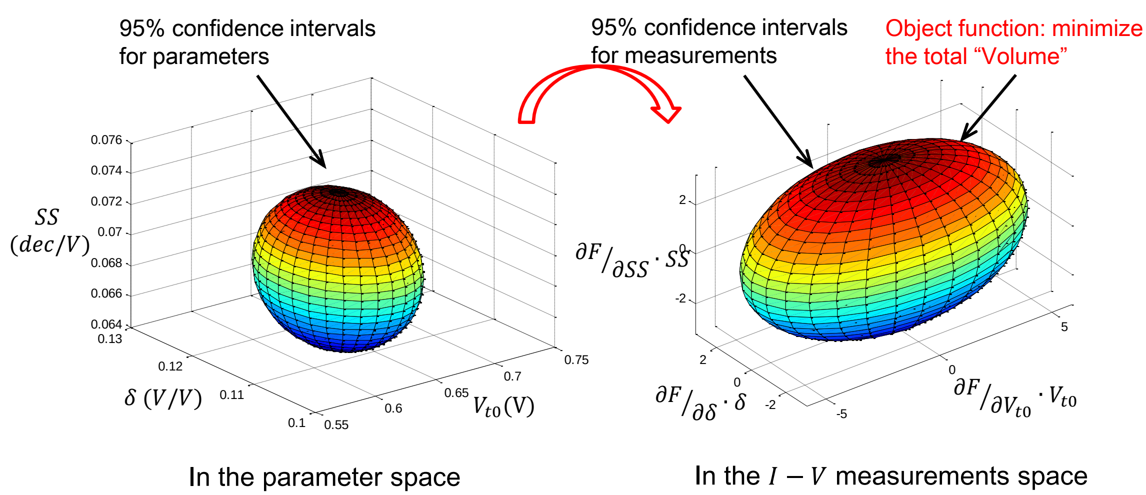


Figure 4-19: An alternative approach for the optimal design of experiment coupled to the MVS model is divided into two steps: (1) estimation of MVS model parameters, and (2) projection from parameter space to output space and minimization of the total "volume."

Chapter 5

Statistical Library Characterization Using Belief Propagation across Multiple Technology Nodes

5.1 Introduction

A standard cell library capturing statistical information of delay and output slew variations is at the core of statistical static timing analysis (SSTA), and cost efficient statistical characterization of such libraries has become essential. The most widely used statistical library cell characterization method is based on the look-up table (LUT) approach where gate propagation delay (t_d), output transition time (S_{out}) and their variations are stored in a look-up table with different combinations of inputs such as cell types, input slew (S_{in}), load capacitance (C_{load}), supply voltage (V_{dd}), and other parameters [89]. The runtime complexity required for such a statistical LUT-based approach is $O(N_{sample} \cdot N_{LUT})$, where N_{sample} is the number of SPICE runs needed to obtain each mean and variance value and N_{LUT} is the number of input vector combinations. This approach will quickly become infeasible as either

N_{LUT} or N_{sample} in a technology increases.

Historically, circuit level Monte Carlo (MC) simulation has been employed to generate a number of samples in the process parameter probability space [60]. Such approach allows variability-aware analysis to be implemented with minor changes on top of existing characterization tools but requires a large number of MC runs. To address this challenge, several approaches based on sensitivity analysis for library characterization have been proposed by EDA vendors. For instance, the Composite Current Source (CSC) approach is adopted by the Synopsys PrimeTime SSTA tool, and the sensitivity-based effective-current-source-model (S-ECSM) is adopted by the Cadence statistical tool. All of these approaches aim at modelling the statistical impact of process parameter variations as a linear superposition of the impact of each parameter in the response model of the affected metric. Several response surface methodologies (RSMs) have also been proposed to exploit the sparsity of the process regression coefficients. An example of such a strategy is least-angle regression (LAR) which uses L_1-norm regularization [46]. One major benefit of regularizing with the L_1-norm is that it results in sample complexity that is logarithmic in the number of features (e.g., principal components). For statistical characterization of standard cells, an error propagation technique using linear sensitivity analysis and response surface methodology (RSM) using Brussel design of experiments (DoE) was proposed for library characterization in [90]. The Brussel DoE performs statistical feature selection keeping only those features that are most relevant to the response under consideration. Then it uses a model selection algorithm to build a suitable regression model for all the responses. More recently, statistical circuit simulators based on uncertainty quantification have been successfully applied to avoid the huge number of repeated simulations in conventional Monte Carlo flows [88, 91, 92, 93, 94, 95].

On the other hand, the expensive simulation cost of the statistical LUT-based approach is not only due to high dimensionality of the process space, but also due to high dimensionality of the cell variable space (e.g., cell type, input slew S_{in}, load capacitance, supply voltage V_{dd}, *etc.*). This problem is further exacerbated as more design options are provided in recent technologies (e.g., multi-V_t, multi-V_{dd}). While most of the existing work focuses on exploiting the sparsity of the regression coefficients of the process space with a reduced process sample size for each variable space vector, correlations between different cells and different variable vectors within the same cell have not been considered in the open literature, to the best of our knowledge. This has been the main motivation of this work, which proposes a novel acceleration method that operates in the library variable space rather than its process space and that can be added to any acceleration used in the process space.

This is achieved through the systematic use of recent advances in statistics and semiconductor metrology that we apply to the development of computationally efficient statistical characterization algorithms for standard cell libraries. We propose two key techniques to explore correlations in the library variable space. The first is a novel ultra-compact, analytical model for gate timing characterisation, and the second is a Bayesian learning algorithm for the parameters of the aforementioned timing model using past library characterizations along with a very small set of additional simulations from the target technology. Bayesian approaches were initially introduced in the area of VLSI design for post-Silicon validation and parameter extraction [48, 50, 83, 96, 97]. The intrinsic simplicity of the proposed timing model combined with the Bayesian learning [44] framework is capable of building very accurate circuit response representations.

The rest of this chapter is organized as follows. Section 5.2 introduces basic notation and formulates the problem of statistical characterization in the *library variable space*. Section 5.3 describes prior work on gate delay modelling and presents our novel ultra-compact analytical model for gate delay and slew [98]. Section 5.4 presents our Bayesian algorithm which learns timing model parameters from past library characterizations and a very small set of additional simulation runs in the target technology. The foundation of this algorithm is the use of maximum a posteriori (MAP) estimation, as applied to other problems in Chapter 3 and 4. In Section 5.5, our new methods are validated on the library characterization in state-of-the-art $14nm$ and $28nm$ technology and compared with the LUT method.

5.2 Problem Formulation

In library characterization, an accurate model for cell delay (T_d) and output slew (S_{out}) is developed given the following variable data: a cell type, input slew (S_{in}), output load capacitance (C_{load}), transition direction (RISE/FALL), and supply voltage (V_{dd}). To formalize the library characterization problem, we consider an individual logic gate with multiple inputs and one output, and for simplicity, we start from the standard assumption that only one timing arc is modelled at a time, which implies that we do not consider simultaneous input switching. For p cell variables ($\xi = \{\xi_1, \xi_2, ..., \xi_p\}$), such as $S_{in}, V_{dd}, C_{load}, etc.$, the cell response is modeled as the following two functions:

$$T_d = f_T(\xi_1, \xi_2, ..., \xi_p) \tag{5.1}$$

$$S_{out} = f_S(\xi_1, \xi_2, ..., \xi_p) \tag{5.2}$$

The problem of nominal library characterisation is to estimate f_T and f_S given k cell variable vectors $\{\xi\} = \{\xi^{(1)}, \xi^{(2)}, ..., \xi^{(k)}\}$ and k output observations $\{T_d^{(1)}, T_d^{(2)}, ..., T_d^{(k)}\}$ and $\{S_{out}^{(1)}, S_{out}^{(2)}, ..., S_{out}^{(k)}\}$, such that the timing prediction error with respect to a baseline case is minimized under the condition that k is very small. The nominal baseline case is defined by SPICE simulations under n different variable vectors ($n >> k$) sampled randomly within the variable space ξ.

We denote by $\{T_d\}$ an ensemble of delay observations. This ensemble has been generated for a given variable vector but under varying process parameters. Now we formulate the problem of statistical library characterisation in *variable space* as that of estimating f_T and f_S given k variable vectors $\{\xi\} = \{\xi^{(1)}, \xi^{(2)}, ..., \xi^{(k)}\}$ and k *ensembles* of output observations $\{\{T_d^{(1)}\}, \{T_d^{(2)}\}, ..., \{T_d^{(k)}\}\}$ and $\{\{S_{out}^{(1)}\}, \{S_{out}^{(2)}\}, ..., \{S_{out}^{(k)}\}\}$, such that the prediction error for the statistical metrics with respect to a statistical baseline case is minimized under the condition that k is very small. The statistical baseline case is defined by statistical SPICE simulations using the same n different variable vectors ($n >> k$) as in the nominal baseline case, where the SPICE simulations are now executed according to the Monte Carlo method in process space.

The metrics of the statistical baseline case include the mean and standard deviation of delay and output slew at each variable vector $i \in \{1, ..., n\}$. They are denoted as $\mu_{T_d}^{(i)}$, $\mu_{S_{out}}^{(i)}$ and $\sigma_{T_d}^{(i)}$, $\sigma_{S_{out}}^{(i)}$ $(i = 1, 2, ..., n)$, respectively.

5.3 Model for Delay and Output Slew

Accurate gate level modeling for delay and slew estimation has become a major challenge for nanometric technologies. Historically, the transistor delay has been simply approximated by $C_{load}V_{dd}/I_{dsat}$, where I_{dsat} is the drain current at $V_{gs} = V_{ds} = V_{dd}$. A more accurate model, named the alpha-power law, was later proposed in the early 1990s [58] where a closed-form expression was derived for the delay of an inverter. A simplified version of the alpha-power law was proposed in [99]. More recently, a simple analytical expression for the intrinsic MOSFET delay, using physics-based models for the effective current and the total gate switching charge, was proposed to better describe nanometric technologies [100].

Although these advanced delay models provide accurate description of transition activity in the cell, they are still quite complex, and detailed process information is required to fit the entire model.

Our first goal therefore is to contribute an ultra compact timing model that is at once a generalisation of older models but whose parameters allow a sparse representation of vell variable space vectors. Fig. 5-1 (a) shows the key factors that affect the delay and output slew of an inverter. In this work, we consider the impact of input slew (S_{in}), output load capacitance (C_{load}), supply voltage (V_{dd}), and driving strength (I_{eff}).

