Statistical Yield Analysis and Design for Nanometer VLSI

by

Javid Jaffari

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2010

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Process variability is the pivotal factor impacting the design of high yield integrated circuits and systems in deep sub-micron CMOS technologies. The electrical and physical properties of transistors and interconnects, the building blocks of integrated circuits, are prone to significant variations that directly impact the performance and power consumption of the fabricated devices, severely impacting the manufacturing yield. However, the large number of the transistors on a single chip adds even more challenges for the analysis of the variation effects, a critical task in diagnosing the cause of failure and designing for yield. Reliable and efficient statistical analysis methodologies in various design phases are key to predict the yield before entering such an expensive fabrication process.

In this thesis, the impacts of process variations are examined at three different levels: device, circuit, and micro-architecture. The variation models are provided for each level of abstraction, and new methodologies are proposed for efficient statistical analysis and design under variation.

At the circuit level, the variability analysis of three crucial sub-blocks of today's system-on-chips, namely, digital circuits, memory cells, and analog blocks, are targeted. The accurate and efficient yield analysis of circuits is recognized as an extremely challenging task within the electronic design automation community. The large scale of the digital circuits, the extremely high yield requirement for memory cells, and the time-consuming analog circuit simulation are major concerns in the development of any statistical analysis technique. In this thesis, several sampling-based methods have been proposed for these three types of circuits to significantly improve the run-time of the traditional Monte Carlo method, without compromising accuracy. The proposed sampling-based yield analysis methods benefit from the very appealing feature of the MC method, that is, the capability to consider any complex circuit model. However, through the use and engineering of advanced variance reduction and sampling methods, ultra-fast yield estimation solutions are provided for different types of VLSI circuits. Such methods include control variate, importance sampling, correlation-controlled Latin Hypercube Sampling, and Quasi Monte Carlo.

At the device level, a methodology is proposed which introduces a variation-aware design perspective for designing MOS devices in aggressively scaled geometries. The method introduces a yield measure at the device level which targets the saturation and leakage currents of an MOS transistor. A statistical method is developed to optimize the advanced doping profiles and geometry features of a device for achieving a maximum device-level yield.

Finally, a statistical thermal analysis framework is proposed. It accounts for the process and thermal variations simultaneously, at the micro-architectural level. The analyzer is developed, based on the fact that the process variations lead to uncertain leakage power sources, so that the thermal profile, itself, would have a probabilistic nature. Therefore, by a co-process-thermal-leakage analysis, a more reliable full-chip statistical leakage power yield is calculated.

Acknowledgements

I would like to express my sincere gratitude to my supervisor Prof. Mohab Anis for his guidance and support during the course of my studies, particularly during the most challenging times. Also, I would like to thank Prof. Dennis Sylvester, Prof. Manoj Sachdev, Prof. Mark Aagaard, Prof. John T.W. Yeow, and Prof. Karim Karim for reviewing this work as well as their valuable comments to make the thesis come out in this current form.

I have been fortunate to work with many wonderful people in the VLSI research group, including: Yasser Azizi, Reza Chaji, Ehsanollah Fathi, Hassan Hassan, Akhilesh Kumar, Minoo Mirsaeedi, Ahmed Nour, and many others. I wish to thank them all for providing a friendly and joyful environment and for their valuable comments and discussions. I would also like to thank Prof. Kumaraswamy Ponnambalam, from Department of Systems Design Engineering for his valuable comments and sharing of his source codes for optimization under uncertainty, Prof. Christiane Lemieux from Department of Statistics and Actuarial Science for her comments on variance reduction methods, Dr. Mohamed Abu-Rahma of Qualcomm Inc for the useful discussions on SRAM cell design for yield, and Dr. Trent McConaghy of Solido Design Automation for comments on analog circuit yield estimation. I would like to thank Dr. Nizar Abdallah and Dr. Julien Dunoyer for their valuable support during my internship at Actel Corporation in 2008.

I am also extremely grateful to my M.Sc. supervisor during my studies at the University of Tehran, Prof. Ali Afzali-Kusha, for his endless kindness and support. Additionally, I want to thank the University of Waterloo administration for providing me with the opportunity to pursue my Ph.D. studies. Particularly, I would like to express my thanks to Wendy Boles, Lisa Hendels, and Annette Dietrich, of ECE Graduate Office. I would like to acknowledge the support of Ontario Graduate Scholarship (OGS) while I was an international student.

I am always grateful to my wife, Neda Nouri, for her invaluable support and inspiration. Her love and understanding played a major role in helping me finish this thesis. I also, would like to thank my parents for their endless love and support.

Dedication

This thesis is dedicated to my wife, Neda Nouri.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Variations: Sources and Impact on Yield

The scaling of the CMOS technology has introduced enormous challenges that must be resolved by the designers. As the silicon industry moves toward nanometer designs, one of the most important design challenges cited is the increasing variability in the device characteristics [6] which threatens the silicon technology, and why CMOS scaling is facing critical yield concerns.

As can be seen in Figure 1.1, the exposure wavelength used for the lithography process to print layouts of different layers has not been scaled as fast as technology minimum feature size [1]. As a result, the printed features will not be exactly the same as the desired shapes.

These lithography-driven variations bring lateral layout variations for gate length (L_g), gate width (W_g), and metal interconnect width (W_M) which affects delay and subthreshold leakage of CMOS transistors and the characteristics of interconnects.

Vertical variations due to Chemical Mechanical Polishing (CMP) is another source of process variations (Figure 1.2) which originates from the difference in the removal rates of materials [7, 8]. The density of a lower layout affects the Inter Layer Dielectrics heights (H_{ILD}), oxide thickness (T_{ox}), and metal thickness (T_M) during CMP processes that consequently impacts the characteristics of interconnects and CMOS gates.

Another source of variation that directly influences the threshold voltage (V_{th}) of MOS transistors comes from ion implantation, chemical vapor deposition (CVD), and thermal annealing processes. This type of variation, called Random Dopant Fluctuations (RDF), causes variability in the number and position of dopant atoms in the channel of MOS devices [9, 10]. Figure 1.3 shows a side and top view of an MOSFET's channel to depict the randomness of atoms in a channel. The shorter the channel, the less dopant atoms are in a channel making the transistor more sensitive to RDF in scaled technologies.

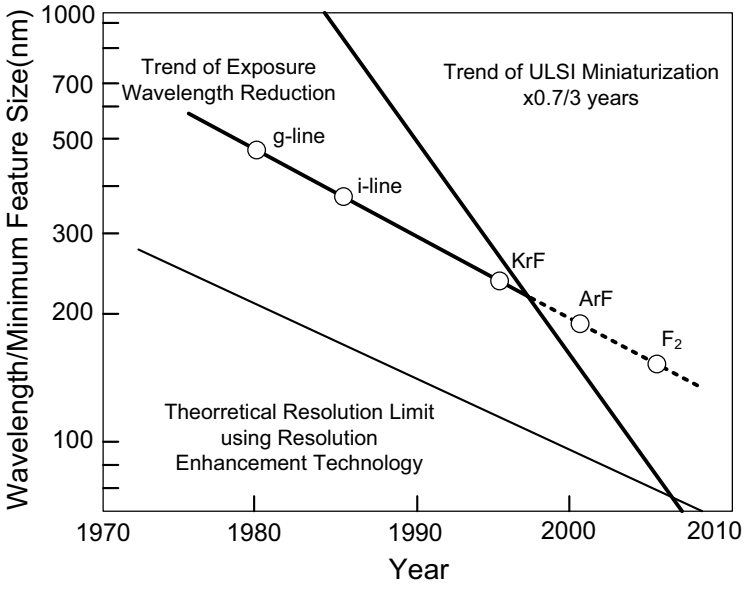


Figure 1.1: Wavelength scaling versus feature size [1].

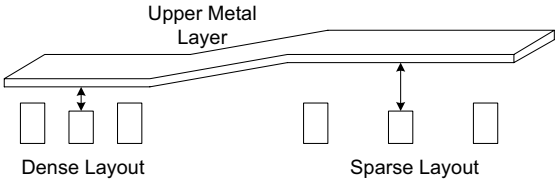
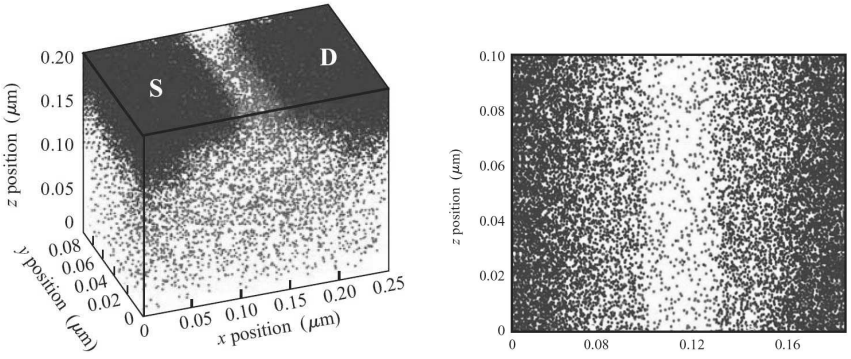


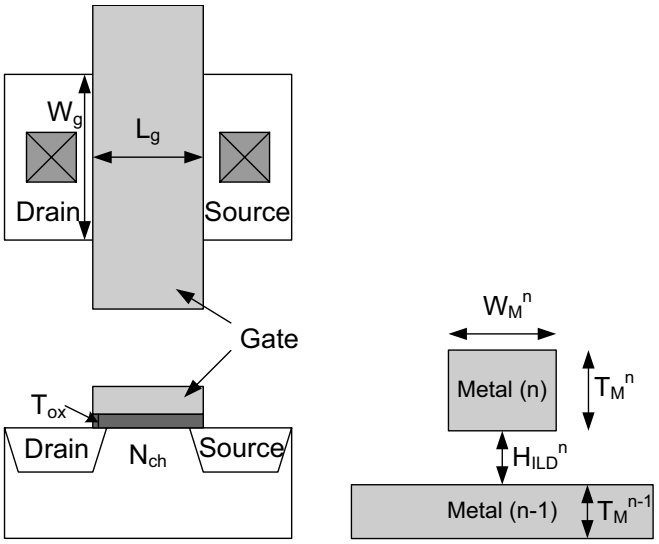
Figure 1.2: The effects of layout density on Post-CMP ILD thickness [2].



(a) Side view (b) Top view

Figure 1.3: Atom fluctuations in a MOSFET's channel[3].

In closing, the physical parameters, experiencing variability and impacting the circuit delay and total leakage current, are depicted in Figure 1.4.



(a) Transistor Parameters (b) Interconnect Parameters

Figure 1.4: Physical parameters experiencing variation

It should be noted that as nominal physical dimensions are rapidly shrunk, more variations are seen in the various physical parameters [11]. The feature size of layouts reduces faster than the lithography wavelength, hence, more aggressive subwavelength effects are seen. Also, due to the increase in the contribution of interconnect to total delay in ultra-DSM CMOS technologies, the CMP-based variations become more critical in each new technology [12]. Finally, the RDF-driven threshold voltage variation increases in each technology as the number of dopant atoms in the shortened channel is rather reduced [3].

The variations on physical parameters cause performance and leakage alteration on a whole chip. Figure 1.5 shows a measured leakage variation as high as 20X for a 30% variation in chip frequency. Consequently, yield of a circuit (probability to meet the desired performance or power specification) is expected to suffer, unless careful statistical design followed by reliable statistical timing and power analysis are performed. In fact, if a circuit does not pass a maximum desired power budget or a minimum clock frequency, it may not be functional and hence reduce the production yield [13, 14]. However, even if a chip could be used in a lower operating frequency (mostly the general purpose processors using frequency binning [15]), the profitability will be reduced as the slower ICs are sold for cheeper.

Besides the mentioned time-invariant variations, there is another type of physical variation which impacts the threshold voltage of devices in time. Negative Bias Temperature Instability

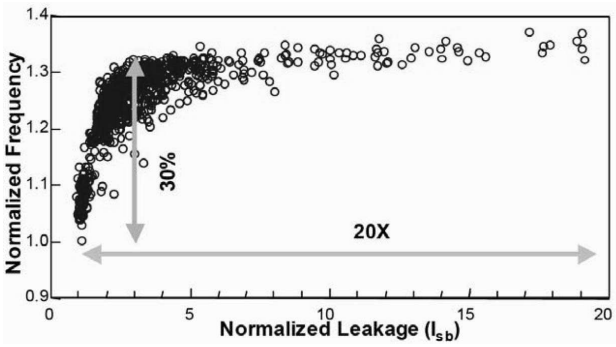


Figure 1.5: Frequency and standby leakage current distribution (0.18μm technology) [4].

(NBTI) increases the absolute value of the MOS transistor's threshold voltage over a period of months or years, depending on the operating conditions of the device. This phenomenon which degrades the performance of circuits gradually over time is worsen with technology scaling [16] and has brought serious reliability concerns in nanoscale technologies as well.

Another type of yield lost is due to process defects causing a short or open in the circuit wires. This type of problems can lead to functional failures and may be addressed by redundancy in design.

In this thesis, only the performance-driven time-invariant variations are considered. Therefore, the usage of the term *yield* is referred to the ratio of the devices/circuits that pass certain performance metric constraints in presence of process variation.

1.2 Motivations

The design of Very Large Scale Integrated (VLSI) systems can be divided into many levels of abstraction. The design process, at each level, requires comprehensive and accurate models of the physical phenomena and the appropriate tools to simulate them in an efficient manner. The VLSI system specifications, such as power consumption, speed (performance), life-time, and thermal behavior must be considered during the various design phases, at each level of abstraction. This is essential to diagnose the sources of mal-functioning as early as possible and to reduce the chance of project failure. To achieve these goals, high performance and high capacity design automation and analysis tools are required.

On top of these traditional design challenges, it is the process variations, introduced during the IC manufacturing process which adds even more challenges to the design process at each level. For example, circuits undergoing variability now may exhibit very high leakage power consumption, pushing them over the power budget. The byproduct of this power consumption

shift is the generation of heat leading to a higher operating temperature that, itself, raises reliability, packaging, and increased leakage power concerns. All are factors that can finally impact the yield. In addition, the performance of a digital circuit is affected by the transistor and interconnect variations, limiting the operating frequency of the circuit, and the yield of a system relying on a high-throughput digital processing. The process variations can also be very harmful for analog blocks and memory cores, directly impacting the yield and the success of a product.

Such an insight has been the motivation to target the process variation at three levels of device, circuit, and micro-architecture. In this thesis, the process variations are first modeled at each level. While the device level parameter variation models are designed to consider the details of channel doping profiles and device geometries, the circuit-level models lump them together as basic transistor-level electrical parameters such as threshold voltage variation. Consequently, the micro-architectural-level models unify the transistor-level parameters into a high-level model of circuit specifications variations, encapsulated in grids of equally varying process parameters. Finally, at each level, a number of computer-aided design and analysis solutions are proposed, each designed to address a gap in design for yield or analysis of variability of nanometer VLSI systems and circuits due to process variations.

1.3 Contributions

At the device-level, the variations of the drive-in capability and leakage currents are considered in order to optimize the geometry and doping profiles of an MOS device. The proposed device optimization methodology incorporates variability-awareness into the device design process by maximally satisfying certain bounds on total leakage, saturation current, and the intrinsic delay of the device undergoing process variations. This approach introduces a new strategy for the design of devices, where traditionally, nominal drive-in and leakage currents have been the objectives of the process design.

However, at circuit-level, the focus of this thesis is on accurate and efficient estimation of the yield for the different types of VLSI circuits: digital, analog, and SRAM cells. The statistical estimation of the circuit yield has been one of the major research areas in electronic design automation in recent years. Despite the considerable progress in this domain, the Monte-Carlo method is still the most reliable method as it can account for any circuit models and their secondary effects. However, the MC method suffers greatly from the lack of efficiency due to its slow convergence rate. In this thesis, a number of advanced sampling and variance reduction methods are developed to enhance the convergence rate of the traditional-MC method for circuit-level process variation-driven yield estimation. The sampling-based yield estimation solutions have been proposed for digital and analog circuits, and SRAM cells.

Finally, at the micro-architectural level, the effects of process variations are studied in a high-level co-thermal-leakage analysis framework. The process variations have traditionally

been studied as a source of full-chip total leakage power variation, leading to an estimation of system-level power consumption yield. However, the generated heat, due to higher leakage power consumption, increases the operating temperature of the chip, that itself in a positive feedback increases the sub-threshold leakage current. In an extreme case, these phenomena can lead to a thermal runaway. A statistical analysis methodology is proposed in this thesis to account for the thermal-leakage loop at micro-architectural-level in presence of process variation.

1.4 Structure of this Thesis

The rest of this thesis is organized as follows:

- **Part I: Device-Level**
  - **Chapter 2**: The MOS device-level models of the various leakage current mechanisms, the saturation current, and the intrinsic delay are presented. Then, a device design methodology is proposed to maximize the yield of MOS devices for a desirable performance and leakage constraints.
- **Part II: Circuit-Level**
  - **Chapter 3**: An overview of advanced sampling and variance reduction methods, used for efficient sampling-based circuit variability analysis, is presented. These methods are the core of the proposed techniques in the later chapters that perform circuit yield estimation with a significantly lower number of samples compared to that of the traditional-MC.
  - **Chapter 4**: Three methods for efficient MC-based timing yield estimation of digital circuits are proposed. The methods are based on Quasi-MC sampling and control variates.
  - **Chapter 5**: The process variation effects on analog circuit performance metrics are studied through the analysis of the response surface of the metrics. The required number of MC samples for sufficiently accurate yield estimation is calculated. Then, an enhanced LHS-based is proposed for the yield analysis of the analog blocks.
  - **Chapter 6**: The failure mechanisms of SRAM cells are investigated. An adaptive importance sampling-based approach is developed for the efficient yield estimation of the SRAM cells with rare failure rate.
- **Part III: Micro Architectural-Level**

  - **Chapter 7**: A high-level model of leakage power uncertainty, due to process variations, is adopted to develop a co-thermal-leakage variation analysis engine. A hotspot formation analyzer and a full-chip leakage power yield analyzer are proposed as two applications of the engine.

- **Part IV: Thesis Closure**

  - **Chapter 8**: The conclusion and future works are presented in the last chapter.

Part I

Device-Level

As CMOS technology is scaling down toward the nano-scale regime, the drastically growing leakage currents and variations in device characteristics are becoming two important design challenges. Traditionally, the device design methodology is based on finding the device parameters which minimize the leakage current while provide enough saturation current for the performance needs. This methodology may change when variations are accounted for design. In this part of the thesis, the process variations are studied in device-level, and a novel device optimization methodology is presented that incorporates variability awareness into the device design flow such that the designed devices satisfy certain bounds on the total leakage, saturation current, and intrinsic delay under parameter variabilities.

Chapter 2

Variability-Aware MOS Device Design

2.1 Introduction

The development of silicon technology has been and will continue to be driven by system needs. These needs have been satisfied by the increase in transistor density and performance, as suggested by "Moore's Law" and guided by CMOS scaling theory. However, the scaling of technology brings up enormous challenges that must be resolved by designers. As the silicon industry moves toward nanometer designs, the two most important design challenges cited are the growing leakage power dissipation [17] and the increasing variability in process dependent device characteristics [18]. Leakage power has been growing at an alarming rate, and constitutes a larger fraction of the total chip power in current and future technology generations. In addition, the manufacturing process of nanometer transistors and structures has introduced several new sources of variation that has made the control of process variation more difficult [19]. Process variations significantly impact chips' performance and power dissipation [18, 20]. The growing leakage power and variability in device characteristics are indeed the two most serious issues that threaten the life time of silicon technology [21].

The leakage power problem is further compounded by its strong dependence on the design parameters and hence on their variations [20]. As a result, circuits experiencing variability, now may exhibit very high leakage power consumption, pushing them over the power budget. In fact, variations in transistor parameters in the 180 nm CMOS technology node causes up to 20X variation in the chip's total leakage and 30% variation in its maximum operating frequency [4] and are worse when the technology scales [22].

Traditionally the device design methodology is based on maximizing the I_{ON}/I_{OFF} ratio, in which a device is designed such that its total leakage current is minimized while it provides a minimum saturation current satisfying the application's performance needs. Typically, the total leakage current consists of three major components, namely: subthreshold, gate direct tunneling,

and reversed biased junction band-to-band-tunneling [17]. However, the analytical models for mean and standard deviation of leakage current components suggest different sensitivity measures to various device parameters [23]. Hence, the variance of the total leakage current depends not only on device's parameter variations, but also on the relative magnitude of the leakage components of the device. Therefore, different devices with relatively equal nominal total leakage current may see considerably different variances on their total leakage current in the presence of variability. This reemphasizes the fact that exclusively minimizing the total leakage may yield a device with a large sensitivity to process parameters and hence less immunity against leakage current variations. Therefore, trading off among the magnitude of leakage components can produce more robust devices in terms of performance and leakage variability.

Motivated by the above challenges, the design of CMOS devices must be revisited to include variability. The objective of this work is to re-design the CMOS device to increase its yield by maximizing its immunity against process variations. To achieve this goal, a Bulk-MOS design methodology is proposed which not only deals with total leakage current reduction but also increases its tolerance to variability, while accounting for the minimum required drive-in current (I_{ON}) and maximum intrinsic delay ($\tau = C_g V / I_{ON}$) of the device.

With the aid of our proposed methodology, the designer would define a targeted technology and three bounds on I_{ON}, intrinsic delay, and total leakage current, and can now exploit the allowable design space for variability to maximize the device's yield. Physical gate length, oxide thickness, and channel doping profile (halo and super steep retrograde well) parameters are considered as the main design variables. These variables form a five-dimensional space where each point represents a device with parameters equal to the coordinates of the point. Then, based on the defined bounds, a problem feasible space is formed where every point (device) in this space satisfies the defined constraints of I_{ON} and the total leakage current. Finally, the yield maximizing step places a cube in the feasible space such that the device lies in the center of that cube has maximum immunity against process variations. It should be noted that to assure compliance of the designed device with the targeted technology, fabrication limitations (e.g. minimum gate length and oxide thickness) and variation parameters of the technology should also be given to the optimizer as technology specific constraints.

The variability has been included into technology optimization by the framework proposed in [24]. The circuit (e.g. V_{DD}, mean repeater sizing and width) and device level variables (e.g. gate length, oxide thickness, and peak halo doping) are optimized such that a design shows a maximum performance-driven yield subject to a maximum average power consumption. Therefore, the variability of the power consumption is simply modeled by the average sub-threshold leakage current based on V_{th} variation. This may lead to a design variable set which shows a satisfying power consumption expected value but high power consumption (leakage) variance. Moreover, the tunneling (gate oxide and BTBT) leakage variations are ignored. Also, the yield is only defined based on the performance which means a fabricated circuit is acceptable if it only pass a minimum performance metric regardless of its leakage current magnitude. Finally, using

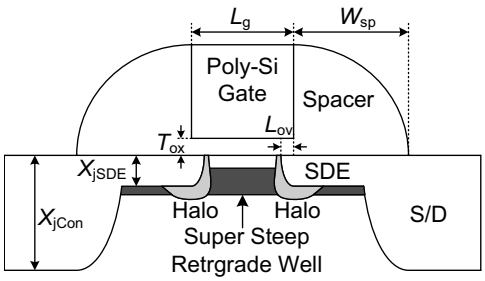


Figure 2.1: Symmetrical Bulk–MOS structure. Parameters: gate length (L_g), oxide thickness (T_{ox}), sidewall spacer width (W_{sp}), gate/SDE overlap (L_{ov}), SDE junction depth (X_{jSDE}), contact junction depth (X_{jCon}), Gaussian Halo, and Super Steep Retrograde Well

simplified device models with numerous fitting curves makes the approach useful for fast general technology variable optimization (as listed before). However, there is still a need to make use of the trade-offs between various leakage components and its effects on a leakage-performance based yield and consider them in a detailed device parameter optimization to build variation immune devices for different technologies.

The rest of this chapter is organized as follows. In Section 2.2, the selected device structure and design parameters are presented, whereas the problem is formulated in Section 2.3. The way the defined constraints on currents are verified is discussed in Section 2.4, and the implementation and results with discussions are given in Section 2.5. Finally, conclusions are presented in Section 2.6.

2.2 Selected Device Structure

As mentioned earlier, the objective of this work is to optimize a device's geometry and doping profiles in order to obtain the highest immunity against variability in the performance and leakage current of the device.

To achieve this goal, a symmetrical Bulk-NMOS device structure as shown in Figure 2.1 is selected. The device with various channel doping implants (Source/Drain Extension (SDE), Gaussian Halo, and vertical Retrograde Well) has been developed to mitigate the short channel effects and improve the leakage characteristics [25]. The parameters of this structure are discussed in two categories: *geometrical* and *doping* parameters.

2.2.1 Geometrical Parameters

The geometrical parameters are physical gate length (L_g), oxide thickness (T_{ox}), sidewall spacer width (W_{sp}), and transistor width (W).

2.2.1.1 Physical Gate Length

The threshold voltage of MOSFET devices decreases with the reduction in gate length. Using depletion approximation, the threshold voltage of a MOS device, V_{th}, can be defined as [26]:

$$V_{th} = V_{fb} + \phi_s + \frac{Q_B}{C_{ox}} \tag{2.1}$$

where V_{fb} is the flat-band voltage; ϕ_s is the surface potential; C_{ox} is the capacitance across the oxide; and Q_B is the depletion charge in the bulk. In short channel devices, the source-drain distance is comparable to the depletion width in the vertical direction under the oxide. As a result, the source and drain depletion regions now penetrate more into the channel, resulting in part of the channel being already depleted. Therefore, less bulk charge (Q_B) is needed for the device to be inverted by the applied gate voltage. The change in the threshold voltage, V_{th}, as a result of channel length scaling can be approximated as [27]:

$$\Delta V_{th} = -\left[2\left(V_{bi} - \phi_s\right) + V_{DS}\right]\left(e^{-L/2l} + 2e^{-L/l}\right) \tag{2.2}$$

where V_{bi} is the potential of the channel/source edge, V_{DS} is the drain-source voltage, L is the effective channel length, and

$$l = \sqrt{\frac{\varepsilon_{si} T_{ox}}{\varepsilon_{ox}} \times \frac{W_{dep}}{\eta}} \tag{2.3}$$

where W_{dep}/η is the average depletion layer width along the channel, and T_{ox} is oxide thickness. Considering Eq. (2.2), in a long channel device ($L >> l$) ΔV_{th} is almost zero, while in the short channel devices, the negative ΔV_{th} causes a reduction in threshold voltage. This Short Channel Effect (SCE) is known as V_{th} roll-off [17].

In addition, subthreshold leakage, I_{sub} of a MOSFET device can be modeled as [28]:.

$$I_{sub} = \mu_0 C_{ox} \frac{W}{L} v_T^2 e^{1.8} e^{(V_{GS} - V_{th})/n v_T} \left(1 - e^{-v_{DS}/v_T}\right) \tag{2.4}$$

where μ_0 is carrier mobility, $\frac{W}{L}$ is width over length ratio of the device, v_T is the thermal voltage, and n is the subthreshold swing coefficient. Considering the exponential dependency between subthreshold leakage and V_{th}, it can be inferred that the gate length as one of the contributors to the threshold voltage variation should be taken into account in a variation driven device design.

It has been shown that V_{th} rolling off can be reduced by applying halo(pocket) implants [29]. However, this improvement may lead to a V_{th} roll-up (Reversed SCE) followed by an abrupt roll-off which can be troublesome for devices beyond the 100 nm regime [30, 31]. By increasing the channel length in the halo implanted device, one can reduce the variation in threshold voltage ($dV_{th}/dL_g \rightarrow$ zero). However, this leads to a penalty in performance because of the reduction in saturation current [22].

Besides to the discussed trade-off role of the gate length between providing enough saturation current and threshold voltage stability, the physical gate length is the main parameter in the hand of device designers to design various devices for different purposes from Low Power (LP) to High Performance (HP) applications [11].

2.2.1.2 Oxide Thickness

The oxide thickness has a considerable effect on threshold voltage [26] since any variation in oxide thickness changes $C_{ox} = \varepsilon_{ox}/T_{ox}$. Hence, it will affect threshold voltage and subthreshold leakage current (as per Eq.2.1 and 2.4). Moreover, the SCE is affected by oxide thickness as given in Eg.2.3, therefore, thinner oxide is needed to overcome V_{th} roll-off in scaled technologies. However, the gate-tunneling leakage cannot be neglected when the oxide thickness is less than 3nm [17]. The gate leakage is due to the tunneling of an electron (or hole) from the bulk silicon through the gate-oxide potential barrier into the gate. Direct tunneling gate leakage density, J_{DT}, is modeled as [32]:

$$J_{DT} = A\left(\frac{V_{ox}}{T_{ox}}\right)^2 \exp\left\{-B\frac{T_{ox}}{V_{ox}}\left[1-\left(1-\frac{V_{ox}}{\phi_{ox}}\right)^{\frac{3}{2}}\right]\right\} \tag{2.5}$$

where V_{ox} is the drop across the thin oxide and ϕ_{ox} is the barrier height for the tunneling particle (electron or hole). A and B are physical parameters depended on barrier height and are given in [32]. It can be seen from Eq. (2.5) that the tunneling current increases exponentially with a decrease in oxide thickness.

In addition, the saturation current and intrinsic delay are also sensitive to variation in T_{ox} due to variations in threshold voltage and gate oxide capacitance.

2.2.1.3 Other Parameters

The transistor width is chosen by the circuit designers to size transistors in order to meet the required specifications for the system. Therefore, it is not considered as a device level design variable in our optimization problem. In addition, sidewall spacers are used to form SDE regions in the two sides of the channel and their width is determined based on the physical gate length

[33]. Hence, their values are determined for every transistor based on its gate length ($W_{sp} = 1.1 \times L_g$) [11], so it is not included in the proposed device design parameter list.

2.2.2 Doping Parameters

Various channel profiles have been developed to overcome short channel effects and improve leakage characteristics [25]. Today's MOS transistors have three profiles in their channel: Source/Drain Extension (SDE), Halo, and Super Steep Retrograde Well (SSRW).

2.2.2.1 Source/Drain Extensions

SDE regions which are traditionally known as Lightly Doped Drain (LDD) are critical for deep sub-micron devices since they suppress the buildup of wide electric fields in the drain and source regions, hence reducing Drain Induced Barrier Lowering (DIBL) and V_{th} roll-off known as short channel effects [34]. The two important aspects associating with SDE region profiles are junction depth and lateral abruptness.

SDE junction depth (X_{jSDE}) plays an important role in deep sub-micron devices. Deeper junctions result in more severe short channel effects due to further spreading potential contours and hence the depletion region into the channel. However, shallower junctions can impose higher series resistance to the transistor's source/drain terminal [35]. This trade-off has pushed designers to find the optimum SDE junction depth which not only reduces the series resistance and hence boosts the drive-in current but also improves short channel effects [35, 36]. Now, it is well understood that in the sub 100 nm regimes the extension junction depth should be scaled more aggressive than the past [11]. Motivated by the needs which are suggested in ITRS (International Technology Roadmap for Semiconductors), the ultra-shallow junctions is now achievable by the new innovations in fabrication techniques [34, 37, 38, 39]. In this work, the existing guidelines reported in ITRS are used for the depth of SDE regions [11].

Another important aspect of the SDE profile is its lateral abruptness. Detailed studies of SDE profiles showed that extension resistance which is an obstacle to achieve high-performance devices is strongly linked to lateral abruptness of the SDE. While more abrupt profile yields less resistivity to the extension, DIBL and threshold roll-off is impacted by too abrupt or too gradual junctions [40]. Based on the above facts, another guideline for optimum lateral abruptness has been reported in ITRS which is used in this work (lateral abruptness in $nm/$decade drop-off in doping concentration $= 0.11 \times L_g$) [11]. It should be noted that, the length of the gate drain overlap (L_{ov}) is correlated with SDE lateral abruptness [35, 41] and is implicitly determined by the lateral abruptness of the SDE.

2.2.2.2 S/D Contacts

Due to existence of extensions, S/D contacts are placed far from the channel. As a result, the short channel effects are independent of the contact junction depth (X_{jCon}), and only the saturation current increases with the increase in X_{jCon} [42]. Therefore, the $X_{jCon} = 1.1 \times L_g$ is determined based on physical gate length as given in ITRS [11].

2.2.2.3 Halo and Super Steep Retrograde Well

In short channel devices, additional non-uniform implants in the lateral and vertical directions are used to improve short channel effects [43, 44]. Halo, a non-uniform lateral doping, has been introduced to improve short channel effects and reduce subthreshold leakage current [45]. Tilt implanting of halo impurities places the pocket regions adjacent to SDE edge which made the profile more useful to suppress punch-through and short channel effects [46]. By proper usage of the profile, a 25 nm CMOS transistor design is feasible without continued scaling of the supply voltage. Therefore, a considerable improvement in device performance is achievable [46, 47].

In addition, to keep acceptable subthreshold leakage current in scaled devices, the channel doping should be increased as the gate length is decreased. However, increasing the channel doping leads to increase in threshold voltage, and consequently degrades device performance. A nonuniform vertical channel doping known as retrograde well can overcome the problem by providing a low surface concentration [17]. Due to suppressing channel impurity scattering, the lower concentration keeps surface channel mobility high while reduces subthreshold current. In fact, Super Steep Retrograde Well (SSRW) is preferred due to the increase in the linear drive current which causes performance improvement for logic gates [48, 49].

The symmetrical 2-D non-uniform channel doping, $N_{CH}(x,y)$, composed by halo and retrograde which is typically assumed to be Gaussian [50] is given as:

$$
\begin{aligned}
&N_{CH}(x,y) = N_{Halo}(x,y) + N_{RW}(y) + N_{Sub} \\
&\text{where} \\
&N_{Halo}(x,y) = P_H \left[\exp\left(\frac{-(x-\alpha_1)^2}{S_{halox}^2} \right) + \exp\left(\frac{-(x-\alpha_2)^2}{S_{halo_x}^2} \right) \right] \exp\left(\frac{-(y-\beta)^2}{S_{halo_y}^2} \right) \\
&\text{and} \\
&N_{RW}(y) = P_{RW} \exp\left(\frac{-(y-Y_{RW})^2}{S_{RW}^2} \right)
\end{aligned}
\tag{2.6}
$$

where P_H and P_{RW} represent the peak halo and retrograde well concentrations, and N_{Sub} is the constant uniform doping of the bulk. S_{halo_y} and S_{halo_x} denote the characteristic decay lengths of the Gaussian halo profile in the vertical and lateral directions, and S_{RW} is the decay length on the vertical retrograde well. Finally, the positions of the halo and retrograde peaks are defined by α_1, α_2, β, and Y_{RW}. α_1 and α_2 are lateral positions of the pocket implant peaks while β and Y_{RW} are

the vertical position of the halo and retrograde peaks, respectively. In this work, the halo(pocket) peaks are placed beside the SDE edge where the extension and background concentrations are equated.

Band To Band Tunneling (BTBT) leakage is strongly linked to the channel and junction profiles [43], and hence, is very sensitive to any channel doping variation [23]. The BTBT current, I_{BTBT} can be estimated as [43]:

$$\begin{aligned} &I_{BTBT} = \left(WX_{jSDE}\hat{A}\Big/E_g^{1/2}\right)\xi V_{DD}\exp\left(-\hat{B}E_g^{3/2}\Big/\xi\right) \\ &\text{where} \\ &\xi = \sqrt{\frac{2qN_{aside}N_{sdside}}{\varepsilon_{si}(N_{aside}+N_{sdside})}\left[V_{DD}+\frac{KT}{q}\ln\left(\frac{N_{aside}N_{sdside}}{n_i^2}\right)\right]} \end{aligned} \tag{2.7}$$

where N_{aside} and N_{sdside} are the p-side and n-side junction doping. E_g is the band-gap of the silicon, and $\hat{A}$ and $\hat{B}$ are physical coefficients given in [10]. Variation on channel peek dopings (halos and retrograde well) and vertical position of the retrograde well affect N_{aside} and hence BTBT leakage [43].

Furthermore, the variation of the peak values and the position of the retrograde well strongly affect threshold voltage and hence subthreshold leakage current, due to the impact on the threshold roll-off and Random Dopant Fluctuation (RDF)-driven threshold voltage variations [10]. In fact, in scaled technologies, RDF is becoming a dominant source of threshold voltage variations as the average number of dopant atoms in the channel is rather reduced. Finally, any change in the threshold voltage impacts the drive–in current and intrinsic delay as well. Moreover, there are no predefined exact values for halo and retrograde peaks and position in ITRS.

Consequently, the following device parameters will be used in forming the device optimization problem.

- L_g: Physical gate length
- T_{ox}: Oxide thickness
- P_H: Halo peak doping concentration
- P_{RW}: Retrograde well peak doping concentration
- Y_{RW}: Vertical position of the retrograde well peak

As shown earlier, each leakage component is a function of the number of five process parameters under consideration. I_{gate}, I_{BTBT}, and I_{sub} are exponentially depends on T_{ox}, N_{aside}, and V_{th}, respectively [17], while V_{th} is a function of all selected process parameters [23]. Therefore, ON drive-in current as well as intrinsic delay is also a function of listed parameters. Hence, the

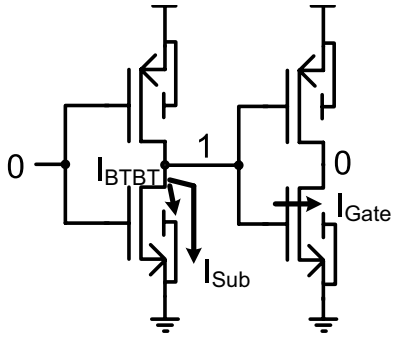


Figure 2.2: Total Leakage (TL) estimation scheme

following representations could be used to show the device characteristics and their dependency to each selected design parameter.

$$\begin{aligned} \tau &= f(L_g, T_{ox}, P_H, P_{RW}, Y_{RW}) \\ I_{ON} &= f(L_g, T_{ox}, P_H, P_{RW}, Y_{RW}) \\ I_{sub} &= f(L_g, T_{ox}, P_H, P_{RW}, Y_{RW}) \\ I_{BTBT} &= f(P_H, P_{RW}, Y_{RW}) \\ I_{gate} &= f(L_g, T_{ox}) \end{aligned} \tag{2.8}$$

2.3 Problem Formulation

2.3.1 General Approach

Considering a five-dimensional space composed by L_g, T_{ox}, P_H, P_{RW}, and Y_{RW}, a yield optimization problem can be represented as follows:

$$\underset{x=\left(L_g, T_{ox}, P_H, P_{RW}, Y_{RW}\right)}{\text{argmax}} \qquad Yield = P_x\{C(x)=1\} \tag{2.9}$$

where $C(x)$ denotes a boolean random variable function defined based on desired bounds on the ON current (I_{ON}), intrinsic delay (τ), and total leakage (TL) and is formulated by Eq. (2.10).

$$\begin{aligned} C(x) = (I_{ON}(x) \geq I_{ON-Min}) AND (\tau(x) \leq \tau_{Max}) \\ AND (TL(x) \leq TL_{Max}) \end{aligned} \tag{2.10}$$

where I_{ON-Min}, τ_{Max}, and TL_{Max} are desirable bounds for device parameters of interest. Therefore, $P_x\{C(x)=1\}$ represents the probability that a device (x) satisfies the currents and

delay constraints in the presence of variations in x elements. This type of problem formulation enables developing different devices for high-performance or low-power applications by assigning various values to I_{ON-Min}, τ_{Max}, and TL_{Max}. The selection criteria for two performance metrics (ON current and intrinsic delay) is based on the fact that the performance improvement is primarily achieved by reduction of gate capacitance and hence reduction of intrinsic delay in every technology node for sub 100nm regime [11], while the I_{ON} is almost constant in scaled technologies and should only meet a minimum to prevent negative impact on the device drivability, critical for driving parasitic/interconnect capacitances.

To have a more realistic indication of the total leakage, TL, in digital circuits, all of the worst case leakage components are added together as given in ITRS [11].

$$\begin{aligned} &TL = \\ &I_{sub}\left(V_{GS}=0, V_{DS}=V_{DD}\right) + I_{BTBT}\left(V_{GS}=0, V_{DS}=V_{DD}\right) + I_{gate}\left(V_{GS}=V_{DD}, V_{DS}=0\right) \end{aligned} \tag{2.11}$$

Figure 2.2 shows a typical scheme where all three leakage components contribute in total leakage power.

2.3.2 Yield Estimation

To solve the optimization problem stated in Eq. (2.9), one should estimate the probability of placing a device in the feasible space defined by the design constraints in the presence of variation in device parameters. This means that the probability which a device with parameters x satisfies the desired constraints on intrinsic delay, leakage, and drive-in current should be estimated. To

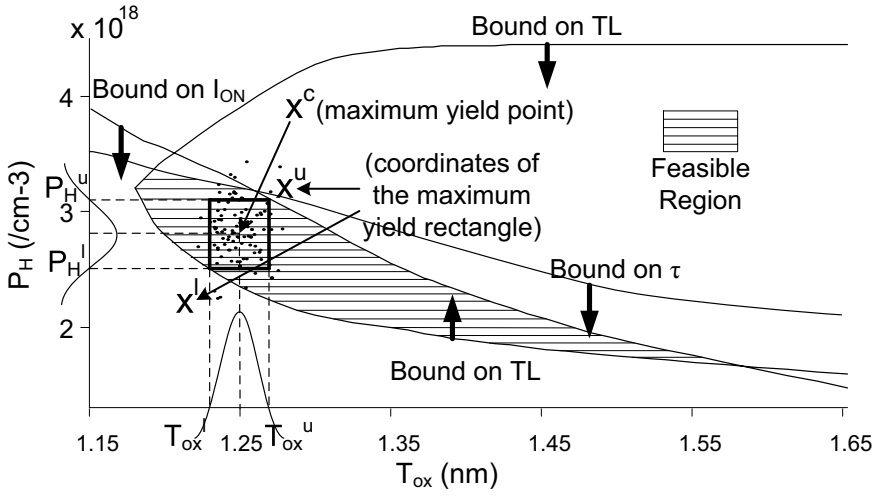


Figure 2.3: Simplified problem in two dimensions

estimate such probability, $P_x(C=1)$, a 5-D cube is formed in the problem space where all points within the cube satisfy the constraints on the I_{ON} and TL bounds.

To clarify this point, a problem with two design variables (T_{ox},P_H) is shown in Figure 2.3. A feasible region is defined based on the problem constraints. A rectangle is figured where its area is in the feasible region (all devices lying in the rectangle have the I_{ON}, τ, and TL within the desired bounds). The center of the rectangle is the max yield point. Now, considering a device placed in the center, the probability of the constraints satisfaction for such device in the presence of independent parameter variations can be estimated as follows:

$$P_{x(2-D)} = P_x\{C(x)=1\} = P\{T_{ox}^l \le T_{ox} \le T_{ox}^u\} \times P\{P_H^l \le P_H \le P_H^u\} \tag{2.12}$$

where T_{ox}^l, T_{ox}^u, P_H^l, and P_H^u are coordinates of the rectangle.

By expanding this 2-D problem to the original 5-D problem given in Eq. (2.9), the 5-D yield probability can be represented as:

$$\text{Assuming}: \begin{cases} x = (L_g, T_{ox}, P_H, P_{RW}, Y_{RW}) \\ x^l = \left(L_g^l, T_{ox}^l, P_H^l, P_{RW}^l, Y_{RW}^l\right) \\ x^u = \left(L_g^u, T_{ox}^u, P_H^u, P_{RW}^u, Y_{RW}^u\right) \end{cases} \tag{2.13}$$

$$Yield\left(x^l, x^u\right) = P_x\{C=1\} = \prod_{i=1}^{5} P\left\{x_i^l \le x_i \le x_i^u\right\} = \prod_{i=1}^{5} \left(CDF_{X_i}\left(x_i^u\right) - CDF_{X_i}\left(x_i^l\right)\right)$$

where x_i is the i^{th} design parameter of device x. x^u and x^l represent the coordinates of the inscribed 5-D cube (instead of rectangle of 2-D problem). Thus, CDF_{X_i} is the cumulative distribution function of the parameter x_i. In this work, the variability of each design parameter is considered to be independent and the distribution is assumed to be Gaussian [20]. But, Gaussian distribution does not have a closed form cumulative distribution function (CDF) which is needed for yield evaluation, so the Kumaraswamy's distribution model is utilized [51, 52]. This double bounded probability density function (DB-PDF), is appropriate for physically bounded variables and provide a simple closed form expression for any probability distribution function [52]. The probability distribution function (PDF) $f(z)$ of this model is in the form of:

$$\begin{aligned} &f(z) = abz^{a-1}(1-z^a)^{b-1} \\ &z = \tfrac{x-x^{lb}}{x^{ub}-x^{lb}} \quad , \qquad x^{lb} \le x \le x^{ub} \end{aligned} \tag{2.14}$$

where x^{ub} and x^{lb} represent upper and lower bounds of double-bounded random variable x. Depending on the values chosen for parameters a and b, DB-PDF can take various shapes. In this work, a truncated Gaussian shape with range $x^{ub} - x^{lb} = 6\sigma_x$ has been used by setting a and b to 3.6 and 8. Therefore x^{ub} and x^{lb} are set to $x^c + 3\sigma_x$ and $x^c - 3\sigma_x$, respectively. However, other forms of distributions such as uniform, triangular, and log-normal can also be used. The

closed-form CDF of this model $F(z)$ which is called DB-CDF is easily available from its integral [52]:

$$F(z) = 1 - (1 - z^a)^b \tag{2.15}$$

Due to the symmetrical nature of design variables, the final optimized device, x^o, is assumed to be in the center of the inscribed 5-D cube. Therefore, its coordinates can be easily calculated as:

$$x^o = x^c = \frac{x^l + x^u}{2} \tag{2.16}$$

By using the closed form of the obtained DB-CDF and Eq. (2.16), the yield function of Eq. (2.13) can be rewritten as follows:

$$\begin{aligned} &Yield\left(x^l, x^u\right) = \prod_{i=1}^{5}\left(F\left(\frac{x_i^u - x_i^{lb}}{x_i^{ub} - x_i^{lb}}\right) - F\left(\frac{x_i^l - x_i^{lb}}{x_i^{ub} - x_i^{lb}}\right)\right) \\ &= \prod_{i=1}^{5}\left(F\left(\frac{x_i^u - (x_i^c - 3\sigma_{x_i})}{6\sigma_{x_i}}\right) - F\left(\frac{x_i^l - (x_i^c - 3\sigma_{x_i})}{6\sigma_{x_i}}\right)\right) = \prod_{i=1}^{5}\left(F\left(\frac{x_i^u - x_i^l + 6\sigma_{x_i}}{12\sigma_{x_i}}\right) - F\left(\frac{x_i^l - x_i^u + 6\sigma_{x_i}}{12\sigma_{x_i}}\right)\right) \end{aligned} \tag{2.17}$$

The gate length and oxide thickness variations are constant for a given technology driven by the lithographic precision. Therefore, their values are set as technology specific parameters. However, the variations of other parameters are defined as percentage of the center point in every yield estimation iteration.

2.3.3 Final Optimization Problem

Till now, the probability of finding a device in a 5-D cube is estimated. However, to solve the optimization problem of Eq. (2.9), a 5-D cube should be inscribed in a feasible region which is defined based on marginal currents values. This 5-D cube is defined as follows:

$$Cube(x^l, x^u) = \left\{x \in \Re^5 | x^l \leq x \leq x^u\right\} \tag{2.18}$$

The cube is inscribed in the feasible region of F_c where every point $x \in F_c$ satisfies the I_{ON} and TL constraints.

$$F_c = \left\{x \in \Re^5 | C(x) = 1\right\} \tag{2.19}$$

The yield maximization objective is to find the 5-D cube inscribed in the F_c such that the portion of points lies in the cube be maximized. Therefore, by using Eq. (2.17), (2.18), and

(2.19), the optimization problem of Eq. (2.9) can be represented as follows:

$$\text{Given}: \begin{cases} \text{Constraints}: \ I_{ON-Min}, \tau_{Max}, \text{and } TL_{Max} \\ \\ \text{Technology} - \text{Specific Variances}: \\ \sigma_x = \left[\sigma_{L_g}, \sigma_{T_{ox}}, \sigma_{P_H}, \sigma_{P_{RW}}, \sigma_{Y_{RW}}\right] \\ \\ \text{Technology} - \text{Specific Limits}: x_i^{min}, x_i^{max} \end{cases} \tag{2.20}$$

$$\begin{cases} \underset{x^l, x^u}{\text{Maximize}} \quad Yield\left(x^l, x^u\right) \\ \\ \text{Subject to}: \begin{cases} Cube\left(x^l, x^u\right) \subseteq F_c \quad \Rightarrow x^o = \frac{x^l + x^u}{2} \\ x^l \leq x^u \\ x^{min} \leq x^c \leq x^{max} \end{cases} \end{cases}$$

To effectively solve this constrained nonlinear optimization problem, a Sequential Quadratic Programming (SQP) optimization engine is used [52, 53]. Technology-specific variances and physical limits are set for to the optimization engine. Three desired margins on delay, drive–in, and total leakage currents are also defined. The engine finds a 5-D cube in the feasible region while it maximizes $Yield$. The actual device parameters will be the center point of the cube which has the largest constraint satisfaction.

2.4 Constraint Verification Scheme

As can be seen in Eq. (2.20), the optimum 5-D cube should be inscribed in the feasible region. Traditionally, the polyhedral approximation was used to linearly model the feasible region [52]. This was done based on the assumption that the performance metrics change linearly with design variables [54]. However, this is not the case for the device design problem where the design constraints mostly behave exponentially with respect to the design variables. In addition, when using linear approximation, the polyhedral region needs to be updated in every iteration which needs expensive MEDICI simulations to find the shortest distance of the center point from the constraints and numerical calculation of the constraints' derivatives over all design variables [52]. Moreover, the design centering and worst-case distant analysis approaches [54] place the optimum point in the center of feasible region which does not necessarily provide maximum yield since the variations of process parameters are not equal. For example, there might be a design variable which is far from constraint borders in comparison to other variables but dominantly impact yield because it has a wide variation. Therefore, maximizing the yield function directly produces better results than centering the design variables or using Maxmin approach.

In this work, the containment condition ($Cube(x^l, x^u) \subseteq F_c$) is verified by checking the worst case scenarios where every x element gets its extreme value. These scenarios can be formed by $2^5 = 32$ combinations of extreme values for every x_i. By inspecting Figure 2.3 of the simple two-dimensional problem, this fact can be observed. It can be seen that locating $2^2 = 4$ corners of the rectangle $\left\{ \left(T_{ox}^l, P_H^l\right), \left(T_{ox}^l, P_H^u\right), \left(T_{ox}^u, P_H^l\right), \left(T_{ox}^u, P_H^u\right) \right\}$ in the feasible region satisfies the containment condition of the problem. Therefore, the containment verification process is reduced to corner cases checking of the design. This condition can be verified by *surface extraction* or *direct evaluation*.

2.4.1 Surface Extraction

In this approach, analytical equations of border curves or surfaces where the constraints are satisfied are extracted. Considering Figure 2.3, it can be seen that there are four curves in the space of T_{ox} and P_{RW} where the devices placed on one of those curves would satisfy the I_{ON}, τ, or TL constraints. Every curve is a border which splits the design space into two regions with respect to drive–in, intrinsic delay, or total leakage current. The intersection of the generated regions forms the feasible space. In a three dimensional problem surfaces rather than curves create the feasible space of the problem [55].

To compute analytical equations for surfaces, at first, various device parameter sets placing over the border of feasible region should be found. Then, the extracted design points should be fitted to some defined nonlinear-equation formats in order to form precise analytical representation for the surfaces. The x points satisfying the constraint borders can be obtained by applying the Gauss-Newton search algorithm [53] to the following equations:

$$\begin{aligned} I_{ON}(x) &= I_{ON-Min} \\ \tau(x) &= \tau_{Max} \\ TL(x) &= TL_{Max} \end{aligned} \tag{2.21}$$

The core of the search algorithm uses the MEDICI 2-D device simulator to calculate the delay, leakage, and drive–in currents of devices [56]. It should be noted that the proposed methodology is flexible to use 3-D or any TCAD engine for the device simulations. Finally, by using the created surfaces, the feasible space is formed and used to direct the optimization engine in order to fit the maximum yield cube in it.

However, after any change on the required I_{ON-Min}, τ_{Max}, or TL_{Max}, the surfaces should be updated to form a new feasible space with respect to the new bounds. Therefore, any attempt to design a new device with different constraints needs numerous MEDICI device simulations to form the new surfaces. Furthermore, due to complexity and imagination concerns of higher than three-dimensional problems, surface fitting and extraction will become a very hard task for our five-dimensional problem. For example, for the three-dimensional device problem there were 10

fitting parameters to describe TL surfaces with non polynomial terms [55]. As a result, in this work, the following approach has been designed as an alternative.

2.4.2 Direct Evaluation

Instead of extracting analytical equations for constraint borders, the delay, drive–in, and total leakage current of corner cases can be evaluated directly during the optimization step. In other words, the containment constraint ($Cube(x^l, x^u) \subseteq F_c$) is split into 2^N triplet constraints using the combinations of x^l and x^u elements where N is the number of design parameters. As a result, in our case, the containment constraint can be rewritten as:

$$\begin{aligned} & Cube(x^l, x^u) \subseteq F_c \\ & \equiv \begin{cases} I_{ON}\left(x_1^l|x_1^u, x_2^l|x_2^u, ..., x_5^l|x_5^u\right) \geq I_{ON-Min} \\ \tau\left(x_1^l|x_1^u, x_2^l|x_2^u, ..., x_5^l|x_5^u\right) \leq \tau_{Max} \\ TL\left(x_1^l|x_1^u, x_2^l|x_2^u, ..., x_5^l|x_5^u\right) \leq TL_{Max} \end{cases} \\ & x_i^l|x_i^u \equiv x_i^l \text{ OR } x_i^u \end{aligned} \tag{2.22}$$

To verify these constraints the MEDICI 2-D device simulator has been used. However, to improve the speed of this approach, these strategies were used: *Redundant Constraints Elimination* and *Reusing Previous Simulation Results*

2.4.2.1 Redundant Constraints Elimination

The SQP numerical optimization engine is an iterative-based algorithm which searches the problem space to find the optimum design point within the constraints. Therefore, in every iteration when a set of design corners (x^l,x^u) is picked, their feasibility should be verified. As elaborated earlier and shown in Eq. (2.22), to verify the feasibility, the containment constraint has been converted to a set of $2^5 = 32$ triplets of inequality constraints. This means that every attempt in picking a new design corner set requires 32 times simulation of devices by MEDICI which produces a long optimization time.

However, by looking through the 32 possible combinations of design corners Eq. (2.22), one can conclude that some of them are redundant and can be eliminated from the list of inequality constraints. For example, if any combination of the upper margin of gate length (L_g^u) satisfies the constraint on delay and drive–in current, others with lower margin of gate length (L_g^l) will also satisfy the constraint since their gate capacitances are lower while their saturation currents are more. Therefore, there is no need to check any combination produced by L_g^l for delay and drive–in current.

In fact, among all of the 32 triplets of corner cases just a few of them represents worst case scenarios with respect to either total leakage or drive-in currents. Consequently, the list of the potentially worst case scenarios for τ, TL, and I_{ON} is given as below:

$$\begin{array}{l} \text{a. } TL\left(L_g^l, T_{ox}^l | T_{ox}^u, P_H^l, P_{RW}^l, Y_{RW}^l | Y_{RW}^u\right) \leq TL_{Max} \\ \text{b. } TL\left(L_g^l, T_{ox}^l | T_{ox}^u, P_H^u, P_{RW}^u, Y_{RW}^l | Y_{RW}^u\right) \leq TL_{Max} \\ \text{c. } I_{ON}\left(L_g^u, T_{ox}^u, P_H^u, P_{RW}^u, Y_{RW}^l | Y_{RW}^u\right) \geq I_{ON-Min} \\ \text{d. } \tau\left(L_g^u, T_{ox}^u, P_H^u, P_{RW}^u, Y_{RW}^l | Y_{RW}^u\right) \geq \tau_{Max} \end{array} \tag{2.23}$$

Eq. (2.23-a) represents the 4 cases where gate tunneling and/or subthreshold leakage are dominant. Shorter gate length (L_g^l) increases subthreshold leakage. Furthermore, a low effective channel doping concentration due to using lower P_H and P_{RW} bounds cannot effectively overcome the short channel effects, hence increases subthreshold leakage. On the other hand, Eq. (2.23-b) represents 4 more cases where the BTBT leakage contributes effectively to the total leakage due to higher side doping concentration (N_{aside}) in the channel.

To verify the I_{ON} constraint, the slow devices among possible design corners should be selected. Such devices have the upper gate length and oxide thickness bound. Furthermore, to achieve a higher threshold voltage, the channel doping should be high as well. As a result of redundant constraint elimination, the number of constraints is reduced to 12 from $32 \times 3 = 96$.

2.4.2.2 Reusing Previous Simulation Results

As mentioned earlier, the optimization procedure is an iteration-based algorithm in which every design variable is repeatedly changed and evaluated to finally converge to the optimum solution. In every iteration, when a single variable is changed, the containment constraint is verified. Suppose that the case where L_g^l is changed, the algorithm can be sped up if simulating the corner devices for I_{ON} is ignored because the I_{ON} constraints Eq. (2.23-c) are independent of L_g^l, and we can make use of previously simulated results instead of running MEDICI redundantly.

Therefore, to speed up the approach, the simulation results (I_{ON} and TL) of every simulation could be saved and reused when needed in next iterations.

2.5 Results and Discussion

To verify the optimization methodology, various MEDICI template files have been developed to simulate Bulk-Si NMOS devices. The templates are designed such that, the value of five design parameters can be changed by the optimization engine during its execution. The terminal voltages of the transistor are set to simulate every worst case leakage current condition.

Table 2.1: Desired bounds and operating supply voltage for designed devices in 90nm technology

	HP1	HP2	LP1	LP2	HP65
I_{ON} $(\mu A/\mu m)$	≥ 1050	≥ 1050	≥ 550	≥ 550	≥ 1150
τ (ps)	≤ 1	≤ 1	≤ 2	≤ 2	≤ 0.85
TL $(nA/\mu m)$	≤ 250	≤ 125	≤ 5	≤ 2.5	≤ 500
V_{DD} (V)	1.2	1.2	1	1	1.1

Table 2.2: Obtained design parameters for each application

	HP1	HP2	LP1	LP2	HP65
L_g (nm)	43.2	41.6	66.5	64	36.5
T_{ox} (nm)	1.41	1.39	1.65	1.7	1.21
P_H $(\times 10^{18}/cm-3)$	2.1	3.4	3.7	2.5	3.8
P_{RW} $(\times 10^{18}/cm-3)$	6.7	5.3	5.9	6.8	4.3
Y_{RW} (nm)	11.3	9.8	5.9	6.9	10.7

MEDICI provides a wide range of models for every physical phenomenon. In this work, LUCMOB has been used to model carrier mobility [57]. LUCMOB is an all-inclusive model accounting for low, high, transverse, and longitudinal field effects. Furthermore, Kane's model has been used to model band to band tunneling current [58]. Finally, to model the gate direct tunneling current a silicon-oxide type insulator has been considered. The net tunneling current across the insulator is numerically calculated using the independent electron approximation [59].

For each high-performance (HP) and low-power (LP) application, two devices have been designed for 90nm technology. The $3\sigma_{T_{ox}}$ and $3\sigma_{L_g}$ are fixed to $4\% \times 1.5nm$ and $12\% \times 90nm$, while for doping parameters, 10% of their center value is assigned to their 3σ in every iteration. The defined bounds on I_{ON} and TL of each device and the corresponding supply voltage are set based on 90nm technology node specifications [11] shown in Table 2.1. To have higher drive-in current and hence better performance, the supply voltage of the HP devices are set higher than the LP ones as suggested by ITRS [11]. The HP2 and LP2 devices are high performance and low power devices with tighter constraints (i.e. the HP2 and LP2 total leakage constraints are lower than the HP1 and LP1). Moreover, a high-performance 65nm transistor (HP65) is also designed to see how the results change when different physical limits and variances are used for another technology with faster while leakier characteristics. New 3σ variations are assigned to L_g and T_{ox} as $12\% \times 65nm$ and $4\% \times 1.2nm$ for 65nm technology as well as shorter lower limit for L_g. The L_g^{min} is set to 28 and 33 nanometer for 65nm and 90nm technologies, respectively while T_{ox}^{min} is kept 1nm for both cases.

Table 2.3: Specifications of designed devices

	HP1	HP2	LP1	LP2	HP65
I_{Sub} $(nA/\mu m)$	48.1	22.5	0.99	0.6	158
I_{BTBT} $(nA/\mu m)$	6.65	21.5	0.4	8e-3	16.9
I_G $(nA/\mu m)$	15.4	19.3	1.07	0.62	93.2
J_G (A/cm^2)	35.6	46.5	1.6	0.97	255
TL $(nA/\mu m)$	70.2	63.4	2.46	1.23	268.1
I_{ON} $(\mu A/\mu m)$	1230	1204	658	648	1280
$\tau = C_g V/I_{ON}$ (ps)	0.78	0.8	1.63	1.55	0.7
V_{th} (mV)	200	212	268	305	157
$DIBL$ (mV/V)	62	48	22	50	60
$Slope$ (mV/dec)	81	81	75	76	79
$L_{Channel}$ (nm)	24.3	24.7	40.8	38.6	22.2

Table 2.2 presents the device parameters of five transistors obtained from the methodology. The HP devices have shorter gate length and thinner oxide thickness in comparison to the LP device. Moreover, to have less impurity scattering and hence more saturation current in HP devices the SSRW peak is located more far from the surface in comparison to LP devices. It should be noted that, in this method, the characteristic decay lengths of halo and SSRW are set based on the fabrication restrictions by the designer. However, the peak and position of the profiles which can be controlled by the ion dosage and the energy during ion implanting process are manipulated as design variables to gain more variation-driven toleration.

The specification of the designed devices are given in Table 2.3. It is evident that the tighter constraints on total leakage, in HP2 and LP2, causes less total leakage for corresponding devices making their drive-in current lower as well. Furthermore, the subthreshold slope factors are better for the LP devices, and the threshold voltages of them are more than the HP devices. Moreover, it can be seen that the device with more BTBT current in each HP or LP group provides more suppression to the depletion region penetration into the channel which produces lower DIBL effects. The I-V characteristic of HP1 and LP1 devices are given in Figure 2.4.

To figure out the effects of process variation on the devices' characteristics, Mont Carlo simulations were done to obtain the actual yield for all devices based on the initial defined bounds on currents and delay (see Table 2.1). To have a more realistic variation analysis and hence fair comparison between the designed devices and industrial ones, the spacer width (W_{sp}) and SDE junction depth (X_{jSDE}) are also varied in Monte Carlo simulations [60]. The W_{sp} and X_{jSDE} variances are set to $12\% \times 90$ AND 65nm and 10% respectively for 90nm and 65nm technologies.

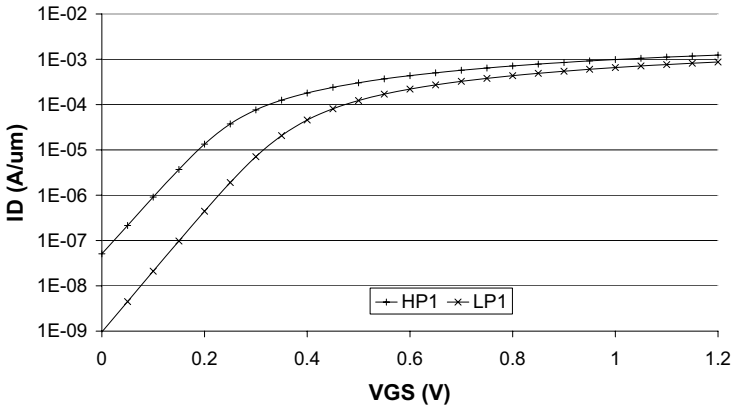


Figure 2.4: I-V characteristics of the HP1 and LP1

The given yield is equal to the percentage of devices satisfying the desired bounds under all parameter variations. The mean and standard deviation of devices' characteristics experimenting Gaussian process variations are listed in Table 2.4. It can be seen that the average speed of the HP65 device is 10% faster in comparison to 90nm devices with higher V_{DD}. However, this would be increased to 25% if the same $V_{DD} = 1.2$ was used. Moreover, the average leakage of the 65nm device is 2.5X times greater than the the 90nm device's average leakage. However, the leakage variance has not been increased with that rate as it is assumed that the absolute values of the gate length and oxide thickness variances are reduced.

Figure 2.5 is depicted to verify the optimization process. In fact, exhaustive search of the whole design space to find the globally maximum yield point by running Monte-Carlo simulation for all feasible design points is not computationally tractable. Therefore, to check if the proposed optimization approach leads a local maximum yield, the first and second derivative test at the obtained optimum point are performed, in which the gradient of the yield function should be zero: $\nabla Yield = \left(\frac{\partial Yield}{\partial L_g}, \frac{\partial Yield}{\partial T_{ox}}, \frac{\partial Yield}{\partial P_H}, \frac{\partial Yield}{\partial PRW}, \frac{\partial Yield}{\partial Y_{RW}}\right) \approx \vec{0}$, and the second derivative of it should be negative. Figure 2.5 depicts the yield curves, obtained for devices around the designed HP1 device (Table 2.2) by running Monte-Carlo simulations. Each sub-figure is extracted by varying one design parameter while keeping others constant and performing Monte Carlo simulations. For example, Figure 2.5(a) depicts the yield and the averages of device characteristics when ($T_{ox} = 1.41nm$, $P_H = 2.1 \times 10^{18}/cm{-}3$, $P_{RW} = 6.7 \times 10^{18}/cm{-}3$, $Y_{RW} = 11.3nm$) while L_g is varied from $36nm$ to $52nm$. As can be seen, the yield is maximum at the designed point and diminishes once a device parameter is moved away from its optimum value.

Finally, Figure 2.6 represents the Monte-Carlo results of the designed devices. It is evident that having both intrinsic delay and I_{ON} constraints in the performance metric constraint list is necessary as can be seen in $\tau - I_{ON}$ figures there are some devices which satisfy τ but not I_{ON} or vice versa.

Table 2.4: The means and standard deviations of devices' characterisitcs

		HP1	HP2	LP1	LP2	HP65
I_{Sub}	μ	85.7	48.3	1.5	1.4	124
$(nA/\mu m)$	σ	112	69.5	2.3	2.1	126
I_{BTBT}	μ	7.4	30.6	0.4	18e-3	34
$(nA/\mu m)$	σ	3.6	19.5	0.18	21e-3	14.5
I_G	μ	15.7	19.5	1.1	0.6	95
$(nA/\mu m)$	σ	2.6	3.5	0.18	0.9	15
TL	μ	109	98.7	3	2	253
$(nA/\mu m)$	σ	111	66	2.3	2.1	120
	σ/μ	102%	67%	75%	103%	47.5%
I_{ON}	μ	1205	1186	660	647	1246
$(\mu A/\mu m)$	σ	88	96	50	58	78
	σ/μ	7.3%	8.1%	7.6%	8.9%	6.25%
τ	μ	0.8	0.82	1.64	1.56	0.73
(ps)	σ	0.11	0.13	0.2	0.21	0.085
	σ/μ	14%	15.5%	12.1%	13.6%	11.6%
Yield		86%	74%	90.5%	75%	85%

It should be noted that the controllability of the process would not allow the L_g and T_{ox} to be optimized continuously. To resolve the issue, after obtaining optimum device parameters the L_g and T_{ox} will be rounded to the nearest achievable values, then other doping parameters will be re-optimized based on the fixed values for L_g and T_{ox}. However, the second optimization would be considerably faster as the number of design variables and verifying constraints are lesser. Also, the resulted profile parameters will not greatly change as the gate length and oxide thickness are also kept very close to the optimized values. To evaluate the yield penalty of such approach we applied this approach to the device with more deviations of L_g and T_{ox} from assumed achievable values (e.g. LP1, assuming 1nm and 0.1nm for L_g and T_{ox} levels of granularity). Therefore, new L_g and T_{ox} would be 66nm and 1.6nm respectively. Having these new fixed values the new optimized doping profiles slightly changed to $P_H = 3.8e18$, $P_{RW} = 6.1e18$, and $Y_{RW} = 6nm$. This reduces the yield from 90.5% to 85.5%.

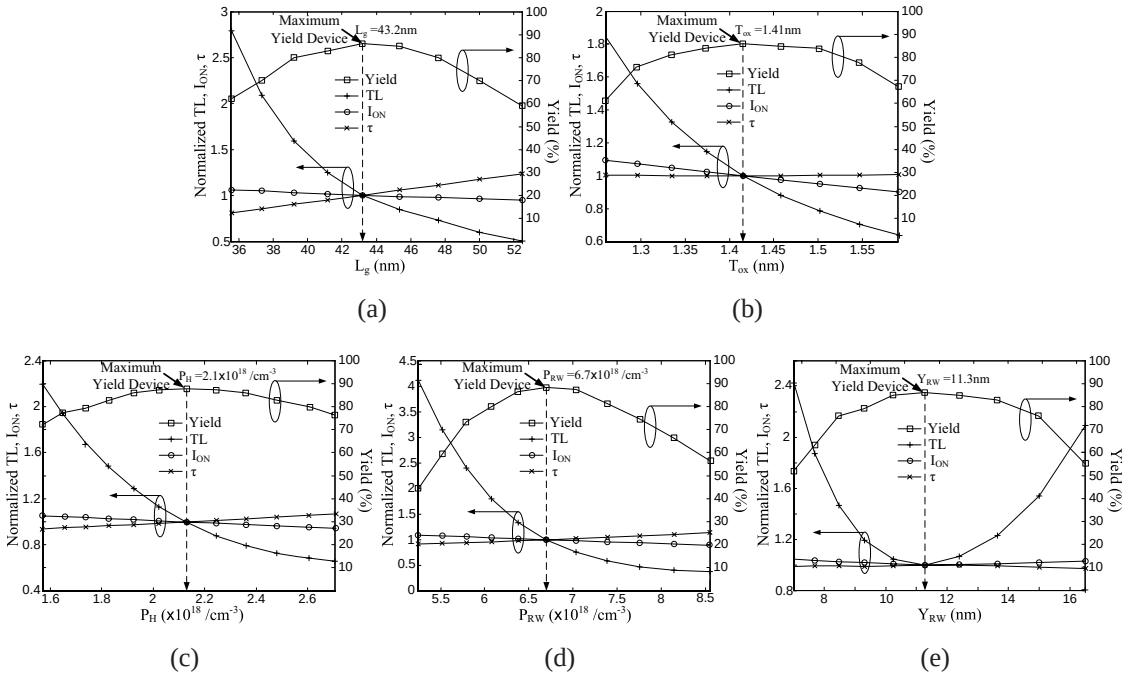
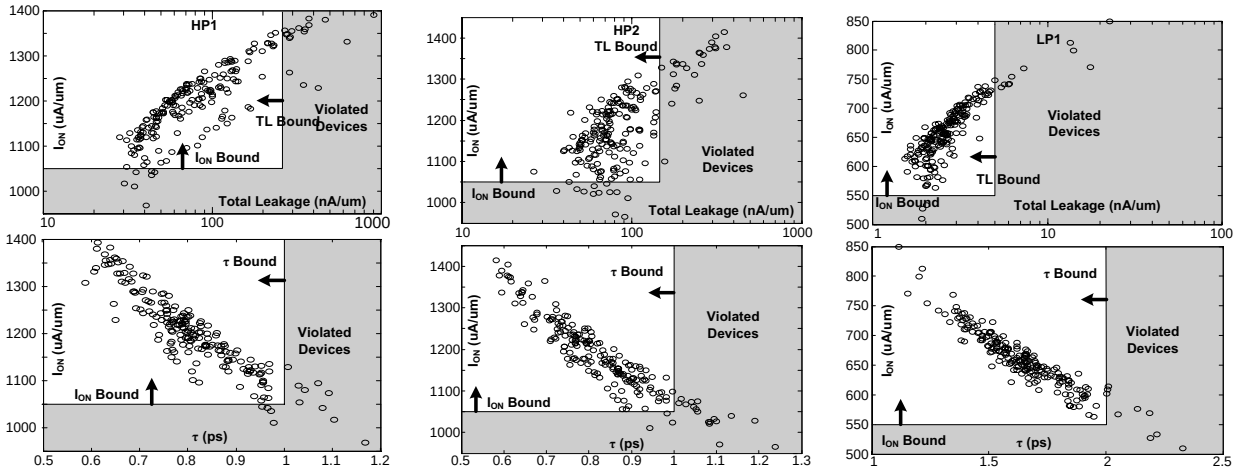


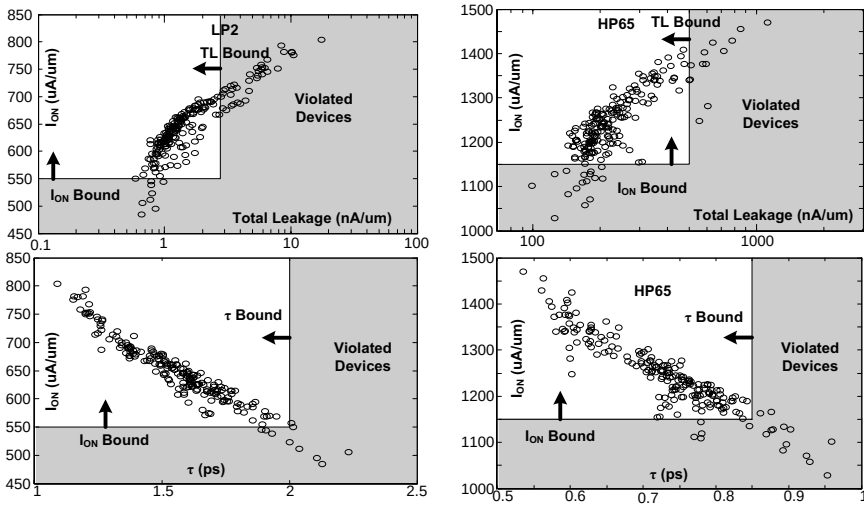
Figure 2.5: Yield, and the average of total leakages, I_{ON}, and τ, obtained by Monte-Carlo simulations for HP1 when the device parameters are shifted from the obtained optimum ones. Each figure is extracted from the cases when one device parameter is swept while others are kept equal to the parameters of HP1.

2.6 Conclusions

In this chapter, a new device design approach is proposed. This method tries to find appropriate values for oxide thickness, gate length, and channel doping profile characteristics (Halo and Retrograde Well) for a known MOS device structure such that the extracted device leads the transistor which maximally satisfies three desired constraints on intrinsic delay, saturation, and total leakage currents, in the presence of variability. The chapter presents a theoretical study of various device parameters and their effects on device characteristics and shows that variability can be considered during device design. The algorithm is based on an optimization technique which places a maximized yield cube in the problem feasible space. The center of this cube is considered as the maximum yield design point. This method takes into account different possible variances on process parameters and desired performance-leakage metrics for a particular application.



(a) HP1 (Relaxed Constraints, Yield=86%) (b) HP2 (Tight Constraints, Yield=74%) (c) LP1 (Relaxed Constraints, Yield=90.5%)



(d) LP2 (Tight Constraints, Yield=75%) (e) HP65 (Yield=85%)

Figure 2.6: Monte Carlo simulations of designed devices

Part II

Circuit-Level

The reliable yet efficient statistical analysis of VLSI circuits is a critical task to diagnose the yield concerns before entering the expensive tape-out phase. The corner-based verification techniques are prone to over-design problem or lack of efficiency due to increasing number of corners. Therefore, the probabilistic-based (non-Monte Carlo) methods have been considered by many researcher as the ultimate solution. Generally, these methods simplify or ignore some second order effects of performance metrics or process variations models, in order to provide tractable solutions for yield estimation of today's large-scale VLSI circuits. As a result, the Monte Carlo (MC) method is still considered as a reliable alternative. However, the major drawback of the MC method is its slow rate of convergence. In this part of the thesis, several solutions are proposed for the efficient sampling-based variability analysis of VLSI circuits through the adoption of advanced sampling and variance-reduction methods. Different solutions are provided for digital and analog circuits, and SRAM cells.

Chapter 3

Overview of Advanced Sampling and Variance Reduction Methods

The traditional Monte Carlo analysis has a very slow convergence rate, so a large number of samples is required to accurately analyze the variabilities, if such a methodology is adopted. Therefore, a number of advanced sampling and variance reduction methods have been adopted for the statistical analysis of different types of VLSI circuits to improve upon the quality of estimations and reduce the number of simulation cycles. In this chapter, an abstract overview of these methods are provided to readers. The detail explanations of the proposed methods for the variability analysis of digital and analog circuits, and SRAM cells are provided in the subsequent chapters of this part.

3.1 Introduction to Monte Carlo method

Suppose $\boldsymbol{x} = \{x^{(1)}, x^{(2)}, \ldots, x^{(d)}\}$ is a set of d-dimensional process parameters with a known Joint Probability Density Function (JPDF), $\varphi(\boldsymbol{x}) : \mathbb{R}^d \rightarrow \mathbb{R}$. Each $x^{(j)}$ represents a process parameter of a circuit element, such as a transistor gate length, width, oxide thickness, threshold voltage, or interconnect dimension. If $h = p(\boldsymbol{x})$ is the performance metric of the circuit under statistical analysis, the following integral can be used to formulate a measure of variability

$$\mu_g = E_\varphi [g(p(\boldsymbol{x}))] = \int_{\mathbb{R}^d} g(p(\boldsymbol{x}))\, \varphi(\boldsymbol{x})\, d\boldsymbol{x}, \tag{3.1}$$

where $E_\varphi[.]$ is the expected value, given φ as the density function of the random parameters. For example, if $g(h)$ is set to the following single-constraint indicator function, $I_\tau(h)$:

$$I_\tau(h) = \begin{cases} 0 & h > \tau \\ 1 & h \leq \tau \end{cases}, \tag{3.2}$$

then the integral of Eq. (3.1) will be equal to the yield of the circuit with respect to the performance metric p and the critical value of τ for the metric, $p(\boldsymbol{x})$. Consequently, $g(h) = h$ and $g(h) = h^r$ lead to the mean and the r-th moment of the metric, respectively.

The MC method suggests a numerical technique to estimate the integral, by N times sampling from the $\varphi(\boldsymbol{x})$ distribution, evaluating the circuit's performance metric in every iteration, and finding the yield or the statistical moments using

$$\hat{\mu}_g = \frac{\sum_{i=1}^{N} g\left(p\left(\boldsymbol{x}_i\right)\right)}{N}. \tag{3.3}$$

Since the samples of consequent simulations are independent and identically distributed, the traditional MC method leads to an estimate with the following variance

$$\mathrm{var}_{\mathrm{MC}}\left(\hat{\mu}_g\right) = N^{-1}\mathrm{var}\left(g\left(p\left(\boldsymbol{x}\right)\right)\right) = N^{-1}\int \left(g\left(p\left(\boldsymbol{x}\right)\right) - \mu_g\right)^2 \varphi\left(\boldsymbol{x}\right) d\boldsymbol{x}. \tag{3.4}$$

The advanced MC methodologies focus on finding alternative estimators or sampling techniques that reduce the variance of the estimation, hence, reduce the required number of samples for a given accuracy.

The proposed methods are compared with the traditional-MC in terms of the estimation bias and standard deviation of the estimation, where the estimation target is mostly the circuit yield or sometimes the statistical moments (mean, standard deviation, and skewness) of a performance metric. Suppose finding yield is the objective of an estimation. If the estimated yield using the proposed is $\hat{y}$, following is the bias of the estimation:

$$\mathrm{bias} = E\left[\hat{y}\right] - y \tag{3.5}$$

where y is the exact yield. The expected value can be estimated by running the proposed method for several times (m times), recording the list of the estimated yield in each run $\{\hat{y}_1, \cdots, \hat{y}_m\}$, and finding the sample expected value by averaging. Note that, each run of the experiment uses N samples advanced sampling technique.

As most of the applied methods in the consequent chapters are unbiased, the bias of a method is reported only if any is observed. However, the major study is conducted by comparing the standard deviation of the estimation against the traditional MC. The standard deviation of an estimation reveals the level of confidence (accuracy) of an estimation. Such a measure can be obtained by calculating the sample standard deviation of $\{\hat{y}_1, \cdots, \hat{y}_m\}$, the experimented yield estimations.

3.2 Latin Hypercube Sampling

Latin Hypercube Sampling (LHS) method [61] partitions each dimension (process parameter) into N disjoint and equi-probable parts, then it draws a sample from each part according to the probability density of the random variable in that part. The samples are then randomly permuted to form N sets of d dimensional samples, guaranteeing a uniform coverage in each dimension.

Suppose the function under expected value analysis, $f(\boldsymbol{x}) = g(p(\boldsymbol{x}))$, is decomposed into the following additive form

$$f(\boldsymbol{x}) = \mu_g + \sum_{j=1}^{d} f_j\left(x^{(j)}\right) + r(\boldsymbol{x}), \tag{3.6}$$

where μ_g is the mean of g (or f) as defined in Eq. (3.1), f_j is a function of $x^{(j)}$ (the j-th process parameter) representing the main effect of the j-th process parameter alone, and r is the residual due to higher order interaction between process parameters. Note that f_j can be formed, as

$$f_j\left(x^{(j)}\right) = \int \left(f(\boldsymbol{x}) - \mu_g\right) \varphi(\boldsymbol{x}) \prod_{\substack{i=1 \\ i \neq j}}^{d} dx^{(i)}. \tag{3.7}$$

If the LHS samples are used to estimate the expected value based on the estimator of Eq. (3.3), the variance of the estimation is [62]

$$\mathrm{var}_{\mathrm{LHS}}(\hat{\mu}_g) = N^{-1} \int r(\boldsymbol{x})^2 \varphi(\boldsymbol{x})\, d\boldsymbol{x} + O(N^{-1}). \tag{3.8}$$

A comparison between equations (3.4) and (3.8) reveals that, the LHS method filters out the main effect parts (or the 1-D ANOVA terms), as a result, the closer f is to the additive forms or the smaller the residual part is, the more the Latin hypercube sampling will help.

Getting back to the VLSI circuit problem, it is usually the variance (e.g. input referred offset of a comparator) and the yield of the circuit that are under investigation. If the variance is needed, then $g(h) = h^2$, therefore, even if $p(\boldsymbol{x})$ (or h), the metric, is composed of major 1-D additive parts, the square of it composes of major 2-D components due to pairwise multiplication of 1-D terms. As a result, a decomposition of the form of Eq. (3.6) will yield a significant residual term due to interaction components. This means that the LHS method does not provide much saving when applied for the the estimation of variance especially for high dimensional cases where the ratio between the pairwise over additive terms increases significantly.

The problem could be even worse for the yield analysis. The indicator function of Eq. (3.2) consists of many higher than one degree terms, especially when yield is close to two extremes.

This can be verified by following approximation. Suppose the error function, a sigmoid function, is used to approximate the indicator function. If a Taylor expansion is used, then

$$g(h) = I_\tau(h) \approx \frac{1+\mathrm{erf}(\alpha(\tau-h))}{2} = \frac{1}{2} + \frac{1}{\sqrt{\pi}}\left(\alpha(\tau-h) - \frac{\alpha^3(\tau-h)^3}{3} + \frac{\alpha^5(\tau-h)^5}{10} - \cdots\right). \tag{3.9}$$

As a result, the closer yield is to 1 or 0 the more far $p(\boldsymbol{x})$ (or h) is from τ, hence, the higher order terms will appear more in the residual portion. The application of LHS for analog circuits yield analysis has been proposed earlier in [63]. However, the authors faced this issue and showed through extensive simulations that the efficiency of their approach significantly drops when the yield reaches the extremes (e.g. over 90% or below 10%). This is a critical issue since the domain of attraction in a VLSI circuit yield analysis problem is actually around the extremes.

In Chapter 5, an advanced LHS-based method for the efficient variability analysis of analog circuits is proposed.

3.3 Quasi Monte Carlo Sampling

An important property of the estimation error, $\hat{\mu}_g - \mu_g$, is that it is related to the equi-distribution (uniformity) of the samples rather than their randomness. This idea strongly suggests that by using a well-spread sequence, which is more uniform than a pseudo-random sequence, a more precise estimation can be achieved [64]. The LHS method, in fact, tries to achieve this goal by increasing uniformity in 1-D projections. However, the discrepancy of the LHS samples is not noticeably better than that of the traditional pseudo random-MC samples in projections higher than 1-D, since the permutation of the samples are performed randomly.

The QMC method utilizes low-discrepancy sequences to provide uniformity in 1-D and higher dimensions projections. However, the convergence rate of the QMC method is dependent to the problem dimension, and it is found to be only asymptotically superior to MC [64], unless the problem is effectively low dimensional in superposition sense [65, 66]. The effective dimension is determined using ANalysis Of VAriance (ANOVA) decomposition of the function f, similar to what has been done in Eq. (3.6) but by continuing the decomposition of the residual term into functions of higher dimensional components, as follows

$$f(\boldsymbol{x}) = \sum_{u\subseteq\ell} f_u(\boldsymbol{x}) = \mu_g + \sum_{i=1}^{d} f_i\left(x^{(i)}\right) + \sum\sum_{i<j} f_{ij}\left(x^{(i)}, x^{(j)}\right) + \cdots + f_{1\cdots d}\left(x^{(1)}, \cdots, x^{(d)}\right), \tag{3.10}$$

where $\ell = \{1, 2, \cdots, \mathrm{d}\}$.

The ANOVA terms are orthogonal under the process parameter JPDF space

$$\int f_u(\boldsymbol{x})\, f_v(\boldsymbol{x})\, \varphi(\boldsymbol{x})\, d\boldsymbol{x} = 0 \quad \text{when u} \neq \text{v}. \tag{3.11}$$

Therefore, the variance of the integrand function, f, can be expressed as the sum of the variances of all of the orthogonal functions, as follows

$$\sigma^2(f) = \sum_{u \subseteq \ell} \sigma^2(f_u). \tag{3.12}$$

If the significant portion of the integrand function's variance is due to ANOVA terms with small dimensions then the problem has low effective dimension in superposition sense [65]. For example, if 90% of the variance of f is due to the functions of single $\boldsymbol{x}$ variables (main effects) and the functions of pairs of variables, the effective dimension in the superposition sense is two. This means that the interactions of more than two random variables have negligible effects (10%) on the function. The superiority of the QMC versus pseudo-random MC method for some of the high-dimensional problems arises from the low-effective dimensionality of such problems and the fact that QMC sequences produce high uniformity in low order projections [67, 66]. However, even for moderate size problems (20 dimension or more), the finite and moderate size (100s) QMC samples can not perfectly cover the high dimensional projections due to the need of exponential number of samples with respect to dimensions [68].

In Chapter 4 the effective dimension and the application of the QMC for digital circuit timing yield analysis are studied, and a solution is proposed to improve the uniformity of the generated samples in high dimensional projections for that application.

3.4 Control Variate Method

Control variate is a promising variance reduction technique for expected value estimation only when a highly correlated auxiliary model (control variable) is available [69]. The amount of the variance reduction is dependent to the magnitude of the correlation between the control variable and the variable of interest, under expected value estimation. The exact expected value of the control variable must also be known.