To find our ultra compact model, we first study gate delay in a simple inverter and generalize it to any combinational logic cell. Recent studies [101, 102, 103] show that the simple $C_{load}V_{dd}/I_{dsat}$ metric follows the experimental inverter delay much better if the on-current in the denominator is replaced with an effective current I_{eff} representing the average switching current. In line with the intrinsic transistor delay

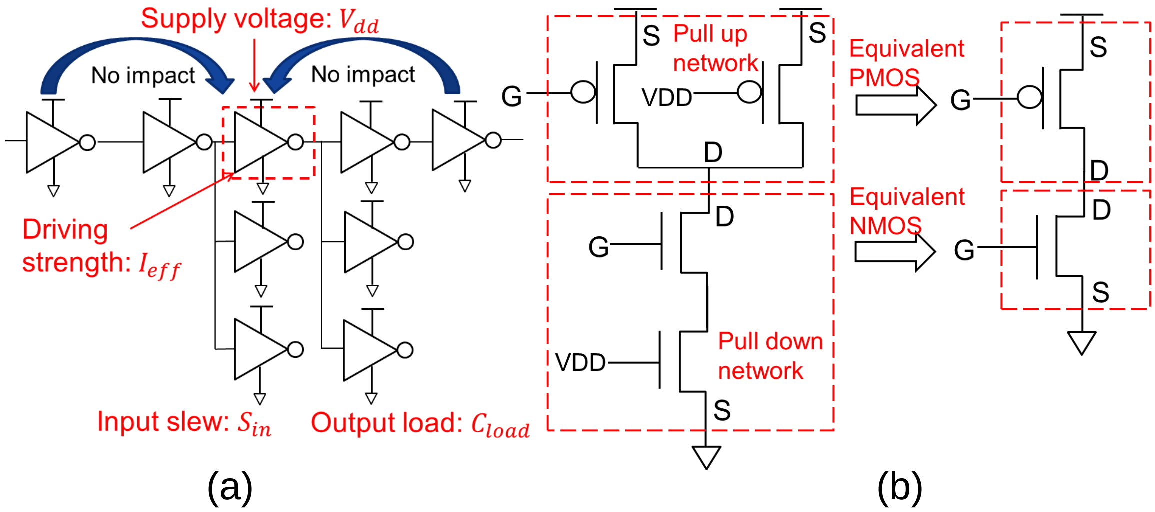


Figure 5-1: (a) Key factors that affect the delay and output slew of an inverter. (b) NAND2 equivalent inverter: The pull-up network is replaced with an "equivalent" PMOS while the pull-down network is replaced with an "equivalent" NMOS device.

defined in [100], we model cell delay as

$$T_d = k_d \frac{\Delta Q}{I_{eff}} \tag{5.3}$$

where k_d is a scaling factor used to obtain a good fit to the actual cell delays. I_{eff} is defined as

$$I_{eff} = \frac{I_d(V_{gs} = V_{dd}, V_{ds} = \frac{V_{dd}}{2}) + I_d(V_{gs} = \frac{V_{dd}}{2}, V_{ds} = V_{dd})}{2} \tag{5.4}$$

and can be evaluated easily through performance modeling or through a circuit simulation that takes into account process variations [100, 55]. Since our focus is to model delay and output slew as functions of cell variables, (5.1) and (5.2), we assume we know I_{eff} for each variable vector. Note that the direct link between process parameters and delay is still preserved in the I_{eff} current. To generalize the above model to any combinational logic cell, we simply replace each gate with an "equivalent inverter" and use the inverter characterization to estimate delays and output slews [104, 105, 106]. Each pull-up or pull-down network is modeled as a four terminal transistor by matching the "$I-V$" curves through SPICE simulation. Fig. 5-1(b) shows the equivalent inverter of a NAND2 where the pull-up network is replaced with a PMOS and the pull-down network is replaced with an NMOS device. The charge transferred to or from the load capacitance during switching is equal to

$$\Delta Q = (V_{dd} + V')(C_{load} + C_{par} + \alpha S_{in}) \tag{5.5}$$

where C_{par}, V' and α are all fitting parameters. Compared with the simple $C_{load}V_{dd}/I_{dsat}$ metric, several effects have been considered: (1) C_{par} is introduced to account for parasitic capacitance, such as those associated with junctions and interconnects, which are not included in C_{load}; (2) V' is introduced to compensate for the inaccuracy of the delay model at low V_{dd}; and (3) a linear coefficient α is introduced to account for input slew S_{in}'s impact on delay. The estimates of f_T and f_S are then converted to parameter extraction problems for $\{k_d, C_{par}, V', \alpha\}$.

A special feature of this simple delay model is that the same format is used to describe not only delay but also output slew S_{out}, albeit with a different set of values for the fitting parameters $\{k_d, C_{par}, V', \alpha\}$.

To validate the proposed model, $T_d \cdot I_{eff}/(V_{dd}+V')$ and $S_{out} \cdot I_{eff}/(V_{dd}+V')$ versus different V_{dd} values are shown in Fig. 5-2, where T_d and S_{out} are simulated through SPICE using a $14nm$ industrial design kit, and two separate V' values are extracted for T_d and S_{out}. For different groups of C_{load} and S_{in} combinations, a constant value of $T_d \cdot I_{eff}/(V_{dd} + V')$ and $S_{out} \cdot I_{eff}/(V_{dd} + V')$ is observed under different V_{dd}.

Fig. 5-3 shows $T_d/(C_{load} + C_{par} + \alpha S_{in})$ and $S_{out}/(C_{load} + C_{par} + \alpha S_{in})$ versus different C_{load} and S_{in} variable combinations. A similar result is observed here, that for different V_{dd} and transition (RISE/FALL) combinations, $T_d/(C_{load} + C_{par} + \alpha S_{in})$ and $S_{out}/(C_{load} + C_{par} + \alpha S_{in})$ are approximately constant.

Table 5.1 shows extracted parameters for the delay model from INV, NAND2 and NOR2 in three different technologies with their fitting errors. Strong similarities in extracted parameters are observed among different cells and technologies from different nodes, which serves as a basis for minimizing the required cell variable combinations in statistical characterization in the next section. Although our proposed

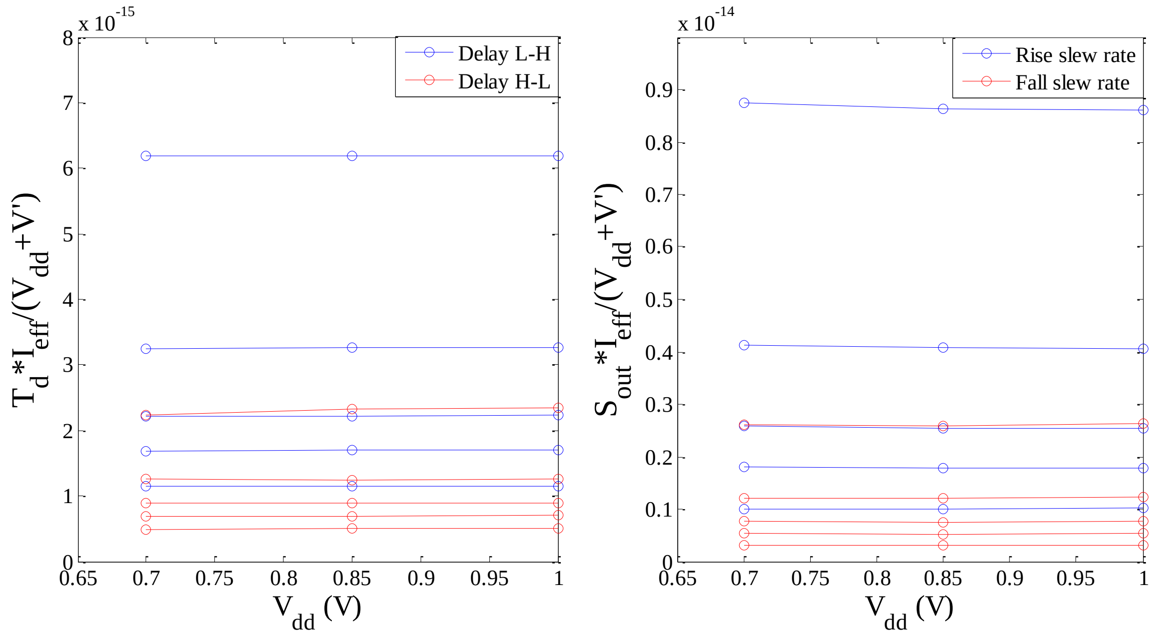


Figure 5-2: For a NOR2 cell designed in a commercial state-of-the-art 14-nm technology, a constant value of $T_d \cdot I_{eff}/(V_{dd}+V')$ and $S_{out} \cdot I_{eff}/(V_{dd}+V')$ is observed versus different V_{dd} and RISE/FALL combinations.

model captures major physical effects, for some technologies there might be an offset between the proposed model and circuit simulations. In those cases, extra fitting terms (e.g., $S_{in} \cdot C_{load}$) might be needed. The optimal model complexity will be given by a trade-off between model accuracy of degree of data compression.

5.4 Bayesian Inference with Maximum A Posteriori (MAP) Estimation

In this section, we present a Bayesian inference approach with maximum a posteriori (MAP) estimation, where instead of computing $\{T_d, S_{out}\}$ at each cell variable condition separately, we will estimate $\{k_d, C_{par}, V', \alpha\}$ globally by maximizing the joint probability of observing $(\xi^{(i)}, T_d^{(i)})$ or $(\xi^{(i)}, S_{out}^{(i)})$, $(i = 1, 2, ..., k)$. The formulation here is analogous to that presented in Chapter 4 for MVS model learning and estimation, but now applied to the problem of statistical timing characterization of a cell.

The first step is to transfer observed training samples $(\xi^{(i)}, T_d^{(i)})$ or $(\xi^{(i)}, S_{out}^{(i)})$,

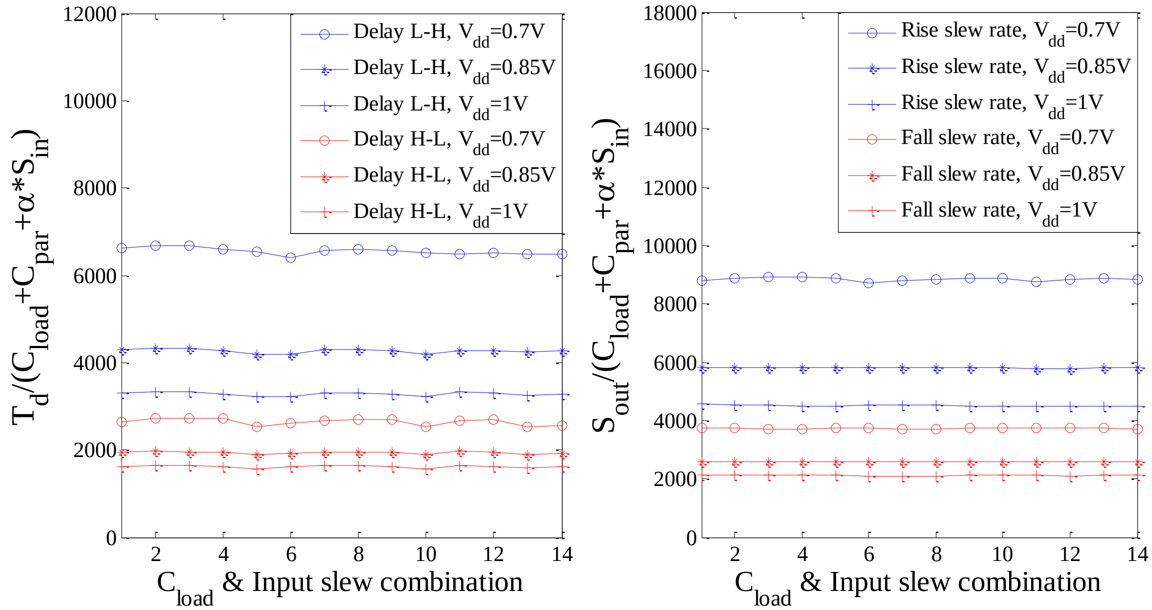


Figure 5-3: For a NOR2 cell designed in a commercial state-of-the-art $14nm$ technology, a constant value of $T_d/(C_{load} + C_{par} + \alpha S_{in})$ and $S_{out}/(C_{load} + C_{par} + \alpha S_{in})$ is observed versus different C_{load}, S_{in} and RISE/FALL combinations.