Suppose f is the random parameter under expected value estimation, if c is the control variable with known expected value of μ_c, then f can be substituted by f^* in computation of $E[f]$

$$f^* = f - \beta(c - \mu_c), \tag{3.13}$$

where β is a constant. The original estimator of Eq. (3.3) can be replaced by

$$\hat{\mu}_g = \frac{\sum_{i=1}^{N} g(p(\boldsymbol{x}_i))}{N} - \beta \left(\frac{\sum_{i=1}^{N} c(\boldsymbol{x}_i)}{N} - \mu_c \right), \tag{3.14}$$

which leads to an estimation variance of

$$\mathrm{var}(f^*) = \mathrm{var}(f) - 2\beta\mathrm{cov}(f,c) + \beta^2\mathrm{var}(c). \tag{3.15}$$

Therefore, a significant variance reduction can be achieved by proper setting of β if f and c are highly correlated. The optimum value of β that minimizes the estimation variance is

$$\beta = \frac{\mathrm{cov}(f,c)}{\mathrm{var}(c)} = \frac{\rho\sigma_f}{\sigma_c}, \tag{3.16}$$

where ρ, σ_f, and σ_c are the correlation coefficient and the standard deviations of f and c, respectively. By using the optimum β, the variance of the new estimator, is reduced to

$$\mathrm{var}_{\mathrm{CV}}(\hat{\mu}_g) = \left(1-\rho^2\right)\mathrm{var}_{\mathrm{MC}}(\hat{\mu}_g). \tag{3.17}$$

However, this classical formulation is not efficient for yield estimation, if it is used directly for the yield indicator function of Eq. (3.2). This is because to obtain a highly correlated control variable with the yield variable formulated as (3.2), an auxiliary model of performance metric, $p(\boldsymbol{x})$, should be found that is not only highly correlated with it but also has the same range and scale. This problem is even worse if the yield approaches the extremes since the scale and range of the model can hardly follow the actual metric in tails. This issue has been observed in the early applications of control variate for circuit yield analysis [70].

To overcome this issue, two different approaches may be taken. **i)** The first few statistical moments can be found efficiently by using the control variate method, then the yield is modeled by fitting a generic distribution (such as Gaussian) to the metric. **ii)** Using an order statistics-base control variate quantile estimator [71]. This technique needs a large number of samples, especially for extreme yield values to eliminate an inherited bias (e.g. more than 230 samples for a 99% yield).

However, it should be noted that in contrast to the LHS and QMC that are black-box sampling methods, the control variate method requires good models from the circuit and the performance metrics. Constructing such models may require circuit analysis and simulations using response surface method after every circuit manipulation, that is a major obstacle toward a practical application of the method for some type of VLSI circuits such as large-scale analog circuits. Moreover, the promised variance reduction of Eq. (3.17) can only be achieved if the optimum β is used which itself requires additional simulations for the estimation of $\mathrm{cov}(f,c)$ in Eq. (3.16).

In Chapter 4, the challenge of the timing yield estimation of digital circuits is studied and two control-variate based solutions are provided for efficient statistical static timing analysis.

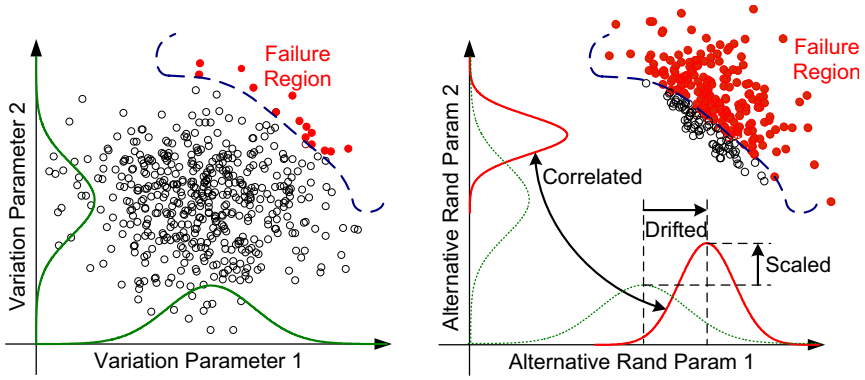


Figure 3.1: An example of importance sampling: capturing more failure cases by using a multivariate correlated-scaled-drifted Gaussian alternative distribution.

3.5 Importance Sampling

Important sampling is another gray-box variance reduction technique that has been adopted for the efficient MC-based yield estimation of circuits with rare failure rate [70, 72].

The problem of the traditional-MC method especially for the yield estimation of circuits with extremely high yield is that most of the generated samples by the $\varphi(\boldsymbol{x})$ distribution reside in the acceptable region. Since these samples do not contribute to the calculation of the failure rate, their simulation is only a waste of runtime. As a result, if an alternative distribution, $\gamma(\boldsymbol{x})$, is chosen to simulate the random parameters such that more failure cases are observed (Fig. 3.1), the variance of the estimation error is reduced. In other words, if the integral in (3.1) is rewritten as

$$\int_{\mathbb{R}^d} \frac{g\left(p\left(\boldsymbol{x}\right)\right)\varphi\left(\boldsymbol{x}\right)}{\gamma(\boldsymbol{x})}\gamma(\boldsymbol{x})\,d\boldsymbol{x} = E_{\gamma}\left[\frac{g\left(p\left(\boldsymbol{x}\right)\right)\varphi\left(\boldsymbol{x}\right)}{\gamma(\boldsymbol{x})}\right], \tag{3.18}$$

then, by simulating the samples from the $\gamma(\boldsymbol{x})$ distribution, the following can be used as an unbiased estimator instead of the original estimator of Eq. (3.3)

$$\hat{\mu}_g = \frac{1}{N}\sum_{i=1}^{N}\frac{g\left(p\left(\boldsymbol{x}_i\right)\right)\varphi\left(\boldsymbol{x}_i\right)}{\gamma(\boldsymbol{x}_i)}. \tag{3.19}$$

Therefore, the variance of the new estimator is

$$\mathrm{var}_{\mathrm{IS}}\left(\hat{\mu}_g\right) = N^{-1}\left[\int \frac{g^2\left(p\left(\boldsymbol{x}\right)\right)\varphi^2\left(\boldsymbol{x}\right)}{\gamma(\boldsymbol{x})}d\boldsymbol{x} - \mu_g^2\right]. \tag{3.20}$$

This variance can ideally reach zero, if

$$\gamma(\boldsymbol{x}) = \frac{g\left(p\left(\boldsymbol{x}\right)\right)\varphi\left(\boldsymbol{x}\right)}{\mu_g}. \tag{3.21}$$

However, finding such an alternative distribution is not always an easy task since μ_g and $g(p(\boldsymbol{x}))$ are unknown prior to analysis.

In summary, the importance sampling technique is mostly useful for the analysis of very rare failure rates. Also, its performance degrades very fast and may even does worse than traditional-MC for even moderate dimension problems ($d > 10$) due to possible missing or less emphasize on some parts of the important regions.

In Chapter 6, the importance sampling for the application of SRAM cell rare failure rate estimation is studied, and an adaptive sampling technique is proposed which updates the alternative distribution toward minimizing the estimation variance.

3.6 Stratified Sampling

In this technique, the problem space is divided into r disjoint partitions and the statistics of interest is estimated in each stratum separately [73]. Suppose Q_i is the i-th stratum and $p_i = \int_{Q_i} \varphi(\boldsymbol{x})\, d\boldsymbol{x}$ is the probability of having a sample in that stratum. Then the expected value of the statistics of interest in that stratum can be found using

$$\hat{\mu}_{g_i} = \frac{\sum_{i=1}^{N_i} g(p(\boldsymbol{x}_i))}{N_i}, \tag{3.22}$$

by sampling $\boldsymbol{x}$ from following distribution

$$\varphi_i(\boldsymbol{x}) = \begin{cases} \varphi(\boldsymbol{x}) / p_i & \boldsymbol{x} \in Q_i \\ 0 & \boldsymbol{x} \notin Q_i \end{cases} \tag{3.23}$$

Then the alternative estimator to Eq. (3.3) will be

$$\hat{\mu}_g = \sum_{i=1}^{r} p_i \hat{\mu}_{g_i}, \tag{3.24}$$

and the variance of this estimator is

$$\mathrm{var}_{\mathrm{ST}}(\hat{\mu}_g) = \sum_{i=1}^{r} p_i^2 \frac{\mathrm{var}_{\mathrm{MC}}(g|Q_i)}{N_i}. \tag{3.25}$$

Therefore, a variance reduction can be achieved if more samples are used for stratum with high variance of g.

Stratified sampling has been adopted for the yield analysis of digital and analog circuits [74, 75]. However, this technique has a limited performance improvement in high dimensional

problems. This is because of the limited number of strata due to limited number of simulation cycles. Note that at least one sample is needed in each stratum, hence in high dimensional problems each stratum covers a very large super-cube. Moreover, in order to gain a variance reduction, the number of samples in each stratum should be determined according to the variance of the statistics under analysis which itself requires knowledge of the circuit response surface and consequently additional characterization step simulations. This issue originates from the gray-box nature of the solution, similar to the control variate and importance sampling methods.

Chapter 4

Digital Circuits: Advanced Monte Carlo-Based Statistical Timing Analysis Methodologies

4.1 Introduction

The reliable yet efficient Statistical Static Timing Analysis (SSTA) is a central task in predicting the yield of a high-performance digital VLSI circuit. The corner-based timing verification techniques are prone to over-design issue and may not lead to an efficient design for a tight power consumption budget. Therefore, several probabilistic-based (non-Monte Carlo) SSTA methods have been proposed to address the challenge of statistical timing analysis for high performance digital circuits. In the probabilistic-based SSTA methods, the signal arrival-times are treated as random variables, and the Probability Distribution Function (PDF) of the circuit's critical delay is extracted by proper statistical analysis. Blaauw et al. [76] provides a recent survey on the state-of-the-art SSTA methods.

Despite the considerable improvement of the recent SSTA methods, there are still concerns on their applications for a reliable and large-scale timing sign-off. The major challenge and drawback of the current probabilistic-SSTA approaches originate from the presence of complex timing and the process variation effects that are partly ignored or simplified in each solution. Such effects include, the nonlinearity of gate delays as a function of the process parameters and capacitive loads, the nonlinearity of the MAX operation due to the arrival time merging, and the resultant non-zero skew signal arrival time PDFs. The interdependency among input/output rise/fall signal transition times and gate delays, interconnect delay models, non-Gaussian process parameters, or the spatial/structural correlations, are some of the other complex issues that have been partially overlooked in the proposed probabilistic-based methods.

Therefore, the Monte-Carlo (MC) method, as a traditional alternative to probabilistic techniques, has recently attracted attentions for a reliable and accurate digital circuit timing sign-off [77, 78, 74, 79]. The major advantage of the MC method is its capability to account for any

timing and process model. Moreover, the development and integration costs of MC-based SSTA tools are minimal, since the available deterministic-STA engines can maximally be reused in developing a new MC-based yield analysis tool. These are in addition to the benefits of breaking of an MC-based timing analyzer into parallel processes and gain from running them on multi-processor systems [74]. However, the most threatening disadvantage of the traditional traditional-MC statistical analysis method is its slow convergence rate. That means to achieve a reasonably precise estimation of the yield, thousands of simulations (samples) might be needed by using the traditional-MC analysis. The precision of the MC-based methods is defined in terms of the statistical confidence interval of the estimation. The traditional-MC's rate of convergence is independent of the problem's dimension, but it decays with the slow rate of $O(N^{-1/2})$ with respect to the number of samples N.

In order to improve upon the convergence rate of the traditional-MC sampling, hybrid sampling methods composed of the Latin Hypercube Sampling (LHS) and a Quasi-MC (QMC) sequence have been proposed recently [74, 79]. LHS method samples every dimension by stratifying its domain into equi-probable subranges, hence it improves the uniformity of the samples in one-dimensional projections. Whereas, the QMC utilizes low-discrepancy sequences for 1-D and higher projections. It is proved that, the estimation error is mitigated by the equi-distribution (uniformity) of the samples rather than their randomness [64]. Therefore, the upper bound of the convergence rate of the QMC method, $O(\log^d N/N)$, is found to be *asymptotically* superior to MC, where d is the problem dimension (e.g. the number of process parameters or the number of principal components). This asymptotic advantage seems to be only achievable if $N >> e^d$ that is absolutely impractical for even moderate size problems. However, the QMC method exhibited a significant advantage over the traditional-MC for the analysis of the high-dimensional computational finance problems during 1990s [80]. This surprising behavior is later justified through the analysis of the variance (ANOVA) of the high-dimensional problems and quantified with the notion of effective dimension [65, 66, 67].

In Section 4.3, this phenomenon is reviewed to provide an insight on how a QMC sampling method can be effectively adjusted for the SSTA problem. The effective dimension of the digital circuit's yield estimation problem is therefore investigated. Inspired by the observations, an algorithm is then proposed to improve the uniformity of the Sobol's [81, 82] QMC samples in high-order projections. By using the proposed optimized-Sobol and LHS method, an SSTA engine is developed to target a high-performance timing yield analysis that requires a fewer number of iterations to estimate the yield with certain confidence than that of a non-optimized QMC sampler.

As will be discussed later, the proposed QMC-based SSTA engine does not significantly outperform the traditional-MC method for moderate number of samples (e.g., $<$2000). As a result, a control variate-based technique is proposed in Section 4.4 to address the concern by providing a mechanism which significantly improves the confidence interval-range when using only a few hundreds of samples. The proposed technique leverages the high accuracy of an

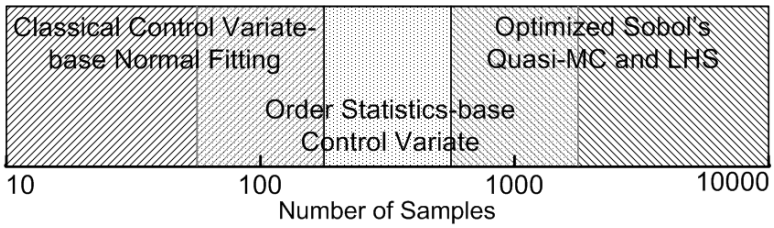


Figure 4.1: The approximate range preferred for each proposed method.

analytically extracted timing model of the nominally-critical path as an auxiliary variable in an order-statistics-based variance-reduced estimator.

However, the drawback of using that estimator is the need for almost a hundred or more samples to avoid an unwanted bias. As a result, another method is developed in Section 4.5, for the cases when only a few (e.g., an order of tens) number of simulations is needed. This is the case for very large circuits or very early design phases where re-design/analysis iterations are run very frequently. The method uses the same auxiliary random variable and applies the classical control variate technique for the estimation of the critical delay's mean and variance. The Gaussian distribution is then used to form a PDF of the critical delay and approximate the yield.

Figure 4.1 illustrates the relative range of the effectiveness for the three proposed method with respect to the number of samples. The scale of the number of samples might vary based on the circuit and its yield, which will be covered in the Section 4.6, where a mechanism is developed to integrate the proposed methods into a single highly efficient MC-based SSTA engine.

4.2 Delay and Process Variation Models, and Simulation Setup

In order to study the performance of the proposed MC-based methods, extensive MC-based timing analysis of digital circuits are performed in this research. The result part of each section that will discuss the advantages and drawback causes of each method is conducted through analysis of number of circuit benchmarks varying in size and logic depth, two critical factors in performance evaluation of an SSTA engine. In addition to examining different types of circuits, e.g. circuits with many short critical paths vs. circuits with a few but long paths, different assumptions are also made for the process variation decompositions into global, spatial, and random components.

Therefore, before actually introducing the MC-based methods in Section 4.3-4.6, in order to avoid presenting repeated simulation setup information in every sections, this section is fully dedicated to provide information on the benchmark circuits, timing, and process models that have been used in analysis throughout this work.

Table 4.1: Benchmark Circuits

Circuit	Cells	Depth	Percentage of critical gates (Slack/D)				
			0	0.01	0.02	0.05	0.1
C432	157	25	15.3	20.3	22.2	50.3	71.3
C499	514	30	15.7	21.2	47.3	69.2	81.1
C880	342	28	8.5	12.0	14.6	16.9	19.0
C1355	483	29	15.1	20.9	46.1	70.6	80.5
C1908	359	37	10.3	10.3	17.5	31.7	46.5
C2670	666	28	5.3	5.3	7.4	10.7	14.5
C3540	733	44	5.7	9.0	10.0	14.5	30.4
C5315	1541	43	3.0	3.8	4.5	5.1	7.5
C6288	2397	121	4.8	10.3	15.7	41.5	65.4
C7552	1924	58	2	3.2	3.4	4.5	5.7
S9234	820	27	4.5	6.1	6.2	7.1	10.6
S13207	1935	29	1.4	1.4	2.0	2.6	3.8
S15850	2735	47	2.3	3.0	3.5	4.2	6.7
S35932	7872	14	4.3	4.3	13.8	28.3	31.4
S38417	8291	37	0.58	0.59	0.63	0.63	0.64
S38584	8249	34	0.47	0.47	0.65	0.67	1.4

ISCAS85 and 89 benchmark circuits [83, 84] are used. The circuits are synthesized using an industrial 65nm CMOS library cell with only inverters, and 2 and 3-inputs NAND and NOR gates. The timing response surfaces of each logic cell is characterized quadratically to deliver a high quality of approximation in terms of process parameters. The output rise/fall and the the propagation delay of each library cell are modeled as functions of input rise/fall time, output load, gate length, and threshold voltage of that cell. The characterized response surface models are composed of constant, linear, quadratic, and linear-linear and linear-quadratic interaction terms. Table 4.1 shows the number of cells and synthesized logic-depth (the depth of the longest path) of each circuit. The registers' input/outputs of the sequential benchmarks are treated as pseudo outputs/inputs during the timing analysis.

The last five columns of the table show the percentage of the logic cells that have timing slack equal or less than zero, 1%, 2%, 5%, or 10% of the critical delay. These values are used in later discussions in order to provide insight on the type of circuits that manifest high or low improvement of yield estimation accuracy using different MC-based methods.

In this work, two types of process parameters are considered, a purely random (the threshold voltage V_T) and a spatially correlation one (the gate length L). The gate length variation itself is decomposed into three distinct components: inter-die or global (ΔL_g), spatially correlated intra-die variation (ΔL_s), and a random residual part (ΔL_r). The spatially correlated behavior of

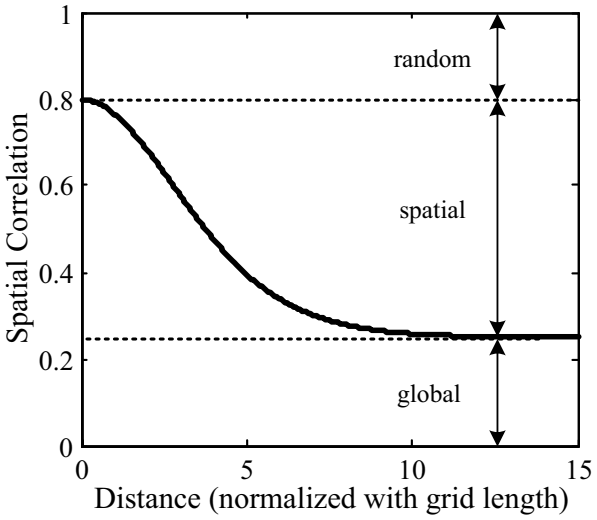


Figure 4.2: Spatial Correlation

the gate length variation is originated from the lithography imperfection that affects the close devices more similarly. This issue has been extensively investigated and modeled in [85, 86, 87]. The area of the conceptual die is divided into rectangular grids, and the gate length variation of transistors sharing a grid are assumed to have equal spatial correlation component, while the random part varies for each transistor separately. The global component is equally added up to all transistors of the circuit. As a result for a circuit with n cells placed in a mesh of m grids, $d = m + 2n$ random variables are used where the $2n$ variables are due to purely random parameters and the residual portion of the correlated parameters, while the m variables are the correlated random variables representing the global and spatial components of the gate length variation in different grids.

The model represented in [85] is used to ensure the positive definiteness of the covariance matrix of the m grid gate length values, as:

$$\text{corr}\left(\Delta L_i, \Delta L_j\right) = \frac{\sigma_G^2 + \rho\left(v_{ij}\right)\sigma_S^2}{\sigma_G^2 + \sigma_S^2 + \sigma_R^2} \tag{4.1}$$

where $v_i j$ is the euclidean distance between the grid i and j, while the variance of each of the three components are σ_G^2, σ_S^2, and σ_R^2. The normalized $\rho\left(v_{ij}\right)$ ratio is computed using the following function:

$$\rho\left(v\right) = 2\left(\frac{bv}{2}\right)^{s-1} K_{s-1}\left(bv\right)\Gamma\left(s-1\right)^{-1} \tag{4.2}$$

where K is the modified Bessel function of the second kind, Γ is the gamma function, and b and s are two real parameter numbers that adjust the shape of the function [85]. Throughout this

chapter various portions of global, spatial, and random variations as well as the shape parameters are used to investigate the effect of the correlated process parameters in the performance of the proposed method.

Throughout this chapter, unless a different setup is mentioned, the magnitude of the global, spatial, and random components of the gate length variation are set to 25%, 55%, and 20% of the total gate length variation ($\sigma_L = 0.12L$) [85]. The grid size is set such that almost every 20 cells are placed in one square-shape grid. As a result, the mesh structure varies from 3×3 to 20×20 for different circuits. The parameters b and s are set to $1/l_{grid}$ and 6, where l_{grid} is the length of the square grid. Figure 4.2 depicts the resultant spatial correlation.

Finally, it should be noted that the Capo [88] placer is used to place the logic cells in order to determine the distance and hence the correlation coefficients of the spatial parameters.

4.3 Efficient QMC/LHS -base SSTA

In this section, the notion of the effective dimension, introduced to explain the unexpectedly high performance of the QMC methods, is reviewed. The effective dimension of the digital circuits' timing yield problem is investigated. Inspired by the effective dimension analysis, an algorithm is developed to minimize the discrepancies of Sobol's QMC samples following the need for low-discrepancy samples in high-dimensional projections for an efficient yield analysis. Finally, a QMC/LHS -base SSTA engine is proposed for the efficient timing yield estimation of digital circuits.

4.3.1 QMC, Effective Dimension and Timing Yield

Suppose $\boldsymbol{p} = \left\{ p^{(1)}, p^{(2)}, \ldots, p^{(d)} \right\}$ is a set of d-dimensional random variables with a known Joint Probability Distribution Function (JPDF), $\varphi(\boldsymbol{p}) : \mathbb{R}^d \rightarrow \mathbb{R}$. Each $p^{(i)}$ represents either a purely random process parameter such as the RDF-driven threshold voltage variation or the residual random component of a spatially correlated parameter such as the gate length variation. If $D(\boldsymbol{p})$ is the critical delay of a circuit, then the following indicator function, I, divides the problem space ($\boldsymbol{p} \in \mathbb{R}^d$) into unacceptable ($I = 0$) and acceptable ($I = 1$) regions, represented as:

$$I_\tau(\boldsymbol{p}) = \begin{cases} 0 & D(\boldsymbol{p}) > \tau \\ 1 & D(\boldsymbol{p}) \leq \tau \end{cases}, \tag{4.3}$$

where τ is the maximum acceptable critical delay. Therefore, the following integral represents the timing yield:

$$y = P(I_\tau = 1) = E_\varphi[I_\tau(\boldsymbol{p})] = \int_{\mathbb{R}^d} I_\tau(\boldsymbol{p})\, \varphi(\boldsymbol{p})\, d\boldsymbol{p}. \tag{4.4}$$

The MC method suggests a numerical technique to solve the integral in (4.4) by N times sampling from the $\varphi(\boldsymbol{p})$ distribution, evaluating the circuit's critical delay, and extracting the mean of $I_\tau(\boldsymbol{p})$ by using the following estimator:

$$\hat{y} = \frac{\#\{i|D_i < \tau\}}{N}, \tag{4.5}$$

where $\#\{.\}$ is the number of elements in a set and N is the total number simulation iterations.

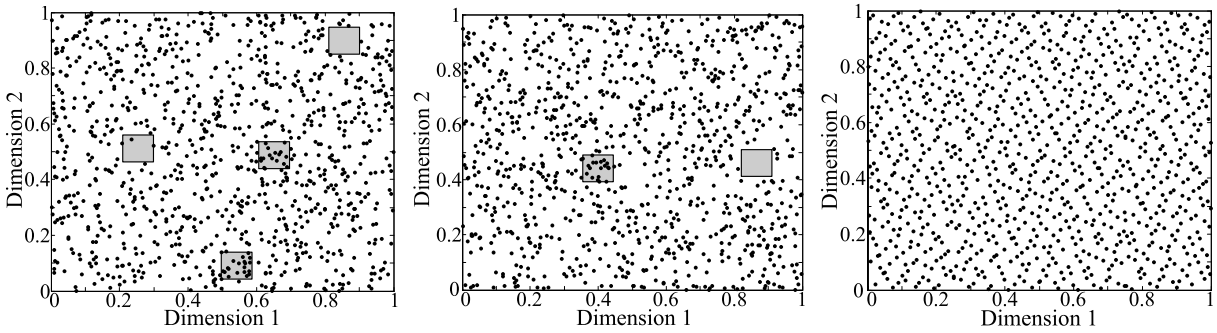
The problem of the traditional-MC is its slow convergence rate ($\sigma_{\hat{y}_\tau} = O(N^{-1/2})$), that is the standard deviation of the estimation's error declines with the inverse of the square root of the number of samples [73]. The following formulation can then be used to determine the number of samples with α-confidence half-range of $\beta(1-y)$ for a yield of y:

$$N = \frac{\left(\Phi^{-1}(0.5+\alpha/2)\right)^2}{\beta^2} \cdot \frac{y}{1-y}, \tag{4.6}$$

where $\Phi^{-1}(.)$ is the inverse of the normal Cumulative Distribution Function (CDF). It is evident that to reduce the interval range (β) by ε, the number of samples must be increased ε^2 times.

However, an important feature of the estimation error, $e = \hat{y} - y$, is that it is related to the equi-distribution (uniformity) of the samples rather than their randomness. This idea strongly suggests that by using a well-spread sequence, which is more uniformly spread than a pseudo-random sequence, a more precise estimation can be achieved [64]. LHS [61] is a sampling technique which increases the convergence rate by providing more uniform samples in 1-D. This is achieved by partitioning the domain of each random variable into equal-probable subranges and generating the same number of samples in each subrange, randomly. A random permutation of the LHS samples are finally adopted to generate the random sample vectors. However, the discrepancy of the LHS samples is not noticeably better than that of the traditional pseudo random-MC samples in projections higher than 1-D, since the permutation of the samples are performed randomly. Figure 4.3(b) shows a 2-D projection of the LHS-based samples. It can be seen that the samples are not much more uniform than the traditional MC-based samples (Fig. 4.3(a)).

Instead of the generating random samples by a pseudo-random number generator, or stratifying each dimension separately as it is done in LHS, the QMC is a technique to produce deterministic low-discrepancy sequences that are more uniformly distributed over the d-Dimensional problem space compare to the former methods. Higher than 1-D uniformity is achievable by using such sequences, that leads to a faster convergence rate than that of the MC or LHS technique. Figure 4.3(c) depicts the 2-D projection of the QMC samples, generated by the Sobol algorithm [82]. Other examples of low discrepancy sequences include Halton [89], Faur [90], and Niederreiter [91]. The error of the QMC technique is given by the Koksma-Hlawka bound, $O\left(log^d N/N\right)$, which promises an asymptotically faster than the MC performance [64]. However, this superiority seems to be unachievable unless $N > e^d$, which is absolutely impractical



(a) Traditional-MC (b) LHS (c) QMC (Sobol)

Figure 4.3: 2-D projections of different sampling approaches. The gray squares represent areas with high or low concentration of samples.

for even moderate size problems ($d > 10$). Surprisingly, the practical applications of QMC on some high-dimensional computational finance problems [80] showed significant advantages over MC's performance. The convergence rate for such problems is roughly $O\left(n^{-1}\right)$, independent of the problem dimension.

Several researches have been conducted to explain this surprisingly good performance [92, 93]. A qualitative explanation is then developed under the notation of effective dimension [65, 67]. Suppose the QMC is used to estimate the following integral:

$$\int_{[0,1]^d} f(\boldsymbol{x})\, d\boldsymbol{x}, \tag{4.7}$$

if the integrand function, $f(\boldsymbol{x})$, is decomposed into a sum of orthogonal functions of the subsets of the problem variables, and a large portion of the total integrand variance comes from a few random variable or orthogonal functions with small dimensions, then the effective dimension is significantly lower than the nominal problem dimension, leading to a high performance QMC estimation.

Consequently, by using the ANalysis Of VAriance (ANOVA) representation, the $f(\boldsymbol{x})$ can be decomposed into a sum of orthogonal functions of all the subsets of $\boldsymbol{x}$, as follows:

$$f(\boldsymbol{x}) = \sum_{u \subseteq \ell} f_u(\boldsymbol{x}) = f_0 + \sum_{i=1}^{d} f_i\left(x^{(i)}\right) + \sum\sum_{i<j} f_{ij}\left(x^{(i)}, x^{(j)}\right) + \cdots + f_{1\cdots d}\left(x^{(1)}, \cdots, x^{(d)}\right), \tag{4.8}$$

where $\ell = \{1, 2, \cdots, \mathrm{d}\}$. The ANOVA terms are orthogonal, therefore, the variance of the integrand function can be expressed as the sum of the variances of all of the orthogonal ANOVA functions, as follows:

$$\sigma^2(f) = \sum_{u \subseteq \ell} \sigma^2(f_u). \tag{4.9}$$

Caflisch, et. al. [65] introduced the notion of the effective dimensions as follows:

1. The effective dimension of f in the superposition sense is d_S, if $\sum_{|u|\leq d_S} \sigma^2(f_u) \geq P\sigma^2(f)$.
2. The effective dimension of f in the truncation sense is d_T, if $\sum_{u\subseteq\{1,2,\cdots,d_T\}} \sigma^2(f_u) \geq P\sigma^2(f)$.

where P is a proportion chosen to be less than, but close to 1. For example if $P = 99\%$ of the variance of f is due to the functions of single $\boldsymbol{x}$ variables, the effective dimension in the superposition sense, d_S, is 1. This means the interactions among the parameters have negligible effects on the function. On the other hand, the truncation sense of effective dimension is related to the list of important variables. Therefore, $d_T = m$ means that the first m variables creates the highest portion of the integrand variance. It should be noted that the variance of any MC-based estimations is directly related to the variance (error) of the integrand function. Therfore, the superiority of the QMC compared to the pseudo-random MC method for some of the high-dimensional problems arises from the low-effective dimensionality of such problems and the fact that QMC sequences produce high uniformity in the first few dimensions (≤ 12) and low order projections (≤ 3) [67, 66].

As a result, first the effectiveness of the QMC method is investigated for the analysis of the timing yield, through the analysis of the effective dimension of the yield function, $I_\tau(\boldsymbol{p})$. For this purpose, a numerical technique [94] is used to estimate the variance of the different ANOVA terms of the indicator-type yield function. This technique utilizes the quasi-regression method [95] which uses shifted Legendre polynomial functions as the bases for orthogonal ANOVA terms.

Table 4.2 lists the relative importance of the 1-D and 2-D ANOVA terms when yield is 0.5 and 0.99. The relative importances are computed using

$$\begin{aligned} 1-\mathrm{D}: &\quad 100\times\sum_{i=1}^{m}\sigma^2(f_i)/\sigma^2(f) \\ \mathrm{Full}1-\mathrm{D}: &\quad 100\times\sum_{i=1}^{d}\sigma^2(f_i)/\sigma^2(f) \\ 2-\mathrm{D}: &\quad 100\times\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\sigma^2(f_i,f_j)/\sigma^2(f) \end{aligned} \tag{4.10}$$

where m is the number of the grids of the mesh structure, while d is the total number of variables including the grids and purely random. The resultant analysis shows a reduction on 1-D terms as the yield increases, meaning that the LHS technique gives a very small improvement over the traditional-MC for a typical yield analysis close to the extreme of the critical delay distribution tail. It also suggests that to benefit more from a QMC sampling, the sampling technique should be carefully optimized to maximize the high-dimensional uniformity. These are important observations, suggesting that the excellent performance of the QMC method seen in computational

Table 4.2: The relative importance of ANOVA terms for of the yield function.

Circuit	Yield = 0.5			Yield = 0.99		
	1-D	Full 1-D	2-D	1-D	Full 1-D	2-D
C432	65	69	7	19.4	20.1	18.5
C499	64	67	7.4	15.2	15.8	19.6
C880	65	70	7	19.6	20.2	18.8
C1355	63	66	6.7	14.8	14.9	24.4
C1908	67	71	8	18.6	19	18.3
C2670	60	64	4.4	11	11.6	21.5
C3540	61	64	4.1	9	9.2	18.6
C5315	61	64	7.1	12	12.3	25.3
C6288	63	64	3.3	9.1	9.3	22.5
C7552	61	64	3.4	9.2	9.6	21.2
S9234	60	64	3.4	8.2	8.4	19.4
S13207	59	63	1.2	7.5	7.9	17.9
S15850	61	65	1.2	6.8	7.5	17.0
S35932	45	59	13	5.4	10.7	36.5
S38417	59	67	25	5.2	7.9	-5.6
S38584	56	64	29	6.1	5	23.5

finance problems may not be easily achieved for digital circuit yield analysis problem. That is due to the fact that, while the aforementioned computational finance problems are found to be effectively very low dimensional, i.e. at most 2-D with significant 1-D portions [94], the yield estimation function is not. Therefore, investigating and possibly improving the high-order discrepancies of QMC samples should be seriously considered if such a method is adopted for SSTA, particularly the yield analysis.

It should be noted that, our analysis reveals strong 1-D ANOVA terms for the mean and strong 2-D terms for the standard deviation of the critical delay as opposed to the yield which also has strong higher order terms. Therefore, both the LHS and QMC are good candidates for the mean estimation [74]. While, for the standard deviation estimation still a carefully designed QMC sampler that produces highly uniform 2-D projections should be considered. Justifying the 1-D and 2-D behavior of the mean and variance of critical delay is not hard. In fact, a promising probabilistic-based SSTA techniques, proposed in [96], approximates the critical delay with a linear additive function of the principal components of the process parameters. This approximation suggests that the critical delay function is effectively 1-D, so the first moment (mean) remains 1-D, while the second moment (variance) includes a significant set of 2-D terms due to the pairwise multiplication of the principal component factors when powering the additive circuit's delay function to two. It is now easier to realize why the yield function is composed

of one, two, and higher dimensional terms. That is because the indicator function of Eq. (4.3) consists of many higher than one degree terms, especially when yield is close to the two extremes. This can be verified by following approximation. Suppose a sigmoid function, the error function, is used to approximate the indicator function. If a Taylor expansion is used, then

$$I_\tau(h) \approx \frac{1+\mathrm{erf}(\alpha(\tau-h))}{2} = \frac{1}{2} + \frac{1}{\sqrt{\pi}}\left(\alpha(\tau-h) - \frac{\alpha^3(\tau-h)^3}{3} + \frac{\alpha^5(\tau-h)^5}{10} - \cdots\right). \tag{4.11}$$

As a result, the closer the yield is to 1 or 0 the more far $h = D(\boldsymbol{p})$) is from τ, hence, the higher order terms will be stronger. In fact, one application of LHS for analog circuits yield analysis that has been proposed in [63] showed a significant efficiency drop when the yield reaches the extremes (e.g. over 90% or below 10%) which can be well justified through the existance of high-dimensional ANOVA terms in those yield ranges. This is a critical issue since the domain of attraction in a digital circuit yield analysis problem is actually close to the high extreme.

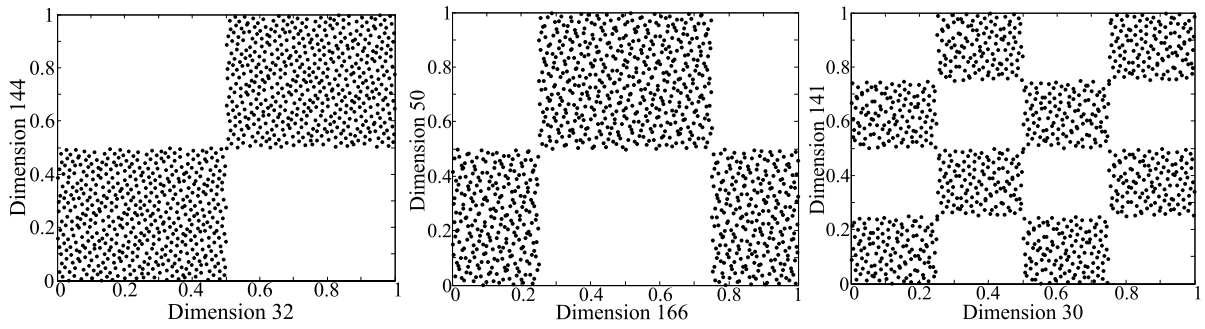
It is now easier to predict that the application of LHS alone does not provide significant improvement on the yield estimation accuracy, and the QMC sampling should be efficiently optimized to obtain the lowest discrepancy in high dimensional projections.

4.3.2 Proposed QMC/LHS -base Yield Analyzer

In this section, the discrepancy of the Sobol's QMC sequence is investigated and a method is proposed which produces low-discrepancy Sobol samples. The proposed method generates Sobol samples such that for a given number of samples, a projection uniformity increases as the dimension indices creating that projection become closer to the first dimension. Therefore, the generated samples can be finally applied to the process parameters and their principal components with an ordering procedure sorted based on the importance (criticality) of them to reduce the estimation error.

4.3.2.1 The Sobol's Sequence Generation and Discrepancy

The Sobol [82] is a low-discrepancy QMC sequence which is preferred over many other QMC sequences [89, 90, 91], especially for high-dimensional estimations, due to its higher uniformity for both 1-D and 2-D projections as a result of it prime base of two [68]. However, due to the finite number of samples, all QMC samples including the Sobol sequence, show low uniformity in many high dimensional projections, which is undesirable for an efficient digital circuit yield analysis. Note that the yield problem is found to be composed of many high dimensional ANOVA terms, so the non-uniformity of samples in each projection increases the variance of the error of that corresponding term in the ANOVA decomposition.



(a) (9,10,2)-net (b) (8,10,2)-net (c) (7,10,2)-net

Figure 4.4: Some bad pairing (high-discrepancy) of Sobol's samples.

The Sobol sequence can be represented as the (*t*,*m*,*s*)-net and (*t*,*s*)-sequence in base 2. The (*t*,*m*,*s*)-net in base 2 is a set of 2^m points in $[0,1]^s$ such that the number of points in every elementary subinterval of volume 2^{t-m} is exactly 2^t, $0 \leq t \leq m$ [97]. Based on the upper bound proposed in [97] for the discrepancy of a general (*t*,*m*,*s*)-net in base *b*, following can be derived as an upper bound of the Sobol's discrepancy:

$$D\left((t,m,s)\text{-net in base } 2\right) \leq \frac{m^{s-1}}{(s-1)!}2^t + O\left(m^{s-2}\right) \tag{4.12}$$

Figure 4.4 illustrates some of the bad 2-D projection pairings for 1024 Sobol samples. Each of the projections depicted in the figure can be considered as a sequence from (*t*,10,2)-nets of base 2, where *t* is 9, 8, and 7 respectively from fig. 4.4(a) to 4.4(c). Therefore, the lower the *t* is, the lower is the discrepancy, so the proposed Sobol sequence should target *t* as an intermediate objective function. Before discussing the discrepancy optimization approach, the general algorithm which generates Sobol samples [82] along with an approach to determine the discrepancy of Sobol samples are reviewed.

To generate $N = 2^m$ samples of a *d*-dimensional Sobol sequence, $x_i^{(j)}$, where $i = 0, \cdots, N-1$ and $j = 1, \cdots, d$, each $x_i^{(j)}$ can be generated from the following equation:

$$x_i^{(j)} = a_1 v_1^{(j)} \oplus a_2 v_2^{(j)} \oplus \cdots \oplus a_m v_m^{(j)}, \tag{4.13}$$

where $\oplus$ denotes a bitwise XOR operation, $v_k^{(j)}$ are binary direction numbers, and the $a_i \in \{0,1\}$ coefficients are extracted from the binary representation of the Gray code of *i*. The Gray code of *i* is defined as $G(i) = i \oplus \text{int}\left[\frac{i}{2}\right]$, where $\text{int}[x]$ represents the largest integer inferior or equal to *x*. For example, to find $x_{25}^{(j)}$, the following steps are taken:

$$\begin{array}{l} i = 25 \rightarrow G(i) = 11001 \oplus 01100 = 10101 \\ \text{and hence, } \; x_{25}^{(j)} = v_1^{(j)} \oplus v_3^{(j)} \oplus v_5^{(j)} \end{array}, \tag{4.14}$$

where each direction number, $v_k^{(j)}$, is a binary fraction that is written as

$$v_k^{(j)} = \nu_k^{(j)} \Big/ 2^k, \tag{4.15}$$

where $\nu_k^{(j)}$ is an odd integer, $0 < \nu_k^{(j)} < 2^k$ for $k = 1, \cdots, m$. For each dimension j, a sequence of integers $\nu_k^{(j)}$ is defined by a q-term recurrence relation as

$$\nu_i^{(j)} = \; 2b_1^{(j)}\nu_{i-1}^{(j)} \oplus 2^2 b_2^{(j)}\nu_{i-2}^{(j)} \oplus \cdots \oplus 2^{q-1}b_{q-1}^{(j)}\nu_{i-q+1}^{(j)} \oplus 2^q \nu_{i-q}^{(j)} \oplus \nu_{i-q}^{(j)} \;, \tag{4.16}$$

where $b_k^{(j)} \in \{0,1\}$, $k = 1, \cdots, q-1$ are the coefficients of a q-degree primitive polynomial [98] specified for each dimension j. Jaeckel [99] offers a collection of more than 8 million primitive polynomials up to degree $q = 27$ to be used for the Sobol generation. It is evident in Eq.(4.16) that in each dimension, there is a great deal of flexibility in choosing the initial values $(\nu_1^{(j)}, \cdots, \nu_q^{(j)})$, whereas the remaining $(\nu_{q+1}^{(j)}, \cdots, \nu_m^{(j)})$ is generated through the q-degree recurrence relation of Eq. (4.16). The constraints on the initial direction values $\nu_k^{(j)}$ for $k = 1, \cdots, q^{(j)}$ are that they must be odd integers and less than 2^k; therefore, for a dimension with a q-degree primitive polynomial, there are $2^{q(q-1)/2}$ possible choices in selecting the initial direction values. Consequently, a random technique is traditionally used to choose the initial $\nu_k^{(j)}$ terms for each dimension in [99].

By referring back to Fig. 4.4, it can be seen that to fill the empty regions and increase the uniformity of the samples, either more samples are needed or the initial direction values of the corresponding dimension should be changed. This is where the proposed technique enters to picture. As a result, the objective of this part of the work is to pick a set of initial direction values which reduces the bad pairings as much as possible. Moreover, this objective should be achieved such that the more uniform projections are generated for the first dimensions and the uniformity becomes worse as the dimension index increases. This is helpful for the fact that not all the process parameters and hence the principal components are highly critical, so a sorting of them can be considered to boost the efficiency of the method.

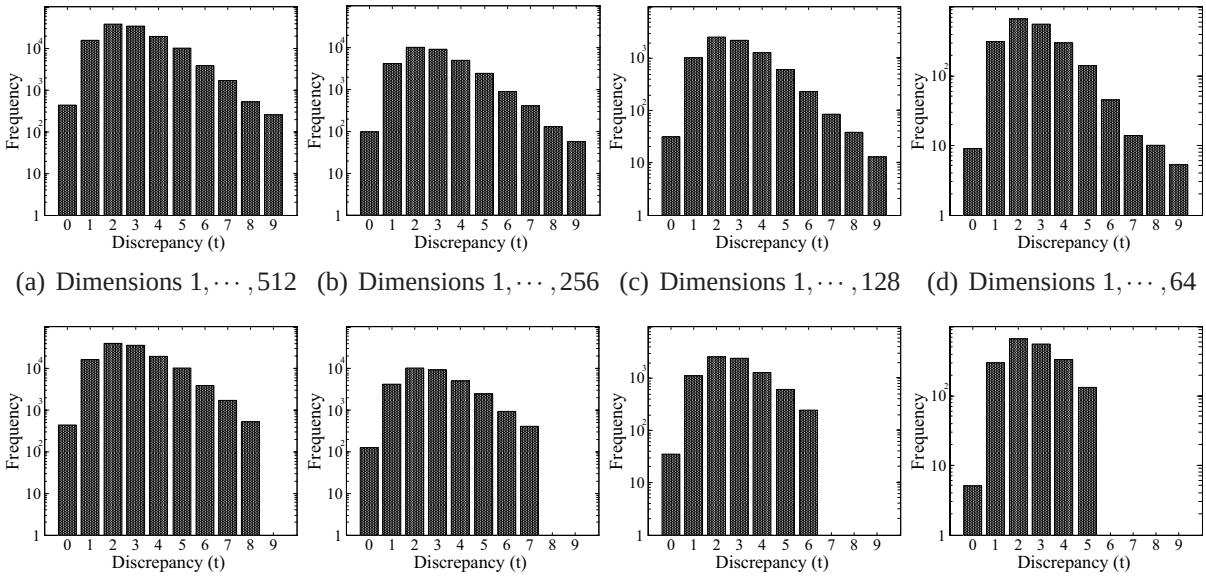
Sobol, himself, has realized the importance of the initial direction values on the quality of the generated sequences, and proposed two properties to increase the uniformity of the samples [100]. However, to satisfy Sobol's proposed properties, 2^{2d} samples are needed that is not practical for even moderate size problems. Cheng and Druzdzel have defined a measure of 2-D uniformity and proposed a search algorithm to find a set of initial direction values with a defined uniformity [101]. The drawback to their technique is that the number of samples and dimensions must be known in advance. Moreover, their technique re-produces Sobol sequences and re-evaluates their defined discrepancy measure in each iteration (after an initial direction value update), substantially increasing the runtime for large number of samples and dimensions. This was due to the assumption that poor dimension pairings cannot be found prior to the generation of sequences [68].

Algorithm 1 Optimize Initial Direction Values (m)

```
Generate random Initial Direction Values (IDV) for 2^(m-1) dimensions;
Generate upto m DVs by using the recursions for each dimension;
{Find pairwise discrepancies and initialize priorities}
for k = m − 1 downto 1 do
  for d2 = 1 to 2^(m−k) do
    for d1 = 2^(m−k−1) + 1 to 2^(m−k) do
      t(d1, d2) = t of the 2D-projection of d1 and d2;
      if t(d1, d2) > m − k − 1 then
        priority(d1) += 2^(t(d1,d2)−(m−k−1)−1);
      end if
    end for
  end for
  Initialize Temperature;
  while There is a bad pairing (any t(d1, d2) > 0) do
    while inner-loop criterion do
      Randomly select a dimension, directed by priorities;
      Randomly select an IDV, in that dimension;
      Randomly change that IDV up to k-th bit;
      Compute the new t matrix and the priority vector;
      if accept(new − old priority, Temperature) then
        Apply the changes to the selected IDV;
        Update t and priority;
        Generate upto m DVs by using the recursion;
      end if
    end while
    Update Temperature;
  end while
end for
```

However, there is no need to actually generate a Sobol sequence to detect the poor pairings or to measure the t. The t value, as a measure of the uniformity, can be found for a pair of dimension by using the definitions given in [102].

Suppose $v_{i,b}^{(j)}$ is the b-th most-significant bit of the binary representation of $v_i^{(j)}$, the i-th direction value of the j-th dimension. If for any integer d_1 and d_2 in the range of $[0,m]$ and $d_1+d_2=l$, the binary system of d_1+d_2 vectors of length m composed of $\{v_{i,b1}^{(j1)} | 1 \le i \le m, 1 \le b1 \le d_1\}$ and $\{v_{i,b2}^{(j2)} | 1 \le i \le m, 1 \le b2 \le d_2\}$ is full-rank, then the 2-D projection of dimensions $j1$ and $j2$ creates a $(m-l,m,2)$-net point set.



(a) Dimensions $1,\cdots,512$ (b) Dimensions $1,\cdots,256$ (c) Dimensions $1,\cdots,128$ (d) Dimensions $1,\cdots,64$

(e) Dimensions $1,\cdots,512$ (f) Dimensions $1,\cdots,256$ (g) Dimensions $1,\cdots,128$ (h) Dimensions $1,\cdots,64$

Figure 4.5: Distribution of t, the measure of discrepancy, for 1024 Sobol samples using (top) random initial direction values and (bottom) optimized initial direction values.

For example, if $\forall i = 1,\cdots,m \quad v_{i,1}^{(j1)} = v_{i,1}^{(j2)}$, then $l = 1$, so $t = m-1$. This means that up to the (2^m)-th sample, the projection of the $j1$-th and $j2$-th dimensions is similar to that of the Fig. 4.4(a). In other words, for all samples $0 \leq s \leq 2^m - 1$, the $x_s^{(j1)} < 0.5 \Leftrightarrow x_s^{(j2)} < 0.5$ inspired by the Eq.(4.13), as the MSB of the m first direction values of the dimension $j1$ and $j2$ are equal.

4.3.2.2 Optimization of the Sobol's Sequence Discrepancy

The timing yield analysis problem is found to be composed of high-dimensional terms. Therefore, the discrepancy of high-dimensional projections affects the estimation error. The discrepancy of a multi-dimensional Sobol sequence can be improved by careful selection of the initial direction values.

However, it is also known that all the random variables of the the yield function are inequally important. For example, the m variables representing the principal components (PC) of the grid random variables in the PCA decomposition contribute the most to the ANOVA decomposition, please refer to Table 4.2. That is due to the fact that a single PC affects to the propagation delay of many gates, hence, affecting the critical delay more strongly. Moreover, the gates with closer to zero timing slack have more chance of becoming critical leading to this conclusion that the PCs with greater contributions to those type of gates have more effects on the circuit timing than the rest.

As a result, if the direction value of the Sobol generator is set such that the lower discrepancy dimensions are obtained first and the higher discrepancy ones come later, the generated Sobol samples can be used in an ordering scheme according to the importance to the PCs for an efficient estimation.

In this part, a simulation annealing optimization algorithm (Algorithm 1) is proposed which produces such direction values. Note that, this is a relatively lengthy process, and can take a day, but it is only an one-time process. The extracted initial direction values are saved and used for future Sobol generation.

For a given m, the objective of the optimizer is to set the initial values such that the maximum t for the pairs of dimensions of $\{1,\cdots,2^{m-k}\}$ is $m-k-1$, where $k=1,\cdots,m-1$. This means, $t=0$ (perfectly uniform) for the first two dimension, $t\leq 1$ for pair of dimensions of one to four, $t\leq 2$ for pair of dimensions of one to eight, and etc. Therefore, any pairing of dimensions $d1=\{2^{m-k-1}+1,\cdots,2^{m-k}\}$ with dimensions $d2=\{1,\cdots,2^{m-k}\}$ should only be verified to satisfy $t\leq m-k-1$ condition, hence, speeding up the optimization process. To help the optimizer to even converge faster, it is only the first k bits of the initial values in the dimensions of ($\{2^{m-k-1}+1,\cdots,2^{m-k}\}$) which are included in the search during the optimization. That is because in that range the maximum t is $m-l=m-k-1$, hence, $l=k+1$, meaning that, at most, up to the k-th bit of these dimensions form the system of independent binary vectors. Moreover, the simulation annealing engine is directed by an initial value selection criterion, giving a high priority to those dimensions that have the worst discrepancies.

Figure 4.5 compares the distribution of t, the measure of discrepancy, before and after the optimization for $m=10$ (1024 samples). As depicted in Fig. 4.5(d), even for the first few dimensions, $(1,\cdots,64)$, and before the optimization, some pairs of dimensions have very high discrepancies $(t=m-1=9)$ and many others have discrepancies higher than the maximum of the optimized version for that dimension range $(t>5)$. However, as shown in Fig. 4.5(e)-4.5(h) for the optimized version, the maximum discrepancy reduces from 8 to 5 as moving down from the dimension 512 toward 64.

4.3.2.3 The Yield Analyzer

The proposed SSTA framework is constructed by combining the obtained low discrepancy Sobol sequence and the Latin Hypercube samples. A similar hybrid approach is also suggested in [74] to leverage from high uniformity of few QMC dimensions for important parameters and use of LHS for the rest.

In this research, for a given number of $N=2^m$ samples, 2^{m-1} dimensions use Sobol samples, whereas the reminder dimensions use LHS samples. The optimum initial direction values of the Sobol generator for a given m is pre-computed and stored by using the algorithm proposed earlier. Since the Sobol samples provide a higher than 1-D uniformity, they are prioritized to be assigned

for the most important PCs of the process parameters. As discussed earlier, the PCs contribute the most to the variance of critical delay, therefore, the more uniform dimensions are assigned for them. However, the LHS samples can be used to provide samples for the non-spatially correlated process parameters (e.g., RDF) or any less important PCs that are not assigned to the Sobol samples.

The number of Sobol dimensions is limited to 2^{m-1} for a given number of samples to limit $t \leq m-2$. However, approaching the first dimension, the uniformities increase. Therefore, it is beneficial to order the PCs, so that the most important components, which contribute more to the circuit's critical delay, use the lower discrepancy dimensions. Consequently, a weight is assigned for each PC as a measure of its criticality. The following is used to derive the criticality of each PC:

$$c_i = \sum_{j=1}^{p} \psi_{i,j} \sum_{k=1}^{N_j} \exp\left\{ \alpha \cdot \left(\frac{Slack_{j,k}}{D_{nom}} \right)^2 \right\}, \tag{4.17}$$

where c_i is the measure of the criticality of the i-th principal component, p is the number of PCs, $\psi_{i,j}$ is the coefficient of the j-th PC in the i-th grid variable (obtained from the PC analysis [96]), N_j is the number of logic cells in the j-th grid, $Slack_{j,k}$ is the slack of the k-th cell in the j-th grid, D_{nom} is the nominal critical delay of the circuit, and $\alpha < 0$ is a constant factor.

As a result, if a grid has many close-to-zero slack cells and/or its neighboring grids have many close-to-zero slack cells, the corresponding PC of that grid has a high criticality.

The PCs are then ordered, based on their criticalities and then assigned to the Sobol dimensions, sequentially. If there are more Sobol dimensions than PCs, the remaining Sobol dimensions are assigned to some of the non-correlated process parameters, according to a simple criticality measure for them, equal to $-1 \times slack_{cell}$. Thus, the smaller the slack of a cell is, the higher the probability that the non-correlated parameters of that cell are assigned to the Sobol samples.

4.3.3 Results

The standard deviations of the estimation errors are investigated for the benchmarck circuits. The values reported in Table 4.3 are the improvement percentage compare to the traditional-MC using the proposed method with and without using the optimized direction values. The maximum acceptable delay is set such that circuits have 95% yield. The standard deviation of the estimated yield is obtained by repeating the MC or LHS/QMC analysis 100 times using a constant number of samples (32, 128, 512, 2048) recording the yield in each run and finally calculating the standard deviation of the 100 estimated yields. Note that, in the QMC-based sampling, rerunning the original Sobol generator does not generate different sequences. Therefore, the scrambled Sobol

Table 4.3: Standard deviation reduction (percentage) of the estimated yield compared to the traditional-MC analysis. The proposed technique (QMC/LHS) is tested with and without applying the optimized direction values.

Samples	32		128		512		2048	
Opt-DV	w/o	w/	w/o	w/	w/o	w/	w/o	w/
C432	11.7	18.6	26.7	34.1	35.8	38.7	42.6	34.0
C499	12.9	28.3	13.4	35.6	21.6	39.8	27.5	47.0
C880	25.6	16.8	25.4	27.5	28.1	42.1	34.6	47.3
C1355	16.3	20.5	21.9	23.3	18.0	34.2	17.7	40.9
C1908	19.0	34.3	30.2	37.0	13.9	37.1	32.4	52.9
C2670	16.8	13.5	19.6	25.8	3.7	29.1	17.4	35.4
C3540	16.0	14.2	16.6	13.7	11.7	35.7	22.2	43.2
C5315	16.2	13.0	23.6	10.1	17.6	27.0	10.8	32.9
C6288	17.1	15.5	7.7	19.6	13.2	22.4	8.7	37.9
C7552	18.4	18.4	12.3	29.5	15.2	22.8	9.9	25.3
S9234	18.0	7.4	2.6	30.6	11.2	19.8	15.8	38.5
S13207	13.1	12.0	12.8	8.8	0.5	36.8	10.1	31.5
S15850	8.4	18.3	16.9	10.5	20.8	20.2	11.9	43.1
S35932	5.9	20.3	3.0	14.9	11.2	20.8	6.8	27.6
S38417	13.9	9.3	5.4	12.6	8.0	20.6	0.4	20.1
S38584	13.8	18.3	23.2	27.6	12.6	20.7	5.4	23.4
Average	15.2	17.4	16.3	22.5	15.3	29.2	17.1	36.3

technique [103] is used in order to generate randomized-QMC samples, so that the generated samples in each run are different and can be used to estimate the variance of the estimation's error. Scrambling adds a degree of randomization into the samples but maintains the structure of samples in terms of discrepancies. Figure 4.6 depicts the standard deviation of the error for the yield of C6288, as an example.

As listed in the table, the non-optimized direction value version shows in average an almost constant 16% improvement for different number of samples. This is because of the fact that the high-dimensional discrepency of very few random variables will be low by the non optimized technique, therefore, mostly it is the 1-D ANOVA terms that contributes to the estimation variance reduction. However, the average improvement of the standard deviation reaches up to 36% as the more low-discrepency random variables are generated by the optimized direction value Sobol sampler. Given the $O(N^{-0.5})$ convergence rate of the traditional-MC, to obtain an estimation of the yield with the same confidence interval as the proposed method, if the improvement of standard deviation is r%, then $(1-r)^{-2}\times$ samples are needed using the traditional-MC method, which translates the 36% improvement to 2.44×2048 samples to get the same accuracy by the

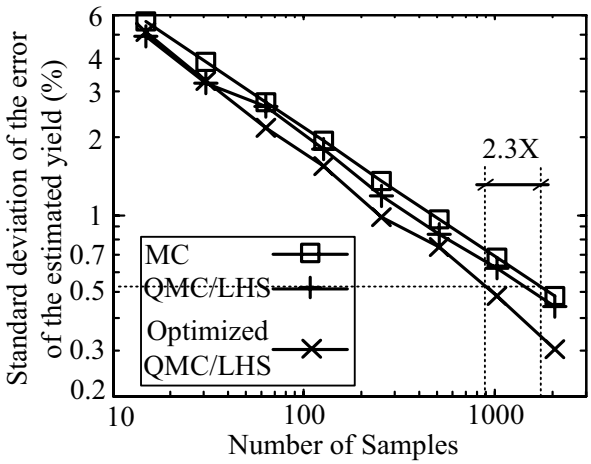


Figure 4.6: Standard deviation of the error of the estimated yield for C6288: comparison of traditional-MC, QMC/LHS method with non-optimized IDV, and proposed QMC/LHS method with optimized IDV.

traditional-MC method.

4.4 Order Statistics-based Control Variate for Yield Estimation

In this section, a timing yield estimation technique is introduced which has higher efficiency than the QMC/LHS-base method for moderate number of samples (few hundred to few thousands). The problem of QMC/LHS method is its negligible variance reduction when small number of samples is used. Moreover, due to the strong high-dimensional ANOVA terms in the yield function, the QMC/LHS is generally not very effective. The classical control variate, a variance reduction technique, is first reviewed in this section. The inefficiency of the direct application of the classical approach for yield function is also investigated through the analysis of the correlation between the actual and control variables of the yield. Finally, an order statistics-base technique is applied for the timing yield estimation using an auxiliary control-variable.

4.4.1 Control Variate and Yield Estimation Problem

Control variate is known as a promising variance reduction technique for expected value estimation when an auxiliary model (control variable) is available. There are two necessary conditions for the control variable: (i) it has to be highly correlated with the parameter under expected

value estimation and (ii) its exact expected value should be known. A rigorous exposition of the classical control variable technique is presented in [69].

As mentioned earlier, in contrast to the advanced sampling techniques (e.g. QMC and LHS), the control variate has shown promising results generally for any type of problems (even high dimensions) and for any range of samples as long as a correlated control variable with a known expected value is available. Following is a quick overview of the classical control variate method: suppose X is the random variable under expected value estimation, if C is the control variable with known expected value of μ_C then X can be substituted by X^* in computing $E[X]$:

$$X^* = X - \beta\,(C - \mu_C)\,, \tag{4.18}$$

where β is a constant. This leads to an estimation variance of:

$$\mathrm{var}\,(X^*) = \mathrm{var}\,(X) - 2\beta\mathrm{cov}\,(C,X) + \beta^2\mathrm{var}\,(C)\,. \tag{4.19}$$

Therefore, a significant variance reduction can be achieved by proper setting of β if X and C are highly correlated.

However, this classical formulation is not effective for yield estimation, if it is applied directly to the yield function of Eq. (4.3). This is because the random variable of yield problem (I_τ) is a crisp function, so no matter how much correlated is a control variable with D, the critical delay, if there is a slight shift (bias or scale) between the random variable D and the control variable model, the I_τ of D will not be highly correlated with I_τ of the control variable. For example, assume $X \equiv N(0,1)$ is a random variable under yield estimation and $C = 0.75(X-1)+0.2N(0,1)$ is a highly correlated control variable where $corr(X,C) = 0.966$. If the threshold value for yield estimation is set to 1.5, the correlation between the yield of X and C will be very low, $corr(I_{1.5}(X), I_{1.5}(C)) = 0.095$, due to the shift and scaling of the controlling variable compared to the original one. In other words, $P(X < 1.5)$ can not be well approximated (modeled) with $P(C < 1.5)$; therefore, no improvement is gained by direct application of the classical approach for the yield problem.

4.4.2 The Proposed Order Statistics-base Control Variate Method

The control variate method has been used for quantile estimation earlier in [71], and a median unbiased order statistics-base [104] estimator has been derived for it. A similar approach is used in this work but to derive the median unbiased estimator for yield.

Suppose an auxiliary control variable, C, is available which is highly correlated with the circuit's critical delay, D. Now suppose that the CDF of C is $\Phi(c)$ and exactly known. Such a control variable will be introduced later. The problem is now reduced to find a quantile point c_q such that the hypothesis test of $P(D < \tau) = P(C < c_q)$ is satisfied, then simply $y = P(D < \tau) =$

$\Phi(c_q)$ will be an estimation of yield based on the knowledge of the control variable and not the simulation of D alone. Such a point, c_q, is located where the following condition is satisfied:

$$\#\left\{i|C_i < c_q\right\} = \#\left\{i|D_i < \tau\right\}. \tag{4.20}$$

This estimator can be used instead of the natural one formulated in Eq.(4.5). For example, suppose 95 out of 100 simulations yield to a critical delay lower than the maximum acceptable delay. If using the natural estimator, Eq. (4.5), 0.95 will be reported for the yield. However, the actual yield might be greater, e.g. 0.97, but due to the random nature of selecting the simulation points, only 95 points have been placed in the the acceptable region (two points less than what it should be). However, because C is correlated with D, it is also most likely that only 95 points (two points less) reside in the acceptable region formed by ($C < c_q$). Therefore, to form such an acceptable region based on the control variables, the c_q should be determined such that the number of points where $C < c_q$ become equal to 95. Obtaining such a c_q results in an estimation of $\hat{y} = \Phi(c_q)$ which is more accurate than the natural estimator based on D alone, since the CDF of C, $\Phi\left(c_q\right)$, is exactly known and the control variable is highly correlated with the critical delay.

In practice, suppose the C and D values are ordered as: $C_1 < C_2 < \cdots < C_N$ and $D_1 < D_2 < \cdots < D_N$. If k is the largest integer such that $D_k < \tau$, then c_q can be set to any value between C_k and C_{k+1} in order to satisfy Eq. (4.20). However, if c_q is set to either C_k or C_{k+1} extremes, the estimation ($\hat{y} = \Phi(c_q)$) will be biased. Therefore, a linear interpolation is used such that the closer the τ is to the D_k the closer will be c_q to C_k. Note that if $k = N$, no simulation with delay higher than threshold (τ) is observed, then c_q is set to C_N. Following expression represents the calculation of exact c_q:

$$\begin{array}{l} \text{if}\,(D_N < \tau) \Rightarrow c_q = D_N \\ \text{if}\,(D_k < \tau < D_{k+1}) \Rightarrow c_q = \frac{(\tau - D_k)(C_{k+1} - C_k)}{D_{k+1} - D_k} + C_k \end{array}. \tag{4.21}$$

As will be seen later in the results part, the first condition ($k = N$) is a source of bias which becomes problematic as the number of samples shrinks.

The derivation of such an estimator is similar to that of the median unbiased quantile point estimator as reported in [71]. The hypothesis $\mathbf{H} : P(D < \tau) = P(C < c_q)$ is equivalent to $\mathbf{H} : P(D < \tau, C < c_q) + P(D < \tau, C > c_q) = P(D < \tau, C < c_q) + P(D > \tau, C < c_q)$ which is equal to $\mathbf{H} : P(D < \tau, C > c_q) = P(D > \tau, C < c_q)$. The test of this hypothesis is achieved by an uniformly most powerful unbiased (UMPU) test and the application of the McNemar's test [71].

Up to this point, it is assumed that an auxiliary variable, C, correlated with critical delay, D, and with known CDF is available. Such a control variable can be determined by extracting the nominal critical path of a digital circuit and deriving a linear expression for its delay with respect to process parameters.

The nominal critical path is defined as the path with the highest delay when all process random variables are set to their nominal value. Linear modeling of the path delay versus process parameters leads to a Gaussian path delay. Therefore, such a control variable has a known CDF. The expression of this control variable is extracted as follows.

Suppose $C(\boldsymbol{p})$, the control variable, is the delay of the nominal critical path which is a function of process parameters $\boldsymbol{p}$, therefore:

$$C(\boldsymbol{p}) = \sum_{i=1}^{\#gates} T^{(i)}\left(\boldsymbol{p}^{(i)}, S^{(i-1)}\right), \tag{4.22}$$

where $T^{(i)}\left(\boldsymbol{p}^{(i)}, S^{(i-1)}\right)$ is the delay of the i-th gate in the critical path as a linear function of process parameters of that gate, $\boldsymbol{p}^{(i)}$, and the signal transition time of the fan-in gate, $S^{(i-1)}$, as follows:

$$T^{(i)}\left(\boldsymbol{p}^{(i)}, S^{(i-1)}\right) = a_0^{(i)} + a_1^{(i)} S^{(i-1)} + \sum_{j=1}^{\#\boldsymbol{p}^{(i)}} a_{j+1}^{(i)} p_j^{(i)}, \tag{4.23}$$

where $a_0^{(i)}$ is the nominal delay of i-th gate, when the process variation parameters are all set to zero and input transition is zero (step function), and

$$S^{(i)}\left(\boldsymbol{p}^{(i)}, S^{(i-1)}\right) = b_0^{(i)} + b_1^{(i)} S^{(i-1)} + \sum_{j=1}^{\#\boldsymbol{p}^{(i)}} b_{j+1}^{(i)} p_j^{(i)}, \tag{4.24}$$

where $b_0^{(i)}$ is the nominal transition time of the i-th gate. The $S^{(0)}$ is set to the constant primary input transition time.

Finally, to model spatially correlated process parameters, the PCA technique adopted for SSTA in [96] is used. The j-th process variable of the i-th gate, $p_j^{(i)}$, is decomposed into a weighted linear sum of a set of independent normal random variables. As a result, C, the delay of the nominally critical path can be formed as a linear function of a set of independent Gaussian random variables. Hence, the PDF of C is Gaussian with a known mean and variance leading to a known CDF.

Some important issues should be considered here: First, the assumption of linearity is only for the control variable, not for the actually estimated critical delay, D. Second, although accounting for only the nominally critical path leads to an underestimated value, there is no problem as long as the control variable remains correlated with D, the actual delay. In fact, the D and C are highly correlated mostly because the underlaying process parameters are globally and spatially correlated. However, even if the process parameters were purely random, there would have been considerable correlation due to sharing critical gates by different paths. Table 4.4 lists the correlation factor between the control variable and the actual critical delays. The first column

Table 4.4: Correlation between the defined control variable and the actual critical delay, with and without considering gate length spatial correlations.

Circuits	Correlated	Random
C432	0.9966	0.9502
C499	0.9806	0.7364
C880	0.9936	0.9175
C1355	0.9707	0.6976
C1908	0.9997	0.9977
C2670	0.9958	0.9709
C3540	0.9984	0.9679
C5315	0.9934	0.8927
C6288	0.9989	0.9777
C7552	0.9996	0.9969
S9234	0.9878	0.8491
S13207	0.9996	0.9960
S15850	0.9964	0.9304
S35932	0.9071	0.8301
S38417	0.9922	0.9068
S38584	0.9895	0.9296

is obtained when the gate length variations are globally and spatially correlated, as modeled in Section 4.2, while the second set of correlation numbers is obtained for purely random gate length variations. As can be seen, the C499 and C1355 circuits show the lowest correlation factors in the purely random case. This is due to the structure of these two circuits where many paths have delays that are equal or very close to the delay of the nominally critical path. In fact, both C499 and C1355 circuits are 32-Bit Single-Error-Correcting circuits in which most of their paths are critical and there are very few gates with non-zero slack (Please refer to Table 4.1). A similar situation is seen for S35932 where the number of gates with low slack is very high compared to other circuits in a same range of gate count. Therefore, using only one nominally critical path to generate the control variable leads to a variable with a low correlation to the actual critical delay, since it is always possible that another path becomes critical and the model almost certainly underestimates the delay.

Finally, it is also recommended to detect different potential critical paths from different process corner analysis. Two approaches are suggested to leverage this information. Firstly, a control variable can be set to weighted sum (or average) of the critical delay of the potential critical paths obtained from limited numbers of corner analysis. This way the control variable will be kept Gaussian but may represent a larger set of critical paths. Secondly, a control variable can be set to the maximum of the limited potential critical paths. This technique models

Table 4.5: Standard deviation reduction (percentage) of the estimated yield compared to the traditional-MC analysis. The order statistics-based control variate technique is tested with and without considering spatially correlated random variables.

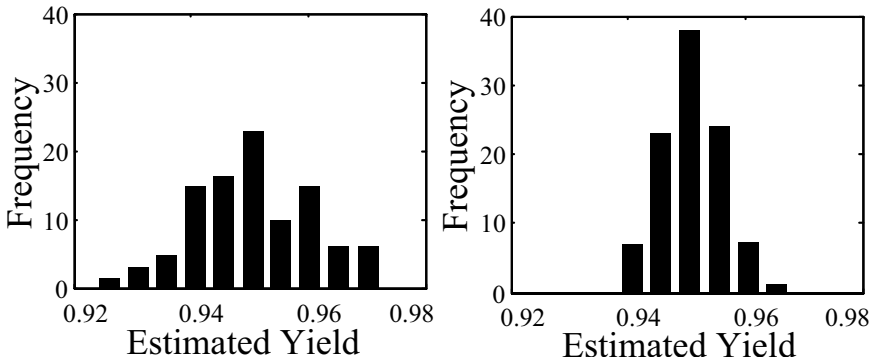
Variations	Correlated			Random		
Samples	128	512	2048	128	512	2048
C432	72.4	65.5	68.3	42.7	42.3	34.7
C499	48.6	47.7	41.6	1.9	9.8	-3.5
C880	65.1	59.4	58.9	28.7	14.7	27.0
C1355	50.2	34.5	42.8	3.9	-4.5	-0.6
C1908	86.9	80.3	76.6	78.1	69.8	62.0
C2670	70.4	65.4	66.9	44.6	43.2	42.4
C3540	80.0	76.0	71.8	53.2	48.2	55.1
C5315	62.3	58.1	56.8	29.3	24.4	16.2
C6288	82.5	73.3	74.6	41.1	42.2	37.8
C7552	87.9	81.7	77.5	74.3	65.2	63.2
S9234	45.7	47.8	51.2	1.6	-1.7	-4.7
S13207	86.7	79.2	75.1	77.1	64.8	64.2
S15850	62.3	65.2	66.0	37.4	21.6	22.7
S35932	16.1	7.2	10.1	7.8	-5.0	4.1
S38417	53.5	53.0	51.5	7.7	15.8	3.9
S38584	53.3	50.6	54.7	48.4	40.0	45.7

the actual critical delay more accurately than the weighted sum method, but the control variable will not be Gaussian anymore. As a result, finding its quantile in the order-statistics estimator requires numerical integration.

4.4.3 Results

Table 4.5 shows the percentage of the standard deviation reduction using the proposed order statistics-based method compared to the traditional-MC sampling. Similar to earlier analysis, the yield of each circuit set to 0.95. As discussed earlier, the efficiency of this approach is highly dependent to the magnitude of the correlation between the critical delay and control variable. Since such a correlation is reduced by assuming less spatially correlated process variations, the efficiency of the method is also reported in an extreme case when all process variations are purely random.

Compared to the QMC/LHS method, the advantage of this method is that the standard deviation reduction is considerable even from moderate number of samples. However, due to faster



(a) Traditional-MC (b) Order statistics-based CV

Figure 4.7: Histogram of 100 estimated yields each obtained from 512 samples of (a) traditional-MC and (b) order statistics-based control variate, for C499 circuit. The proposed method's estimation is unbiased with E[y]=0.95 but shows 48% standard deviation reduction.

convergence rate of the QMC/LHS method, it can outperform this method for some benchmark circuits (e.g. S35932) even with low number of samples.

Figure 4.7 shows 100 yield estimations each obtained from 512 samples using the traditional-MC and the order statistics base control variate for the C499 circuit. The average of both method is 0.95, but the standard deviation of the proposed method shows 48% reduction.

Finally, as discussed for Eq. (4.21), there is always a possibility with probability of y^N to detect no failure circuit out of N timing simulation. This means no linear interpolation can be used to extract c_q and its value should be set to the largest C entry, as formulated in Eq. (4.21). This is a source of biasness which increases as N, the number of samples, reduces. The bias of an estimation is the deviation of the expected value of that estimation, $\hat{y}$, from the actual yield, y. Ideally a bias of zero ($E[\hat{y}] - y = 0$) is desired, but as seen in Fig. 4.8 this is not the case for the proposed method when number of samples reduces. As can be seen, a negative bias is introduced as the number of samples is reduced, due to the possible underestimated approximation of c_q by C_N in Eq. (4.21). Moreover, for a fix N, the bias is higher for a 99% yield than that of the 95% yield since the probability of no failure (y^N) is higher.

4.5 Classical Control-Variate and Gaussian Modeling for Yield Estimation

As reported in the previous section, the order statistics-base control variate technique outperforms the optimal QMC/LHS method especially for moderate number of samples; however, it is observed that the method is prone to underestimation of the yield. This leads to a negatively

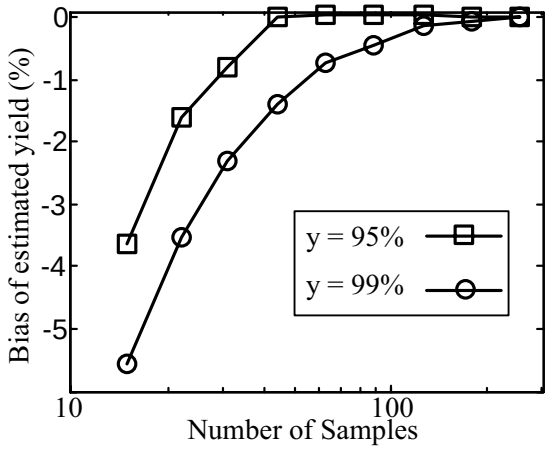


Figure 4.8: Bias of the estimated yield for C6288: comparison of 99% and 95% yield.

biased estimation, when the yield analysis is performed with low number of samples. The magnitude of the bias increases rapidly as the number of samples reduces beyond a threshold.

In this section, the Gaussian PDF is used to approximately model the probability distribution function of the critical delay. Assuming a Gaussian distribution, the mean and variance of the critical delay can be estimated by MC simulations and be used for yield estimation. This technique is suitable for large circuits and early stages of design phases when a quick estimation of yield is required with small number of samples. The hybrid QMC/LHS and the order statistics-base control variate have been shown to be inefficient for such cases due to a large error variance or unwanted negative bias.

However, there are two problems associated with such a technique. First, the variance of the error can still be very large since the low number of samples leads to less confident estimation of critical delay's mean and variance. To address this issue, the classical control variate technique will be applied to provide a highly accurate estimation of mean and variance.

The second problem of this method arises due to the error originated from the Gaussian model approximation. The Gaussian assumption may lead to an error that can not be fixed by increasing the number of samples or improving the confidence interval using a control variate technique, simply because the actual PDF of the critical delay is not exactly normal. However, a designer may live with this level of inaccuracy, and give credit to the traditional approach of yield analysis in terms of $\hat{\mu}+k\hat{\sigma}$ quantiles as long as the estimated mean and variance are highly accurate. This is especially the case for early stages of design phases where other previously reviewed MC techniques are inefficient. Other solutions would be to employ PDF models that capture the higher number of moments such as skew-normal distribution [105] or the asymptotic probability extraction method [106].

Inspired by the classical control variate equation (4.18), the following formula is used to

derive a variance reduced estimation of the critical delay mean, $\hat{\mu}_{D^*}$:

$$\hat{\mu}_{D^*} = \frac{\sum_{i=1}^{N} \left(D_i - \beta_\mu C_i\right)}{N} + \beta_\mu \mu_C, \tag{4.25}$$

where D_i is the critical delay at the i-th iteration, C_i is the control variable value at the i-th iteration, β_μ is a constant, and μ_C is the exact expected value of the control variable, C. It should be noted that, the control variable is the same as what has been used for the order-statistics method, which is the delay of the nominally critical path expressed in terms of linear gate delay equations in (4.22-4.24). Using the Eq. (4.19) and its derivative over β, the optimum β_μ which minimizes the estimated mean variance is:

$$\beta_\mu = \frac{\mathrm{cov}\,(D,C)}{\mathrm{var}\,(C)} = \frac{\rho\sigma_D}{\sigma_C}, \tag{4.26}$$

where ρ, σ_D, and σ_C are the correlation coefficients and the standard deviations of D and C, respectively.

If similar regression is used for the variance estimation, the following expression is derived for the variance reduced estimation of the critical delay variance, $\hat{\sigma}^2_{D^*}$.

$$\hat{\sigma}^2_{D^*} = \frac{\sum_{i=1}^{N} \left((D_i - \hat{\mu}_D)^2 - \beta_\sigma \left(C_i - \hat{\mu}_C\right)^2 \right)}{N-1} + \beta_\sigma \sigma^2_D, \tag{4.27}$$

where $\hat{\mu}_D = \sum_{i=1}^{N} D_i \Big/ N$ and $\hat{\mu}_C = \sum_{i=1}^{N} C_i \Big/ N$.

The term, $(N-1)$, in the denominator eliminates the bias of the variance estimation as the $E[(x-\hat{\mu}_x)^2] = \frac{N-1}{N}\sigma_x^2$ for N samples.

In order to achieve a variance reduction for the introduced critical delay variance estimator, the two variables: $(D-\hat{\mu}_D)^2$ and $(C-\hat{\mu}_C)^2$ must be correlated. The covariance between them are obtained as:

$$\mathrm{cov}\left((D-\hat{\mu}_D)^2, (C-\hat{\mu}_C)^2\right) = 2\left(\frac{N-1}{N}\right)^2 \rho^2 \sigma_C^2 \sigma_D^2, \tag{4.28}$$

if the critical delay is approximated with a Gaussian PDF. As a result, the optimum β_σ is:

$$\beta_\sigma = \frac{\rho^2 \sigma_D^2}{\sigma_C^2}, \tag{4.29}$$

since $\mathrm{var}\left((C-\hat{\mu}_C)^2\right) = 2\left(\frac{N-1}{N}\right)^2 \sigma_C^4$.

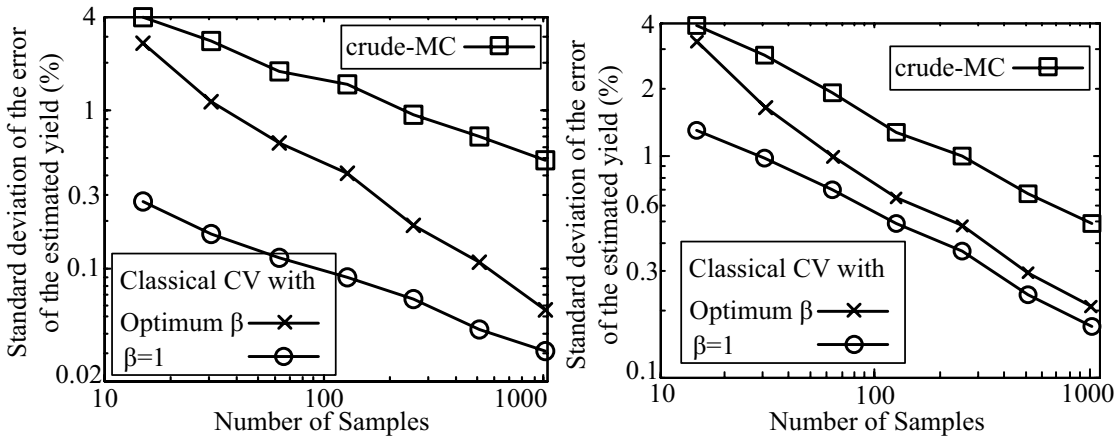
The problem of Eq. (4.26 and 4.29) is that they are functions of ρ and σ_D that are unknown themselves. One option is to use the MC simulation data to estimate ρ and σ_D and use them to calculate the optimum β factors. However, this causes a bias in the estimation of the mean and variance. Another option would be to use a portion of the samples to estimate ρ and σ_D and to obtain an approximate optimum-β, while use the rest of the samples to actually estimate the variance reduced mean and variance using Eq. (4.25 and 4.27). Fortunately, this is an unbiased method, but the variance reduction would not be as big as what is reported in Eq. (4.19) since the β itself would be a varying parameter. The variance of the estimated β increases when the number of samples reduces, which impacts the variance reduction of the estimation for very small number of samples (e.g. 10).

Looking back to the Eq. (4.19), it is evident that the estimation variance follows a quadratic function with respect to β, hence it can be concluded that by setting β to a value not too far from the optimum point, there would not be a huge estimation variance penalty. In fact, as long as $\beta < 2 \times \beta_{optimum}$ still some variance reduction is achievable (please refer to Eq. (4.26 and 4.19)). Therefore one may intuitively assume that σ_D is very close to σ_C since former is the standard deviation of the circuit's delay and the latter is that of its nominal critical path. Also, ρ can be assumed to be very close to one given the highly correlated behavior of the two random variables. Therefore, β_μ and β_σ can be roughly set to one.

Figure 4.9 compares the standard deviation of the estimated yield by using Gaussian approximation for the traditional-MC and classical control variate method. Here, by the traditional-MC, the author means the extraction of mean and variance by traditional-MC and calculating yield using a Gaussian fit. This is different from the traditional-MC method previously noted which was based on finding the expected value of the yield function as a Bernouli distribution (Eq. 4.5). Two options are considered for the β calculation in the control variable technique. In one approach, the optimum β is determined using the 1/3rd of the sample populations, and the rest (2/3rd) are used for the critical delay's mean and variance estimation. In another approach, both β_μ and β_σ are set to 1, as discussed earlier. As can be seen, the magnitude of variance reduction is lower for C1355 than that of the C6288, same as the order-statistics base method. That is due to the lower correlation between control variable and actual circuit delay in the C1355 circuit. As it was also expected, the variance reduction is higher for a fixed $\beta = 1$ than that of the estimated optimum β, but the difference gradually vanishes as the number of samples increases until the optimum *beta* outperforms the fixed value.

Please note that the reported standard deviation does not reflect the intrinsic error (bias) due to the Gaussian approximation. That error is affected by the factors which are involved in producing higher than second order moments (non-Normal terms) of the critical delay PDF (e.g. circuit graph and topology, and technology parameters).

Table 4.6 lists the standard deviation reduction using the proposed technique for fixed $\beta = 1$. The number of samples are 16. As listed the standard deviation is significantly higher compared to the two earlier methods; however, the estimation is biased due to the intrinsic error of model-



(a) C6288 (b) C1355

Figure 4.9: Standard deviation of the error of the estimated yield using Gaussian approximation: comparison of the traditional-MC and the proposed classical control variate method using optimum-β and constant-β (= 1).

ing the critical delay with a Gaussian PDF. Compared to the order-statistics-based method, two conclusions can be made. Firstly, due to the fact that a generic PDF is used in the classical method, the method has generally a lower estimation variance compared to the yield estimation by the Bernouli-based estimator. Secondly, the order-statistics-based method requires high correlation between the control variable and critical delay around the critical delay threshold value, that seems to be harder to achieve with a single nominal path.

4.6 Putting Them All Together

In this section, the proposed timing yield analysis techniques are integrated together to form a unified engine. So far, three MC-base timing analysis methods are reviewed: **i-** The low-discrepancy QMC/LHS engine: this method is not efficient for small number of samples (e.g. < 500). However the magnitude of the variance reduction achieved by this method is almost constant for different types of circuits, as it is only dependent to the relative importance of yield function ANOVA terms which is pretty close for various circuits (please refer to Table 4.2). **ii-** The order-statistics control variate engine: this method is highly biased for small number of samples, the number of samples required to disappear bias is yield-dependent (please refer to Fig. 4.8). Moreover, the magnitude of variance reduction is highly circuit topology dependent. **iii-** The classical control variate engine: this method models the critical delay with a generic distribution, so it is inherently biased but the bias never vanishes.

Since the QMC/LHS method is a robust method (circuit-independent) but not very effi-

Table 4.6: Standard deviation reduction (percentage) and bias ($100E[\hat{y}]-95$) of the estimated yield compared to the traditional-MC analysis. The classical control variate technique is tested with and without considering spatially correlated random variables.

Variations	Correlated		Random	
	Bias	Std.	Bias	Std.
C432	-0.19	91.0	0	66.9
C499	-0.22	78.6	0.19	33.8
C880	0.05	90.4	-0.04	65.1
C1355	-0.16	75.9	-0.33	21.8
C1908	-0.12	96.2	0.19	93.6
C2670	-0.11	89.8	0.20	78.9
C3540	-0.07	93.4	0.02	77.9
C5315	-0.04	88.7	0	61.5
C6288	-0.11	94.8	0.04	81.4
C7552	-0.21	96.3	0.11	91.6
S9234	-0.07	84.7	-0.33	50.5
S13207	-0.18	96.5	0.15	93.3
S15850	-0.24	88.6	-0.36	57.5
S35932	-0.32	55.4	-0.43	43.5
S38417	-0.27	84.9	0.10	59.1
S38584	-0.45	68.7	0.27	77.9

cient for low number of samples, one may combine it with the order-statistics control variate method. That means the samples should be generated using the QMC/LHS method, while the order-statistics-base control variable estimator should be utilized for the yield estimation. Consequently, the combined engine still works fine even if the circuit's topology does not produce a highly correlated control variable. The simulation results verifies the benefit of such a combination. The standard deviation of the error has been found to be slightly better than the maximum of variance reductions achieved by applying each method alone. However, the bias issue for small samples will be inherited to this engine.

In order to fix that issue, the classical control variate estimator can be first used to provide an estimate of yield. However, the sampling technique should be kept QMC/LHS so that the simulation results can later be used for the combined QMC/LHS and order-statistics control variate estimator. Therefore, suppose y is found to be the estimated yield using the Gaussian approximation. The bias of the order-statistics method is originated from the cases when no failure circuit is observed using N simulations (please refer to Eq.(4.21)). Intuitively, if the number of samples be large enough that the probability of observing no failure is less than 0.1, one may assume that the bias is negligible. As a result, such a threshold would be $N > -1/\log(y)$.

Such a threshold is 230 and 45 for 99% and 95% yields. The validity of these numbers can be verified from the Fig. 4.8. In conclusion, given an approximate estimation of *y* using the classical control variate, if simulating with more samples than this threshold is timely affordable, one may continue sampling and simualting but use the combined QMC/LHS with order statistics-base control variate estimator.

4.7 Conclusions

In this chapter, three MC-based timing yield estimation techniques for digital circuits are introduced. The major drawback of Monte-Carlo techniques is the slow convergence rate. Advanced sampling techniques and the control variate method are applied to reduce the number of simulations. Following three methods are investigated: **i-** An optimized-discrepancy QMC/LHS engine: this method provides greater variance reduction than the non-optimized QMC/LHS method. The quality of the results are almost circuit-independent but it is not significantly better than the traditional-MC especially for low number of samples (e.g. < 500). **ii-** An order-statistics based control variate engine: the number of samples reduction achieved by this method can reach to an order of magnitude, however, in circuits where there are many paths with zero slack, the saving could drop significantly. Moreover, this method is highly biased when using low number of samples, the number of samples required to disappear bias is also yield-dependent. **iii-** A classical control variate engine: this method models the critical delay with a generic distribution, so it is inherently biased, and the bias never vanishes. However, it is a good candidate for early stage timing yield estimation when the first two techniques are either inefficient or highly biased.