$(i = 1, 2, ..., k)$ to the parameter subspace $\{k_d, C_{par}, V', \alpha\}$ and use both to derive a probability distribution on the parameter space. The *pdf*'s on $\{k_d, C_{par}, V', \alpha\}$ for delay and output slew can then be calculated and the parameter extraction problem solved using maximum a posteriori (MAP) estimation.

Without loss of generality, we describe the MAP estimation for delay parameters group $\mathbf{P_T} = \{k_d, C_{par}, V', \alpha\}$. Parameters for output slew are estimated in a similar manner.

First, we assume that $\mathbf{P_T}$ follows a multivariate Gaussian distribution $\mathbf{P_T} \sim \mathcal{N}(\mu_{\mathbf{P_T}}, \mathbf{\Sigma_{P_T}})$:

$$pdf(\mathbf{P_T}) = \frac{1}{4\pi^2\sqrt{|\Sigma_{P_T}|}} \cdot exp[-\frac{1}{2}(\boldsymbol{\mu}_{P_T} - P_T)^{\mathbf{T}} \mathbf{\Sigma}_{P_T}^{-1} (\boldsymbol{\mu}_{P_T} - P_T)] \tag{5.6}$$

where $\mu_{\mathbf{P_T}}$ and $\mathbf{\Sigma_{P_T}}$ are the mean vector and covariance matrix of the parameter subgroup $\mathbf{P_T}$, respectively. Next, we assume that the $\boldsymbol{\mu}_{P_T}$ follows a conjugate Gaussian prior distribution $\boldsymbol{\mu}_{P_T} \sim \mathcal{N}(\boldsymbol{\mu}_{t0}, \mathbf{\Sigma}_{t0})$.

Table 5.1: Extracted parameters for delay model from INV, NAND2 and NOR2 in three different technologies with their fitting error.

Tech	Cell	k_d	$C_{par}(fF)$	$V'(V)$	α	% error
A	INV	0.389	0.951	-0.266	0.092	1.56%
A	NAND2	0.372	1.328	-0.209	0.034	1.98%
A	NOR2	0.356	1.186	-0.241	0.102	0.91%
B	INV	0.416	1.046	-0.287	0.103	1.50%
B	NAND2	0.403	1.471	-0.228	0.034	2.05%
B	NOR2	0.374	1.276	-0.253	0.104	1.12%
C	INV	0.389	0.978	-0.272	0.107	1.84%
C	NAND2	0.383	1.12	-0.258	0.050	1.94%
C	NOR2	0.368	1.225	-0.264	0.117	1.47%

$$pdf(\boldsymbol{\mu}_{P_T}) = \frac{1}{4\pi^2\sqrt{|\boldsymbol{\Sigma}_{\mathbf{t0}}|}} \cdot exp[-\frac{1}{2}(\boldsymbol{\mu}_{P_T} - \boldsymbol{\mu}_{t0})^{\mathbf{T}}\boldsymbol{\Sigma}_{t0}^{-1}(\boldsymbol{\mu}_{P_T} - \boldsymbol{\mu}_{t0})] \tag{5.7}$$

where $\mu_{\mathbf{t0}}$ and $\boldsymbol{\Sigma}_{\mathbf{t0}}$ are the mean vector and covariance matrix of $\mu_{\mathbf{P_T}}$, respectively. We also define the delay model precision as $\beta_{f_{T_d}}$, which equals the inverse variance of modeling errors across different technologies. Given $\boldsymbol{\mu}_{P_T}$ and $\beta_{f_{T_d}}$, we calculate the likelihood of observing the delay at the ith cell variable condition $T_d^{(i)}$ associated with subspace distribution $pdf(\boldsymbol{P_T})$ as

$$\begin{aligned} pdf(T_d^{(i)}|\boldsymbol{\mu}_{P_T}, \beta_{f_{T_d}}(\xi^{(i)})) = &\sqrt{\frac{\beta_{T_d}(\xi^{(i)})}{2\pi}} \\ &\cdot exp[-\frac{1}{2}(T_d^{(i)} - f_T(\xi^{(i)}, \boldsymbol{\mu}_{P_T}))^2\beta_{f_{T_d}}(\xi^{(i)})] \end{aligned} \tag{5.8}$$

As in Chapter 4, the learning of precision $\beta_{f_{T_d}}$ is a key step in this method. In practice, $\beta_{f_{T_d}}$ represents our "uncertainty" on the proposed delay model at different cell variable conditions due to its inability to capture certain physical effects. While they depend on the details of the technologies, these precisions show a strong systematic trend across different cell variable conditions ξ. In this work, extracted parameters $\mu_{\mathbf{P_T}}$ from past technologies are used to learn the systematic precision $\beta_{f_{T_d}}$ at different cell variable conditions. Characterizations from a variety of technology nodes enable us to propagate our historical belief to a new technology node. While generic or broad historical technologies can be used to learn approximate precisions, in order to achieve the highest applicable prior precision, the best historical technologies would be those with the same design or process choices as the target technology. For example, if we intend to fit a library in a low power process, appropriate historical technologies would also be technologies in low power processes. Therefore a bias-variance tradeoff is needed in the selection of historical libraries.

The detailed learning process proceeds as follows. First, a full set of standard cell libraries in N_{tech} fabrication processes and technology nodes ($N_{tech} = 6$ in Chapter 4, including technologies from $14nm$ to $45nm$, with both bulk-Silicon and SOI technologies and non-FINFET and FINFET technologies) are employed as "historical data" to improve our confidence in predicting $\beta_{f_{T_d}}$ on an unknown library. After selection of a group of historical libraries, each cell is fitted into the proposed delay model with different cell variable conditions ξ. $\beta_{f_{T_d}}$ is then calculated by the inverse variance of the relative difference between measurements and delay model predictions using extracted parameters.

$$\beta_{f_{T_d}} = \frac{1}{\frac{1}{N_{tech}} \sum_{j=1}^{N_{tech}} (\frac{T_d^{(j)} - f_T(P_T^{(j)})}{T_d^{(j)}})^2 - (\frac{1}{N_{tech}} \sum_{j=1}^{N_{tech}} \left| \frac{T_d^{(j)} - f_T(P_T^{(j)})}{T_d^{(j)}} \right|)^2} \tag{5.9}$$

As in Chapter 4, the maximum-a-posteriori (MAP) estimation finds optimal estimates of $\boldsymbol{\mu}_{P_T}$ that maximize the log likelihood of the posterior distributions $\ln pdf(\boldsymbol{\mu}_{P_T}|T_d)$. This can be mathematically formulated as an optimization problem

$$\underset{\boldsymbol{\mu}_{P_T}}{maximize} \ln pdf(\boldsymbol{\mu}_{P_T}) + \sum_{i=1}^{k} \ln pdf(T_d^{(i)}|\boldsymbol{\mu}_{P_T}, \beta_{f_{T_d}}(\xi^{(i)})) \tag{5.10}$$

Substituting (5.7) and (5.8) into (5.10) and removing the constant items yield:

$$\begin{aligned} \underset{\boldsymbol{\mu}_{P_T}}{minimize} \; & \frac{1}{2}(\boldsymbol{\mu}_{P_T} - \boldsymbol{\mu}_{t0})^{\mathbf{T}} \boldsymbol{\Sigma}_{t0}^{-1} (\boldsymbol{\mu}_{P_T} - \boldsymbol{\mu}_{t0}) \\ & + \frac{1}{2} \sum_{i=1}^{k} (T_d^{(i)} - f_T(\xi^{(i)}, \boldsymbol{\mu}_{P_T}))^2 \beta_{f_{T_d}}(\xi^{(i)}) \end{aligned} \tag{5.11}$$

where (5.11) is the summation of a concave quadratic function. Hence the optimization problem in (5.11) is also a convex programming problem and can be solved both efficiently and robustly.

So far we have achieved individual library cell characterization (no statistical characterization included). The detailed efficient statistical library cell characterization proceeds as follows. N_{sample} different seeds for each cell under process variation are generated through Monte Carlo (MC) simulation or design of experiments (DoE) [90]. For jth seed in each cell, $\{T_d\}$ and $\{S_{out}\}$ under k cell variable conditions are simulated through a SPICE simulation using the .ALTER statement. $P_T^{(j)}$ and $P_S^{(j)}$ are extracted through proposed Bayesian inference with maximum a posteriori (MAP) estimation for the jth seed. For a targeted cell variable condition ξ, the probability distribution of delay and output slew are calculated as $pdf(f_T(\xi, P_T))$ and $pdf(f_S(\xi, P_S))$.

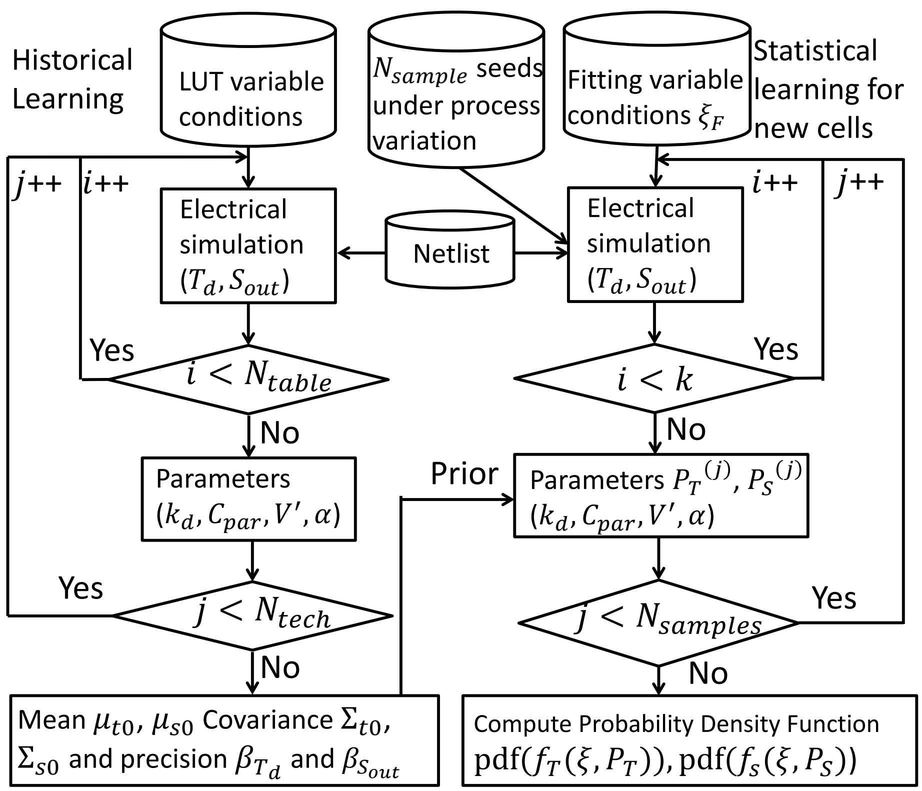


Figure 5-4: Proposed flow for statistical characterization with both old and new libraries interacting, and priors being passed from an old library to a new library.