Chapter 5

Analog Circuits: Correlation Controlled Sampling for Efficient Variability Analysis

5.1 Introduction

Variability analysis is an important step toward design of robust analog circuits in scaled CMOS technologies. The MOS transistors are susceptible to electrical and physical parameter variations, due to ionization, chemical-mechanical polishing, and lithography variations leading to threshold voltage, oxide thickness, effective width and length mismatch of identically sized transistors. The worst-case (guard-banding) design approach does not lead to an optimum design for a tight power and area budget, therefore, statistical analysis is an essential step toward designing a robust VLSI circuit and trading-off among performance, power, noise, and accuracy [107]. Examples of analog and mixed-signal circuits, extremely vulnerable to transistor mismatches are flash ADCs [108], current steering digital-to-analog converters [109], SRAM sense amplifiers [110], ring oscillators [111], and bandgaps [112].

Mismatch analysis can be performed by either sensitivity-based [113, 114] or Monte-Carlo (MC) based methods. In the sensitivity-based methods, first, a linear model is derived for the performance metric under variability analysis, then the variance of the metric is calculated. Even though this method is fast, it requires human supervision and circuit analysis. Moreover, its accuracy is simply compromised by neglecting high-order (nonlinear) effects of the analog circuit's performance metrics. In contrast to the sensitivity-based methods, the MC simulation method is straightforward; it is easy to be employed for different circuit topologies and produces reliable results for the mismatch analysis of analog circuits. The MC method can be utilized for any form of circuit analysis, such as dc, ac, and transient with any number of process parameters as the convergence rate of the MC technique is independent of the problem dimension (number of the process and mismatch parameters). However, the negative aspect is that the MC analysis requires a large number of samples/simulations, typically thousands, to produce a reasonably accurate sta-

tistical estimation. The convergence rate of the MC sampling method is $O(N^{-1/2})$ meaning that to achieve an estimation with ε times higher accuracy, the number of samples should be increased by ε^2 times [73]. The accuracy of an estimation is defined in terms of the statistical confidence interval of an estimation.

In order to tackle the poor performance of the MC-based variability analysis method, a number of sampling and variance reduction techniques has recently been developed for various VLSI circuits, such as digital circuits, as seen in the previous chapter, and SRAM cells, as will be presented in next chapter, where they utilize Latin Hypercube Sampling (LHS) [61], low-discrepancy Quasi-MC sequences [64], control variate method [69, 71], importance [115] and adaptive sampling. Early studies of the variance reduction and sampling techniques for yield analysis of analog circuits can be found in [70, 63, 75]. However, these techniques either have a limited performance improvement or face practical concerns due to the curse of dimensionality, if they are used for the variability analysis of large-scale analog integrated circuits.

Analytical expressions are derived in Section 5.2 to examine the performance of the traditional-MC technique and answer the question of, *"how many samples are needed for a precise MC-based circuit variability analysis?"*. The analysis provide estimations of the number of samples needed for performance metric's standard deviation and yield estimations for given accuracies. It is a common practice in yield analysis to model a performance metric with a generic probability density function, e.g. Gaussian, after estimating its mean and standard deviation rather than actually estimating the yield by finding the ratio of the failed cases over the total of simulations. Equations are derived to compare the accuracy of these two methods in terms of their confidence interval. Finally, the error introduced by neglecting the skewness in a Gaussian fit is studied.

A sampling method is proposed in Section 5.3 that significantly reduces the yield estimation error (confidence interval) by minimizing the error of the mean and variance estimations in analog circuits. The sampling method generates samples with controlled linear and quadratic cross correlations that are suitable for an efficient variance and mean estimation of functions with significant linear and quadratic terms. This is a major improvement compared to the traditional-LHS and QMC methods where the estimation of variance is inefficient due to the poor uniformity of samples in two-dimensional projections. The motivating factor in the development of such a method originates from the strong presence of first and second order terms in the decomposition of the performance metric functions into functions of circuit's process parameters. In fact, the method answers the following question: *"if the performance metric functions are linear enough (or quadratic) to be analyzed by sensitivity-based methods, how can this fact be employed for generating samples that are highly efficient for mean and variance estimation?"*

Finally, the proposed method is verified by the yield analysis of an Operational Transconductance Amplifier (OTA). The developed engine is shown to be superior to the traditional-LHS in terms of the mean square error of the yield estimation.

5.2 Traditional Monte Carlo Analysis and the Required Number of Samples

In this section, the performance of the traditional MC-based mean, variance, and yield estimations are investigated. The required numbers of samples are then calculated for a certain level of estimation confidence. For the yield analysis, two types of estimators are considered: the expected value estimator based on the indicator function of Eq. (3.2) and the yield estimator based on modeling performance metric with a Gaussian distribution. Gaussian modeling of the process parameter is a common practice in yield analysis. This methods is shown to be superior to the former method in terms of the estimation variance; however, it suffers from estimation bias due to neglecting higher order moments, especially in skewed-distribution cases.

5.2.1 Estimation of the Mean

As formulated earlier in Chapter 3, Eq. (3.3), the traditional-MC method can be used to estimate the mean of the performance metric, p, in presence of process parameters $\boldsymbol{x}$. If N sets of d-dimensional samples are simulated from the $\varphi(\boldsymbol{x})$ JPDF and the following estimator is used for the mean estimation

$$\hat{\mu} = \frac{\sum_{i=1}^{N} p(\boldsymbol{x}_i)}{N}, \tag{5.1}$$

then the estimator is unbiased ($E[\hat{\mu}] = \mu$), and its variance (the variance of the estimation error) is $\mathrm{var}(\hat{\mu} - \mu) = \mathrm{var}(\hat{\mu}) = \mathrm{var}(f)/N$ [73]. For $N > 30$, the following equation can be derived based on the z-test to determine the required number of samples to achieve a half confidence interval-range of c with $(\alpha \times 100)\%$ confidence

$$N_\mu = \frac{\left(\Phi^{-1}\left(0.5 + \frac{\alpha}{2}\right)\right)^2 \sigma^2}{c^2}, \tag{5.2}$$

where σ is the standard deviation of p, and $\Phi^{-1}(.)$ is the inverse of the standard normal cumulative distribution function. It is evident that to reduce the interval range (c) by ε times, the number of samples must be increased ε^2 times. For example, 664 simulation iterations are needed to estimate the open loop dc-gain of an OTA with 99% half confidence interval range of 0.2db, if the standard deviation of the dc-gain is 2db. In other words, if the mean dc-gain is exactly 70db, after using 664 simulations, in 99% of times, the estimated mean resides within the [69.8db,70.2db] bound.

However, it should be noted that since p itself is under inspection, most likely, there is not enough prior knowledge about its standard deviation to determine the required number of samples in advance. Therefore, an estimation of the standard deviation should be updated while the simulation iterations are proceed in order to obtain an approximate number of required iterations.

5.2.2 Estimation of the Standard Deviation

The standard deviation of a performance metric may be needed to be used in modeling the performance metric's PDF to estimate the yield or it can be directly used as a design specification (e.g. the standard deviation of the input-referred offset voltage of a comparator in a flash analog-to-digital converter). The unbiased estimator of standard deviation is

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{N} \left(p\left(\boldsymbol{x}_i\right) - \hat{\mu}\right)^2}{N-1}}. \tag{5.3}$$

In order to find the confidence interval of an estimator, its distribution is needed first. In contrast to the mean estimator that converges to Gaussian due to the central limit theory, the distribution of the standard deviation estimator is a chi-distribution for a Gaussian p [116], as

$$\mathrm{PDF}\left(\hat{\sigma}\right) = \sqrt{2}\hat{\sigma}^{\mathrm{N}-2}\mathrm{e}^{-\frac{(\mathrm{N}-1)\hat{\sigma}^2}{2\sigma^2}} \frac{\left(\frac{\mathrm{N}-1}{2\sigma^2}\right)^{\left(\frac{\mathrm{N}}{2}-1\right)}}{\Gamma\left(\frac{\mathrm{N}-1}{2}\right)}, \tag{5.4}$$

where σ is the actual standard deviation of p, and Γ is the gamma function [117]. Therefore, following is the α-confidence interval of the standard deviation estimator,

$$\sqrt{\frac{2}{N-1}}\frac{\Gamma\left(\frac{N}{2}\right)}{\Gamma\left(\frac{N-1}{2}\right)}\sigma \pm \sqrt{\frac{2P^{-1}_{(N-1)/2}\left(\frac{\alpha}{2}\right)}{N-1}}\sigma, \tag{5.5}$$

where $P_K^{-1}\left(\alpha\right)$ is the solution (x) of $\alpha = P\left(K,x\right)$, and P is the regularized gamma function.

It is apparent that finding a closed-form expression which determines the required N for the standard deviation estimation with an α-confidence is not an easy task. Therefore, a simpler closed-form expression is derived as follows. Let's start with the distribution of the variance rather than the standard deviation. The exact distribution of the sample variance of a Gaussian random variable follows a chi-square distribution [118]. However, the sample variance is formed by sum of N independent and identically distributed random variables, hence it can be approximated with a Gaussian distribution if N is large enough. Please note that the "independent" condition is not exactly valid since the estimator of the variance also contains the sample mean, hence there will be a small correlation between every two random variables of $\left(p\left(\boldsymbol{x}_i\right) - \hat{\mu}\right)^2$ and $\left(p\left(\boldsymbol{x}_j\right) - \hat{\mu}\right)^2$ that are about to be added. Fortunately, such a source of error vanishes as N increases. The variance of the Gaussian approximation of the sample variance is $\mathrm{var}(\hat{\sigma}^2) = 2\sigma^4/(N-1)$ [116, 118].

Now suppose x^* is a zero-mean Gaussian random variable defined as: $x^* = (\hat{\sigma}^2 - \sigma^2)/\sigma^2$, then the standard deviation of x^* is $\sqrt{2/(N-1)}$. The sample standard deviation can now be

written as $\hat{\sigma} = \sigma\sqrt{1+x^*}$. Since N is large (e.g. $>$100) and the mean of x^* is zero, it can be assumed that $|x^*| << 1$, therefore, using only the first Taylor series term, $\hat{\sigma} = \sigma(1+0.5x^*)$. As a result, the standard deviation can also be approximated with a Gaussian distribution with mean and standard deviation of σ and $\sigma/\sqrt{2(N-1)}$, respectively. Therefore, the following equation can be derived based on the z-test to determine the required number of samples to achieve a half confidence interval-range of c with $(\alpha \times 100)\%$ confidence for the standard deviation estimation

$$N_\sigma = \frac{\left(\Phi^{-1}\left(0.5+\frac{\alpha}{2}\right)\right)^2 \sigma^2}{2c^2} + 1. \tag{5.6}$$

For example, 2065 simulations are needed to estimate the standard deviation of the input-referred offset voltage of a comparator with 99% half confidence interval of 1mV, if the actual standard deviation is 25mV. In other words, after using 2065 simulations, in 99% of times, the obtained estimation of the stanadrd deviation resides within the [24mV,26mV] bound. Same as the previous results, in this case also the standard deviation of p, σ, in unknown perior to simulations, therefore, its estimation should be used to approximate the number of iterations and finally to triger the MC analysis stop criteria condition.

5.2.3 Estimation of the Yield

The yield estimation can be either performed by the expected value estimation of a Bernoulli distribution formed by the identifying function of Eq. (3.2) or calculated by approximating the performance with a Gaussian distribution. In this section, the required number of samples of each method is extracted and a comparison between the two techniques in terms of the estimation variance and bias are given.

An unbiased estimator of yield can be formed using the estimator of Eq. (3.3) and setting g equal to the identifying function of Eq. (3.2). Therefore, the following formula can be used to determine the number of samples with α-confidence half-range of $\beta(1-y)$, for a yield of y

$$N_{Y_B} = \frac{\left(\Phi^{-1}(0.5+\frac{\alpha}{2})\right)^2}{\beta^2} \cdot \frac{y}{1-y}. \tag{5.7}$$

For example, to estimate a circuit yield of $y = 95\%$ with 99% confidence interval in the range of [93.71%,96.29%], which means $\alpha = 0.99$ and $\beta = 0.1$, 12606 samples are needed.

Although modeling the yield as a Bernoulli distribution and estimating it directly using the ratio of the number of acceptable cases over the total number of simulations is an unbiased technique, the example shows that it is a very inefficient method and requires very high number of samples for even a non-extreme quantile point (e.g. 95%).

Therefore, the efficiency of an alternative method is examined where the performance metric is modeled with a Gaussian distribution using the estimated mean and variance of the metric. To derive the required number of samples using the Gaussian modeling approach, it is assumed that p, the performance metric, can be well approximated with a normal distribution, with unknown mean and standard deviation of μ and σ. Now suppose the mean and standard deviation of p is estimated using N simulations of the MC method as $\hat{\mu}$ and $\hat{\sigma}$. As derived earlier in this section, these two estimations can be modeled with a normal distribution as follows

$$\begin{array}{l} \hat{\mu} \sim \mathcal{N}\left(\mu, \sigma_m^2\right) \\ \hat{\sigma} \sim \mathcal{N}\left(\sigma, \sigma_s^2\right) \end{array}, \tag{5.8}$$

where

$$\begin{array}{l} \sigma_m = \frac{\sigma}{\sqrt{N}} \\ \sigma_s = \frac{\sigma}{\sqrt{2(N-1)}} \end{array}. \tag{5.9}$$

Suppose $Z = \frac{\tau - \hat{\mu}}{\hat{\sigma}}$ is a random variable that produces a quantile factor at the threshold value of τ based on the Gaussian assumption for p. The yield will then be equal to $\Phi(Z) = \frac{1}{2}(1 + \text{erf}(\frac{Z}{\sqrt{2}}))$ using the standard normal cumulative distribution function.

The mean and standard deviation of Z can be derived as follows. Lets define $\sigma^* = \frac{(\hat{\sigma} - \sigma)}{\sigma}$, then $\sigma^* \sim \mathcal{N}(0, \frac{\sigma_s^2}{\sigma^2})$, and $Z = \frac{\tau - \hat{\mu}}{\sigma(1+\sigma^*)}$. However $|\sigma^*| << 1$, therefore, by using the first order Taylor approximation, $Z \approx \frac{\tau - \hat{\mu}}{\sigma}(1 - \sigma^*)$.

As a result, $\mu_Z = E[Z] \approx \frac{\tau - \mu}{\sigma}$, and

$$\begin{array}{l} \sigma_Z = \sqrt{E[Z^2] - E[Z]^2} \approx \frac{\sqrt{E\left[(\tau - \hat{\mu})^2 (1-\sigma^*)^2\right] - E[(\tau - \hat{\mu})(1-\sigma^*)]^2}}{\sigma} \\ \approx \frac{\sqrt{\sigma_m^2\left(\sigma^2 + \sigma_s^2\right) + \sigma_s^2 (\tau - \mu)^2}}{\sigma^2}. \end{array} \tag{5.10}$$

Substituting σ_m and σ_s with Eq. (5.9) results in

$$\sigma_Z \approx \frac{\sqrt{\sigma^2\left(N - \frac{1}{2}\right) + \frac{N}{2}(\tau - \mu)^2}}{\sqrt{N(N-1)}\sigma}, \tag{5.11}$$

and since $N >> 1$,

$$\sigma_Z \approx \sqrt{\frac{1 + \frac{\mu_Z^2}{2}}{N}}. \tag{5.12}$$

These values can then be used to form the Gaussian method's $(\alpha \times 100)\%$ confidence interval range, CI_G

$$\text{CI}_\text{G} = \Phi(\mu_Z + b\sigma_Z) - \Phi(\mu_Z - b\sigma_Z) = \frac{1}{2}\left(\text{erf}\left(\frac{\mu_Z + b\sigma_Z}{\sqrt{2}}\right) - \text{erf}\left(\frac{\mu_Z - b\sigma_Z}{\sqrt{2}}\right)\right), \tag{5.13}$$

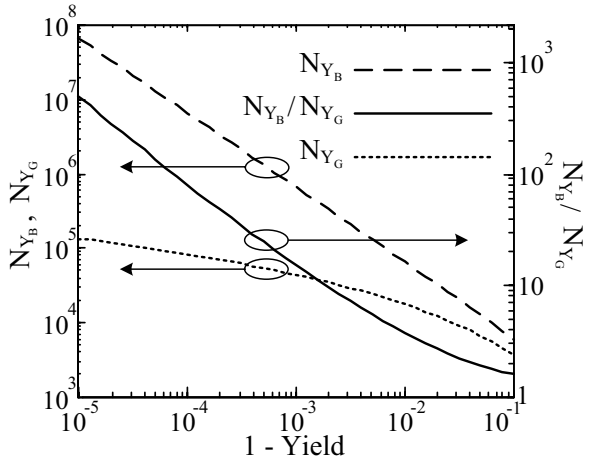


Figure 5.1: Required number of samples to obtain 99% confidence inerval range with $\beta = 0.1$. Comparison between the Bernoulli and Gaussian assumptions.

where

$$b = \Phi^{-1}\left(0.5 + \frac{\alpha}{2}\right). \tag{5.14}$$

An explicit formula for the required number of samples can then be derived by using the first derivative of the error function, $\frac{\mathrm{d}}{\mathrm{d}x}\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}e^{-x^2}$, at $x = \frac{\mu_Z}{\sqrt{2}}$, as follows

$$N_{Y_G} = \frac{\left(\Phi^{-1}\left(0.5 + \frac{\alpha}{2}\right)\right)^2 e^{-\mu_Z^2}\left(1 + \frac{\mu_Z^2}{2}\right)}{2\pi\beta^2\left(1-y\right)^2}, \tag{5.15}$$

so that an estimation of yield with α-confidence half-range of $\beta(1-y)$ is obtained. In a same scenario as the example of the Bernoulli-based estimation, Eq. (5.7), where a yield of 95% is to be estimated with $\beta = 0.1$, only 6642 simulations are needed, which is almost half of the number of samples needed for the unbiased Bernoulli-based method. Note that to evaluate Eq. (5.15), μ_Z should be substituted by $\Phi^{-1}(y)$.

Figure 5.1 depicts the required numbers of samples with respect to yield for two methods with 99% confidence interval and $\beta = 0.1$. It shows that the ratio of the samples needed for the Bernoulli estimation over that of the Gaussian approximation method increases as the yield approaches one. This is a good inspiration to use such an approximation method for rare events analysis, e.g. the SRAM cell yield analysis. However, in contrast to the unbiased Bernoulli method, the Gaussian approximation is prone to bias ($E[\hat{y}] \neq y$) especially when the actual distribution is skewed.

In summary, it can be concluded that to significantly reduce the yield estimation's confidence interval range, one may fit a generic density function to the performance metric using the esti-

mated statistical moments. However, the density function should be selected carefully to avoid an unwanted bias due to possible asymmetry of the distribution function.

5.3 The Proposed Method

The most promising non-MC based variability analysis methods is based on analyzing the linear sensitivity of performance metrics with respect to mismatch parameters and calculating the total metric variance as the sum of the square of linear coefficients. However, performing extensive circuit simulations is inevitable considering the complex secondary effects in the scaled MOS transistor characteristics to obtain accurate estimation of such sensitivity measures. Moreover, the linear models may not capture the whole variation effects of the process variations on performance metrics.

In this section, an LHS-based sampling method is proposed which improves the confidence interval of the estimation compared to the traditional LHS. This is achieved by a supervised permutation step in LHS generation. The traditional-LHS permutes the samples from disjoint intervals randomly and has no control over the permutation process. Whereas, the proposed permutation minimizes the linear and quadratic cross correlations between pairs of the random process variables. This is inspired by the presence of considerable linear and quadratic components in the decomposition of the performance metrics' functions with respect to mismatch parameters.

5.3.1 Assessing the Performance Metrics' Response Surface

If a metric can be well modeled by an additive-form function, a controlled permutation sampling method can improve the estimated yield accuracy. In this part, formulations and circuit examples are provided to justify the strong presence of linear and quadratic additive terms in the response surface of analog circuit's performance metrics.

Suppose $p(\boldsymbol{x})$ is the performance metric under statistical analysis. Let's assume a least square second-degree response surface model is constructed for p as

$$p_S(\boldsymbol{x}) = p_0 + \sum_{i=1}^{d} a_i x^{(i)} + b_i x^{(i)^2}, \tag{5.16}$$

that minimizes the error of $\int_{\mathbb{R}^d} \left(p\left(\boldsymbol{x}\right) - p_S\left(\boldsymbol{x}\right)\right)^2 \varphi\left(\boldsymbol{x}\right) d\boldsymbol{x}$ for the process parameters with the given probability density of $\varphi\left(\boldsymbol{x}\right)$. The following measure represents the ratio of the total metric's variance that has been captured by the model. In other words, it quantifies how well p is modeled by p_S.

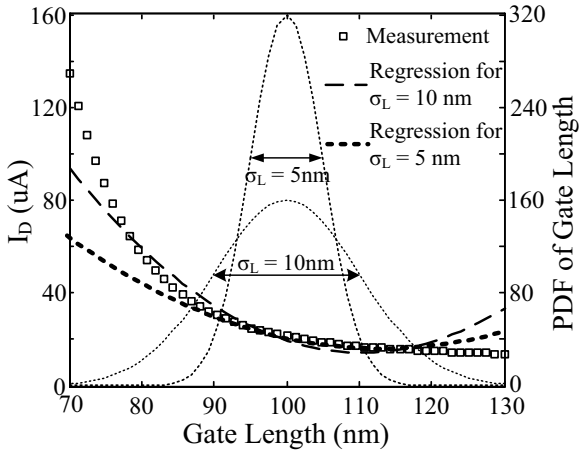


Figure 5.2: Assessing the quality of a quadratic response model for the drain current with respect to gate length variation.

$$q = \frac{\mathrm{var}\,(p_S)}{\mathrm{var}\,(p)}, \quad 0 \leq q \leq 1 \tag{5.17}$$

By considering the fact that the standard deviation of each process parameter is around or less than 10% of its nominal value [119], one may consider the second order taylor approximation of the performance metric around the nominal design point a sufficiently accurate approximation of the metric, leading to a $q \approx 1$. For illustrative purposes, the drain current of an industrial 90nm-technology NMOS is depicted with respect to its gate length variation in Fig. 5.2. Regression models are obtained for two gate length variances of $\sigma_L = 10nm$ and $\sigma_L = 5nm$. While the regression model for $\sigma_L = 10nm$ models a wider region around the nominal point, $L = 0.1\mu m$, fairly accurately, the fitting for $\sigma_L = 5nm$ has a more accurate prediction of the drain current in a shorter distance from the center point. The calculated q measures are 0.822 and 0.992, respectively.

It should be noted that when dealing with real-world large-scale circuits with several process parameters, there exists terms due to the interaction of process parameters that has not been included in this modeling scheme. Therefore, q may be practically lower than this example. However, one of the important source of variation in analog circuits is due to the mismatches of the identically-sized transistors, that imbalances the DC drain-source currents or gate-source voltages of two symmetrical transistors as depicted in Fig. 5.3. This type of current and voltage mismatches are, in fact, traditionally formulated using the first order Taylor approximation, as follows [107, 120]:

$$\begin{aligned} V_{OS,in} &= \frac{I_{DS}}{g_m}\frac{\beta_2-\beta_1}{\beta} - (V_{TH_2} - V_{TH_1}) \\ I_{OS} &= I_{DS}\frac{\beta_2-\beta_1}{\beta} - g_m\,(V_{TH_2} - V_{TH_1}) \end{aligned} \quad . \tag{5.18}$$

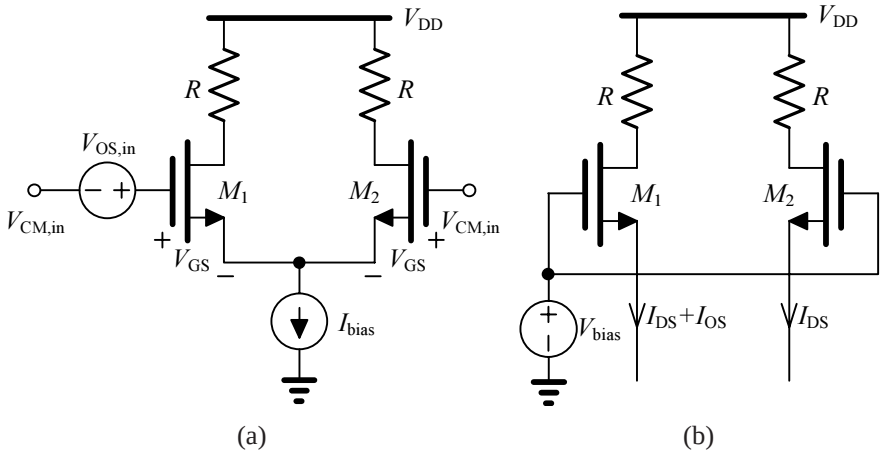


Figure 5.3: Transistor pairing arrangements and the effects of mismatches on the DC operating points: (a) current biasing: the mismatches causes the V_{GS} to vary and (b) voltage biasing: the mismatches causes the I_{DS} to vary.

Table 5.1 lists the *q*-measures of the performance metrics of three different circuits designed in a 180 nm industrial technologies. The gate length, width, threshold voltage, and the oxide thickness of all transistors are varying, and their standard deviation are set as suggested by the technology. As can be seen, the last two circuits have a highly additive performance metric. The offset voltage of a comparator follows a close-to-linear relation with respect to transistor mismatches, similar to that of the Eq. (5.18), while the period of the oscillations of the ring oscillator is just the sum of inverter's propagation delays. The propagation delay of an inverter is itself nothing but the average (weighted sum) of the low to high and high to low delays that each can be fairly well modeled with quadratic functions of pull-up and pull-down transistors drive-in mismatches. However, the OTA circuit is the case that shows lower *q* especially for gain bandwidth product and phase margin metrics where a product of two performance metrics are present.

5.3.2 Permutation Controlled LHS

As discussed in Section 3.2, Latin hypercube sampling does not provide a variance reduction when the function under analysis has major interaction terms. Therefore, it does not provide a considerable variance reduction for high order statistical moment estimations. For example, the second central moment, variance, consists of major pairwise interactions of underlying process parameters. However, as seen earlier in this section, circuits' performance metrics consist of major linear and quadratic 1-D ANOVA terms. Therefore, by using the decomposition form of Eq. (3.6) for second moment $f(\boldsymbol{x}) = p(\boldsymbol{x})^2$, the variance of the residual term, $r(\boldsymbol{x})$ will

Table 5.1: Sample circuits and their q-measures

Circuit	Metric	q-measure
Two Stage Folded Cascade OTA 24 Transistors	DC Gain	0.980
	Bandwidth	0.984
	GBW-Product	0.954
	Phase Margin	0.948
	Power	0.999
Regenerative Comparator 9 Transistors	Offset Voltage	0.998
Seven-stage inverter ring oscillator 14 Transistors	Period	0.999

mainly be due to terms in the following forms: $x^{(i)}x^{(j)}$, ${x^{(i)}}^2x^{(j)}$, and ${x^{(i)}}^2{x^{(j)}}^2$. Therefore, the new decomposition replacing Eq. (3.6) is

$$f(\boldsymbol{x}) = \mu_g + \sum_{j=1}^{d} f_j\left(x^{(j)}\right) + \sum_{i<j} c_{ij}{x^{(i)}}^2{x^{(j)}}^2 + \sum_{i\neq j} b_{ij}{x^{(i)}}^2x^{(j)} + \sum_{i<j} a_{ij}x^{(i)}x^{(j)} + r'(x) \tag{5.19}$$

The added terms are essentially the sample covariances between $(x^{(i)},x^{(j)})$, $({x^{(i)}}^2,x^{(j)})$, and $({x^{(i)}}^2,{x^{(j)}}^2)$.

In the traditional-LHS, d-dimensional vectors of uniform samples are generated by randomly permuting the order of samples in each dimension. That is, the i-th sample of the j-th process parameter, $x_i^{(j)}$ is

$$x_i^{(j)} = \frac{\pi_j(i) - U_{ij}}{N} \tag{5.20}$$

where $i = \{1,\cdots,N\}$, $j = \{1,\cdots,d\}$, $U_{ij} \sim \mathcal{U}(0,1)$ is a uniform sample, and $\pi_j(1)\cdots\pi_j(N)$ is a random permutation of integers from 1 to N. The random permutation means that there is no control on d permutations that create the s-th vector sample, $\pi_1(s)\cdots\pi_d(s)$, leading to unwanted correlation between $x^{(i)}$ and $x^{(j)}$.

In this section, a permutation algorithm is proposed that minimizes the linear and quadratic covariance between pairs of process parameters. As a result, the estimation of $E[f]$ using the controlled permutation samples filters out not only the main effect parts (1-D ANOVA terms), but also the 2-D terms due to linear and quadratic interactions. The idea of removing (reducing) the correlations of Latin hypercube samples has been previously studied in [121], which only

minimizes the linear covariances, hence, may not provide a great saving in case of performance metrics that could behave quadratically.

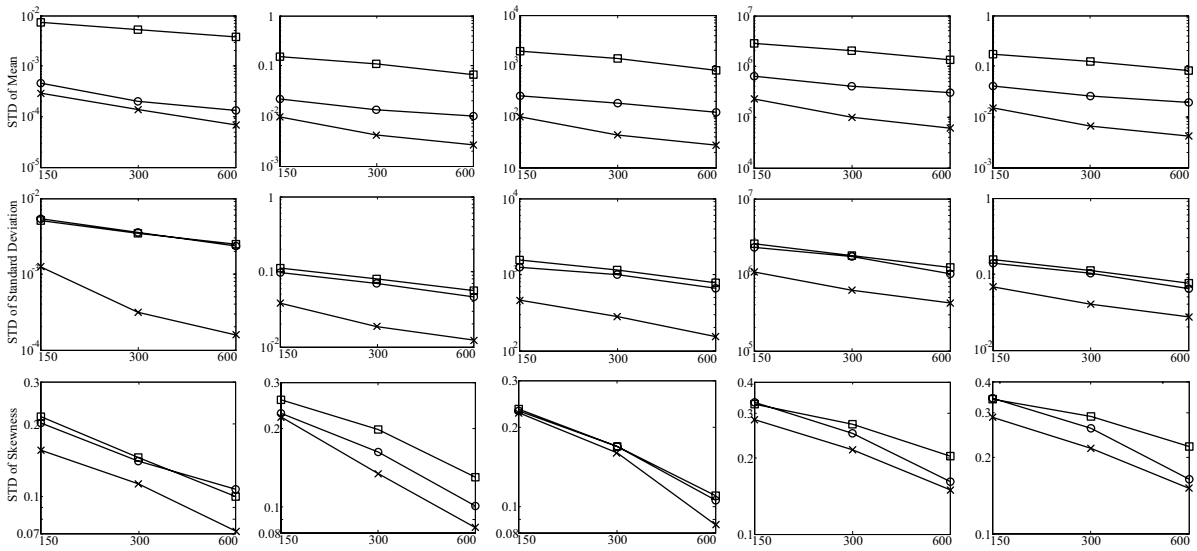
Our proposed method is a simulation annealing-based method that randomly permutes $\pi_j(i)$ values $(1,\cdots,N)$ in each column $(1,\cdots,d)$ and minimizes the following cost function:

$$cost = \alpha_1 \sum_{j<k} \text{corr}^2\left(\kappa_j,\kappa_k\right) + \alpha_2 \sum_{j\neq k} \text{corr}^2\left(\kappa_j^2,\kappa_k\right) + (1-\alpha_1-\alpha_2) \sum_{j<k} \text{corr}^2\left(\kappa_j^2,\kappa_k^2\right) \quad (5.21)$$

where $0<\alpha_1,\alpha_2<1$ are the coefficients determining the relative importance of each type of covariance, and for $j=1,\cdots,d$,

$$\kappa_j = \left[\Phi^{-1}\left(\frac{\pi_j(1)-0.5}{N}\right),\cdots,\Phi^{-1}\left(\frac{\pi_j(N)-0.5}{N}\right)\right]^T \quad (5.22)$$

The resultant $\pi_j(i)$ permutations are then used in Eq. (5.20) to create uniform random samples. The inverse of the normal cumulative distribution, Φ^{-1}, can then be used to transform the uniform samples into Gaussian. Algorithm 2 shows the pseudo-code of the method. The



(a) Power (mW) (b) DC Gain (dB) (c) Bandwidth (Hz) (d) Gain-BW Product (Hz) (e) Phase Margin (degree)

Figure 5.4: Standard deviation (STD) of the estimations of the mean, standard deviation, and skewness with respect to the number of samples (x-axis) for the OTA's performance metrics. Square: Monte Carlo, Circle: traditional LHS, X-mark: the proposed permutation controlled-LHS.

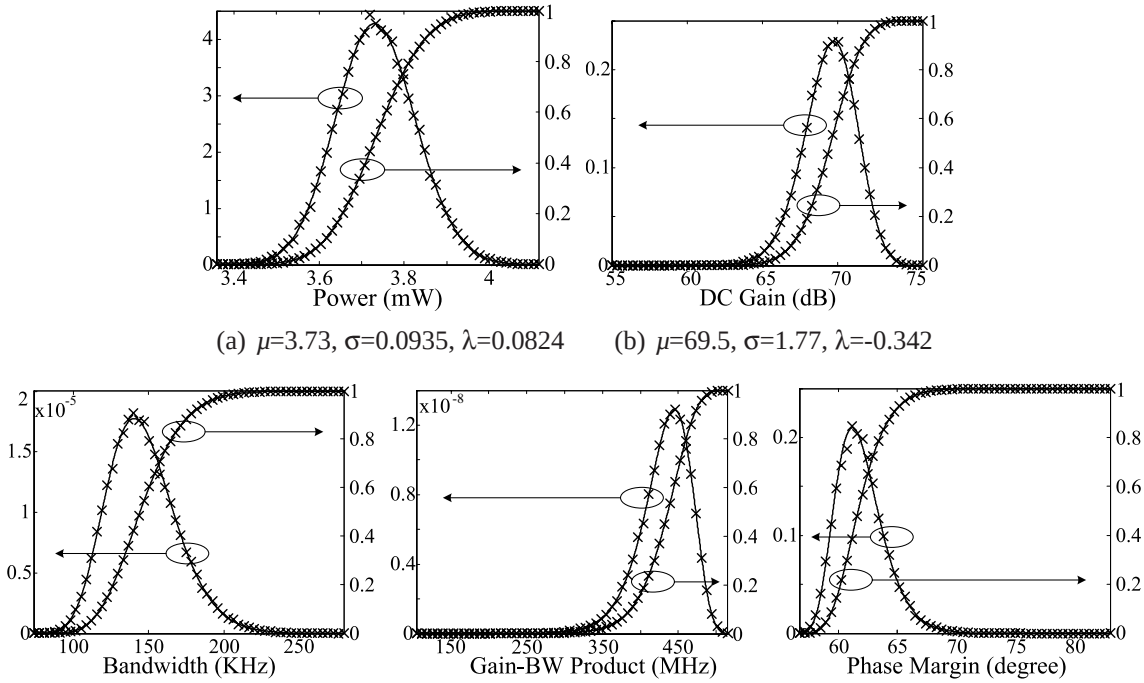
Algorithm 2 ControlledPermutationNormalLHS(N,d)

1: Generate a permutation matrix $\Pi_{N\times d}$
2: Find *cost* using Eq. (5.21 and 5.22)
3: $avgCost = cost$
4: $T = T_{init}$
5: **while** (($T > T_{final}$) OR (No *cost* improvement)) **do**
6: $sumCost = 0$
7: **for** $cnt = 1$ to $innerIter$ **do**
8: Select a column $1 \leq j \leq d$
9: Select two rows $1 \leq i1 \leq N$ and $1 \leq i2 \leq N$
10: Swap $\pi_j(i1)$ and $\pi_j(i2)$
11: Find change of the cost, $\Delta cost$
12: **if** $\left(\exp\left(-\frac{\Delta cost}{avgCost\times T}\right) < (u \sim \mathcal{U}(0,1))\right)$ **then**
13: Swap $\pi_j(i1)$ and $\pi_j(i2)$
14: **else**
15: $cost = cost + \Delta cost$
16: **end if**
17: $sumCost = sumCost + cost$
18: **end for**
19: $avgCost = sumCost/innerIter$
20: $T = T \times coolingRate$
21: **end while**
22: return $\Phi^{-1}(\frac{\Pi - (U\sim\mathcal{U}_{N\times d}(0,1))}{N})$

technique is a simulation annealing-based routine, therefore to improve its runtime, several implementation considerations should be followed. For example, the selection of a column, at line 8, must give higher priority to columns that contribute more to the cost. Also, the evaluation of the cost in each iteration should be limited to finding the change of the cost due to the swap of the two rows only in the corresponding column.

The single-threaded C implementation of the algorithm on an Intel Xeon 3GHz PC takes 17 seconds to produce 300 samples of 100-dimension with inner and outer loop count of 10000 and 500, respectively. This runtime overhead is much lower than the actual time taking to simulate a circuit of 100 process parameter-size, for 300 times on the same machine. Moreover, the simulation does not need to use these extremely large number of inner and outer iteration counts, as a significant reduction of the cost function (from 35 to 1) is achieved within the very first outer iterations (the fist 50), while the rest of the cost minimization (1 to 0.1) is achieved later.

An OTA is designed in an industrial 180 nm CMOS technology. The traditional MC, LHS, and the proposed permutation controlled LHS methods are used to estimate the first three mo-



(a) μ=3.73, σ=0.0935, λ=0.0824 (b) μ=69.5, σ=1.77, λ=-0.342

(c) μ=146, σ=23.3, λ=0.542 (d) μ=432, σ=33.9, λ=-0.85 (e) μ=61.9, σ=2.07, λ=0.806

Figure 5.5: Fitting shifted-lognormal distributions to histogram of 60000 samples of Monte Carlo simulation of the OTA. X-mark: MC histogram and cumulative distribution, Solid line: the fitted shifted lognormal PDF and CDF.

ments of its various performance metrics. The estimation procedure is repeated for 100 times, and the standard deviation of the 100 estimations are calculated for each method and compared in Figure 5.4 with respect to the number of samples.

Many important issues can be observed in the figure. It can be seen that the traditional-LHS performs significantly better than MC only for the mean estimation. That is due to the more than 1-D effective dimensionality of the standard deviation and skewness functions. It is also noticeable that the proposed method performs better than traditional-LHS even for the mean estimation, which is due to filtering out the linear and quadratic interaction terms as well as the 1-D main effect terms. It is also seen for the power consumption analysis, that the difference of the traditional and proposed LHS is not significant for the mean estimation. This is because the power consumption model has few interaction terms in it (Please refer Table I). That is why there is a significant improvement of the standard deviation estimation of the power consumption when using the proposed method compared to the traditional LHS. In contrast, a lower gain is achieved by using the proposed method for the standard deviation estimation of the gain-bandwidth and phase margin compared to the rest of the metrics due to lower q-measure. However, the magnitude of improvement is still large enough, e.g. 4X less samples are required to estimate the

standard deviation of the phase margin with the same accuracy of the traditional-LHS. Finally, it can be seen that since the ANOVA decomposition of the skewness composes of major 3-D as well as 2-D terms, the proposed method managed to reduce the estimation variance, hence the number of needed samples reduces up to 50% in some cases.

5.3.3 Finding Yield from the Statistical Moments

Inspired by the analysis in Section 5.2.3, finding the yield through modeling the performance metric with a generic distribution can lead to a lower estimation variance compared to that of the yield estimation and using the natural estimator of Eq.(3.3). However, the distribution of the underlying performance metric is not known, instead they are the statistical moments that can be estimated from the Monte-Carlo (or the controlled permutation LHS) analysis. Method of moments is an estimation technique to construct a probability distribution function by matching its first few moments with the estimations.

Although it is easy to work with, considering only the first two moments (mean and variance) and using the Gaussian distribution introduces a significant bias in skewed distributions. In fact, for skewed data, if the higher is the yield (the lower is the failure rate), the more would be the error due to ignoring the skewness. That is the reason behind the inaccuracy of modeling the SRAM failure mechanisms with Gaussian distribution as it is an extremely rare-event.

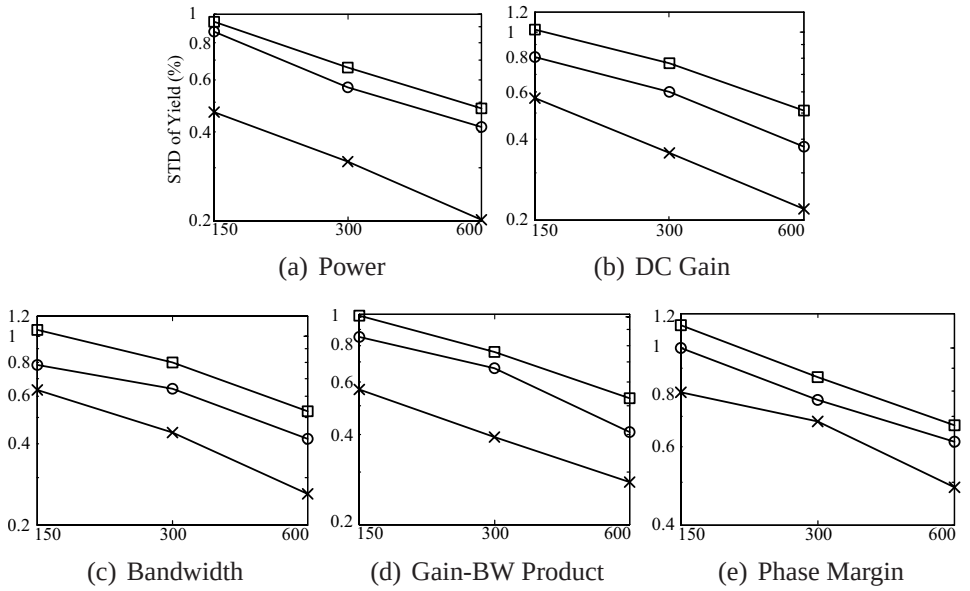
An asymptotic probability extraction methodology is proposed in [106] to construct a PDF given a number of moments. The technique is not trivial to implement and requires high number of moments (e.g. around 8 order) to produce a stable CDF function. However, the accuracy of the moment estimations decline with the order of the moment for a MC-based technique, suggesting that the estimated moments used for the matching method might be very off from its actual value, for high order moments. In addition, it is not clear how this technique can be adopted for a multivariate PDF formation, for the purpose of multi-parameter yield estimation.

In this work, a generalization of the lognormal PDF, named shifted-lognormal distribution, is used to fit the first three moments, mean, variance, and skewness into this generic PDF [122]. This a simple technique to implement, yet it can efficiently model variety of the performance metrics from highly skewed to almost-symmetrical cases (please refer to Figure 5.5). The multivariate extension of this distribution can be used to calculate the yield with respect to several performance metrics.

A single variate shifted-lognormal distribution is a three-parameter distribution, which can be generated from a normal distribution as follows:

$$Y \sim \mathcal{LN}(\alpha_1, \alpha_2, \alpha_3) = \alpha_1 + \exp\left\{X \sim \mathcal{N}\left(\alpha_2, \alpha_3^2\right)\right\} \tag{5.23}$$

Given the MC estimations of the mean, standard deviation and skewness of a design performance metric, as $\hat{\mu}$, $\hat{\sigma}$, and $\hat{\lambda}$, respectively, the three parameters of the shifted-lognormal



(a) Power (b) DC Gain

(c) Bandwidth (d) Gain-BW Product (e) Phase Margin

Figure 5.6: Standard deviation of 100 runs of yield estimation using the lognormal fitting model with respect to number of samples. P(power<3.92mW) = P(gain>65.7dB) = P(BW>106.5KHz) = P(GBW>353.4MHz) = P(PM>58.65) = 0.975. Square: Monte Carlo, Circle: traditional LHS, X-mark: the proposed permutation controlled-LHS.

distribution are

$$\begin{aligned}\alpha_3 &= \sqrt{\ln(t)} \\ \alpha_2 &= \ln\left(\frac{\hat{\sigma}}{\sqrt{e^{\alpha_3^2}-1}}\right) - \frac{\alpha_3^2}{2} \\ \alpha_1 &= \text{sign}\left(\hat{\lambda}\right)\hat{\mu} - e^{\alpha_2+\frac{\alpha_3^2}{2}},\end{aligned} \tag{5.24}$$

where t is the real root of

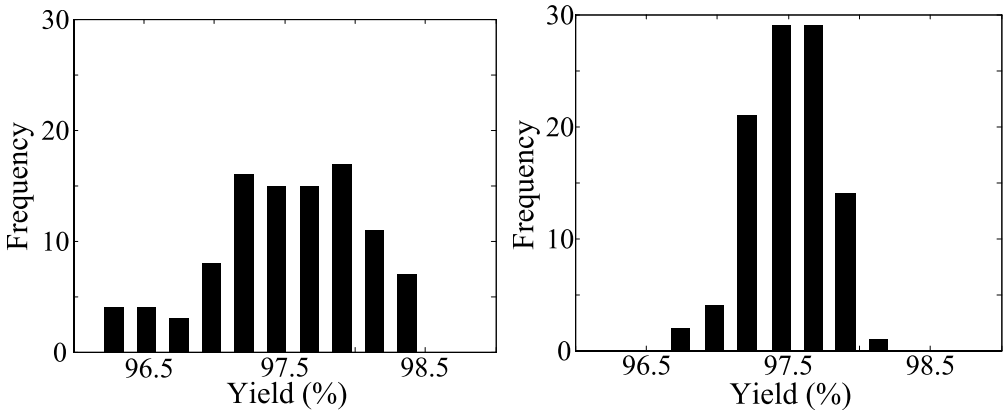
$$t^3 + 3t^2 - 4 - \hat{\lambda}^2 = 0; \quad (t > 1). \tag{5.25}$$

The resultant PDF of the process parameter will then be

$$f(y) = \frac{\exp\left\{-\frac{(\ln(y-\alpha_1)-\alpha_2)^2}{2\alpha_3^2}\right\}}{\sqrt{2\pi}\alpha_3(y-\alpha_1)}, \tag{5.26}$$

and the CDF, the yield for a given maximum value of τ is

$$P(Y<\tau) = \frac{1}{2}\left(1+\text{erf}\left(\frac{\ln(\tau-\alpha_1)-\alpha_2}{\sqrt{2}\alpha_3}\right)\right). \tag{5.27}$$



(a) MC (b) Permutation Controlled LHS

Figure 5.7: Histograms of 100 estimated gain-bandwidth yields using the MC and proposed methods for 600 samples.

Please note that this CDF model only captures a positive skewness, however, to model the negative skewness one should simply consider the negative of the performance metric, and reformulate Eq. (5.27) as

$$P(Y<\tau)=\frac{1}{2}\left(1-\text{erf}\left(\frac{\ln(-\tau-\alpha_1)-\alpha_2}{\sqrt{2}\alpha_3}\right)\right). \tag{5.28}$$

Figure 5.6 shows the standard deviation of the error obtained from 100 runs of yield estimation using the MC, traditional and proposed LHS. The shifted-lognormal model is used to fit the first three moments obtained from each method in each run. The yield with respect to each performance metric is set to 97.5%. Therefore, the standard deviation of 0.5% means that the estimated yield is within $97.5\% \pm 2.57 \times 0.5\%$ in 99% of times.

Figure 5.7 shows histograms of the 100 estimated gain-bandwidth yields using the MC and proposed method for 600 samples. The histogram confirms the reduction of estimation standard deviation, as also reported in Fig. 5.6. The standard deviations of the MC method and the proposed permutation sampling method are 0.53% and 0.275%, respectively, meaning that to gain an estimation with the same accuracy as the proposed method almost $(0.53/0.275)^2 \approx 3.7$ times samples are needed using the MC method.

Finally, for the case of multi-performance metric yield analysis, the following multivariate lognormal JPDF is suggested:

$$f(\boldsymbol{y})=\frac{\exp\left\{(\boldsymbol{x}-\boldsymbol{\alpha_2})^T\Sigma(\boldsymbol{x}-\boldsymbol{\alpha_2})\right\}}{(2\pi)^{k/2}|\Sigma|^{0.5}\prod_{i=1}^{k}x^{(i)}}, \tag{5.29}$$

where $\boldsymbol{y} = [y^{(1)}, \cdots, y^{(k)}]$ is the row vector of k performance metrics, $\boldsymbol{x} = [\ln(y^{(1)} - \alpha_1^{(1)}), \cdots, \ln(y^{(k)} - \alpha_1^{(k)})]$ is the transform to normal of $\boldsymbol{y}$, and Σ is the $k \times k$ covariance matrix of $\boldsymbol{x}$ such that $\sigma_{ii} = \alpha_3^2$ and

$$\sigma_{ij} = \ln\left(\frac{\text{cov}\left(y^{(i)}, y^{(j)}\right)}{\exp\left\{\alpha_2^{(i)} + \alpha_2^{(j)} + \frac{\alpha_3^{(i)^2} + \alpha_3^{(j)^2}}{2}\right\}} + 1\right), \tag{5.30}$$

where $\text{cov}\left(y^{(i)}, y^{(j)}\right)$, the covariance between the i-th and j-th performance metrics is estimated through the sampling based analysis.

The multi-performance metric yield can then be calculated by transforming the performance metrics into normal variables then using a numerical method [123] to evaluate the multivariate normal distribution CDF. The estimated yield is

$$P\left(Y^{(1)} < \tau_1, \cdots, Y^{(k)} < \tau_k\right) = \Phi_{(\boldsymbol{\alpha_2}, \Sigma)}\left(\ln\left(\tau_1 - \alpha_1^{(1)}\right), \cdots, \ln\left(\tau_k - \alpha_1^{(k)}\right)\right). \tag{5.31}$$

5.4 Conclusions

In this chapter, the variability (mean, standard deviation, and yield) analysis of analog circuits by the means of sampling-based methods is investigated. The formulations, derived for the required number of samples, quantify the reduction of number of samples, due to modeling yield with a generic distribution such as Gaussian. They also show how much samples are needed for sufficiently accurate standard deviation estimation. The LHS method, a more practical solution compared to other variance reduction and sampling methods, is shown to be not efficient enough for the standard deviation and yield estimations due to presence of high order terms in the ANOVA decomposition. However, because of the strong one-dimensional linear and quadratic correlations between the performance metrics of analog circuits and process parameters, a proposed permutation-controlled LHS sampling which minimizes the cross-linear and quadratic correlations is shown to be highly efficient for both the standard deviation and yield analysis of analog circuits. Finally, a multi-variate shifted log-normal distribution is used to fit simulation data with a generic JPDF that reduces the bias originated from neglecting skewness in a Gaussian fit.

Chapter 6

SRAM Cells: Adaptive Sampling for Failure Probability Estimation

6.1 Introduction

Manufacturing variability has become an issue in the design of sub-100 nm VLSI circuits and memory cells [19]. SRAM cells are designed under a very tight area constraint. Therefore, due to their scaled transistor channel area, they undergo significant random variations [124, 10, 125]. Also, in a memory block of millions of cells, the failure of only one (or few) cells may lead to chip failure. This is the most challenging element of any SRAM cell yield analysis method undermining either its accuracy or efficiency. However, to preserve sufficient variability margin yet prevent over-design, it is critical to follow a methodology which efficiently provides an accurate yield estimation during the design cycles.

The SRAM cell yield analysis has been widely studied by analytical techniques [126, 127, 128]. However, in order to analytically calculate the yield, various modeling simplifications are involved, such as, the first-order Taylor approximation of the models, trivial current-voltage modeling of MOS transistors, and finally, determining the yield through statistical Gaussian fitting of the performance metrics. Since the statistical domain of attraction in SRAM cell failure analysis is extremely far (5-6 sigma) from the mean, any minor linearization and Gaussian assumption error can introduce a significant error in the extreme quantile and yield estimations. Therefore, to perform a reliable, yet non-pessimistic stability sign-off of an SRAM cell, Spice-accurate mismatch simulations are still inevitable, despite the significant improvement of the analytical approaches.

Recently, the variance reduction Monte-Carlo (MC)-based methods, as alternatives to analytical methods, have attracted attention by addressing the shortcomings of the statistical analysis of VLSI circuits for digital and analog circuits as seen in the previous chapters. The yield estimation of SRAM cell has not been an exception in this trend [72]. The advantages of the MC-based

methods are their capability to perform Spice-accurate simulations and cut development, integration, and modeling costs. However, the most threatening disadvantage, inherent in the crude traditional MC method, is the slow convergence rate, $O(N^{-0.5})$. Therefore, the Importance Sampling (IS), a variance reduction method for rare-event statistical estimation problems, has been adopted to reduce upon the required number of iterations for SRAM cell analysis [72]. This is achieved by determining an alternative but fixed Joint Probability Distribution Function (JPDF) to simulate mismatch samples, such that faulty (important) cells are simulated more frequently than that of the crude-MC. Therefore, the mean square error of the estimation can be reduced leading to possibly more accurate results even with fewer number of simulations. However, it is not a trivial task to determine such a JPDF even for a low dimensional problem [70]. In fact, the cost of a poor selection of a JPDF can be huge and lead to a significant increase in the estimation error even worse than that of the crude-MC [129]. This risk also exists in the mixture IS (MixIS) method [72]. Its development was based on the early research of Hesterberg [130] whose proposal (MixIS) introduced an insurance against performing much worse than crude-MC by using a mixture of several PDFs. However, the cost of using a mixture, is a much worse performance improvement than that of the non-mixed IS with a good choice of an alternative JPDF [129]. Moreover, no systematic way of calculating the mixture of several PDFs is reported to guarantee a reasonable performance [72].

In Section 6.2, the SRAM yield estimation is formulated and a background on the adaptive sampling techniques are given. Then, **(a)** the behavior of SRAM cell failure mechanisms (read stability, write failure, and read access failure) is studied with respect to threshold voltages' mismatches in Section 6.3. By using the results, a general form of the multivariate Gaussian JPDF is chosen as the alternative sampling JPDF format. **(b)** Instead of fixing a multivariate Gaussian JPDF from the beginning of the simulations, an adaptive method is proposed in Section 6.4. The adaptive method manipulates (improves) the JPDF after each MC iteration by learning from the previous simulation results. The JPDF evolution is directed toward further minimization of the estimation variance by using a high-order Householder's method [131] to provide a faster convergence rate than that of the Newton's method. This process eliminates the risk associated with the IS method while provide a high performance engine. **(c)** Finally in Section 6.5, to achieve an even faster convergence, a method is proposed to analytically calculate an initial JPDF that is very close to the optimum one, instead of starting from an arbitrary one.

6.2 Background

6.2.1 Problem Formulation

As also formulated earlier in Chapter 3 and 5, suppose $\boldsymbol{x}$ is a vector of d process/mismatch parameters, and $f(\boldsymbol{x})$ is a performance metric of interest. The following indicator function, I,

divides the problem space ($\boldsymbol{x} \in \mathbb{R}^d$) into acceptable ($I = 0$) and unacceptable ($I = 1$) regions, represented as:

$$I_\tau(\boldsymbol{x}) = \begin{cases} 0 & f(\boldsymbol{x}) > \tau \\ 1 & f(\boldsymbol{x}) \leq \tau \end{cases}, \tag{6.1}$$

where τ is the threshold value of the performance metric. If $\varphi(\boldsymbol{x})$ is the JPDF of $\boldsymbol{x}$, then the following integral represents the failure probability:

$$P(I_\tau = 1) = E_\varphi[I_\tau(\boldsymbol{x})] = \int_{\mathbb{R}^d} I_\tau(\boldsymbol{x})\,\varphi(\boldsymbol{x})\,d\boldsymbol{x}. \tag{6.2}$$

The crude-MC method suggests a numerical technique to solve the integral in (6.2) by sampling from the $\varphi(\boldsymbol{x})$ distribution and extracting the mean of $I_\tau(\boldsymbol{x})$. Therefore, the required number of simulation iterations to estimate a failure rate of P with α-confidence for a half-length of βP is

$$N = \frac{\left(\Phi^{-1}(0.5+\alpha/2)\right)^2}{\beta^2} \cdot \frac{1-P}{P}, \tag{6.3}$$

where $\Phi^{-1}(.)$ is the inverse of the normal Cumulative Distribution Function (CDF). It is evident that for a rare event, where P approaches to zero, N increases inversely with P.

The problem with the crude-MC method is that most of the generated samples by the $\varphi(\boldsymbol{x})$ distribution reside in the acceptable region because the failure probability is low. Since these samples do not contribute to the calculation of the failure rate, their simulation is only a waste of runtime. As a result, if an alternative distribution, $h(\boldsymbol{x})$, is chosen to simulate the random parameters such that more failure cases are observed, the variance of the estimation error is reduced, i.e, if the integral in (6.2) is rewritten as

$$\int_{\mathbb{R}^d} \frac{I_\tau(\boldsymbol{x})\,\varphi(\boldsymbol{x})}{h(\boldsymbol{x})} h(\boldsymbol{x})\,d\boldsymbol{x} = E_h\left[\frac{I_\tau(\boldsymbol{x})\,\varphi(\boldsymbol{x})}{h(\boldsymbol{x})}\right], \tag{6.4}$$

then, by simulating the samples from the $h(\boldsymbol{x})$ distribution, the following can be used as an unbiased estimator for the failure probability:

$$\hat{P}_h = \frac{1}{N}\sum_{k=1}^{N} \frac{I_\tau\left(\boldsymbol{x}^{(k)}\right)\varphi\left(\boldsymbol{x}^{(k)}\right)}{h\left(\boldsymbol{x}^{(k)}\right)}, \tag{6.5}$$

where $\boldsymbol{x}^{(k)}$ is the k-th set of the mismatch samples. Therefore, the variance of the new estimator is

$$\mathrm{Var}\left(\hat{P}_h\right) = \frac{1}{N}\left[\int_{\mathbb{R}^d} \frac{I_\tau^2(\boldsymbol{x})\,\varphi^2(\boldsymbol{x})}{h(\boldsymbol{x})} d\boldsymbol{x} - P^2\right]. \tag{6.6}$$

If the alternative distribution is determined carefully, the given variance should be lower than the crude-MC estimator variance, $(P-P^2)/N$. This method is called the Importance Sampling (IS). Given (6.6), a zero-variance estimator is theoretically achieved, if $h(\boldsymbol{x})$ is set to:

$$h(\boldsymbol{x}) = \frac{I(\boldsymbol{x})\,\varphi(\boldsymbol{x})}{P}. \tag{6.7}$$

This fact establishes that the IS is a very promising variance reduction method. Many researchers continue to improve and adapt it for their applications [132]. By inspecting the Eq. (6.7), it is evident that $I(\boldsymbol{x})$ and P are inexplicit or unknown a priori. As a result, such a "perfect" alternative distribution is not available for a problem. However, two conclusions can be drawn here: **(a)** As seen in (6.7), to reasonably gain from any IS method, the alternative distribution should produce more samples in the spaces where both I and φ are high. In SRAM analysis, this means simulating the samples that fail the cell and have a relatively high probability in silicon realization. **(b)** It can be concluded from (6.6) that there is no guarantee that the IS always leads to a better performance, especially for multivariate cases, where a careless choice of h (fix alternative distribution) can easily lead to $h(\boldsymbol{x}) << I(\boldsymbol{x})\varphi(\boldsymbol{x})$ in some regions of $\mathbb{R}^d$. Missing or less emphasized important regions [129] can be catastrophic.

Several recent studies have been conducted to improve the IS method. The Large Deviation Theory (LDT) is used to improve the IS for rare events [133, 134]; however, the necessity of the computationally expensive condition checking and the asymptotic efficiency of the method limit LTD's empirical applications. Another approach is to use an adaptive method to improve the alternative distribution during the simulations. This, in fact, eliminates the burden of selecting a good and fixed JPDF from the beginning, and provides a mechanism to avoid poor behavior by directly focusing on the estimation variance minimization. In the next section, a review of the history of this idea is provided to create a background for the adaptive sampling method developed for the application of the SRAM yield estimation.

6.2.2 Adaptive Sampling Method

In an adaptive sampling approach, the previous simulation outputs are used to iteratively adjust the alternative distribution in order to generate more samples in the domains of interest, which is the ultimate goal of the IS. The adaptive sampling is not a new approach and its history dates back to the early 1990s where it was first introduced for structural safety analysis [135]. Lately, much has done to improve the method. In [136], two (parametric and non-parametric) adaptive techniques are proposed. The simulation runs are divided into smaller groups. Then, the alternative distribution is tuned by extracting statistical behavior of the last group's results. The problem with these methods is that if the event is extremely rare an the initial distribution does not set properly, there might be no detection of the failure event during the relatively few runs

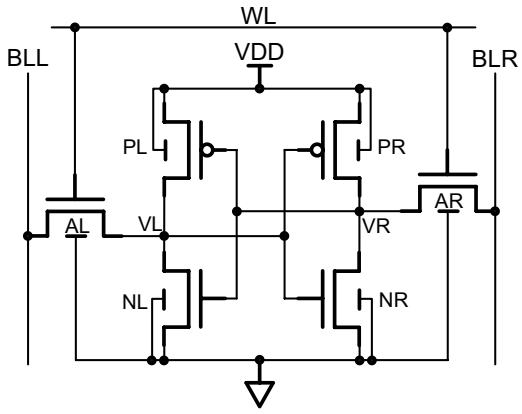


Figure 6.1: A 6T SRAM cell

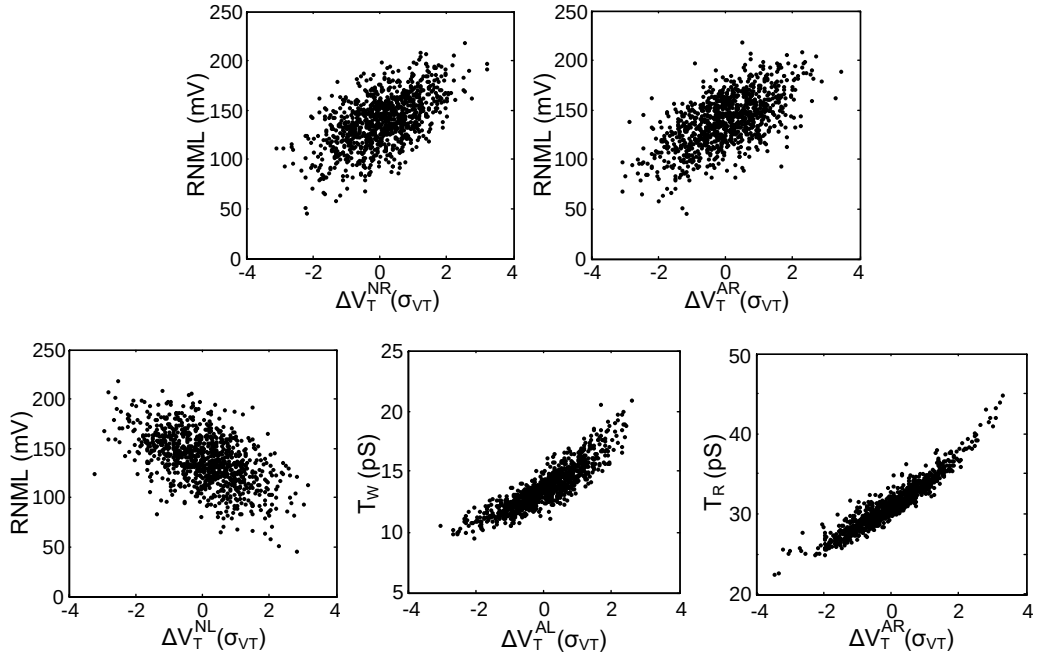


Figure 6.2: Mismatch simulation of Read Noise Margin Low (RNML), V_L zero write time (T_W), and V_R zero read time (T_R). Process parameters are normalized over their standard deviation.

in each group, leading to no improvement of the distribution. Moreover, the performance gain is limited due to the need of, at least, hundreds of runs in each group. In [137], another adaptive method is proposed that partitions the d-dimensional problem space into M^d hypercubes and performs N simulations in each of them in iteratively. Then, based on the estimated variance in each hypercube, the method continues with refining and repartitioning each partition. This approach is very expensive for even a moderate dimension problem ($d > 4$). Moreover, in a

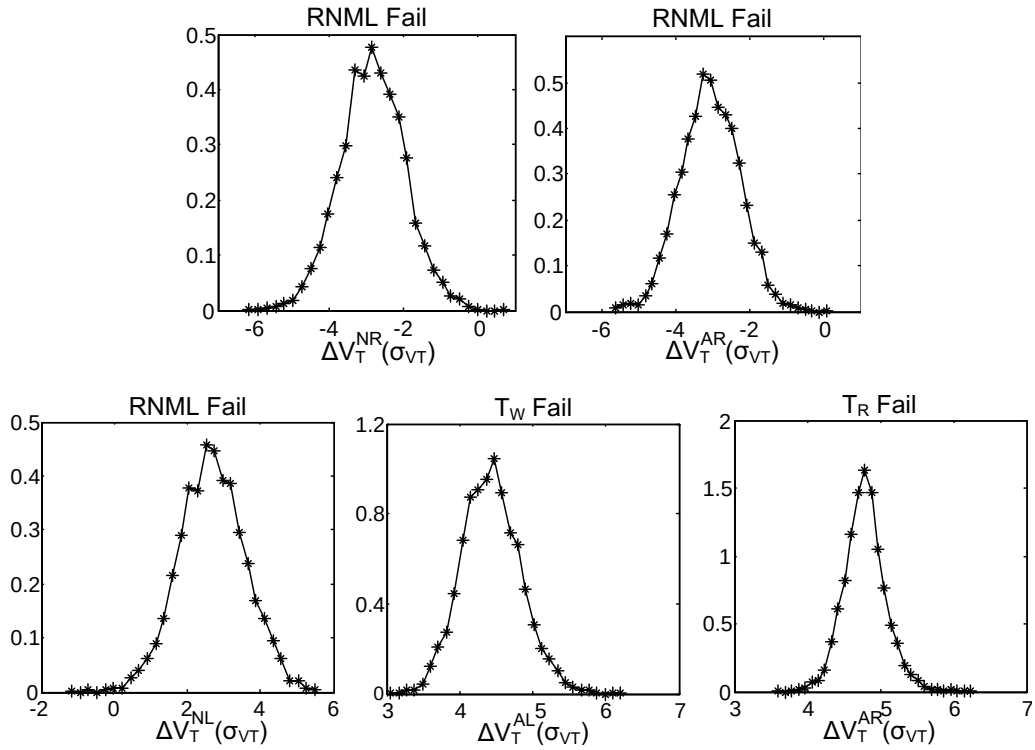


Figure 6.3: PDF of mismatch parameters for following failure conditions (RNML$<$2mV, $T_W >$30pS, and $T_R >$50pS).

yield estimation problem, where the estimated function is an identity function, the variance in most of the partitions is estimated to be zero, unless many samples are used for each partition which contradicts with the reason behind using the IS. Finally, to overcome the problem of so many runs in each group or each partition, a stochastic approximation-centric [138] method is proposed in [139]. Here, the Robbins-Monro algorithm [140] is used to direct the drift vector of a multivariate normal IS to minimize the estimation variance. However, no systematic way of selecting the coefficients of the Robbins-Monro algorithm is proposed, a definite obstacle for achieving a robust method. Moreover, this solution faces the same problem as the others for a rare-event identity-type function. This is due to the stochastic approximation of variance, typically zero, after each iteration. Consequently, no update of the drift or improvement of the sampling distribution in each iteration.

In this work, firstly, a method is proposed that updates the drifts based on a direct estimation of the variance derivatives, unlike the Robbins-Monro algorithm. This not only removes the need of Robbins-Monro's sequence coefficients settings, but also adds a degree of freedom to apply the high-order Householder's method by computing high-order derivatives, which eventually increase the convergence rate. In addition, a mechanism is proposed to address the commonly

mentioned problem of the zero estimation of the variance in rare-event identity-type functions. Secondly, an analytical framework is developed to calculate the initial close-to-optimal multivariate Gaussian distribution parameters (drifts and covariances) so that the estimations converges effectively faster.

6.3 SRAM Failure Mechanisms

Before introducing the yield estimation method, various failure types of the popular 6T-CMOS SRAM cell in Fig. 6.1 are explored. This study is conducted by extensive mismatch simulations with a 65nm industrial CMOS technology. The objective is to examine the behavior of the failure mechanisms with respect to each transistor's threshold voltage variation.

There are three sources of cell failure.

1. **Read Failure**: flipping of the cell state during the read access. This is also referred to the Read Static Noise Margin (RNM)-based failure [126].
2. **Write Failure**: inability to change the state of a cell during writing in a given time frame [124].
3. **Access Time Failure**: inability to provide enough differential voltage to saturate the sense amplifier in a given time frame during the read access [125].

A recent analytical study of these failure mechanisms suggests a strong linearity of the performance metrics with respect to the threshold variations [128]. The RNM is found to have a highly linear relation with the mismatch factors. Also, both the inverse of write time (T_W) and read time (T_R) exhibit a strong linearity. These are circuit facts that can also be verified qualitatively. For example, in the case of reading a zero-state from the right side of a cell, the saturated access transistor has a close to linear I_{DS} in relation to its threshold voltage. This leads to a linear relation between $1/T_R$ and ΔV_T^{AR}, since $T_R = C_L \Delta V / I_{DS}^{AR}$.

However, due to the simplifications that inevitably bring inaccuracies, *no* linear relation is assumed to perform the statistical analysis. In contrast, the existing linearity is exploited to establish a Spice-accurate adaptive MC method which works over a drifted (non-zero mean) multivariate normal distribution with a non-identity covariance matrix.

In this section, the reason for choosing a general drifted normal distribution for the alternative and adaptive distribution is demonstrated. In fact, the contents of this section provide only visual and quantitative justifications, while the corresponding mathematical analysis is given in Section 6.5.

Figure 6.2 depicts the actual mismatch simulation of three performance metrics, RNML, T_W, and T_R, to illustrate their behavior in relation to some of the transistors' threshold variations. Since the remaining mismatch parameters have almost no effect on the corresponding performance metric, they have not been plotted. It is evident that, positive and negative linear cross-correlations, among pairs of performance metrics and mismatch parameters, exist.

Now, refer to the conclusion derived from Eq. (6.7). It is stated that in order to reduce the estimation variance by using the IS, the alternative distribution should generate more samples in the failure region. As a result, by examining Fig. 6.2(a), to observe more RNML failure cases (e.g. RNML$<$2mV) the alternative distribution should simulate ΔV_T^{NR} with high negative values. This is also the case for ΔV_T^{AR}, but opposite for ΔV_T^{NL} which are in agreement with circuit analysis. Similar observations can be made for the $T_W > T_W^{max}$ and $T_R > T_R^{max}$ failure regions.

Therefore, if a properly set alternative non-zero mean distribution is used to generate mismatch samples, there is a higher chance of capturing more failure samples. However, it is critical to remember that not each overly-drifted distribution, which creates many failure samples, is necessarily a good candidate. By looking at Eq. (6.7), the condition for gaining from an alternative distribution is that the generated samples should have a relatively high probability in reality (or large $\varphi(\boldsymbol{x})$). The trade-off in drifting the distribution is to reach a point, where not only are many failure cases observed, but also they have the highest probability in the actual silicon realization.

Figure 6.3 depicts the empirical distribution, obtained by performing extensive (tens of millions) MC simulations, and extracting only the failure cases. Figures 6.3(a) and 6.3(b) demonstrate that the distributions of ΔV_T^{NR} and ΔV_T^{AR} are negatively drifted for the RNML failure cases, that is in agreement with the positive correlation, plotted in the Monte Carlo graphs in Fig. 6.2(a) and 6.2(b) suggesting the need for negative delta-mismatches in order to obtain a low RNML. However, to reduce the RNML, the ΔV_T^{NL} should be increased which is confirmed in Fig. 6.3(c). Note that the drift magnitude seems to be proportional to the correlation. For example, ΔV_T^{AL} and ΔV_T^{AR} show very high drifts in the simulations of T_W and T_R (Fig. 6.3(d),6.3(e)) because they are highly correlated to the two mismatch parameters (Fig. 6.2(d),6.2(e)).

Besides the drifts, the variance of the normalized delta-mismatches of the failure cases slightly deviates from 1, according to Fig. 6.3. It is also evident that the higher the correlation between the performance metric and the mismatch parameter, the lower the failure distribution's variance.

Moreover, the correlation between the failed mismatch parameters are also portrayed in Fig. 6.4. As seen in Fig. 6.4(a), ΔV_T^{NR} and ΔV_T^{AR} are negatively correlated, and ΔV_T^{NR} and ΔV_T^{NL} are positively correlated. Due to the fact that, if in a case, ΔV_T^{NR} is largely negative, there is a good chance that the RNML failure occurs even with a large positive ΔV_T^{AR} or a large negative ΔV_T^{NR}. Note that Fig. 6.4(a) does not depict the correlation between the actual mismatch parameters, since they can have a very small or no correlation, which is the case for Random Dopant Fluctuations. Figure 6.4(a) shows the correlation between the delta mismatches that produce failed SRAM cells. For analysis related to these observations refer to Section 6.5.

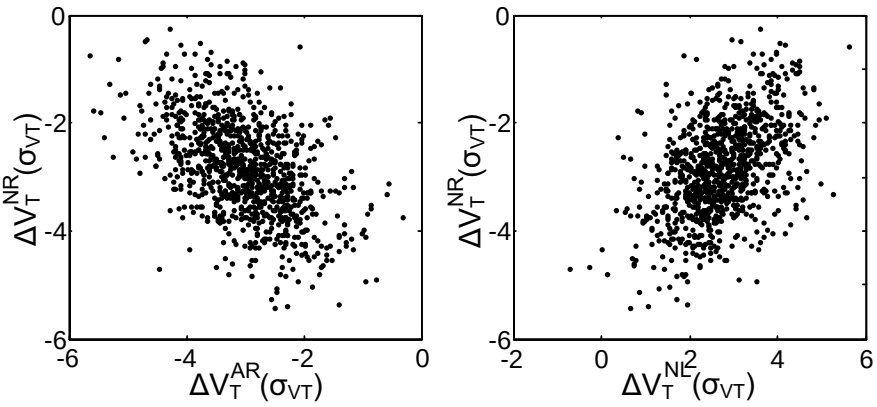


Figure 6.4: Positive and negative cross-correlation among the failure (RNML$<$2mV) mismatch parameters.

It is finally implied that a drifted multivariate normal distribution with a non-identity covariance matrix results in a fairly good choice for an alternative distribution to mimic the SRAM failure region. However, no prior knowledge of the magnitude of the drift and the covariance matrix is available up to this point. Note that over-drifting or a poor covariance formation can lead to a performance worse than that of the crude-MC method. The next two sections provide establish the foundation to adaptively and analytically achieve the optimum distribution parameters to build an efficient MC method.

6.4 Adaptive Multivariate Normal Sampling

6.4.1 The Algorithm

In this section, an adaptive method is developed to iteratively update the drifts of a multi-variate normal distribution with any arbitrary covariance matrix. The drift-updating process is directed toward minimizing the estimation variance. In contrast to the Robbins-Monro-based method, reported in [139], the derivatives of the estimation's variance are estimated directly, so that no risk is associated with a poor selection of the Robbins-Monro's sequences. This also adds the flexibility to apply high-order Householder's method to further increase the convergence rate. Lastly, a mechanism is proposed to address the challenge of variance estimation for rare event identity-type functions, in dealing with SRAM yield estimation (refer to Section 6.2).

The interest is in estimating the following probability:

$$P = P\left(I_\tau = 1\right) = E_\varphi\left[I_\tau\left(\boldsymbol{x}\right)\right] = E_h\left[\frac{I_\tau\left(\boldsymbol{x}\right)\varphi\left(\boldsymbol{x}\right)}{h\left(\boldsymbol{x}\right)}\right]. \tag{6.8}$$

The variance of the estimation determines the confidence interval for a given number of iterations, but the focus is on finding an alternative distribution function, $h(\boldsymbol{x})$, which provides a lower estimation variance. Therefore, the problem is formulated as

$$\arg\min_{h} \mathrm{var}_h\left(\frac{I_\tau(\boldsymbol{x})\varphi(\boldsymbol{x})}{h(\boldsymbol{x})}\right) = \arg\min_{h} E_h\left[\frac{I_\tau(\boldsymbol{x})\varphi^2(\boldsymbol{x})}{h^2(\boldsymbol{x})}\right] = \arg\min_{h} E_\varphi\left[\frac{I_\tau(\boldsymbol{x})\varphi(\boldsymbol{x})}{h(\boldsymbol{x})}\right]. \tag{6.9}$$

Suppose the drifted and correlated multivariate normal distribution is chosen as the alternative distribution for the following vector of six normalized mismatch parameters: $\boldsymbol{x} = \left[\Delta V_T^{PR}, \Delta V_T^{NR}, \Delta V_T^{AR}, \Delta V_T^{PL},\right.$ so

$$h(\boldsymbol{x},\boldsymbol{\mu},\Sigma) = \frac{1}{(2\pi)^3|\Sigma|^{1/2}} e^{\frac{\sum_{i=1}^{6}\sum_{j=1}^{6} C_{ij}\left(x_j-\mu_j\right)(x_i-\mu_i)}{-2|\Sigma|}}, \tag{6.10}$$

where Σ is the arbitrary covariance matrix, C_{ij} are the covariance matrix's cofactors, and $\boldsymbol{\mu} = [\mu_1, \mu_2, \cdots, \mu_6]$ represents the drift vector.

$$\begin{aligned}
\frac{\partial E_\varphi\left[\frac{I_\tau(\boldsymbol{x})\varphi(\boldsymbol{x})}{h(\boldsymbol{x},\boldsymbol{\mu},\Sigma)}\right]}{\partial \mu_l} &= E_h\left[\frac{I_\tau(\boldsymbol{x})}{2} e^{\frac{\sum_{i=1}^{6}\sum_{j=1}^{6} C_{ij}\left(x_j-\mu_j\right)(x_i-\mu_i)}{|\Sigma|}} - \sum_{i=1}^{6} x_i^2 \left(\sum_{i=1}^{6}\left(C_{il}+C_{li}\right)\left(\mu_i - x_i\right)\right)\right] \\
\frac{\partial^2 E_\varphi\left[\frac{I_\tau(\boldsymbol{x})\varphi(\boldsymbol{x})}{h(\boldsymbol{x},\boldsymbol{\mu},\Sigma)}\right]}{\partial \mu_l^2} &= E_h\left[\frac{I_\tau(\boldsymbol{x})}{2} e^{\frac{\sum_{i=1}^{6}\sum_{j=1}^{6} C_{ij}\left(x_j-\mu_j\right)(x_i-\mu_i)}{|\Sigma|}} - \sum_{i=1}^{6} x_i^2 \left(\frac{\left(\sum_{i=1}^{6}\left(C_{il}+C_{li}\right)(\mu_i-x_i)\right)^2}{2|\Sigma|} + 2C_{ll}\right)\right]
\end{aligned} \tag{6.11}$$

Now, suppose $\boldsymbol{\mu}^{(k)}$ is the drift used to generate the samples at the k-th iteration, and is updated after each iteration. Even though the sampling is no longer identically distributed, due to the independent sampling property, Eq.(3.19) is still valid in producing an unbiased estimator, as follows:

$$\hat{P} = \frac{1}{N}\sum_{k=1}^{N} \frac{I_\tau\left(\boldsymbol{x}^{(k)}\right)\varphi\left(\boldsymbol{x}^{(k)}\right)}{h\left(\boldsymbol{x}^{(k)},\boldsymbol{\mu}^{(k)},\Sigma\right)}. \tag{6.12}$$

Without the loss of generality, assume a zero-drift and identity covariance JPDF for the original mismatch parameters. Therefore, the following equation can be derived for the weight function:

$$\frac{\varphi(\boldsymbol{x})}{h\left(\boldsymbol{x},\boldsymbol{\mu}^{(k)},\Sigma\right)} = |\Sigma|^{1/2} e^{\frac{\sum_{i=1}^{6}\sum_{j=1}^{6} C_{ij}\left(x_j-\mu_j^{(k)}\right)\left(x_i-\mu_i^{(k)}\right)}{2|\Sigma|} - \frac{\sum_{i=1}^{6} x_i^2}{2}}. \tag{6.13}$$

To numerically solve Eq.(6.9) and find the optimum drifts, one should find a solution for:

$$\frac{\partial E_{\varphi}\left[\frac{I_{\tau}(\boldsymbol{x})\varphi(\boldsymbol{x})}{h(\boldsymbol{x},\boldsymbol{\mu},\Sigma)}\right]}{\partial \boldsymbol{\mu}} = 0, \tag{6.14}$$

and, by using Newton's method,

$$\boldsymbol{\mu}^{(k+1)} = \boldsymbol{\mu}^{(k)} - \frac{\frac{\partial}{\partial \boldsymbol{\mu}^{(k)}} E_{\varphi}\left[\frac{I_{\tau}(\boldsymbol{x})\varphi(\boldsymbol{x})}{h\left(\boldsymbol{x},\boldsymbol{\mu}^{(k)},\Sigma\right)}\right]}{\frac{\partial^2}{\partial \boldsymbol{\mu}^{(k)2}} E_{\varphi}\left[\frac{I_{\tau}(\boldsymbol{x})\varphi(\boldsymbol{x})}{h\left(\boldsymbol{x},\boldsymbol{\mu}^{(k)},\Sigma\right)}\right]}, \tag{6.15}$$

where, for $l = \{1,2,\cdots,6\}$, the first and second derivatives are derived in terms of $E_h[.]$ as Eq.(6.11).

Since the covariance matrix is positive definite, the diagonal elements of the cofactor matrix (C_{ll}) are all positive; therefore, the second derivative is positive definite. This implies that the root of the first derivative resides on the global minimum of the estimation's variance.

At first look, by the naive substitution of the derivatives in Eq.(6.15), it can be seen that to update $\boldsymbol{\mu}^{(k)}$, two expected values need to be estimated such that each of them requires several circuit simulations (notice the presence of $I_{\tau}(\boldsymbol{x})$ in both derivatives). However, a solution is adopted here from the concepts, introduced in the stochastic approximation field [138], i.e., instead of an accurate estimate for the expected value in the first derivative (numerator) by performing several simulation, only the last simulation is used to provide a rough estimate. The progress of this process generates the same effect as the averaging needed at the first point for the expected value estimation of the numerator.

Moreover, to estimate the denominator, a set of (e.g., 100) last circuit simulation results are used. This approach leads to a biased estimation of the second derivate, since each of the previous simulations is performed by using a different drift ($\boldsymbol{\mu}$). However, the unbiasness of the coefficient is not a requirement in Newton's method and might impact only the effective convergence rate. Note that since the second derivative is positive definite, the first derivate is a non-decreasing function. Therefore, if the first derivate of Eq.(6.11) is modeled by a non-decreasing function $g(\boldsymbol{\mu}^{(k)},\Sigma)$, the Newton's method of Eq.(6.15) can be substituted with a simpler but less efficient one as follows: $\boldsymbol{\mu}^{(k+1)} = \boldsymbol{\mu}^{(k)} - \varepsilon g\left(\boldsymbol{\mu}^{(k)},\Sigma\right)$, where $\varepsilon > 0$ has a fixed (biased), but sufficiently small value [138]. In fact, the proposed method replaces the fixed ε with an approximation to $1\Big/\frac{\partial}{\partial \boldsymbol{\mu}^{(k)}} g\left(\boldsymbol{\mu}^{(k)},\Sigma\right)$. This is accomplished by averaging the last few simulation results, and hence, is more efficient than the fixed ε form.

Since the estimates of the derivatives are used directly and the high-order derivatives exist, the Householder's method [131] is used as an alternative to Newton's method to further improve the

convergence rate. For example, the third order Householder's method increases the convergence rate from the second order to the fourth order, unlike the Newton's method. In this case, the following equation should be used instead of the denominator in Eq.(6.15):

$$\left(\frac{6g'^3 - 6gg'g'' + g^2g'''}{6g'^2 - 3gg''}\right)\left(\boldsymbol{\mu}^{(k)}, \Sigma\right), \tag{6.16}$$

where $g\left(\boldsymbol{\mu}^{(k)}, \Sigma\right)$ is the first derivative, reported in Eq.(6.11). Therefore, g' is equal to the second derivate derived in Eq.(6.11). It should be emphasized that the same strategy is adopted again in estimating this alternative denominator by using the last few simulation results.

Up to this point, the problem, associated with the rare and identity-type functions, has not been addressed. The problem occurs with an arbitrary $\boldsymbol{\mu}^{(0)}$. It is very likely that most of the samples reside in the acceptable region, $I_\tau(\boldsymbol{x}) = 0$. That is due to the rare nature of the failure event. As a result, the rough estimate of the first derivate (the numerator in Eq.(6.15)), which is calculated according to the last simulation result, is zero in most cases. This causes no change in the drift, $\boldsymbol{\mu}$, as the simulations proceed. It is even more problematic if the denominator equals zero or become very small due to the low possibility of $I_\tau(\boldsymbol{x}) = 1$. To overcome these issues, instead of the actual threshold value (τ), a secondary fake one, T, is used for the purpose of the derivative estimation only. The value of T is determined by the mean and standard deviation of the last few, (e.g., 100) performance metrics such that a considerable portion (e.g., 20%) of the simulations are considered as failures. However, to estimate the yield itself by using Eq.(6.12), the original τ is used, so no error exists in the estimation itself. It should be noted that T is only an intermediary parameter in the calculations to form a factor and to determine the amgnitute of drift after each iteration.

Algorithm 3 presents the pseudo code of the proposed method. The method starts with an initial drift, $\boldsymbol{\mu}$ and a covariance matrix, Σ. *FCnt* is the number of the last performance metrics that are used to estimate the fake threshold, T. *DenCnt* is the number of simulation results, required to estimate the expected value of the denominator in Eq.(6.16). Lines 9-20 establish the value of the fake threshold, T. The factor of -0.5 in line 17 affects the fraction of the simulations that resides in the fake-failure region. Line 27 constructs a simple form of the weight function, Eq.(6.13), that is used in the derivative and yield calculations. Lines 28-30 compute the yield, based on Eq.(6.12) and (6.13). The first four derivatives of $E_h[.]$ are calculated in lines 32-35 and used to form the denominator in line 36, according to Eq.(6.16). Finally, the average of the last *DenCnt* estimations of the denominator is used to find the new drift. Note that, in computing the first derivative, the last estimation results are used to compute $w \times g$, instead of the expected value. However, by accounting for only the last sample to find the expected value, a very noisy estimation of the expected value is produced. This can result in a large drift change, casing a convergence problem. Therefore, the experimental factor of 0.01 is used in lines 38 and 40 to avoid large changes of the drift. Observe that increasing this factor improves the convergence speed but mitigates the robustness. The choice of 0.01 is small enough that the robustness is

Algorithm 3 Yield = EstimateYield(τ,μ,Σ,*IterCnt*,*FCnt*,*DenCnt*)

Require: Σ is positive definite.