Fig. 5-4 summarizes the major steps of the proposed statistical library cell characterization method, with both old and new libraries interacting, and priors being passed from an old library to a new one. If we assume that library cell characterizations have been done in previous technologies, the total computation cost is $O(k \cdot N_{sample})$, which is at least one order of magnitude smaller compared with $O(N_{LUT} \cdot N_{sample})$ in prior work and several order of magnitudes smaller than $O(N_{LUT} \cdot N_{MC})$ in the standard method. The total computation cost is $O(k \cdot N_{sample} + N_{Tech} \cdot N_{LUT})$ if we need to re-run characterization for old technologies, which is still a moderate speed up compared to traditional approaches.

5.5 Validation

In this section, two library cell characterization examples in several cutting-edge CMOS technologies are used to demonstrate the efficiency of our proposed method. All test cases as well as the historical library cell characteristics are generated using different BSIM based industrial design kits reflecting real measurements. To test and compare with traditional approaches, we have also implemented both deterministic extraction and statistical extraction using a look-up table (LUT) approach.

The baseline characterization is defined in this work by a 1000 point Monte Carlo simulation sampled randomly within the whole cell variable space $\xi = \{S_{in}, C_{load}, V_{dd}\}$. Note that these points only represent different operating conditions for a target cell, while the effects of process variation are not included. Fig. 5-5 shows a scatter plot for 1000 points among the cell variable space, where we will compare our characterization results with standard methods.

The first example is to conduct a nominal delay and output slew characterization for a library designed in a commercial state-of-the-art $14nm$ FINFET technology. Both fitting and testing samples are generated through SPICE simulation using a well calibrated compact transistor model. Fig. 5-6 shows average prediction error compared with the baseline characterization using the proposed model with Bayesian inference, the proposed model with least-square error (LSE) optimization, and a look-up table approach. To achieve the same characterization accuracy on delay T_d, our proposed method achieves up to 15X sample size reduction compared to a traditional lookup table approach, where 6X reduction is contributed by our proposed timing

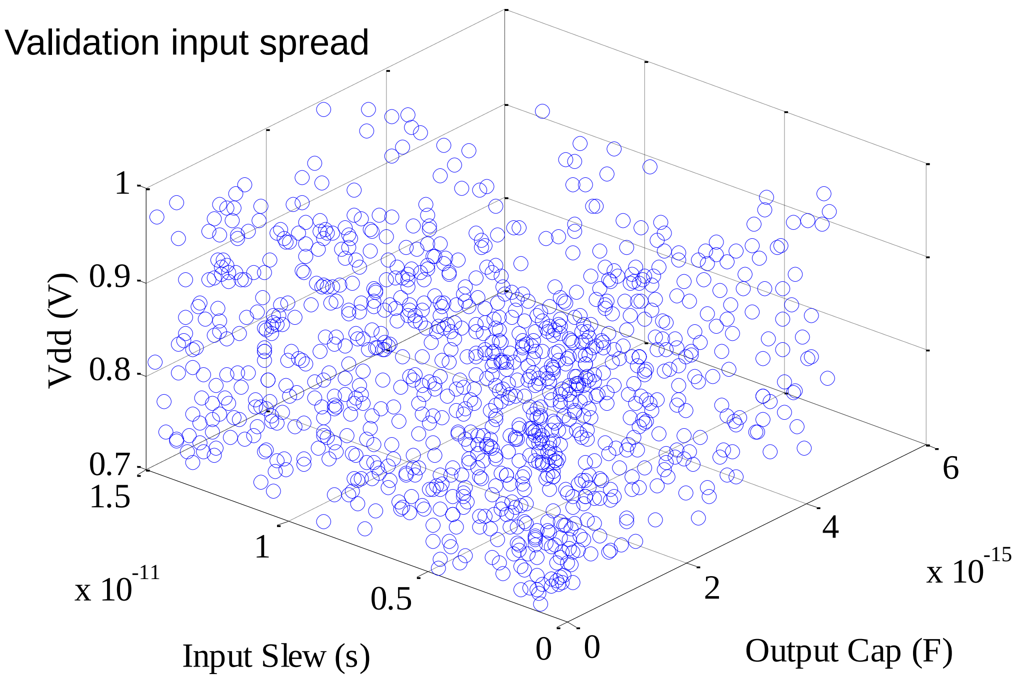


Figure 5-5: A scatter plot of 1000 points among the cell variable space $\xi = \{S_{in}, C_{load}, V_{dd}\}$ used for comparing our characterization results with standard methods.

model and an extra reduction of 2.5X is contributed by the Bayesian inference. Given the prior and two additional fitting cell variable combinations, a 4.3% average error compared with the baseline characterization is achieved for all combinations of C_{load}, S_{in} and V_{dd}. This demonstrates the sparsity of effects across cell variable vectors and the validity of the proposed delay model.

The second example is to conduct statistical delay and output slew characterization for a library designed in a commercial state-of-the-art $28nm$ bulk-Silicon technology, which is different from the model used in the first example. The baseline characterization is defined similar to the previous example, where 1000 cell variable combinations are sampled randomly within the whole space $\xi = \{S_{in}, C_{load}, V_{dd}\}$. In this case 1000 Monte Carlo simulations under process variation are generated for each of 1000 cell variable combinations to obtain statistical distributions for delay and output slew with different variable combinations.

The error functions for statistical characterization of $\mathcal{E}(\mu_{T_d})$, $\mathcal{E}(\mu_{S_{out}})$, $\mathcal{E}(\sigma_{T_d})$ and

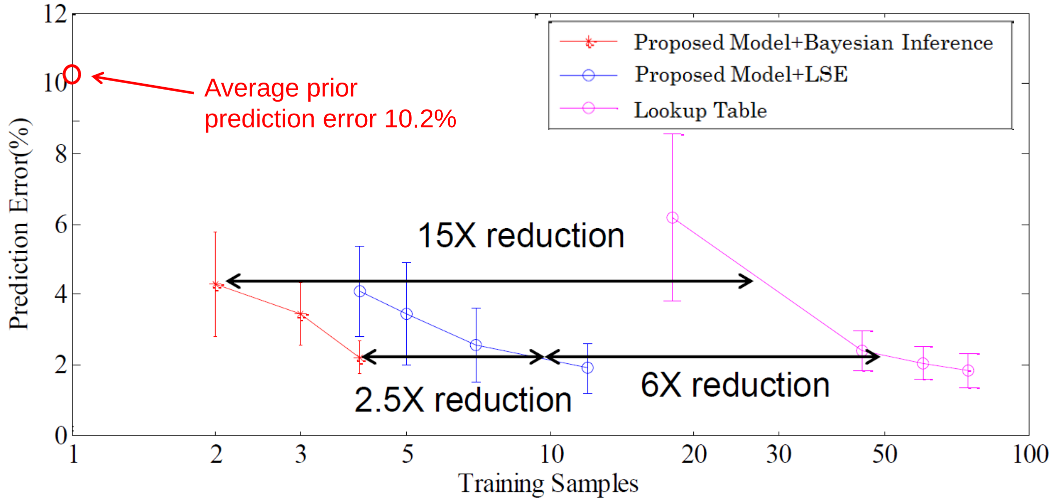


Figure 5-6: Average testing error for delay T_d characterizing a library designed in a commerical state-of-the-art 14nm technology. Error bars show one standard deviation of testing error for different cell and RISE/FALL combination.

$\mathcal{E}(\sigma_{S_{out}})$ are defined as

$$\mathcal{E}(\mu_{T_d}) = \frac{1}{n}\sum_{i=1}^{n}\left|\mu(f_T(\xi^{(i)}, P_T))) - \mu_{T_d}^{(i)}\right| \tag{5.12}$$

$$\mathcal{E}(\mu_{S_{out}}) = \frac{1}{n}\sum_{i=1}^{n}\left|\mu(f_S(\xi^{(i)}, P_S)) - \mu_{S_{out}}^{(i)}\right| \tag{5.13}$$

$$\mathcal{E}(\sigma_{T_d}) = \frac{1}{n}\sum_{i=1}^{n}\left|\sigma(f_T(\xi^{(i)}, P_T))) - \sigma_{T_d}^{(i)}\right| \tag{5.14}$$

$$\mathcal{E}(\sigma_{S_{out}}) = \frac{1}{n}\sum_{i=1}^{n}\left|\sigma(f_S(\xi^{(i)}, P_S)) - \sigma_{S_{out}}^{(i)}\right| \tag{5.15}$$

Fig. 5-7 and Fig. 5-8 show average prediction error for mean and standard deviation of delay and output slew characterizing a library designed in a commercial state-of-the-art 28nm technology using the proposed method and a look-up table approach, for the 1000 baseline combinations shown in Fig. 5-5. Up to 20X training set size reduction is observed to achieve the same characterization accuracy in mean value and standard deviation of T_d and S_{out}.

Fig. 5-9 shows delay probability density simulated for variable combination $V_{dd} = 0.734V$, $S_{in} = 5.09ps$, $C_{load} = 1.67fF$, the proposed method with seven training variable combinations, and an interpolation of look-up tables with 60 training cell variable combinations together with baseline distribution using SPICE Monte Carlo simulation. The proposed method shows a much better prediction for delay distribution that correctly predicts the non-Gaussian distribution for low V_{dd}.

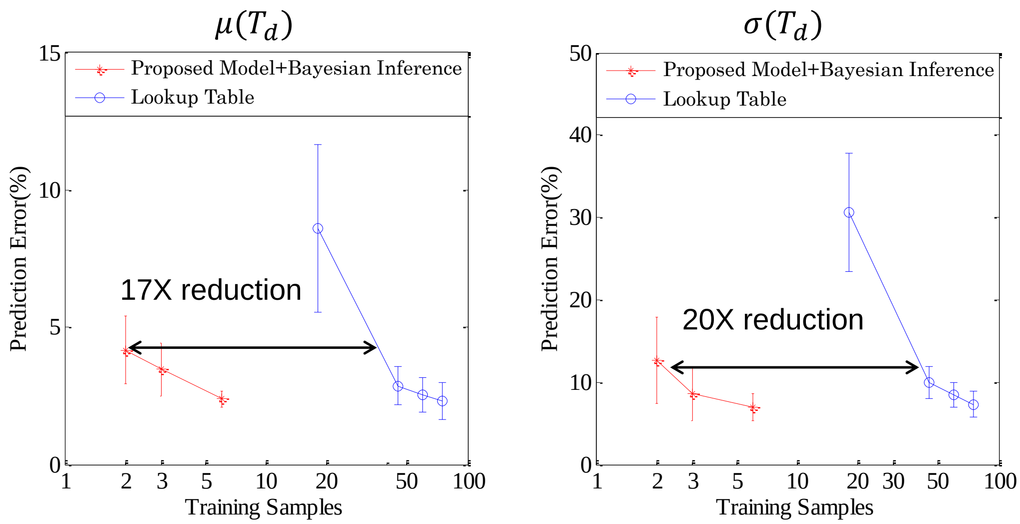


Figure 5-7: Average testing error for mean and standard deviation of delay T_d characterizing a library designed in a commerical state-of-the-art $28nm$ technology. Error bars show one standard deviation of testing error for different cell types and RISE/FALL combinations.

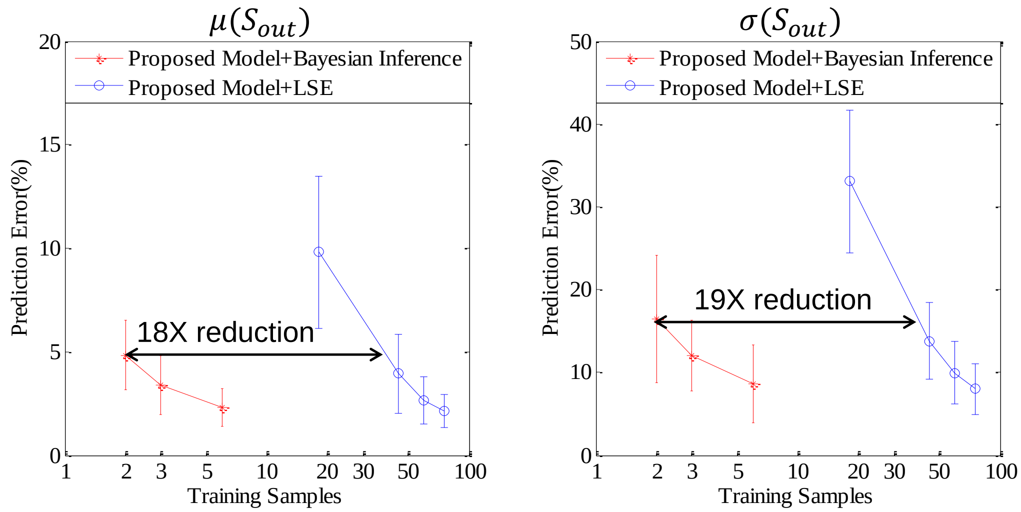


Figure 5-8: Average testing error for mean and standard deviation of output slew S_{out} characterizing a library designed in a commerical state-of-the-art $28nm$ technology. Error bars show one standard deviation of testing error for different cell types and RISE/FALL combinations.

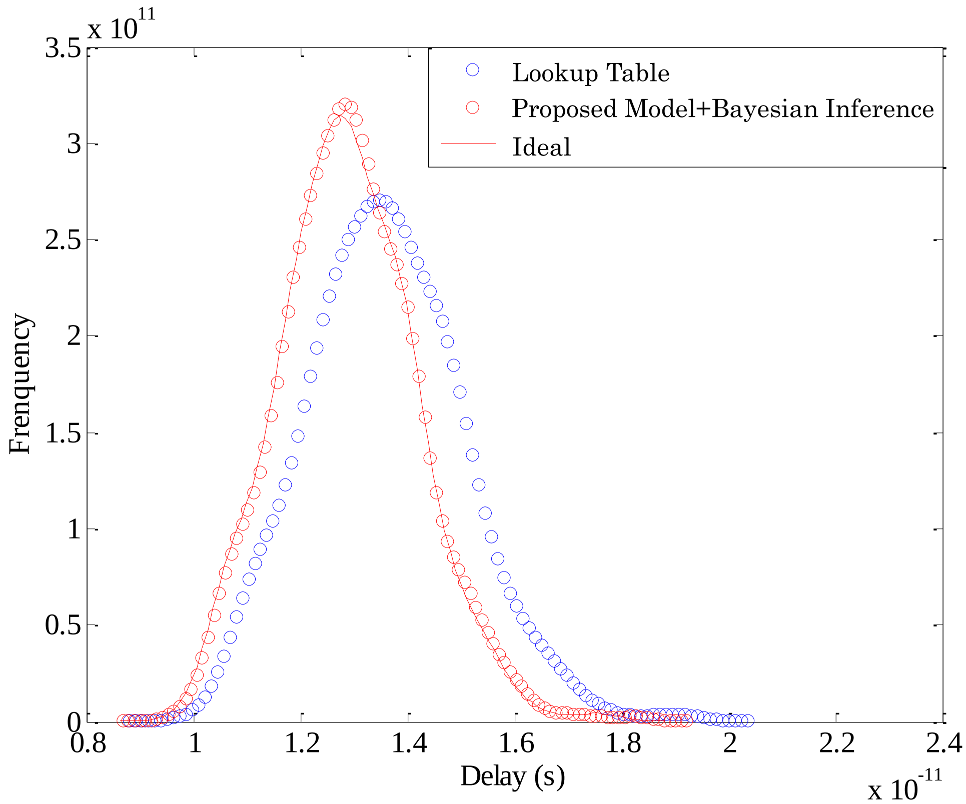


Figure 5-9: Delay probability density simulated for cell variable combination $V_{dd} = 0.734V$, $S_{in} = 5.09ps$, $C_{load} = 1.67fF$, using proposed method and an interpolation of look-up tables, together with baseline distribution using SPICE Monte Carlo simulation.

Chapter 6

Thesis Summary and Future Work

The contributions of this thesis, and possibilities for future work in this area are discussed in this chapter.

6.1 Thesis Contributions

In this thesis, we propose accurate and efficient statistical techniques to solve the transistor compact model parameter estimation and post-Silicon performance estimation problems. These techniques facilitate yield control throughout the product lifecycle, from early technology evaluation to process monitoring during mass-production, which is vital to rapidly improving yield.

More specifically, Fig. 6-1 shows several key issues in device variation and statistical compact modeling which are addressed by this thesis. The major technical contributions of this thesis are summarized below:

• The existing MIT virtual source (MVS) model with necessary statistical formulation is extended to support circuit variation analysis and extraction using backward propagation of variance (BPV). Variability parameters from the statistical MVS model have been derived directly from the nominal MVS model. Accurate statistical circuit performance using the statistical MVS model is demonstrated, including statistical characterization for standard cells, SRAMs, and D flip-flops.

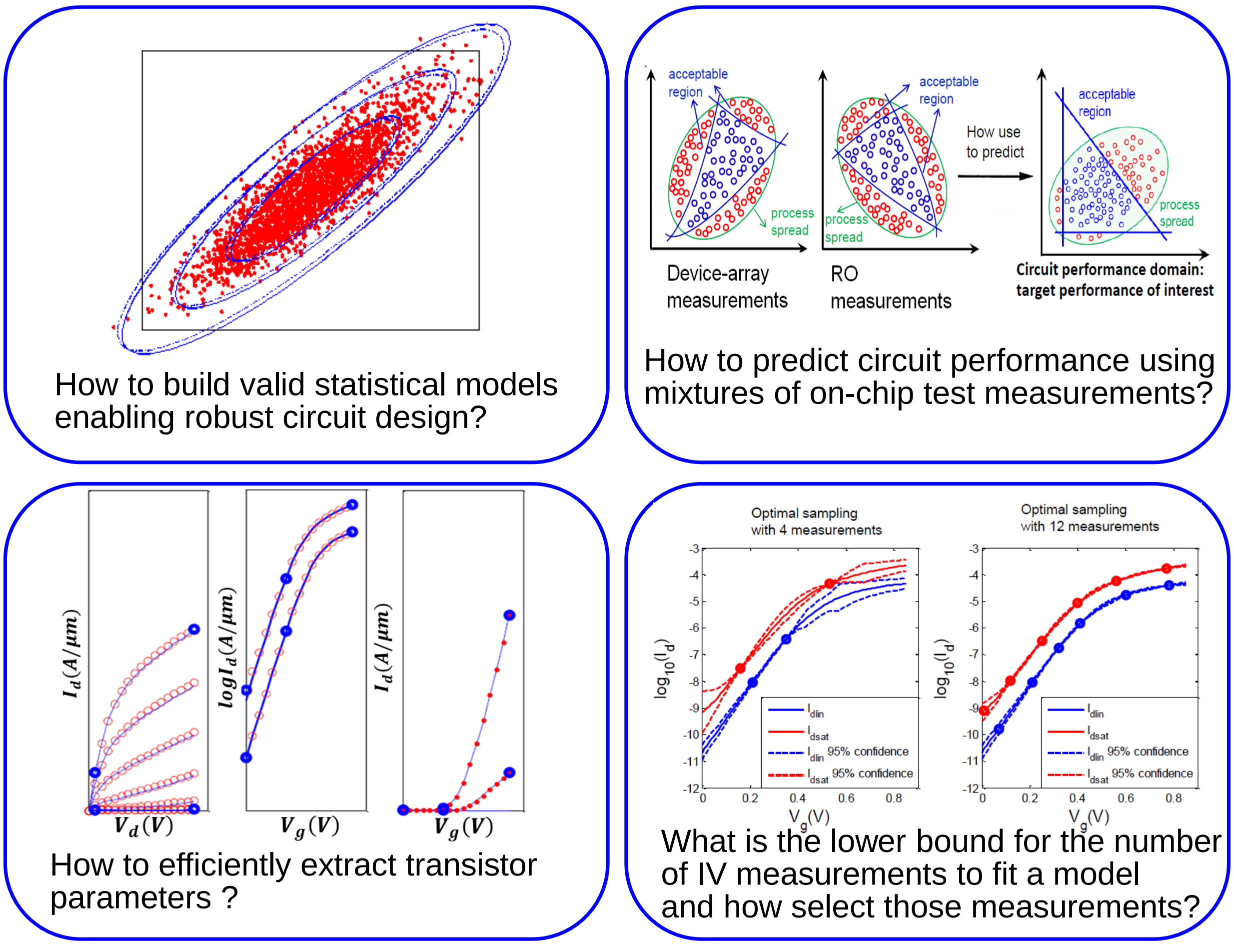


Figure 6-1: Key issues in device variation and statistical compact modeling.

- A novel methodology for integrated circuit performance estimation is proposed by joint modeling of measurements from different on-chip test structures. The fact that data arising from a variety of test structures are typically physically correlated under different circuit configurations and topologies is exploited. This is first achieved by a *physical subspace projection* technique to project different groups of on-chip measurements to a unique subspace likelihood map spanned by a set of physical variables. Then a Bayesian treatment is developed by introducing prior distributions over these subspace variables. Furthermore, an expectation-maximization (EM) algorithm is applied for maximum a posteriori (MAP) estimation in circuit performance.

- A novel MOSFET parameter extraction method to enable early technology evaluation is proposed. The distinguishing feature of the proposed method is that it enables the extraction of an entire set of MOSFET $I-V$ model parameters using limited and incomplete $I-V$ measurements from on-chip monitor circuits. An important step in this method is the use of maximum a posteriori estimation where past measurements of transistors from various technologies are used to learn a prior distribution and its uncertainty matrix for the parameters of the target technology. The proposed extraction can also be used to characterize the statistical variations of MOSFETs, with the significant benefit that some constraints required by the backward propagation of variance (BPV) method are relaxed.

- The lower bound requirement for number and selection of transistor measurements to extract the full set of $I-V$ parameters for a compact model is studied, and an efficient algorithm is proposed and demonstrated for selecting the optimal measurement biases by minimizing the average uncertainty.