C = Matrix cofactor of Σ
lastF = Vector allocation to save last *FCnt* performance metrics
lastDen = Matrix allocation to save last *DenCnt* denominators (6 cols)
DenIdx = 0
FailureProb = 0
for *iter*=0 **to** *IterCnt*-1 **do**
  x = Generate a vector of 6 Gaussian samples from N(μ,Σ)
  f = Simulate circuit with x and return the performance metric
  lastF(*iter* **mod** *FCnt*) = f
  if *iter*<*FCnt* **then**
    meanLastF = Average of {*lastF*(0), $\cdots$, *lastF*(iter)}
    stdLastF = Standard deviation of {*lastF*(0), $\cdots$, *lastF*(iter)}
  else
    meanLastF = Average of {*lastF*(0), $\cdots$, *lastF*(*FCnt*-1)}
    stdLastF = Standard deviation of {*lastF*(0), $\cdots$, *lastF*(*FCnt*-1)}
  end if
  T = *meanLastF* $-0.5\times$ *stdLastF*
  if $T < \tau$ **then**
    $T = \tau$
  end if
  if $f < T$ **then**
    I = 1
  else
    I = 0
  end if
  if $I == 1$ **then**
    $w = \exp(\sum_{i=1}^{6}\sum_{j=1}^{6} C_{ij}(x_i-\mu_i)(x_j-\mu_j)/|\Sigma| - \sum_{i=1}^{6} x_i^2)$
    if $f < \tau$ **then**
      FailureProb = *FailureProb* + $\sqrt{w\times|\Sigma|}/$*IterCnt*
    end if
    for l=1 **to** 6 **do**
      $g = 0.5\times\sum_{i=1}^{6}(C_{il}+C_{li})(\mu_i-x_i)$
      $g' = g^2/|\Sigma| + 2C_{ll}$
      $g'' = g^3/|\Sigma|^2 + 3gC_{ll}/|\Sigma|$
      $g''' = g^4/|\Sigma|^3 + 6g^2C_{ll}/|\Sigma|^2 + 3C_{ll}^2/|\Sigma|$
      lastDen(*DenIdx* **mod** *DenCnt* , l) = $w\times\frac{(6g'^3-6gg'g''+g^2g''')}{(6g'^2-3gg'')}$
      if *DenIdx* < *DenCnt* **then**
        $nmu_l = \mu_l - \frac{0.01\times g\times w}{\text{Average }\{lastDen(0,l),\cdots,lastDen(DenIdx,l)\}}$
      else
        $nmu_l = \mu_l - \frac{0.01\times g\times w}{\text{Average }\{lastDen(0,l),\cdots,lastDen(DenCnt\text{-}1,l)\}}$
      end if
    end for
    μ = *nmu*
    DenIdx = *DenIdx* + 1
  end if
end for
return 1 - *FailureProb*

found to be not an issue in the extensive tests described in this work. Other methods, such as ignoring the large change drifts, can also be applied to eliminate the sudden drift changes.

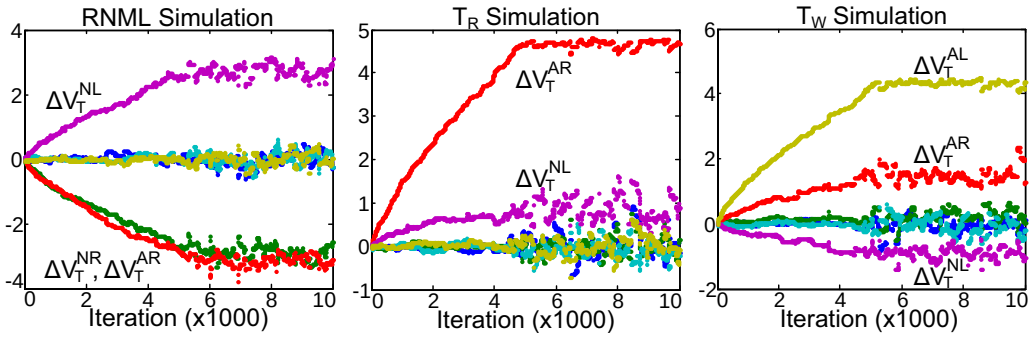


Figure 6.5: Adaptive updates of the alternative distribution's drifts.

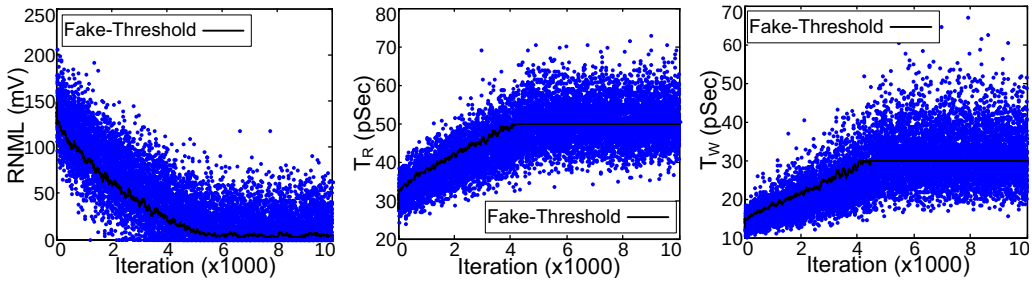


Figure 6.6: Performance metrics and the fake-thresholds.

6.4.2 Results

By naively applying the zero-drift as the initial $\boldsymbol{\mu}$, and using the identity covariance matrix, the algorithm is run for the following three specifications (RNML$<$2mV, $T_W >$30pS, and $T_R >$50pS). Figure 6.5 shows how the drifts are altered, when the algorithm is run for 10,000 simulations. $\boldsymbol{\mu}$s are drifted along the direction that is expected by the circuit analysis and simulations. For example, for T_W and T_R, the $\mu_{\Delta V_T^{AR}}$ and $\mu_{\Delta V_T^{AL}}$ increase as the simulations proceed. The magnitudes of the final drifts match the mean points of the PDFs, depicted in Fig. 6.3. These drifts increase the chance of failure, reducing the estimation variance. That is verified in Fig. 6.6, where the performance metric values are moving toward failure regions by changes of the drifts. It is evident that as the adaptive engine runs, the failure chance increases. In addition, the figures show the fake-threshold that is used in the algorithm. Figure 6.7 portrays a comparison of the convergence rate of the Householder's method and that of the Newton's method.

6.4.3 Determining the Number of Iterations, the Stop Criteria

Algorithm 3 runs with a fixed number of iterations, however, to achieve an estimation with a certain confidence-range, the number of iterations should be set accordingly. In the crude-

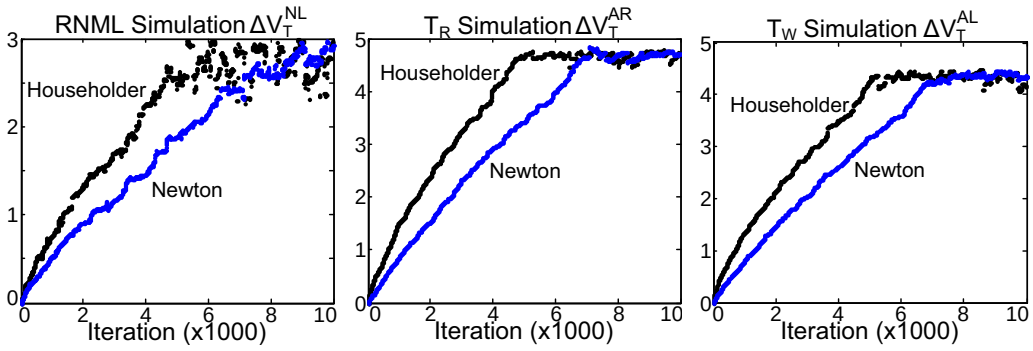


Figure 6.7: Convergence comparison between the Newton's and the Householder's method.

MC method, this goal can be obtained easily by using Eq. (6.3), while continuing with the simulations. In contrast to the crude-MC method, the adaptive method does not work with a fixed number of iterations for a certain confidence-range. The number of samples can vary in each run of the engine. This occurs because of the random and dynamic nature of the algorithm which dictates an uncertain effective convergence rate.

However, the good news is that the variance of the estimation, which leads to the confidence interval, can be estimated during the process. If, at any point, the confidence interval reaches a certain threshold, the algorithm can be stopped. Such a threshold is determined in the form of a ratio over the estimated failure probability. For example, it might be tempting to stop once the 99% confidence interval becomes smaller than the 1/10 of the failure probability itself.

To find the variance of the estimated failure at the Nth iteration, the following is derived, based on Eq. (6.12):

$$\mathrm{Var}\left(\hat{P}\right) = \frac{1}{N^2}\left(\sum_{k=1}^{N} \frac{I_\tau\left(\boldsymbol{x}^{(k)}\right)\varphi^2\left(\boldsymbol{x}^{(k)}\right)}{h^2\left(\boldsymbol{x}^{(k)}, \boldsymbol{\mu}^{(k)}, \Sigma\right)} - N\hat{P}^2\right). \tag{6.17}$$

Therefore, the α% confidence interval at the Nth iteration is

$$\hat{P} \pm \left(\Phi^{-1}(0.5+\alpha/2) \times \sqrt{\mathrm{Var}\left(\hat{P}\right)}\right). \tag{6.18}$$

However, a number of supervision techniques should be considered to control the stop criteria, including, selecting a minimum N, ignoring the estimated confidence-interval when $\hat{P}$ is zero, and restarting the engine when N rises to a very high number without obtaining the required confidence accuracy.

6.5 The Analytical Framework for Optimum Drift and Covariance Matrix Extraction

6.5.1 The Analysis

Based on the statistical simulations, reported earlier, it has been shown that the drift can be manipulated to achieve a reduced-variance estimation. Also, it has been discussed that the covariance matrix can be considered as a factor to increase the chance of failure (Fig. 6.4). In this section, the problem is approached analytically. As a byproduct of the analysis, an approximation of the optimum drift and covariance matrix are derived. These approximations are fed into the adaptive engine as the initial guesses. So instead of starting from naively chosen initial drifts and an arbitrary covariance matrix, the simulation starts with a closer to optimum guess, and consequently, converges faster.

Suppose a performance metric (e.g., RNM, T_W, and T_R) is expressed with the following linear function with respect to the normalized mismatch parameters, x_i:

$$f(\boldsymbol{x}) = a_0 + \sum_{i=1}^{6} a_i x_i + \varepsilon, \tag{6.19}$$

where ε is an independent zero-mean Gaussian error. Since it models the non-linear portion of the performance metric with an independent Gaussian noise, this is not an accurate model. However, given the close-to-linear behavior of the problem and the fact that an approximate optimization of the drift and covariance matrix is targeted, the model is a fairly good one for the proposed method. Note that the approximate results used as initial guess is eventually modified in the adaptive engine anyways.

The objective of this section is to derive the optimal drifts and the covariance matrix. However, the derived equations are only given without the extensive algebraic steps. As shown in Eq(3.21), an ideal alternative distribution should simulate only the failure cases and follow the original distribution in that region. In other words, the perfect and ideal alternative distribution is nothing but $\mathrm{PDF}(x_i \,|\, f<\tau) = \mathrm{PDF}(x_i, f<\tau)\big/\mathrm{Prob}(f<\tau)$. By using the bivariate Gaussian distribution,

$$\mathrm{PDF}(x_i, f<\tau) = \frac{\int_{-\infty}^{\tau} \exp\left\{ \dfrac{x_i^2 + \frac{(f-a_0)^2}{\sigma_f^2} - \frac{2x_i\rho_{x_i}(f-a_0)}{\sigma_f}}{-2\left(1-\rho_{x_i}^2\right)} \right\} df}{2\pi\sigma_f\sqrt{1-\rho_{x_i}^2}}. \tag{6.20}$$

Hence, the following is the alternative distribution:

$$\mathrm{PDF}\left(x_i \,|\, f<\tau\right)=\frac{\mathrm{erf}\left(\frac{\tau-a_0-x_i\sigma_f\rho_{x_i}}{\sigma_f\sqrt{2\left(1-\rho_{x_i}^2\right)}}\right)+1}{2\pi\left(\mathrm{erf}\left(\frac{\tau-a_0}{\sigma_f\sqrt{2}}\right)+1\right)}e^{-\frac{x_i^2}{2}}, \tag{6.21}$$

where $\mathrm{erf}\left(x\right)=2\int_0^x e^{-t^2}dt\Big/\sqrt{\pi}$ and

$$\begin{cases}\sigma_f=\sqrt{\sum_{i=1}^{6}a_i^2+\sigma_\varepsilon^2}\\ \rho_{x_i}=a_i\big/\sigma_f\end{cases} \tag{6.22}$$

As it can be seen in Eq.(6.21), the resultant ideal alternative distribution is not an exact Gaussian. However, as suggested by Fig. 6.3(e) it can be approximated by a Gaussian, given the existence of the bell-shape tail-decaying term, $e^{-\frac{x_i^2}{2}}$, in Eq. (6.21). Therefore, a normal distribution is fitted by matching the mean and the variance. The first two moments of the alternative distribution are derived as follows:

$$\begin{aligned}E\left[x_i \,|\, f<\tau\right]&=-\frac{\sqrt{2}\rho_{x_i}e^{-\frac{\left(\tau-a_0\right)^2}{2\sigma_f^2}}}{\sqrt{\pi}\left(\mathrm{erf}\left(\frac{\tau-a_0}{\sigma_f\sqrt{2}}\right)+1\right)}\\ E\left[x_i^2 \,|\, f<\tau\right]&=1-\frac{\sqrt{2}\rho_{x_i}^2(\tau-a_0)e^{-\frac{\left(\tau-a_0\right)^2}{2\sigma_f^2}}}{\sqrt{\pi}\sigma_f\left(\mathrm{erf}\left(\frac{\tau-a_0}{\sigma_f\sqrt{2}}\right)+1\right)}.\end{aligned} \tag{6.23}$$

A significant observation is that the drift, $E\left[x_i \,|\, f<\tau\right]$, is proportionally related to ρ_{x_i}. This is justifiable by noting that if the mismatch parameter, x_i, and the performance metric, f, have a high positive correlation, the alternative distribution requires a high negative drift to cover the failure region. Moreover, the optimum variance reduces quadratically with correlation. These facts have been already verified through the extensive SRAM simulations in Section 6.3.

The last step in forming the Gaussian alternative JPDF is to complete the covariance matrix by computing the covariance coefficients. The following is the derived JPDF of the alternative distribution:

$$\mathrm{JPDF}\left(x_i,x_j \,|\, f<\tau\right)=\frac{\mathrm{erf}\left(\frac{\tau-a_0-\sigma_f\left(\rho_{x_i}x_i+\rho_{x_j}x_j\right)}{\sigma_f\sqrt{2\left(1-\rho_{x_i}^2-\rho_{x_j}^2\right)}}\right)+1}{2\pi\left(\mathrm{erf}\left(\frac{\tau-a_0}{\sigma_f\sqrt{2}}\right)+1\right)e^{\frac{x_i^2+x_j^2}{2}}}. \tag{6.24}$$

Therefore, the cross-correlation is

$$E\left[x_i x_j \,|\, f < \tau\right] = -\frac{\sqrt{2}\rho_{x_i}\rho_{x_j}(\tau - a_0)\, e^{-\frac{(\tau-a_0)^2}{2\sigma_f^2}}}{\sqrt{\pi}\sigma_f\left(\mathrm{erf}\left(\frac{\tau-a_0}{\sigma_f\sqrt{2}}\right)+1\right)}. \tag{6.25}$$

6.5.2 Results by Integrating the Analytical Framework with the Adaptive Engine

To integrate the provided analysis into the simulation-centric adaptive yield analysis engine developed based on the method of Section 6.4, an initial training step is needed to characterize the approximate model of Eq.(6.19). In the experiments described in this chapter, only a few (e.g., 50) SRAM simulations are conducted. Eq.(6.23) and (6.25) are used to compute the near-optimal drift and covariance matrix. After, the data are fed to the adaptive engine, and the rest of the yield estimation process is performed.

Figure 6.8 and 6.9 convey the simulation results, starting with the calculated initial drifts and covariance matrix by using a model trained by 50 simulations. By comparing these figures with Fig. 6.5 and 6.6, it is evident that the performance metrics reach the failure region much faster. One observation in comparing Fig. 6.6 and 6.9 is the narrower spread of the performance metric due to the use of a non-identity covariance matrix in the latter case. This suggests that the border of the failure region is sampled more frequently than that of the identity covariance matrix case eventually improving the estimation error. It should be reminded that the simulation of the non-failure regions is a waste of runtime, also, it is not worthy to simulate the deep of the failure region since their silicon appearance probability is extremely low. Therefore, observing more samples around the threshold (the narrower spread) is a good sign in terms of the method performance.

Lastly, Fig. 6.10 depicts the histogram of 1,000 estimated yields by applying the adaptive technique and the initial computations and using the stop criteria. The stop criteria is set such that the ratio of 99% confidence interval over the failure rate is less than 0.2. The average of the required number of samples are 3444, 7343, and 6862 for RNML, T_R, and T_W, respectively. To achieve the same confidence interval with the crude-MC method, millions of simulations would be needed, several orders of magnitude runtime improvement. It should be also noted that, in contrast to the crude-MC, the number of samples does not grow if the failure probability decreases because of the adaptive nature of the engine and the initialization phase. In fact, it is more the linearity of the problem that determines the quality of this technique, rather than the failure probability itself. That is why RNML, the most linear performance metric according to analysis in Section 6.3, requires fewer iterations than the rest even though it has the lowest failure probability.

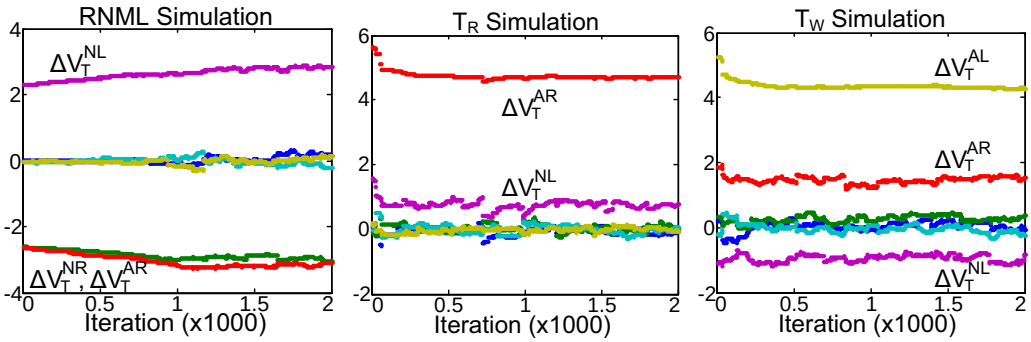


Figure 6.8: Drifts started from analytically calculated initial values.

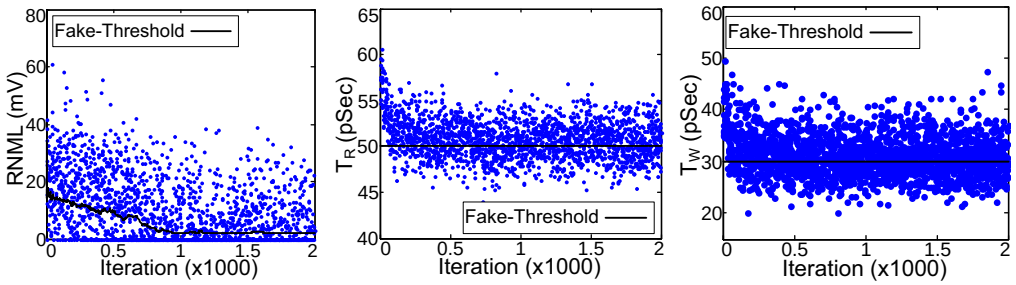


Figure 6.9: Performance metrics for non-identity covariance matrix and starting from non-zero drift.

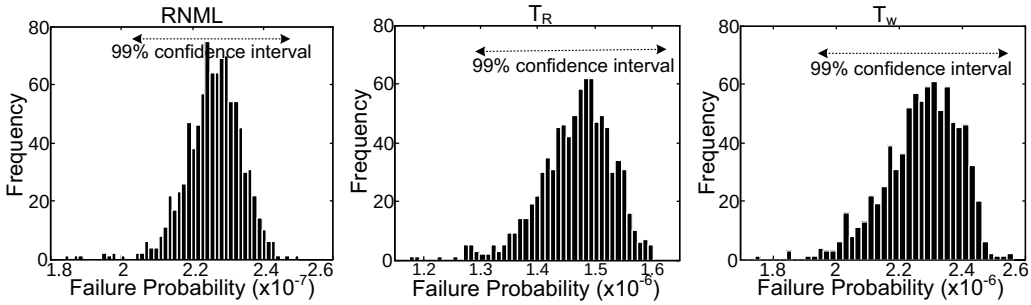


Figure 6.10: The histogram and 99% confidence interval of the estimated yield for 5,000 simulations.

6.6 Conclusions

An adaptive sampling method is proposed for the yield analysis of SRAM cells. The multi-dimensional, rare-event and identity format of the SRAM yield analysis problem make it a challenging problem. We have employed the *almost*-linear relation of SRAM cells' performance metrics with process parameters to establish an adaptive sampling method. The drift vector of

an alternative distribution is updated after each SRAM simulation in order to concentrate the samples into the failure region of the problem space. The near-optimal drift vector and covariance matrix are also computed and integrated into the adaptive sampler in order to improve the efficiency of the method. The range of a few thousand (3,000 to 8,000) samples is found to be enough, in average, to confidently estimate the failure rates of performance metrics around 1e-7 to 1e-6, leading to several orders of magnitude runtime saving. Unlike crude-MC, the required number of samples does not grow with the failure probability decline, instead it is related to the magnitude of the linearity of the performance metric under estimation.

Part III

Micro-Architectural-Level

The nonuniform substrate thermal profile and process variations are two major concerns in todays ultra-deep sub micron designs. To correctly predict performance/leakage/reliability measures and address any yield losses during early stages of design phases, it is desirable to have a reliable estimation of the thermal profile. However, the leakage power sources vary greatly due to process variations and temperature which results in significant variations in the hotspot and thermal profile formation. Traditionally, no leakage variations have been considered during full-chip thermal analysis. In this part of the thesis, the dependency among the process variability, leakage power consumption, and thermal profile are considered at the micro-architectural-level to effectively extract a reliable statistical thermal profile of a working large-scale chip. Knowledge of this is key for proper identification of the hotspot locations and determining a leakage/thermal-based yield.

Chapter 7

Statistical Thermal Profile under Process Variations: Analysis and Applications

7.1 Introduction

As CMOS technology scales down toward sub-90nm regimes, a reliable temperature analysis has become inevitable in early stages of the design process. The increase in the power density has elevated the junction temperature of chips and brought serious reliability concerns into future designs. Moreover, the uneven power consumption profile and anisotropic heat conduction of the die's sidewalls generate local hotspots. The non-uniform high temperature profile over a substrate causes a range of design challenges, as it affects the gate and interconnect delays [141, 142], introduces new timing faults [143], increases the leakage power [144], and accelerates the chip failure due to electro-migration and thermal runaway [145, 146]. Therefore, to achieve a robust design which guarantees satisfaction of system constraints (performance, power, and reliability), knowing the average temperature of a system is not sufficient, and the reliable thermal data should be fed into design automation tools during design phases.

In response to this need, various efforts have been made to extract the temperature profile of silicon substrate [5, 147, 148, 149]. Finite Difference Method is the most popular approach for thermal analysis in which the chip and its packaging materials are discretized to rectangular cubes, and hence, the thermal extraction problem is mapped to a linear circuit simulation. Hotspot [5], a publicly available micro-architectural level IC temperature modeling tool [150], uses the fix meshing technique to extract the thermal profile. ILLIADS-T [147] is an electro-thermal timing simulator in which a developed FDM-based thermal simulator is connected to a circuit simulator, so the thermal-aware timing behavior of a circuit can be performed to detect new thermal-driven timing faults. The meshing process is done adaptively to reduce the size of the problem while providing acceptable accuracy. Another enhanced FDM-based IC thermal analyzer is proposed in [148] based on the multigrid technique for large sparse system of linear

equations which is suitable for large number of meshes in detailed thermal analysis. Finally, as an alternative to the FDM method, green functions are used to analytically extract the thermal profile [149].

All of the current thermal analysis approaches consider deterministic power sources. However, the power dissipation of a circuit (composed of the dynamic and leakage parts) can no longer be deterministically defined in presence of process variations. This is because the leakage power, a key contributor to the total power consumption in scaled technologies [151], has exponential relations to physical parameters, and hence, exhibits significant variations due to process variations. The leakage power has been shown to have a lognormal distribution as it is exponentially dependent to the variation parameters [152, 153]. Analysis over some circuit benchmarks showed that in presence of gate length variations the mean of the subthreshold leakage current is 30% more than its nominal value [152]. Empirical measurements of fabricated chips also showed a 20X total leakage variation in 0.18μm technology [4]. This effect becomes even more critical when considering the exponential increase in the total leakage power over each CMOS technology generation [154] and its dominancy in high-performance circuits for recent CMOS technologies [151].

The earlier works on statistical power analysis (e.g. [152, 153]) assume that the temperatures are kept at nominal values. However, the temperature and leakage of circuits are coupled together which brings a need for an integrated and self-consistent statistical thermal/power analyzer. In fact, the subthreshold leakage power increases nonlinearly with temperature which consequently generates more heat and boosts the temperature in a loop until the generated power is equated with the removed power from the die. Throughout this chapter the term: '*leakage-thermal loop*' is used to refer to this phenomenon. A recent study showed that some parts of a POWER-4 like microprocessor at 130 nm technology will have up to 7 degree thermal difference if the leakage-thermal loop is considered during thermal analysis [155]. The existing thermal extraction tools simply leave this dependency unaddressed, so the designers need to iteratively run the thermal analyzer to account for leakage-thermal loop which is not definitely a runtime efficient way, as many same time-consuming initializations and mathematical calculations should be performed redundantly in each run.

The probabilistic nature of leakage power consumptions caused by process-driven physical parameter fluctuations directly generates uncertain (statistical) thermal profiles since the value of temperature over any location of a die is a function of power consumptions over the whole substrate. However, the subthreshold leakage portion of power consumption increases nonlinearly with temperature, so the resulted variations on thermal profile will be intensified as a wider statistical thermal distribution generates a wider leakage distribution. Therefore, the statistical leakage power analysis results based on nominal temperatures may lead to underestimations in power variations when the leakage-thermal loop is ignored. It has been recently concluded that a circuit which is designed to meet its thermal requirement, without taking into account process variation aware thermal analysis, may fail after fabrication [156]. It has been also shown, by

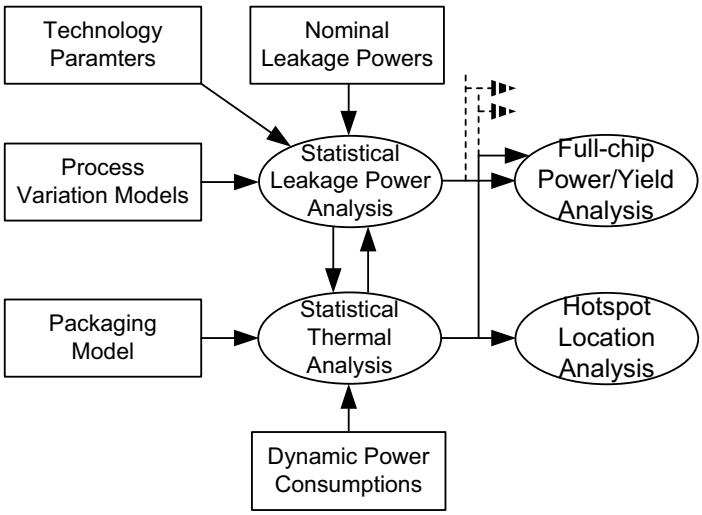


Figure 7.1: Dependency among parameters and models in the developed framework

applying three random sample scenarios for blocks' leakage consumptions, that the locations of hotspots significantly vary in a sample die [157] which makes the hotspot locations indeterministic, as well. Finally, any power estimation by ignoring process and temperature variations can lead to significant error in estimated yield which is expected to degrade further with technology scaling [158]. Therefore, by using reliable statistical thermal information rather than deterministic ones, one can perform more robust system analysis in terms of power/performance/reliability requirements.

In this chapter, a statistical temperature profile analyzer is constructed to estimate the probability density functions (PDF) of the temperatures over each location of a die. It is also shown how the expected value and variance of temperature vary over a sample die. In addition, a method which quantifies the relocations of thermal hotspots is developed which provides a study of how hotspots form while considering variabilities. Finally, a full chip statistical total power estimation technique is proposed by using the statistical information obtained from the analyzer to report a reliable power-constraint yield. Figure 7.1 depicts the diagram of the developed framework and the dependency of the models and parameters.

To have a comprehensive scheme from the sources of variabilities, the gate length and oxide thickness variations are considered in modeling the leakage variation. To realistically model the process variabilities, both inter and intra-die sources of variations are considered. It is assumed that the parameter variations are spatially correlated due to the lithography and chemical mechanical polishing imperfections, so closer gates are more likely to have similar physical characteristics.

The contributions of this work are summarized as follows: 1- Statistical modeling of the temperature, in presence of major process variation parameters. 2- Considering leakage-thermal loop

during the statistical thermal analysis and studying the importance of considering it for full-chip power-constraint yield analysis. 3- Quantifying and analyzing the temperature profiles formations and hotspots relocations. 4- Providing a robust statistical full-chip total power estimation using the obtained statistical thermal analysis data.

The chapter is organized as follows: In Section 7.2, the preliminaries for deterministic thermal profile extraction method, physical parameter variations and leakage power models are presented. The statistical thermal profile analyzer is proposed in Section 7.3, while two applications of the analyzer in hotspot's relocation evaluation and total power estimation along with prior work studies are presented in Section 7.4. By using the developed analyzer, the profile of temperature statistical moments will be derived for a sample die and verified by Mont-Carlo simulations in Section 7.5 where the extracted power consumption probability density function and the result of a sample hotspot location analysis are also verified. Finally, the chapter is concluded in Section 7.6.

7.2 Preliminaries

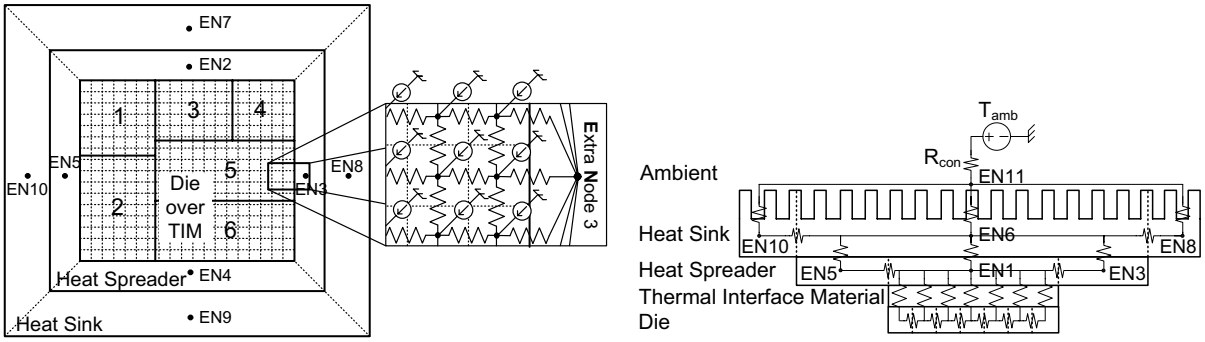
7.2.1 Deterministic Thermal Profile Extraction

The steady state thermal profile over a die is governed by following heat conduction equation [159]:

$$k(x,y,z) \cdot \nabla^2 T(x,y,z) + p(x,y,z) = 0 \tag{7.1}$$

where k is the thermal conductivity of the material ($W/m^{\circ}C$), T is the temperature ($^{\circ}C$), and p is the power density of the heat sources (W/m^2).

As mentioned in the introduction, the numerical approach of solving the Poisson equation of (7.1) is by using the Finite Difference Method (FDM), in which the area of the die is discretized and modeled as a lumped circuit network [5]. Using the well-known duality between thermal and electrical models, each node in the equivalent electrical model corresponds to a grid on a die. The node voltage is the temperature of the grid and the power dissipation of that grid is modeled by a current source flowing into that node. The thermal conduction paths between each grid and its neighboring grids or surrounding packaging structures are modeled by electrical resistances. Therefore, a Kirchhoff's Current Law (KCL)-based admittance matrix is formed for the equivalent electrical model. Solving such linear system of equations for the node voltages produces the temperature profile of the die [160]. To solve that sparse linear problem, either an iterative or direct (LU factorization) [161] method can be applied that is used to construct the inverse of the admittance matrix. It should be noted that as the micro-architectural level designs are targeted for early stage thermal analysis, coarse meshing of a die area will be sufficient [5]. Therefore, having at most few thousands grids allows us to use matrix inversion efficiency. Hence, the temperature of grids can be obtained by the following matrix multiplication:



(a) Top view (b) Lateral view

Figure 7.2: The views of a 6-core sample die with its packaging structure (dimensions are not scaled) [5]

$$t_{m\times 1} = A_{m\times m} \times p_{m\times 1} \tag{7.2}$$

where t and p are the vectors of node temperatures and power consumptions, respectively, and A is the inverse of the equivalent admittance matrix. It should be noted that $m = n + 11$ where n is the number of die grids, and 11 extra nodes represent packaging components and ambient nodes [5]. In which, the heat spreader and sink layer, each has five nodes: one corresponds to the area over the underlaying layer and four trapezoids correspond to periphery that is not covered by the lower layer. Therefore, $[p_1, \cdots, p_n]$ are power consumptions of n grids, $[p_{n+1}, \cdots, p_{n+10}]$ are all zero since no power is generated in packaging nodes. Finally, p_m is a current source used to model chip to ambient removing power which can be determined by the Norton equivalent ($p_m = T_{amb}/R_{con}$) of the ambient temperature voltage source, T_{amb}, and its serially connected convective heat resistance from the heat sink to the air, R_{con}. For illustrative purposes, a chip composed of six numbered cores is depicted in Figure 7.2. The grids' borders are defined by dashed lines. 10 extra nodes (EN1..10) are used to model the heat conduction paths from the heat spreader and heat sink.

7.2.2 Physical Parameter Variation Model

The typical scheme in modeling process variations is by partitioning the surface of a die into rectangular grids [87]. The gates placed in the same grid are assumed to have perfect correlation on their physical parameters since adjacent devices are more likely to have similar physical characteristics after fabrication. Therefore, it is assumed that in a single grid, the variations of any process parameter is constant. In this work, the gate length (L_g) and oxide thickness (T_{ox}) are considered as the sources of physical variabilities. Let X_i be the physical parameter of interest

in grid i, so:

$$X_i = X_0 + \Delta X_i \tag{7.3}$$

where X_0 is the nominal value of the X physical parameter, and ΔX_i is its variation from the nominal value in grid i. The Gaussian zero-mean random distributions is assumed for $\Delta X = \{\Delta L, \Delta T_{ox}\}$ with the standard deviations given for a particular technology as σ_L, and $\sigma_{T_{ox}}$.

Due to spatial imperfection of CMP and lithography processes the gate length and oxide thickness variations are spatially correlated [85, 86, 87, 162]. As a result, their variations are modeled as a random variable, ΔX, decomposed into three distinct components: the inter-die variation ΔX_{inter}, spatially correlated intra-die variation ΔX_{cor} , and a residual part ΔX_{res} that models the purely independent random variation that is not explainable by other variation components. Hence, ΔX can be expressed as [85]:

$$\Delta X = \Delta X_{inter} + \Delta X_{cor} + \Delta X_{res} \tag{7.4}$$

where ΔX_{inter}, ΔX_{cor}, and ΔX_{res} are zero-mean independent Gaussian random variables [96, 85]. The inter-die variation models the variation that is shared for all devices within a die, so it will be the same for all devices in a same chip while the intra-die variations may be different for different grids within the same chip. In fact, the intra-die variations are composed of spatially correlated (ΔX_{cor}) and purely random (ΔX_{res}) components.

To model the spatial correlation, the two $n \times n$ covariance matrix ($\Psi_X = \{\Psi_L, \Psi_{T_{ox}}\}$) which represent the covariance between gate lengths and oxide thicknesses are used. The diagonal elements $\psi_X(i,i)$ of such matrices are the variances of X parameters in the grid i, and the covariances between X parameters of grid i and j are determined in $\psi_X(i,j)$. By applying mathematical random field techniques which assure the positive semi-definiteness of Ψ_X, the necessary condition of any covariance matrix, such matrices can be formed as follows [85]:

$$\begin{cases} \psi_X(i,i) = \sigma_X^2 = \sigma_{\Delta X_{inter}}^2 + \sigma_{\Delta X_{cor}}^2 + \sigma_{\Delta X_{res}}^2 \\ \\ \psi_X(i,j) = cov(\Delta X_i, \Delta X_j) = \sigma_{\Delta X_{inter}}^2 + \rho\left(\upsilon_{ij}\right) \cdot \sigma_{\Delta X_{cor}}^2 \end{cases} \tag{7.5}$$

where
υ_{ij} : Euclidean distance between grid i and j
$0 \leq \rho\left(\upsilon_{ij}\right) \leq 1$: Decreasing function of υ_{ij}(e.g. $\rho\left(\upsilon_{ij}\right) = e^{b\upsilon_{ij}}; b < 0$)

where the shape of $\rho\left(\upsilon_{ij}\right)$ and the values of $\sigma_{\Delta L_{inter}}$, $\sigma_{\Delta L_{cor}}$, and $\sigma_{\Delta L_{res}}$ are defined from statistical measurement data of the technology of interest [85, 86, 87].

In this work, it is assumed that there is no inter-correlation between gate length and oxide thickness variations, as each source of variation is originated from different fabrication process. However, the formulations given in Section 7.3 is flexible enough to consider such correlations as well.

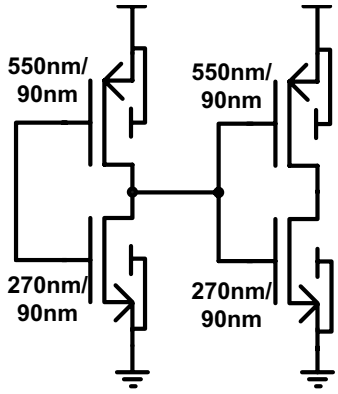


Figure 7.3: A sample inverter used for demonstrating the leakage model

7.2.3 Leakage Power Model

Once the area of a die is discretized into rectangular grids, the total leakage power of each grid has to be modeled in order to provide sufficient flexibility to include thermal and variability awareness into the analysis. The following architectural level leakage power model is used:

$$P_{leak-i} = \lambda_i \cdot \left(1 + \alpha_i \left(T_i - T_{ref}\right) + \beta_i \left(T_i - T_{ref}\right)^2\right) \cdot \hat{I}_{leak-i} \tag{7.6}$$

where P_{leak-i} denotes a random variable used to represent leakage power of grid i in presence of physical variations, and λ_i is the nominal total leakage power of grid i at the reference temperature T_{ref}. The nominal power consumption of grids which share more than one core are calculated based on the weighted sum of sharing cores power density components. Throughout this chapter, the term '*nominal*' is used to indicate the situation when no process variation has been taken into account, and all parameters have their own nominal X_0 value. λ_i is determined by the grid's circuit topology and accounts for effects like logic style, transistor sizing, transistor stacking, ratio and number of NMOS and PMOS transistors in the grid, and the technology used for the circuit implementation. This value can be obtained by using circuit level leakage simulations or given by the core provider at T_{ref}, when no process variation is taken into account. For example, in 90nm technology ($V_{DD} = 1.2, L_{eff} = 35nm$), the Spice simulation of the sample circuit depicted in Figure 7.3 using PTM BSIM4 models [163] shows $\lambda_i = 114.4nW$ total average leakage power at $T_{ref} = 100^{\circ}C$.

The total leakage power in CMOS circuits composed of three components namely subthreshold, gate direct tunneling, and reversed biased band-to-band tunneling currents [17]. However, the subthreshold leakage currents which contributes the largest portion of the total leakage power in high performance chips due to low threshold voltage and high operating temperature is strongly temperature sensitive, as below:

$$I_{sub} \propto \mu(T)\, v_T^2 e^{-\frac{V_{th}(T)}{m v_T}} \tag{7.7}$$

where $v_T = kT/q$ is the thermal voltage, $\mu(T) \propto T^{-1.5}$ is the charge mobility, and V_{th} is the temperature-dependent threshold voltage which drops when the temperature increases.

Therefore, the total leakage power is modeled by a quadratic approximation around the nominal value [144]. In fact, the second order leakage-temperature model fits better to the measured points than a first order exponential model, $\lambda_i \cdot e^{\alpha_i \left(T_i - T_{ref}\right)}$, as it has one more fitting parameter which provides less fitting error. For example, the comparison between the first order exponential model and the quadratic model are depicted in Figure 7.4(a) while fitted to BSIM measurements of the sample two inverters circuit. The first order exponential model has up to 5% absoulute fitting error while the quadratic model has less than 0.06% error.

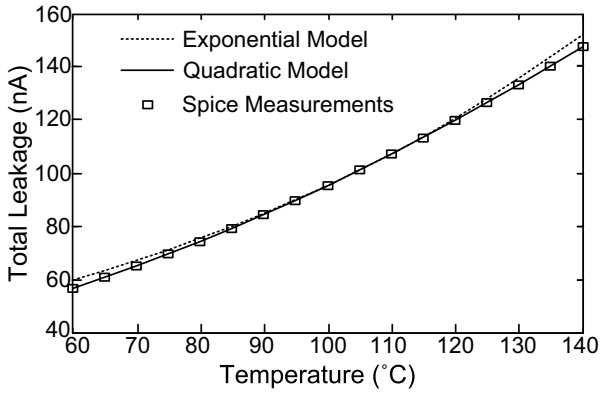
Finally, $\hat{I}_{leak-i}$ is the normalized (dimensionless) total leakage current of grid i including process dependent effects. In contrast to the wide temperature distribution, the magnitude of process variations is observed to be less than 15% in practice, hence $\hat{I}_{leak-i}$ can be well-approximated by using an exponential of a first-order Taylor expansion at the nominal values of process parameters. Therefore, the normalized leakage current can be written as an exponent of linear weighed sum of process parameters around the nominal values [23, 153]. This is because the subthreshold leakage current is exponentially related to the threshold voltage which varies with gate length and oxide thickness, also the gate direct tunneling varies with oxide thickness exponentially. Therefore, the normalized leakage current can be represented as:

$$\hat{I}_{leak-i} = e^{\beta_{L_i} \cdot \Delta L_i + \beta_{Tox_i} \cdot \Delta T_{ox_i}} \tag{7.8}$$

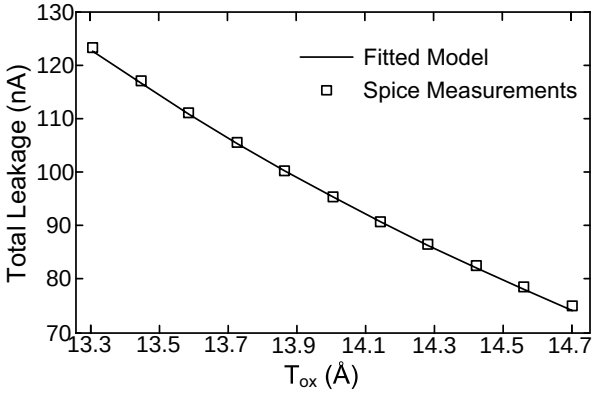
where ΔX_i is the variation of the parameter X from its nominal value in grid i, and β_{X_i} is the first order derivative of the grid's i leakage current logarithm:

$$\beta_{X_i} = \left. \frac{\partial \left(\ln \hat{I}_{leak-i}\right)}{\partial X} \right|_{X=X_0} \tag{7.9}$$

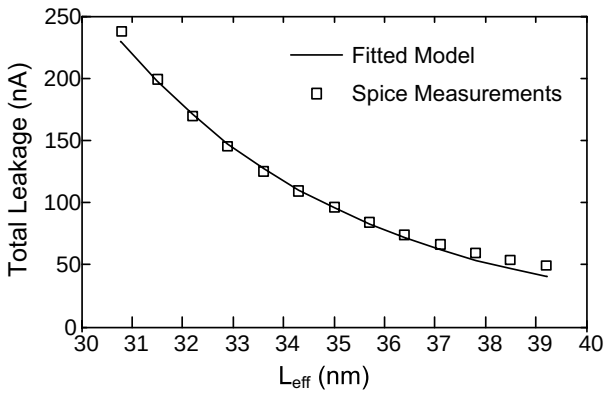
It should be noted that the correlated V_{th} variation has been considered through spatially correlated models of gate length and oxide thickness which both affect V_{th}. However, if there are any other sources of correlated variations in V_{th}, the proposed methodology is flexible enough to account for them through adding extra terms on the power of the exponent in Eq. (7.8) and updating the consequent equations. β_{X_i} factors can be calculated either analytically [158] or numerically by fitting the total leakage simulation results of the circuit around the nominal point with Eq. (7.6). Figures 7.4(b) and 7.4(c) show the fitted total leakage currents of the sample circuit (Figure 7.3) with the actual BSIM measurements when the oxide thickness and effective gate length are varied around the nominal points. β_{L_i} and β_{Tox_i} are set to $-7.35/L_{eff_0}$ and $-5.02/T_{ox_0}$, respectively.