- A novel flow to enable computationally efficient statistical characterization of delay and slew in standard cell libraries is proposed. This is traditionally modeled by a look-up table (LUT) approach. While existing work was focused on exploiting the sparsity of the regression coefficients of the process space with a reduced process sample size for each cell variable space vector, correlations between different cells and different cell variable vectors within the same cell are exploited in our proposed approach. Two key techniques are proposed to exploit correlations in library variable space. The first is a novel ultra-compact, analytical model for gate timing characterization, and the second is a Bayesian learning algorithm for estimating the parameters of the aforementioned timing model using past library characterizations along with a very small set of additional simulations from the target technology.

6.2 Future Work

Future work in this area could involve the refinement of the statistical techniques developed in this work to predict circuit performance and provide more accuracy through a deeper learning of fabrication process and device physics. Trends in process development and physical modeling need to be modeled in a quantitative way using a general method.

6.2.1 Extension of MIT Virtual Source (MVS) Model as a Predictive Statistical Compact Model

In Chapter 2 and Chapter 4, we have shown that the MVS model is a physically-based compact model and that model parameters can be correctly extracted individually and statistically from measurements, even if they are strongly correlated. However, to accurately predict the characteristics of nanoscale CMOS devices for early circuit design, it is critical to develop a predictive statistical ultra-compact model that includes process variations and correlations among process model parameters. The predictive MOSFET models should be ultra-compact, reasonably accurate, scalable with main process and design knobs, and capable of correctly capturing emerging device physical effects that are strongly influenced by these process trends. The MIT virtual source (MVS) model is one natural candidate for extension to meet such requirements.

However, such predictive MOSFET models significantly rely on empirical extrapolations, in combination with the understanding of physical principles involved. One approach is to integrate a methodology to correctly capture the correlations among process model parameters into the model, as presented in [28]. For example, the Predictive Technology Model (PTM) identifies and integrates critical correlations among L_{eff}, V_{th}, μ, and v_{sat}. Then the scaling trend of key physical parameters can be derived, as a function of geometry and process dependencies. For example, Fig. 6-2 shows the trend of effective oxide thickness (EOT) scaling from 250 to 32nm nodes [28]; such trends could be captured in an extended compact model. However, such extensions inevitably introduce extra dependencies in the model, and add

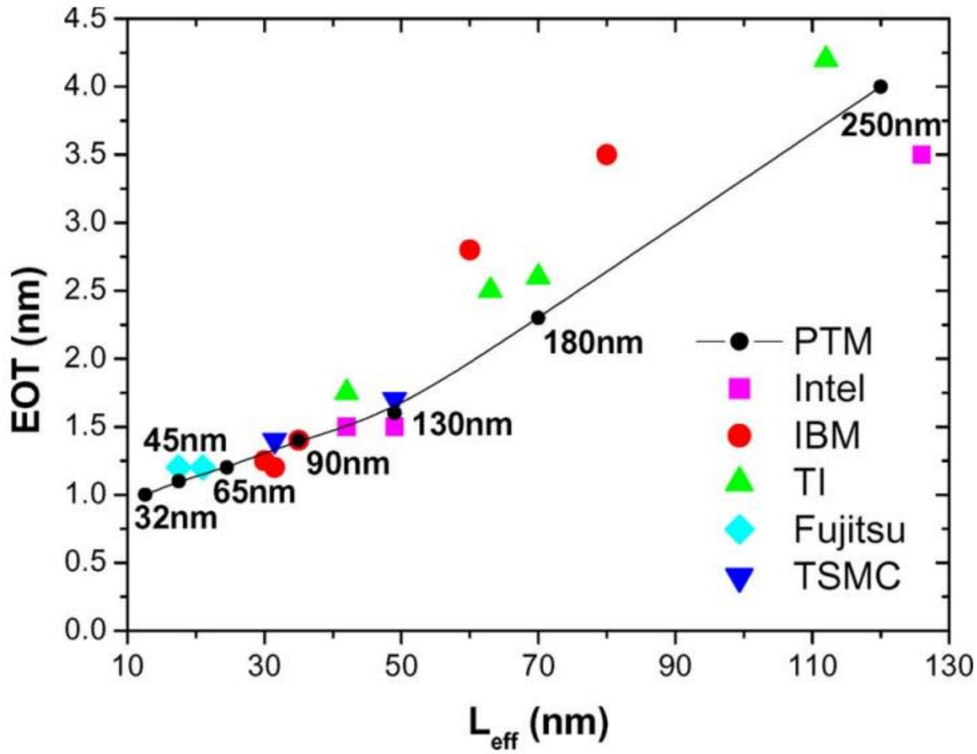


Figure 6-2: Trend of effective oxide thickness (EOT) scaling from 250- to 32-nm nodes [28].

complexities to the parameter extraction procedure.

A better approach is to include such process and device correlations in the parameter extraction procedure, rather than directly include the correlations in the model. The Bayesian extraction framework proposed in Chapter 4 provides such benefits. A new objective function can be described as

$$\begin{aligned}\mathcal{E}(\mathbf{P_{sub}}) = &\frac{1}{2}(\boldsymbol{\mu}_{L_{eff}} - \boldsymbol{\mu}_{s0})^{\mathbf{T}}\boldsymbol{\Sigma}_{s0}^{-1}(\boldsymbol{\mu}_{L_{eff}} - \boldsymbol{\mu}_{s0}) \\ &+\frac{1}{2}\sum_{n=1}^{N}\beta_{\mathbf{ln}F_n}\{\mathbf{ln}(F_n) - \mathbf{ln}(f(\mathbf{V_n}, \mathbf{P_{sub}}, \mathbf{P_{above}}))\}^2\end{aligned} \quad (6.1)$$

The difference between $\mu_{P_{sub}}$ in (4.13) and $\mu_{L_{eff}}$ in (6.1) is illustrated conceptually in Fig. 6-3, where $\mu_{P_{sub}}$ is the mean vector of extracted parameters from historical technologies, and $\mu_{L_{eff}}$ is the empirical extrapolations of the extracted parameters. The calculation of $\boldsymbol{\Sigma}_{s0}$ needs to be adjusted accordingly, and the uncertainty of $\mu_{L_{eff}}$ is significantly smaller than that of $\mu_{P_{sub}}$. This method allows us to preserve the physical correlation among parameters, without needing to include any explicit correlation equations in the MVS model.

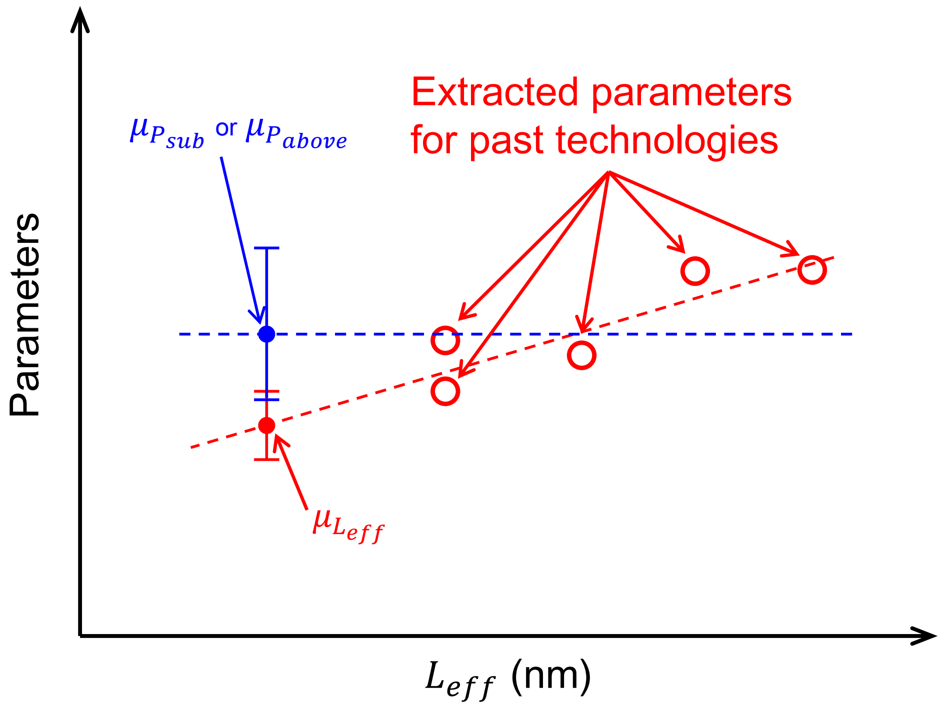


Figure 6-3: Difference between $\mu_{P_{sub}}$ and $\mu_{L_{eff}}$: $\mu_{P_{sub}}$ is the mean vector of extracted parameters from historical technologies, and $\mu_{L_{eff}}$ is the empirical extrapolations of the extracted parameters. Thus trend for process parameters such L_{eff} enable us to extrapolate impact of process trends on device performance.

6.2.2 Variation Prediction in a Process Development Cycle Using Bayesian Inference

For the Bayesian extraction framework described in Chapter 4, we made an important assumption that the uncertainty matrix of each $I-V$ measurement is fixed given a detailed technology. We learn this uncertainty matrix for the parameters of the target technology from historical technologies. 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 better product to the market ahead of the competition, and to continually improve 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 one follows every microarchitectural change with a die shrink of the process technology [107]. For a ”Tick” period, a new line of processors are released shortly after a shrinking of the process technology. However, simulations with an early version of the design kit may have large differences with realistic or later manufacturing output of the technology. 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. That is to say, extrapolations of the shrinking covariance matrices for model parameters over time as the process matures, could be used to predict yield improvement trends or expectations.

An example is shown for hypothetical parameter variations for several process development cycles in Fig. 6-4. 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 is able to achieve tighter tolerances that typically at the beginning of the technology introduction. Another observation is the trend that more variation is associated with each generation in technology scaling. For example, the device saturation current depends inversely on the channel length of the device; a percent deviation from a smaller nominal channel length will result in a larger percent deviation in saturation current than that caused by the same percent deviation from a larger nominal channel length. A methodology could be introduced using the Bayesian framework and uncertainty analysis proposed in Chapter 4 to learn the evolution of process tolerances over the lifetime of a manufacturing technology. At the same time, incorporation of major technology changes that substantially change the variation trend could also be accommodated (e.g., the reduction of line edge roughness impact arising from multiple patterning approaches for FINFETs [108]).

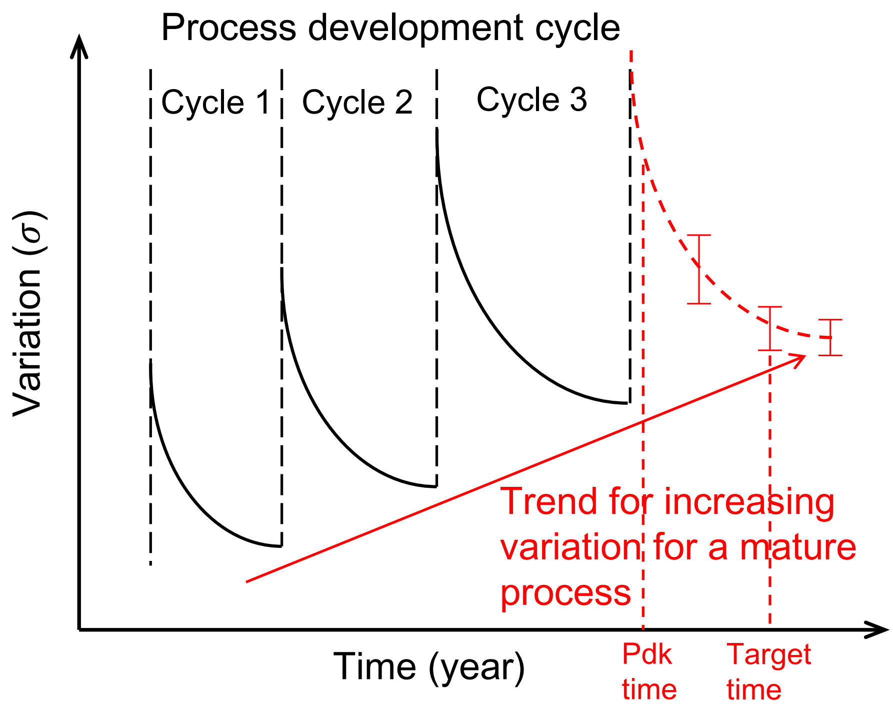


Figure 6-4: Hypothetical parameter variations for several process development cycles. The goal is to dynamically predict parameter variations for a later targeted release date with only early stage process information for the targeted technology and historical information on how past technologies have evolved for a complete process development cycle.

Bibliography

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

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

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

[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.

[5] R. Keyes, "The effect of randomness in the distribution of impurity atoms on FET thresholds," *Applied physics*, vol. 8, no. 3, pp. 251–259, 1975. [Online]. Available: http://dx.doi.org/10.1007/BF00896619

[6] E. Colgan, R. Polastre, M. Takeichi, and R. Wisnieff, "Thin-film-transistor process-characterization test structures," *IBM Journal of Research and Development*, vol. 42, no. 3.4, pp. 481–490, May 1998.

[7] M. G. Buehler, "Microelectronic test chips for VLSI electronics," in *VLSI Electronics Microstructure Science*. New York: Academic Press, 1983, vol. 6, ch. 9.

[8] 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.

[9] 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.

[10] 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.

[11] K. Gettings and D. Boning, "Study of CMOS process variation by multiplexing analog characteristics," *IEEE Transactions on Semiconductor Manufacturing*, vol. 21, no. 4, pp. 513–525, Nov. 2008.

[12] K. Agarwal, F. Liu, C. McDowell, S. Nassif, K. Nowka, M. Palmer, D. Acharyya, and J. Plusquellic, "A test structure for characterizing local device mismatches," in *VLSI Circuits, 2006. Digest of Technical Papers. 2006 Symposium on*, June 2006, pp. 67–68.

[13] K. Agarwal, J. Hayes, and S. Nassif, "Fast characterization of threshold voltage fluctuation in mos devices," *IEEE Transactions on Semiconductor Manufacturing*, vol. 21, no. 4, pp. 526–533, Nov. 2008.

[14] R. Rao, K. Jenkins, and J.-J. Kim, "A completely digital on-chip circuit for local-random-variability measurement," in *Solid-State Circuits Conference, 2008. ISSCC 2008. Digest of Technical Papers. IEEE International*, Feb. 2008, pp. 412–623.

[15] S. Mukhopadhyay, K. Kim, K. Jenkins, C.-T. Chuang, and K. Roy, "Statistical characterization and on-chip measurement methods for local random variability of a process using sense-amplifier-based test structure," in *Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers. IEEE International*, Feb. 2007, pp. 400–611.

[16] A. Chang, "A test structure for the measurement and characterization of layout-induced transistor variation," Master's thesis, Massachusetts Institute of Technology, Department of Electrical Engineering, 2009.

[17] L. Yu, "A study of through-silicon-via (TSV) induced transistor variation," Master's thesis, Massachusetts Institute of Technology, Department of Electrical Engineering, 2011.

[18] N. Drego, "Characterization and mitigation of process variation in digital circuits and systems," Ph.D. dissertation, Massachusetts Institute of Technology, Department of Electrical Engineering, 2009.

[19] K. Balakrishnan and D. Boning, "Measurement and analysis of contact plug resistance variability," in *Custom Integrated Circuits Conference, 2009. CICC '09. IEEE*, Sept. 2009, pp. 415–422.

[20] 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 Transactions on Semiconductor Manufacturing*, vol. 17, no. 1, pp. 2–11, Feb. 2004.

[21] N. Drego, A. Chandrakasan, and D. Boning, "A test-structure to efficiently study threshold-voltage variation in large mosfet arrays," in *Quality Electronic Design, 2007. ISQED '07. 8th International Symposium on*, March 2007, pp. 281–286.

[22] J. S. Panganiban, “A ring oscillator based variation test chip,” Master’s thesis, Massachusetts Institute of technology, Department of Electrical Engineering, 2002.

[23] L.-T. Pang and B. Nikolic, “Impact of layout on 90nm CMOS process parameter fluctuations,” in *VLSI Circuits, 2006. Digest of Technical Papers. 2006 Symposium on*, 2006, pp. 69–70.

[24] M. Bhushan, A. Gattiker, M. Ketchen, and K. Das, “Ring oscillators for CMOS process tuning and variability control,” *IEEE Transactions on Semiconductor Manufacturing*, vol. 19, no. 1, pp. 10–18, Feb. 2006.

[25] L. Yu, W.-Y. Chang, K. Zuo, J. Wang, D. Yu, and D. Boning, “Methodology for analysis of TSV stress induced transistor variation and circuit performance,” in *Quality Electronic Design (ISQED), 2012 13th International Symposium on*, March 2012, pp. 216–222.

[26] 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.

[27] 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.

[28] W. Zhao and Y. Cao, “New generation of predictive technology model for sub-45 nm early design exploration,” *IEEE Transactions on Electron Devices*, vol. 53, no. 11, pp. 2816 –2823, Nov. 2006.

[29] J. D. Trimmer, *Response of Physical Systems*. New York, Wiley, 1950.

[30] M. Sharma and N. Arora, “Optima: A nonlinear model parameter extraction program with statistical confidence region algorithms,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 12, no. 7, pp. 982–987, Jul. 1993.

[31] K. Krishna and S. Director, “A novel methodology for statistical parameter extraction,” in *Computer-Aided Design, 1995. ICCAD-95. Digest of Technical Papers., 1995 IEEE/ACM International Conference on*, Nov. 1995, pp. 696–699.

[32] 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.

[33] 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.

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

[35] I. Jolliffe, *Principal Component Analysis.* Springer Verlag, 1986.

[36] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes: the art of scientific computing.* Cambridge Univ. Press, New York, 1986.

[37] L. Daniel and J. White, "Automatic generation of geometrically parameterized reduced order models for integrated spiral RF-inductors," in *Behavioral Modeling and Simulation, 2003. BMAS 2003. Proceedings of the 2003 International Workshop on*, Oct. 2003, pp. 18–23.

[38] Z. Zhang, I. Elfadel, and L. Daniel, "Model order reduction of fully parameterized systems by recursive least square optimization," in *Computer-Aided Design (ICCAD), 2011 IEEE/ACM International Conference on*, Nov. 2011, pp. 523–530.

[39] Z. Zhang, Q. Wang, N. Wong, and L. Daniel, "A moment-matching scheme for the passivity-preserving model order reduction of indefinite descriptor systems with possible polynomial parts," in *Design Automation Conference (ASP-DAC), 2011 16th Asia and South Pacific*, Jan. 2011, pp. 49–54.

[40] B. Bond and L. Daniel, "Parameterized model order reduction of nonlinear dynamical systems," in *Computer-Aided Design, 2005. ICCAD-2005. IEEE/ACM International Conference on*, Nov. 2005, pp. 487–494.

[41] L. Daniel, O. C. Siong, L. S. Chay, K. H. Lee, and J. White, "A multiparameter moment-matching model-reduction approach for generating geometrically parameterized interconnect performance models," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 23, no. 5, pp. 678–693, Nov. 2006.

[42] B. Bond and L. Daniel, "A piecewise-linear moment-matching approach to parameterized model-order reduction for highly nonlinear systems," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 26, no. 12, pp. 2116–2129, Dec. 2007.

[43] 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.

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

[45] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, pp. 407–499, 2004.

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

[47] W. Zhang, T. Chen, M. Ting, and X. Li, "Toward efficient large-scale performance modeling of integrated circuits via multi-mode/multi-corner sparse regression," in *Design Automation Conference (DAC)*, 2010, pp. 897–902.

[48] F. Wang, W. Zhang, S. Sun, X. Li, and C. Gu, "Bayesian model fusion: large-scale performance modeling of analog and mixed-signal circuits by reusing early-stage data," in *Design Automation Conference (DAC)*, 2013, pp. 64:1–64:6.

[49] L. Yu, S. Saxena, C. Hess, I. Elfadel, D. Antoniadis, and D. Boning, "Efficient performance estimation with very small sample size via physical subspace projection and maximum a posteriori estimation," in *Design, Automation and Test in Europe (DATE)*, March 2014, pp. 1–6.

[50] S. Sun, F. Wang, S. Yaldiz, X. Li, L. Pileggi, A. Natarajan, M. Ferriss, J. Plouchart, B. Sadhu, B. Parker, A. Valdes-Garcia, M. Sanduleanu, J. Tierno, and D. Friedman, "Indirect performance sensing for on-chip analog self-healing via Bayesian model fusion," in *Custom Integrated Circuits Conference (CICC)*, Sep. 2013, pp. 1–4.

[51] W. Zhang, X. Li, T. Liu, E. Acar, R. Rutenbar, and R. Blanton, "Virtual probe: A statistical framework for low-cost silicon characterization of nanoscale integrated circuits," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 30, no. 12, pp. 1814–1827, Dec. 2011.

[52] A. Khakifirooz, O. Nayfeh, and D. Antoniadis, "A simple semiempirical short-channel MOSFET current-voltage model continuous across all regions of operation and employing only physical parameters," *IEEE Transactions on Electron Devices*, vol. 56, no. 8, pp. 1674 –1680, Aug. 2009.

[53] C. Jeong, D. Antoniadis, and M. Lundstrom, "On backscattering and mobility in nanoscale silicon MOSFETs," *IEEE Transactions on Electron Devices*, vol. 56, no. 11, pp. 2762 –2769, Nov. 2009.

[54] L. Wei, O. Mysore, and D. Antoniadis, "Virtual-source-based self-consistent current and charge FET models: From ballistic to drift-diffusion velocity-saturation operation," *IEEE Transactions on Electron Devices*, vol. 59, no. 5, pp. 1263 –1271, May 2012.

[55] S. Rakheja and D. Antoniadis, "MVS 1.0.1 nanotransistor model (silicon)," Nov 2013. [Online]. Available: https://nanohub.org/resources/19684

[56] M. Lundstrom and D. Antoniadis, “Compact models and the physics of nanoscale FETs,” *IEEE Transactions on Electron Devices*, vol. 61, no. 2, pp. 225–233, Feb. 2014.

[57] E. Consoli, G. Giustolisi, and G. Palumbo, “An accurate ultra-compact I-V model for nanometer MOS transistors with applications on digital circuits,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 59, no. 1, pp. 159 –169, Jan. 2012.

[58] T. Sakurai and A. Newton, “Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas,” *IEEE Journal of Solid-State Circuits*, vol. 25, no. 2, pp. 584–594, Apr. 1990.

[59] D. Boning, K. Balakrishnan, H. Cai, N. Drego, A. Farahanchi, K. Gettings, D. Lim, A. Somani, H. Taylor, D. Truque, and X. Xie, “Variation,” in *International Symposium on Quality Electronic Design (ISQED)*, Mar. 2007, pp. 15 – 20.

[60] L. Yu, L. Wei, D. Antoniadis, I. Elfadel, and D. Boning, “Statistical modeling with the virtual source MOSFET model,” in *Design, Automation Test in Europe Conference Exhibition (DATE)*, 2013, pp. 1454–1457.

[61] P. Drennan and C. McAndrew, “A comprehensive MOSFET mismatch model,” in *International Electron Devices Meeting (IEDM)*, 1999, pp. 167 – 170.

[62] M. Lundstrom, Z. Ren, and S. Datta, “Essential physics of carrier transport in nanoscale MOSFETs,” in *International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, Mar. 2000, pp. 1–5.

[63] G. Wright, “Threshold modelling of MOSFETs for CAD of CMOS-VLSI,” *Electronics Letters*, vol. 21, no. 6, pp. 223 –224, 14 1985.

[64] L. Yu, O. Mysore, L. Wei, L. Daniel, D. Antoniadis, I. Elfadel, and D. Boning, “An ultra-compact virtual source FET model for deeply-scaled devices: Parameter extraction and validation for standard cell libraries and digital circuits,” in *Asia and South Pacific Design Automation Conference(ASPDAC)*, 2013, pp. 521–526.

[65] “BSIMSOI compact MOSFET model.” [Online]. Available: http://www-device.eecs.berkeley.edu/bsim/?page=BSIMSOI

[66] Y. Ye, S. Gummalla, C. Wang, C. Chakrabarti, and Y. Cao, “Random variability modeling and its impact on scaled CMOS circuits,” *J. Comput. Electron.*, vol. 9, no. 3 - 4, pp. 108–113, Dec. 2010.

[67] A. Khakifirooz and D. Antoniadis, “Transistor performance scaling: The role of virtual source velocity and its mobility dependence,” in *International Electron Devices Meeting*, Dec. 2006.

[68] A. Lochtefeld and D. Antoniadis, "On experimental determination of carrier velocity in deeply scaled NMOS: how close to the thermal limit?" *IEEE Electron Device Letters*, vol. 22, no. 2, pp. 95 – 97, Feb. 2001.

[69] M. Pelgrom, A. Duinmaijer, and A. Welbers, "Matching properties of MOS transistors," *IEEE Journal of Solid-State Circuits*, vol. 24, no. 5, pp. 1433 – 1439, Oct. 1989.

[70] W. Zhao, Y. Cao, F. Liu, K. Agarwal, D. Acharyya, S. Nassif, and K. Nowka, "Rigorous extraction of process variations for 65nm CMOS design," in *European Solid State Device Research Conference (ESSDERC)*, Sept. 2007, pp. 89 – 92.

[71] A. Agarwal, D. Blaauw, and V. Zolotov, "Statistical timing analysis for intra-die process variations with spatial correlations," in *International Conference on Computer Aided Design (ICCAD)*, Nov. 2003, pp. 900 – 907.

[72] *International Technology Roadmap for Semiconductors.* Semiconductor Industry Association, 2011.

[73] Y. Cao and L. Clark, "Mapping statistical process variations toward circuit performance variability: An analytical modeling approach," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 26, no. 10, pp. 1866 – 1873, Oct. 2007.

[74] S. Saxena, C. Hess, H. Karbasi, A. Rossoni, S. Tonello, P. McNamara, S. Lucherini, S. Minehane, C. Dolainsky, and M. Quarantelli, "Variation in transistor performance and leakage in nanometer-scale technologies," *IEEE Transactions on Electron Devices*, vol. 55, no. 1, pp. 131–144, 2008.

[75] N. Drego, A. Chandrakasan, and D. Boning, "Lack of spatial correlation in MOSFET threshold voltage variation and implications for voltage scaling," *IEEE Transactions on Semiconductor Manufacturing*, vol. 22, no. 2, pp. 245–255, May 2009.

[76] D. Boning, J. Panganiban, K. Gonzalez-Valentin, S. Nassif, C. McDowell, A. Gattiker, and F. Liu, "Test structures for delay variability," in *Timing Issues in the Specification and Synthesis of Digital Systems*, 2002, p. 109.

[77] C. Gu, E. Chiprout, and X. Li, "Efficient moment estimation with extremely small sample size via Bayesian inference for analog/mixed-signal validation," in *Design Automation Conference (DAC)*, 2013, pp. 65:1–65:7.

[78] A. Ortiz-Conde, F. G. Snchez, J. Liou, A. Cerdeira, M. Estrada, and Y. Yue, "A review of recent MOSFET threshold voltage extraction methods," *Microelectronics Reliability*, vol. 42, no. 45, pp. 583 – 596, 2002.

[79] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm." *J. Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[80] S. Yao, T. Morshed, D. Lu, S. Venugopalan, W. Xiong, C. Cleavelin, A. Niknejad, and C. Hu, "Global parameter extraction for a multi-gate MOSFETs compact model," in *IEEE International Conference on Microelectronic Test Structures (ICMTS)*, 2010, pp. 194–197.

[81] Q. Zhou, W. Yao, W. Wu, X. Li, Z. Zhu, and G. Gildenblat, "Parameter extraction for the PSP MOSFET model by the combination of genetic and Levenberg-Marquardt algorithms," in *IEEE International Conference on Microelectronic Test Structures (ICMTS)*, 2009, pp. 137–142.

[82] C. Mcandrew, X. Li, I. Stevanovic, and G. Gildenblat, "Extensions to backward propagation of variance for statistical modeling," *IEEE Design and Test of Computers*, vol. 27, no. 2, pp. 36–43, Mar. 2010.

[83] S. Reda and S. Nassif, "Analyzing the impact of process variations on parametric measurements: Novel models and applications," in *Design, Automation Test in Europe (DATE)*, April 2009, pp. 375–380.

[84] D. MacKay, "Information-based objective functions for active data selection," *Neural Computation*, vol. 4, no. 4, pp. 590–604, July 1992.

[85] 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.

[86] M. H. DeGroot, "Uncertainty, information, and sequential experiments," *The Annals of Mathematical Statistics*, pp. 404–419, 1962.

[87] A. Narayan and D. Xiu, "Stochastic collocation with least orthogonal interpolant Leja sequences," in *SIAM Conference on Computational Science and Engineering*, March 2013.

[88] Z. Zhang, T. El-Moselhy, I. Elfadel, and L. Daniel, "Stochastic testing method for transistor-level uncertainty quantification based on generalized polynomial chaos," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 32, no. 10, pp. 1533–1545, Oct 2013.

[89] L. Lavagno, L. Scheffer, and G. Martin, *EDA for IC Implementation, Circuit Design, and Process Technology.* Addision-Wesley, 2006.

[90] L. Brusamarello, P. Wirth, G.and Roussel, and M. Miranda, "Fast and accurate statistical characterization of standard cell libraries," *Microelectronics Reliability*, vol. 51, no. 12, pp. 2341 – 2350, 2011.

[91] Z. Zhang, I. Elfadel, and L. Daniel, "Uncertainty quantification for integrated circuits: Stochastic spectral methods," in *International Conference on Computer-Aided Design (ICCAD)*, Nov 2013, pp. 803–810.

[92] Z. Zhang, T. El-Moselhy, I. Elfadel, and L. Daniel, "Calculation of generalized polynomial-chaos basis functions and gauss quadrature rules in hierarchical uncertainty quantification," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 33, no. 5, pp. 728–740, May 2014.

[93] Z. Zhang, X. Yang, I. V. Oseledets, G. E. Karniadakis, and L. Daniel, "Enabling high-dimensional hierarchical uncertainty quantification by ANOVA and tensor-train decomposition," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, to appear 2015.

[94] Z. Zhang, X. Yang, G. Marucci, P. Maffezzoni, I. Elfadel, G. Karniadakis, and L. Daniel, "Stochastic testing simulator for integrated circuits and MEMS: Hierarchical and sparse techniques," in *Custom Integrated Circuits Conference (CICC)*, Sep. 2014, pp. 1–8.

[95] Z. Zhang, T. El-Moselhy, P. Maffezzoni, I. Elfadel, and L. Daniel, "Efficient uncertainty quantification for the periodic steady state of forced and autonomous circuits," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 60, no. 10, pp. 687–691, Oct 2013.

[96] 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.

[97] S. Sun, X. Li, H. Liu, K. Luo, and B. Gu, "Fast statistical analysis of rare circuit failure events via scaled-sigma sampling for high-dimensional variation space," in *International Conference on Computer-Aided Design (ICCAD)*, Nov. 2013, pp. 478–485.

[98] L. Yu, S. Saxena, C. Hess, I. Elfadel, D. Antoniadis, and D. Boning, "Statistical library characterization using belief propagation across multiple technology nodes," in *Design, Automation Test in Europe Conference Exhibition (DATE)*, 2015, p. To appear.

[99] V. Stojanovic, D. Markovic, B. Nikolic, M. Horowitz, and R. Brodersen, "Energy-delay tradeoffs in combinational logic using gate sizing and supply voltage optimization," in *European Solid-State Circuits Conference (ESSCIRC)*, Sept. 2002, pp. 211–214.

[100] A. Khakifirooz and D. Antoniadis, "MOSFET performance scaling - part I: Historical trends," *IEEE Transactions on Electron Devices*, vol. 55, no. 6, pp. 1391–1400, June 2008.

[101] M.-H. Na, E. Nowak, W. Haensch, and J. Cai, "The effective drive current in CMOS inverters," in *International Electron Devices Meeting (IEDM)*, Dec. 2002, pp. 121–124.

[102] J. Deng and H. Wong, "Metrics for performance benchmarking of nanoscale Si and carbon nanotube FETs including device nonidealities," *IEEE Transactions on Electron Devices*, vol. 53, no. 6, pp. 1317–1322, June 2006.

[103] E. Yoshida, Y. Momiyama, M. Miyamoto, T. Saiki, M. Kojima, S. Satoh, and T. Sugii, "Performance boost using a new device design methodology based on characteristic current for low-power CMOS," in *International Electron Devices Meeting (IEDM)*, Dec. 2006, pp. 1–4.

[104] N. Weste and K. Eshraghian, *Principles of CMOS VLSI Design: A Systems Perspective.* Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1985.

[105] X. Lin, Y. Wang, and M. Pedram, "Joint sizing and adaptive independent gate control for FinFET circuits operating in multiple voltage regimes using the logical effort method," in *International Conference on Computer-Aided Design (ICCAD)*, Nov. 2013, pp. 444–449.

[106] X. Lin, Y. Wang, S. Nazarian, and M. Pedram, "An improved logical effort model and framework applied to optimal sizing of circuits operating in multiple supply voltage regimes," in *International Symposium on Quality Electronic Design (ISQED)*, Mar. 2014, pp. 249–256.

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

[108] K. Patel, T.-J. K. Liu, and C. J. Spanos, "Gate line edge roughness model for estimation of FinFET performance variability," *IEEE Transactions on Electron Devices*, vol. 56, no. 12, pp. 3055–3063, Dec. 2009.