(a)



(b)



(c)

Figure 7.4: Comparison between Spice measured total leakage current of the circuit depicted in Figure 7.3 and the fitting models when process parameters and temperature vary around the nominal value

Consequently, by using Eq. (7.6 and 7.8) following total leakage power model will be obtained for grid *i*:

$$P_{leak-i} = \lambda_i' \left(1 + \alpha_i' T_i + \beta_i' T_i^2\right) e^{\beta_{L_i} \cdot \Delta L_i + \beta_{Tox_i} \cdot \Delta T_{ox_i}} \tag{7.10}$$

where $\acute{\lambda}_i$, $\acute{\alpha}_i$, and $\acute{\beta}_i$ can be simply obtained from λ_i, α_i, β_i, and T_{ref}.

7.3 Statistical Thermal Analysis

In this section, the statistical thermal analyzer is proposed where the probability density functions (PDF) of grids' temperatures are estimated. The problem is to solve Eq. (7.1) when $p(x,y,z)$ is a function of $T(x,y,z)$ while it has spatially statistical behavior. To formulate the problem, the

leakage power uncertainties have to be integrated into the thermal extraction. Hence, instead of using deterministic power consumption elements, p_i in vector p of Eq. (7.2), correlated random variable power sources, P_i, are used. P_i, the random variable of the power consumption in grid i, is defined as:

$$1 \leq i \leq n: \quad P_i = p_{dyn-i} + P_{leak-i} \tag{7.11}$$

where P_{leak-i} is the random variable representing the leakage power of grid i defined in Eq. (7.10), and p_{dyn-i} is the dynamic power consumption of the grid i. It should be noted that since the switching power consumption is not as sensitive as leakage power to variations [164], it is assumed to be a constant variable for a grid.

Having a statistical form for power consumptions of grids, a random variable can be assigned for the temperature of node i based on Eq. (7.2), as T_i:

$$T_i = \sum_{j=1}^{n} a_{ij} \cdot P_j + a_{im} \cdot p_m \tag{7.12}$$

where a_{ij} coefficients and p_m have been defined in Section 7.2.1.

However, the estimation of T_is' PDFs is not straightforward since there is a relation between leakage power and temperature of each grid (Eq. 7.10), while the leakage power consumption sources over a die are all spatially correlated due to spatially correlated gate length variations (Eq. 7.5 and 7.8). Therefore, to estimate PDFs of T_is, the problem is broken into the following steps:

Step 1: In the first step, deterministic nominal thermal extraction is performed iteratively. The iterations are done to take into account the leakage-thermal loop effect during temperature extraction. The following set of equations are evaluated iteratively until no significant change on thermal profile could be seen in the new iteration:

$$\begin{aligned} p_{leak-j}^{(k)} &= \lambda'_j \cdot \left(1 + \alpha'_j T_j^{(k)} + \beta'_j T_j^{(k)^2}\right) \\ T_i^{(k+1)} &= \sum_{j=1}^{n} a_{ij} \left(p_{dyn-j} + p_{leak-j}^{(k)}\right) + a_{im} p_m \end{aligned} \tag{7.13}$$

The extracted thermal profile is named nominal temperature profile ($T_i^{nom} = T_i^{(l)}$), where l is the number of iterations until the thermal profile convergences. Typically $l = 4$ iterations are enough to extract the nominal thermal profile if the initial temperature is set to the ambient temperature [155] and can be reduced down to 2, if a more reasonable initial temperature is used [144].

Step 2: After nominal thermal profile extraction, the parameter variations are added into the calculations. In this step, the expected value vector and covariance matrix of the grids' temperatures are calculated considering correlated power sources due to correlated physical parameter

variations. As can be seen in Eq.(7.12), T_i is defined by the summation of a set of finite random variables (P_js). Therefore, to calculate the moments of T_is, the following property of the sum of random variables is needed.

Property 1: Given a vector of random variables $Z^T_{n\times 1} = [Z_1, Z_2, \cdots, Z_n]$ with a known expected value vector and covariance matrix of $\left(M_{Z_{n\times 1}}, S_{Z_{n\times n}}\right)$, if there is another vector of random variables $Y^T_{n\times 1} = [Y_1, Y_2, \cdots, Y_n]$ such that: $Y_i = \sum_{j=1}^{n} c_{ij} \cdot Z_j$, where c_{ij}s are constant coefficients, then [165]:

$$\begin{aligned} E\left[Y_i\right] &= \sum_{j=1}^{n} c_{ij} \cdot E\left[Z_j\right] \\ \operatorname{cov}\left(Y_i, Y_j\right) &= \sum_{k=1}^{n}\sum_{l=1}^{n} c_{ik} \cdot c_{jl} \cdot \operatorname{cov}\left(Z_k, Z_l\right) \end{aligned} \tag{7.14}$$

If matrix multiplication is used, the given linear equations can be represented as follows:

$$\begin{aligned} M_{Y_{n\times 1}} &= C_{n\times n} \times M_{Z_{n\times 1}} \\ S_{Y_{n\times n}} &= C_{n\times n} \times S_{Z_{n\times n}} \times C^T_{n\times n} \end{aligned} \tag{7.15}$$

where $\left(M_{Y_{n\times 1}}, S_{Y_{n\times n}}\right)$ are expected value vector and covariance matrix of random variables Y_is, and C is the matrix representation form of the constant coefficients c_{ij}s. In fact, the matrix multiplication scheme reduces the computational complexity of the Y's covariance extraction from $O(n^4)$ to $O(n^3)$ naive matrix multiplication. This computational complexity reduction is achieved by storing the intermediate calculation results into $C \times S_Z$ and reusing them in future multiplication with C^T. Also, by one matrix multiplication, both variances and covariance are extracted, which both are necessary in the future steps of the analysis. In fact, most of the computations are mapped to the form of matrix multiplication which significantly increases the efficiency of the approach. The implementation of the naive matrix multiplication has been intensely optimized for various processor architectures (e.g. AMD, Apple, IBM, Intel, Sun) using Basic Linear Algebra Subprogram (BLAS) libraries which provide orders of magnitude speed-up over naively coded routines [166]. In addition, the runtime complexity may be further reduced down to $O(n^{2.376})$ if the Coppersmith–Winograd fast square matrix multiplication technique is used [167].

$$\begin{aligned} &E\left[P_{leak-i}P_{leak-j}\right] = \lambda'_i\lambda'_j \times \\ &\left(\begin{array}{l} E\left[\hat{I}_{leak-i}\hat{I}_{leak-j}\right] + \alpha'_i E\left[T_i\hat{I}_{leak-i}\hat{I}_{leak-j}\right] + \alpha'_j E\left[T_j\hat{I}_{leak-i}\hat{I}_{leak-j}\right] + \\ \beta'_j E\left[T_j^2\hat{I}_{leak-i}\hat{I}_{leak-j}\right] + \alpha'_i\alpha'_j E\left[T_iT_j\hat{I}_{leak-i}\hat{I}_{leak-j}\right] + \alpha'_i\beta'_j E\left[T_iT_j^2\hat{I}_{leak-i}\hat{I}_{leak-j}\right] + \\ \alpha'_j\beta'_i E\left[T_jT_i^2\hat{I}_{leak-i}\hat{I}_{leak-j}\right] + \beta'_i\beta'_j E\left[T_i^2T_j^2\hat{I}_{leak-i}\hat{I}_{leak-j}\right] + \beta'_i E\left[T_i^2\hat{I}_{leak-i}\hat{I}_{leak-j}\right] \end{array} \right) \end{aligned} \tag{7.16}$$

By using Property 1 and the matrix-based representation, if the expected value vector and covariance matrix of grid's power are given, the expected value vector and covariance matrix of temperatures $\left(M_{T_{n\times1}}, S_{T_{n\times n}}\right)$ can be extracted, as:

$$\begin{aligned} M_{T_{n\times1}} &= A_{n\times n} \times M_{P_{n\times1}} + p_m \cdot a_{n\times1} \\ S_{T_{n\times n}} &= A_{n\times n} \times S_{P_{n\times n}} \times A^T_{n\times n} \end{aligned} \tag{7.17}$$

where $A_{n\times n}$ is the first left/upper $n\times n$ square matrix portion of the inverse admittance matrix defined in Eq. (7.2) and $a_{n\times1} = [a_{1m}, \cdots, a_{nm}]^T$ is the vector of ambient temperature coefficients (a_{im}). Therefore, the expected values and covariances of temperatures are $E[T_i] = M_T(i)$, $\text{cov}(T_i, T_j) = S_T(i,j)$. However to estimate the first power consumption's statistical mean and covariances $\left(M_{P_{n\times1}}, S_{P_{n\times n}}\right)$, the following property needs to be defined:

Property 2: Given a normal random variable X with mean and variance of $\left(\mu, \sigma^2\right)$, if $Y = e^{\beta\cdot X}$, then the expected value of Y can be calculated as [165]:

$$E[Y] = \exp\left\{\beta\mu + \frac{\beta^2\sigma^2}{2}\right\} \tag{7.18}$$

By using this property, leakage power model (Eq. (7.10), and Ψ_Xs (covariance matrices of process parameters), the expected value vector and covariance matrix of grid's power consumptions $\left(M_{P_{n\times1}}, S_{P_{n\times n}}\right)$ are extracted as follows:

$$\begin{aligned} &M_P(j) = E\left[P_j\right] = p_{dyn-j} + E\left[\lambda'_j\left(1+\alpha'_j T_j^{nom} + \beta'_j T_j^{nom^2}\right)\hat{I}_{leak-j}\right] \\ &= p_{dyn-j} + \lambda'_j\left(1+\alpha'_j T_j^{nom} + \beta'_j T_j^{nom^2}\right)\eta_j \\ &\text{and} \\ &S_P(i,j) = \text{cov}\left(P_i, P_j\right) = E\left[P_i\cdot P_j\right] - E[P_i]\cdot E\left[P_j\right] \\ &= \lambda'_i\lambda'_j\eta_i\eta_j\left(e^{\beta_{L_i}\beta_{L_j}\psi_L(i,j)+\beta_{Tox_i}\beta_{Tox_j}\psi_{Tox}(i,j)} - 1\right)\cdot \\ &\left(1+\alpha'_i T_i^{nom} + \beta'_i T_i^{nom^2}\right)\left(1+\alpha'_j T_j^{nom} + \beta'_j T_j^{nom^2}\right) \\ &\text{where } \eta_i = \text{E}\left[\hat{\text{I}}_{\text{leak}-\text{i}}\right] = \text{e}^{\frac{\beta^2_{\text{L}_\text{i}}\sigma^2_{\text{L}}+\beta^2_{\text{Tox}_\text{i}}\sigma^2_{\text{Tox}}}{2}} \end{aligned} \tag{7.19}$$

However, in this step, the calculation of the given statistical moments of grid's leakage consumptions was performed based on the nominal temperature values obtained from Step 1, as no statistical information is available for temperatures by this point.

Step 3: In this step, the new expected value vector and covariance matrix are extracted for temperatures by considering the statistical information of the temperatures from the previous step as well as the process variability data at the same time. Therefore, first the expected value of power consumptions is re-evaluated as follows:

$$E\left[P_j\right] = p_{dyn-j} + \lambda'_j\left(\eta_j + \alpha'_j E\left[T_j\hat{I}_{leak-j}\right] + \beta'_j E\left[T_j^2\hat{I}_{leak-j}\right]\right) \tag{7.20}$$

in which, $E\left[T_j\hat{I}_{leak-j}\right]$ and $E\left[T_j^2\hat{I}_{leak-j}\right]$ are needed. First, we find $E\left[T_i\hat{I}_{leak-j}\right]$:

$$\begin{aligned}
&E\left[T_i\hat{I}_{leak-j}\right] = \sum_{k=1}^{n} a_{ik}\lambda'_k\left(1+\alpha'_k T_k^{nom}+\beta'_k T_k^{nom^2}\right)E\left[\hat{I}_{leak-j}\hat{I}_{leak-k}\right]\\
&+\sum_{k=1}^{n} a_{ik}p_{dyn-k}E\left[\hat{I}_{leak-j}\right]+a_{im}p_m E\left[\hat{I}_{leak-j}\right]\\
&=A_{n\times n}\times\left(L_{n\times n}\times M_{n\times n}+N_{n\times n}\times O_{n\times n}\right)+P_{n\times n}
\end{aligned}$$

where

$$\begin{aligned}
&L_{n\times n}:\begin{cases}\text{if } i\neq j & l_{ij}=0\\ \text{else} & l_{ii}=\lambda'_i\left(1+\alpha'_i T_i^{nom}+\beta'_i T_i^{nom^2}\right)\end{cases}\\
&N_{n\times n}:\begin{cases}\text{if } i\neq j & n_{ij}=0\\ \text{else} & n_{ii}=p_{dyn-k}\end{cases}\\
&M_{n\times n}: m_{ij}=\eta_i\eta_j e^{\beta_{L_i}\beta_{L_j}\psi_L(i,j)+\beta_{Tox_i}\beta_{Tox_j}\psi_{Tox}(i,j)}\\
&O_{n\times n}: o_{ij}=\eta_j\\
&P_{n\times n}: p_{ij}=a_{im}p_m\eta_j
\end{aligned}\tag{7.21}$$

The sparsity of L and N are used during matrix multiplication to speed up this step. Now, $E\left[T_j^2\hat{I}_{leak-j}\right]$ should also be calculated. Therefore, we need to define following property of multiplication of lognormal random variables.

Property 3: Given a set of lognormal correlated random variables $\{X_1,\cdots,X_k\}$. If m_{X_i} and s_{X_i} are the expected value and standard deviation of X_i's logarithm, and $\rho_{X_iX_j}$ is the correlation coefficient between X_i's and X_j's logarithms, the random variable $Y=\prod_{i=1}^{k}X_i^{n_i}$ is lognormal with the expected value of:

$$E[Y]=e^{\sum_{i=1}^{k}n_i m_{X_i}+\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}n_i n_j s_{X_i}s_{X_j}\rho_{X_iX_j}+\frac{\sum_{i=1}^{k}n_i^2 s_{X_i}^2}{2}}\tag{7.22}$$

The expected value and standard deviations of T_j logarithm's can be extracted from the last step, and to find the correlation coefficients between logarithms of random variables (T_j and $\hat{I}_{leak-j}$) when $E\left[T_j\hat{I}_{leak-j}\right]$ is known, the following property should be defined:

Property 4: Assume two correlated lognormal random variables $X_1=e^{Z_1}$ and $X_2=e^{Z_2}$ with given $E[X_1X_2]$, where Z_1 and Z_2 have the mean and standard deviation of (μ_1, σ_1) and (μ_2, σ_2), respectively, then:

$$\rho_{Z_1Z_2}=\frac{\ln\left(E[X_1X_2]\right)-\left(\mu_1+\mu_2+\frac{\sigma_1^2+\sigma_2^2}{2}\right)}{\sigma_1\sigma_2}\tag{7.23}$$

By using these properties, $E\left[T_j^2\hat{I}_{leak-j}\right]$ can be calculated and used to find $E\left[P_j\right]$ in Eq. (7.20). Next, the covariances of power sources can be extracted as follows:

$$\operatorname{cov}\left(P_i,P_j\right)=E\left[P_{leak-i}P_{leak-j}\right]-E\left[P_{leak-i}\right]E\left[P_{leak-j}\right] \tag{7.24}$$

where the $E\left[P_{leak-i}\right]=E\left[P_i\right]-p_{dyn-i}$ has already been extracted from Eq. (7.20), so only $E\left[P_{leak-i}\cdot P_{leak-j}\right]$ is needed to be found from Eq. (7.16).

As can be seen in Eq. (7.16), the expected values of more combinations of T_i, T_j, $\hat{I}_{leak-i}$, and $\hat{I}_{leak-j}$ are needed which all can be extracted using the property 3. Finally, by using the new statistical moments for power consumptions, the updated expected value vector and covariance matrix of temperature can be re-extracted by re-evaluating Eq. (7.17).

Step 4: In this step, the moments of power consumptions and temperatures are updated iteratively. In every iteration, new moments are derived for power consumptions using Eq.(7.20, 7.24, 7.16), and the temperatures' moments are re-evaluated using Eq. (7.17). However, in this step, the $E\left[T_i\hat{I}_{leak-j}\right]$ are derived using the following equation, rather than previous Eq. (7.21).

$$\begin{aligned}
&E\left[T_i\hat{I}_{leak-j}\right]=\sum_{k=1}^{n}a_{ik}\lambda'_kE\left[\left(1+\alpha'_kT_k+\beta'_kT_k^2\right)\hat{I}_{leak-j}\hat{I}_{leak-k}\right]\\
&+\sum_{k=1}^{n}a_{ik}p_{dyn-k}E\left[\hat{I}_{leak-j}\right]+a_{im}p_mE\left[\hat{I}_{leak-j}\right]\\
&=\left(A_{n\times n}\times\left(L_{n\times n}\times M_{n\times n}+N_{n\times n}\times O_{n\times n}\right)\right)+P_{n\times n}
\end{aligned} \tag{7.25}$$

where

$$L_{n\times n}:\begin{cases}\text{if } i\neq j & l_{ij}=0\\ \text{else} & l_{ii}=\lambda'_i\end{cases}$$

$$M_{n\times n}:m_{ij}=E\left[\hat{I}_{leak-i}\hat{I}_{leak-j}\right]+\alpha'_iE\left[T_i\hat{I}_{leak-i}\hat{I}_{leak-j}\right]+\beta'_iE\left[T_i^2\hat{I}_{leak-i}\hat{I}_{leak-j}\right]$$

N, O, and P are the same as Eq. (7.21)

It should be noted that $E\left[\hat{I}_{leak-i}\hat{I}_{leak-j}\right]$, $E\left[T_i\hat{I}_{leak-i}\hat{I}_{leak-j}\right]$, and $E\left[T_i^2\hat{I}_{leak-i}\hat{I}_{leak-j}\right]$ have been already extracted when Eq. (7.16) was evaluated.

Step 5: In the last step, after extracting the temperatures' expected value $(E[T_i]=M_T(i))$ and variance $(\operatorname{var}(T_i)=S_T(i,i))$, a probability density function, f_{T_i}, is formed for the random variable T_i.

As suggested by the Eq. (7.12), the temperature of a grid can be written as a linear weighed sum of grids' power consumptions, and since the multiple of lognormal random variables is still lognormal, integrating the polynomial leakage vs. temperature relation into the problem keeps the leakage probability distribution lognormal. This is because the sum of lognormal

distributions can be modeled as another lognormal random variable [168]. The lognormal density function is estimated using Wilkinson's method [169] based on matching the first two moments. Hence, the lognormal probability density function of temperature at grid i, f_{T_i}, has the following general 3-parameter format:

$$f_{T_i}(T_i = t) = \frac{\exp\left\{-\frac{\left(\ln(t-t_{min-i})-m_{T_i}\right)^2}{2s_{T_i}^2}\right\}}{(t-t_{min-i})\sqrt{2\pi}s_{T_i}} \tag{7.26}$$

where m_{T_i} and s_{T_i} are the mean and standard deviation of the T_i's logarithm, and can be determined by matching with the obtained values from step 3, as:

$$\begin{aligned} s_{T_i}^2 &= \ln\left(1+\frac{\text{var}(T_i)}{(E[T_i]-t_{min-i})^2}\right) \\ m_{T_i} &= \ln\left(E[T_i]-t_{min-i}\right) - \frac{s_{T_i}^2}{2} \end{aligned} \tag{7.27}$$

where t_{min-i} is the mathematically minimum possible temperature of grid i in presence of variability. This minimum point can be found by deterministic thermal extraction when all process parameters (ΔX_i) are set to their worst case value ($3\sigma_X$) which provide the minimum leakage.

7.4 Applications

In this section, two applications of the developed analyzer are proposed in which the extracted moments of temperatures are used in power and hotspots formation analysis.

7.4.1 Early Stage Statistical Thermal and Process Aware Full-Chip Power Estimation

Due to the nonlinear dependency between leakage power and temperature, higher than average temperature spots contribute over-proportionally to the total power dissipation of a chip. Therefore, predicting the leakage power and thus the total system power require detailed and accurate knowledge of the temperature distribution and its statistical behavior. Hence, ignoring them might lead to an inaccuracy in power consumption estimations and yield analysis [156, 158].

Su et al. [144] estimated the full chip leakage considering uneven voltage and temperature profile. They have used the heat conduction relation (Eq. 7.1) to accurately model the thermal profile while using the polynomial leakage-thermal model. However, they have not taken into account the variability of process parameters, so their approach only provides a crisp value of nominal leakage and cannot be used for estimating yield. Process variations are accounted during leakage estimation in [158], but it uses a simple average temperature model for die-to-die

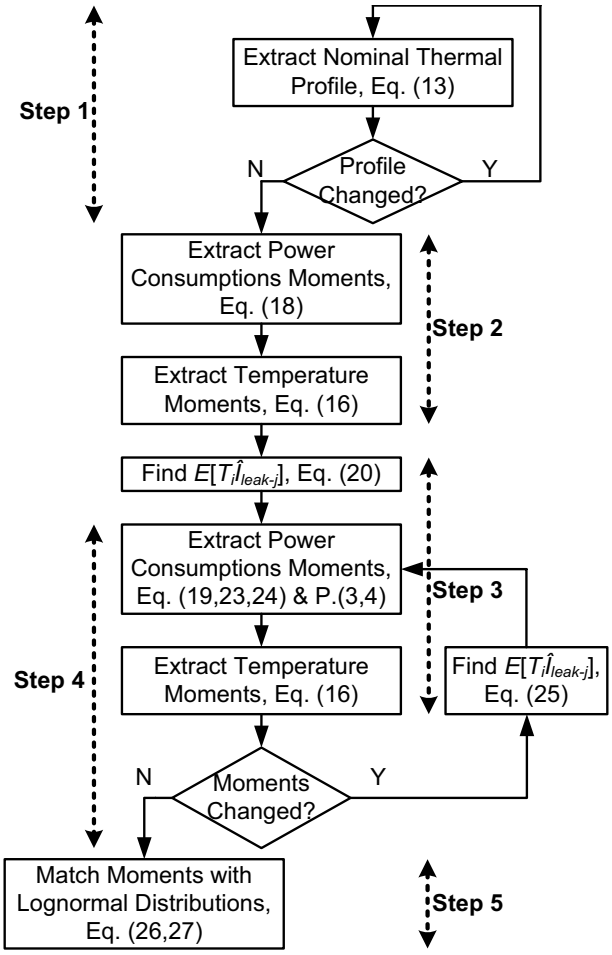


Figure 7.5: Flowchart of the proposed statistical thermal analyzer

temperature variation. They have also considered an arbitrary and constant value for temperature variance over a die which is not the case in practice. Moreover, they have ignored the correlation between temperatures of different locations and temperature-process parameters covariances. In fact, ignoring these two types of correlations leads to an underestimation in the magnitude of leakage uncertainty and hence the estimated yield.

If the estimated means and covariances of temperatures are used to find the probability density function of the full chip total power consumption, a more reliable power-driven yield analysis can also be performed. This is because the thermal statistical moments have been obtained by considering all the placement-driven power consumption information, process variabilities, and leakage-thermal loop.

However, the obtained temperature statistical moments should be utilized carefully in the estimation of the total power to avoid any intense computations. Since both the process variations

and leakage-thermal loop have been addressed once before during the statistical thermal analysis, a following fast approach is developed to estimate the PDF of total power consumption based on the extracted thermal moments.

The average temperature of a chip when it reaches the thermal equilibrium – the state of which the removed heat is equal to the total generated heat (dynamic + thermal-dependent leakage) – can be derived from the following equation [170]:

$$T_{avg} = T_a + R_\theta \cdot \frac{P_{tot}}{Ar} \tag{7.28}$$

where T_{avg} is the average chip temperature, T_a is the ambient temperature, P_{tot} (W) is the total power consumption, Ar (cm^2) is the chip area, and R_θ $(cm^{2\circ}C/W)$ is the equivalent thermal resistance of the substrate layer plus the packaging and heat sink.

As can be seen in Eq. (7.28), if the probability density function of the average temperature is known, the probability density function of the total power which ended up with such average temperature can be determined. As a result, the mean and standard deviation of T_{avg} are calculated by using the statistical information from the developed statistical thermal analyzer, as follows:

$$\begin{aligned} E\left[T_{avg}\right] &= \frac{\sum_{i=1}^{n} M_T(i)}{n} \\ \mathrm{var}\left(T_{avg}\right) &= \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} S_T(i,j)}{n^2} \end{aligned} \tag{7.29}$$

By using the obtained T_{avg} moments and Eq. (7.28), the mean and variance of P_{tot} can be calculated as follows:

$$\begin{aligned} E\left[P_{tot}\right] &= \left(E\left[T_{avg}\right] - T_a\right) \cdot \frac{Ar}{R_\theta} \\ \mathrm{var}\left(P_{tot}\right) &= \mathrm{var}\left(T_{avg}\right) \cdot \left(\frac{Ar}{R_\theta}\right)^2 \end{aligned} \tag{7.30}$$

Finally, the probability density function of total power, $f_{P_{tot}}$, can be determined from the general 3-parameters lognormal PDF as below:

$$\begin{aligned} &f_{P_{tot}}\left(P_{tot} = p\right) = \frac{\exp\left\{-\frac{\left(\ln\left(p - p_{dyn-tot}\right) - m_{P_{tot}}\right)^2}{2s_{P_{tot}}^2}\right\}}{\left(p - p_{dyn-tot}\right)\sqrt{2\pi}s_{P_{tot}}} \\ &\text{where} \\ &s_{P_{tot}}^2 = \ln\left(1 + \frac{\mathrm{var}(P_{tot})}{\left(E[P_{tot}] - p_{dyn-tot}\right)^2}\right) \\ &m_{P_{tot}} = \ln\left(E\left[P_{tot}\right] - p_{dyn-tot}\right) - \frac{s_{P_{tot}}^2}{2} \end{aligned} \tag{7.31}$$

where $p_{dyn-tot} = \sum_{i=1}^{n} p_{dyn-i}$ is the total dynamic power consumption.

By having the PDF of P_{tot}, one can construct a power-constrained yield, Y_P, by integrating over the PDF and solve it numerically:

$$Y_P = P\left(P_{tot} \leq p_{max}\right) = \int_{p_{dyn-tot}}^{p_{max}} f_{P_{tot}}\left(p\right) d_p \tag{7.32}$$

7.4.2 Evaluation of Hotspots Relocations

In this section, another application of the developed analyzer is presented in which the extracted moments of temperatures are used in the analysis of hotspots formation. In fact, the probabilistic nature of the temperature, driven by the variable process-dependent circuit leakages, may cause some areas of a chip to show peak temperature while they have not been the hottest part of the die when no variability has been taken into account during simulations. This phenomenon brings uncertainty to the hotspot formation.

By applying three arbitrary random leakage scenarios, Link et al. [157] showed that the location of hotspots vary significantly from die to die due to process variation. Therefore, even in the presence of highly accurate predictive deterministic thermal modeling, process variation will prevent accurate localized modeling of power distribution. However, the authors have not quantitatively analyzed the relocation problem, so the designer does not have a measure of how probable is a location to show a higher temperature than the traditional hotspot. This information is key in considering a location as hotspot and guiding the designer in applying further thermal management solutions, such as: more precise on-die temperature sensors placement for adaptive hotspot avoiding mechanisms and efficient design of advanced cooling systems which requires placement of localized cooling solutions (e.g. local spray cooling, thin-film thermoelectric coolers) to eliminate the hot-spots [171].

To evaluate the hotspots relocations, consider the grids i and j, such that: $T_j^{nom} > T_i^{nom}$ when no variation is accounted. In the presence of process variations and hence thermal variations, it is possible that grid i experiences higher temperature than j. This might happen when the variation in leakage distribution causes considerably more temperature elevation in location i than j which may produce a relocation of the hotspot from where it is originally expected to be seen. This effect can be troublesome if it has not been addressed and considered during design processes.

In this section, the probability that the temperature of grid i exceeds the temperature of grid j, $P(T_i > T_j)$, is estimated. Generally, having two dependent random variables of T_i and T_j, such a probability is:

$$P\left(T_i > T_j\right) = \int_{-\infty}^{+\infty} \int_{y}^{+\infty} f_{ij}\left(T_i = x, T_j = y\right) d_x d_y \tag{7.33}$$

where $f_{ij}\left(T_i, T_j\right)$ is the joint probability density function (JPDF) of T_i and T_j. However, one needs an analytical JPDF for $f_{ij}\left(T_i, T_j\right)$ to estimate the desired probability. As shown in Section

7.3, the PDFs of T_i and T_j can be approximated as lognormal distributions. Therefore, a bivariate lognormal distribution [172] can be assumed for the $f_{ij}\left(T_i, T_j\right)$:

$$f_{ij}\left(T_i = x, T_j = y\right) = \frac{\exp\left\{-\frac{q}{2-2\rho_{ij}^2}\right\}}{2\pi(x-t_{min-i})(y-t_{min-j})s_{T_i}s_{T_j}\sqrt{1-\rho_{ij}^2}}$$

where

$$q = \left(\frac{\ln(x-t_{min-i})-m_{T_i}}{s_{T_i}}\right)^2 + \left(\frac{\ln\left(y-t_{min-j}\right)-m_{T_j}}{s_{T_j}}\right)^2 - 2\rho_{ij}\left(\frac{\ln(x-t_{min-i})-m_{T_i}}{s_{T_i}}\right)\left(\frac{\ln\left(y-t_{min-j}\right)-m_{T_j}}{s_{T_j}}\right) \tag{7.34}$$

where $\{m_{T_i}, s_{T_i}, m_{T_j}, s_{T_j}\}$ are the mean and standard deviation of $ln(T_i)$ and $ln(T_j)$, obtained from Eq. (7.27), after the calculation of the expected value and variance of T_i and T_j. Also, ρ_{ij} is the correlation coefficient between the random variables $ln(T_i)$ and $ln(T_j)$ which can be obtained using property 4, as:

$$\rho_{ij} = \frac{1}{s_{T_i}s_{T_j}} \ln\left(1 + \frac{\text{cov}\left(T_i, T_j\right)}{\exp\left\{m_{T_i} + m_{T_j} + \frac{s_{T_i}+s_{T_j}}{2}\right\}}\right) \tag{7.35}$$

where $\text{cov}(T_i, T_j) = S_T(i, j)$ has been previously obtained from the statistical thermal analysis part. Finally, The desired probability $P(T_i > T_j)$ can be numerically estimated from the given integral of Eq. (7.33).

7.5 Implementation, Results, and Discussions

To validate the analyzer, a power model based on the Alpha 21364 microprocessor is used [173]. The power consumption parameters of the processor's blocks were set based on the average power consumptions in 90*nm* technology when running MCF application [150, 157]. The nominal total leakage power is 33% of the total power consumption in this sample. It should be noted that the constant time of changes on temperature is orders of magnitude slower than the input vector transition rate (millisecond vs. nanosecond). As a result, the average power consumption of each block during a moderately long time can be used without worrying about the temporarily short transitions on instantaneous power. The general purpose Alpha processor benchmark is only used as a sample in which the power consumptions of its blocks are given when running an application (a long run of an instruction set). If another application is being used for such processor the new values of power consumptions should be fed into the model to generate the new statistical thermal and power information.

The ev6-like floorplan provided by the publicly available HotSpot tool [150] was considered. The packaging structure shown in Figure 7.2(b) was used which consists of a 50μm thermal

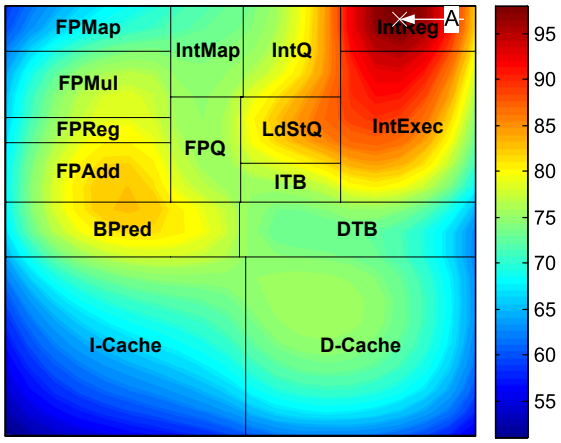
interface material over the $300\mu m$ die thickness. The aluminum heat spreader and heat sink have the dimensions (height×width×depth) of 30×30×1 and 60×60×6.9 millimeters. 35°C was assigned to the ambient air temperature of the case where the chip is supposed to work. The $3\sigma_L$ and $3\sigma_{T_{ox}}$ are set to 12% and 5%, respectively, of which the inter-die, spatially correlated intra-die, and residual variations consitute 25%, 55%, and 20% of total variations [85]. The elements of the covariance matrix (Ψ_X) are defined such that the correlation between grid *i* and *j* parameters follows the diminishing rate of the $exp(-bv_{ij})$ [85].

The analysis was done for the sample structure by meshing the area of the microprocessor into $n = 50 \times 50 = 2500$ grids. The runtime of the method developed in Matlab and executed on a Pentium IV, 3.4GHz, 2GB RAM PC was 158 seconds including admittance matrix inversion and all initialization steps for five iterations. Adding one extra iteration increases the runtime 33 seconds out of which 22.5 seconds are for evaluating Eq.(7.17 and 7.25) where the matrix multiplications are performed. The naive matrix multiplication approach $O(n^3)$ was used as the standard Matlab multiplication method. However, if the method had been developed in C with application of any of the fast matrix multiplication mentioned earlier, the runtime could have been improved more since the Eq.(7.17, 7.21, and 7.25) are the runtime bottleneck. The runtime complexity of the method for $m > 1$ iterations is $O((3m-1)n^3)$ based on the number of non-sparse matrix multiplications. The memory usage is $O(8n^2)$ to load Ψ_L, $\Psi_{T_{ox}}$, A matrices and update S_T, S_P, $E\left[\hat{I}_{leak-i}\hat{I}_{leak-j}\right]$, $E\left[T_i\hat{I}_{leak-i}\hat{I}_{leak-j}\right]$, and $E\left[T_i^2\hat{I}_{leak-i}\hat{I}_{leak-j}\right]$ matrices in each iteration.

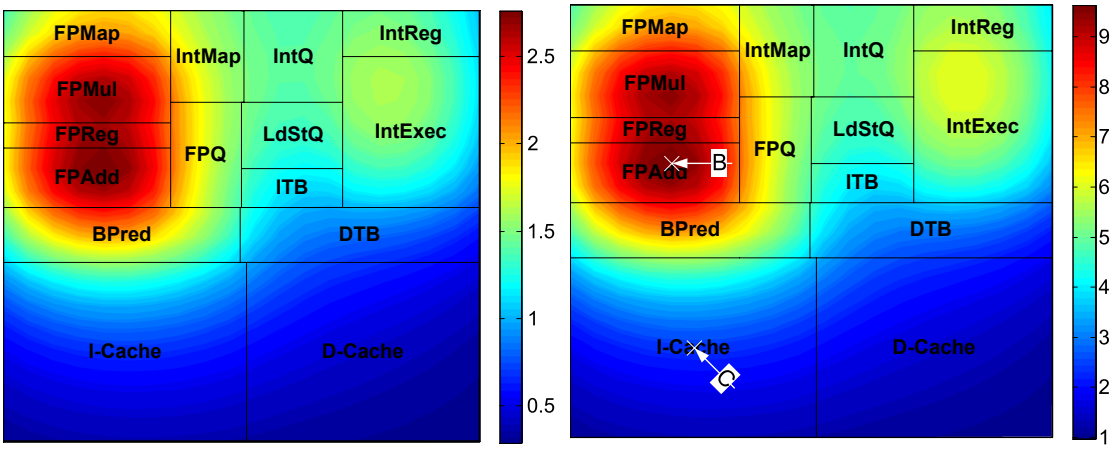
To verify the technique, Monte-Carlo iterative simulations considering leakage-thermal coupling have been done with 10000 samples over the HotSpot tool, which took almost 4 days using the same computer for the 2500 nodes case. Therefore, if the inverse admittance matrix is reused for the sequence of the Monte-Carlo simulations to avoid redundantly reconstruction of the inverse admittance matrix for each new sample, which reduced the Monte-Carlo runtime down to almost 42 minutes. However, the Monte-Carlo simulation runtime is still too high in comparison to our developed approach. We also performed the Monte-Carlo simulations with lower number of samples, to investigate how much it affects the accuracy of the results. It has been observed that a lower number (e.g., 1000) produces more than 14% error in standard deviation.

Figure 7.6 shows the results obtained from the analyzer for the sample core. The nominal thermal profile of the chip (considering leakage-thermal loop) is shown in Figure 7.6(a). This profile is obtained without considering any parameter variations (nominal). However, after considering process variability, the deviation profile of the expected value from the nominal temperatures, obtained from our method, is depicted in Figure 7.6(b) after five iterations. It can be seen that the level of increase in the expected value is up to 2.7°C, while, as shown in Figure 7.6(c), the standard deviations of the grids' temperature are widely varied from 1 to 9.6 °C. This indicates how much the temperature of each location in a die can vary from chip to chip after fabrication due to process-induced leakage variations.

It should be noted that, increasing the number of grids provides slight change in results only



(a) Nominal temperature

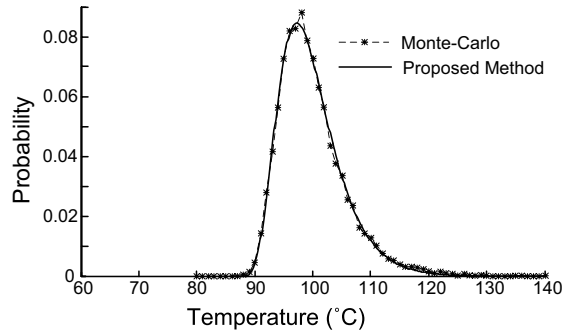


(b) Deviation of the expected value from the nominal temperature

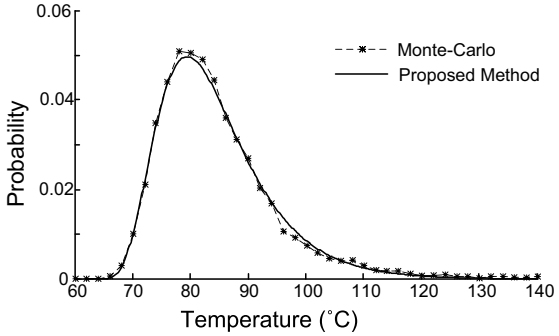
(c) Standard deviation of temperature

Figure 7.6: Statistical thermal profile of Alpha 21364 CPU core

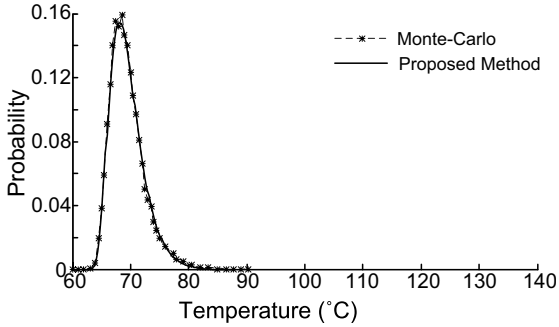
if the power density of the grids are given based on the grid-size resolution. Otherwise, for the case of the early stage analysis in which the power consumption of blocks is the highest information resolution in hand, increasing the number of grids does not lead to a significant difference. In this case, first the experiments with a 40×40 and then 50×50 grid structure are performed which ended up with up to 3% contrast in the standard deviation profile over blocks' borders, but not showing considerable benefit from going to 60×60 structure. However, another option for performing simulations with higher effective resolution while keeping the runtime tractable is to disctritize the die area non-uniformly in the block level. Since there is no need to discritize the large blocks (Caches) into many grids, therefore, different blocks would have the same number of grids and can provide the reasonable accuracy with lower number of grids.



(a) Node A: The nominally hottest point



(b) Node B: the highest thermal variable point



(c) Node C: A sample cold point (cache)

Figure 7.7: The obtained PDF from our method compared with the Monte-Carlo simulations

This idea was called hybrid-sized grid cells in the literature [5]. The proposed methodology can simply account for such structure by using the appropriate process covaraince matices (Ψ_X) based on the new structure.

It can be inferred, by investigating Eq. (7.19, 7.20, and 7.16), that the variance of temperature in grid i is a function of the nominal temperature (T_i^{nom}) and nominal leakage (λ_i') at that node. Therefore, both nominally high temperature die parts (parts with more activity) and high leakage die parts (usually high-performance parts) show more temperature variance. This fact can be seen by comparing the PDF of nodes A, B, and C. The estimated PDF of sample nodes A, the nominally hottest point, B, the highest thermal variable point which is over one of the high performance blocks, and C, the nominally very cold cache are depicted in Figure 7.7 along with the Monte-Carlo simulation results. Figure 7.7(b) shows a temperature range of 65-140°C for a node, which can bring a source of huge uncertainty in the power grid noise, circuit reliability, and timing/power characteristic of the designed circuit. Therefore, a design which shows satisfactory behavior during traditional (nominal) thermal analysis may fail after fabrication, considering the wide variation in the die temperature due to process variation. Finally, in terms of the accuracy, the error of the estimated expected values in comparison to Monte-Carlo simulations is less than

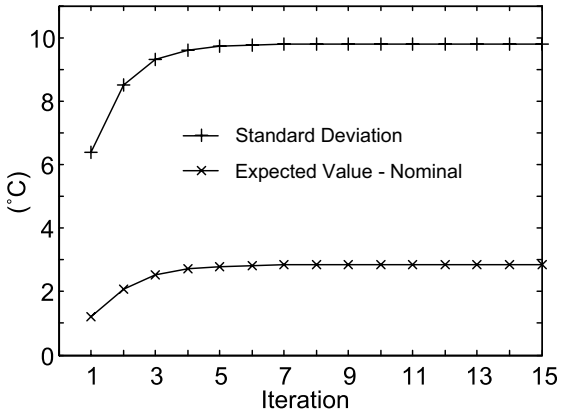


Figure 7.8: The standard deviation and expected value of node B's temperature in each iteration.

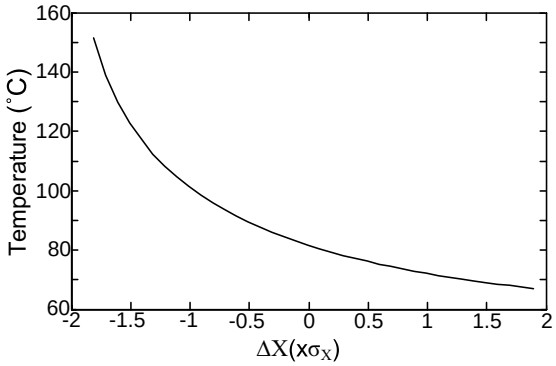


Figure 7.9: Corner based thermal extraction of node B. All ΔL_i and ΔT_{ox_i} variations are from $(-1.8\sigma_L, -1.8\sigma_{T_{ox}})$ to $(1.8\sigma_L, 1.8\sigma_{T_{ox}})$ simultaneously.

0.03% over the die. Also, the error of the estimated standard deviation is less than 2%. It should be noted that, at first, the random dopant fluctuation was also considered during thermal analysis. However, the results were the same as the case when it is ignored due to its uncorrelated nature. This is because adding large number ($n \to \infty$) of uncorrelated random variables, each with standard devation over mean of $\sigma/\mu = k$, leads to a random variable with zero standard devation over mean $\sigma/\mu = k/\sqrt{n} \to 0$.

Figure 7.8 shows the standard deviation and deviation of the expected value from nominal temperature of point B for each iteration. Results shown in the first iteration are the output of step 2, and the outputs of the step 3 are shown as the second iteration. The rests are the thermal statistical moments obtained from the step 4, iteratively. It can be seen that the results converge after almost five iterations with an acceptable accuracy.

Corner based thermal extraction was also performed to avoid possible pessimism in sta-

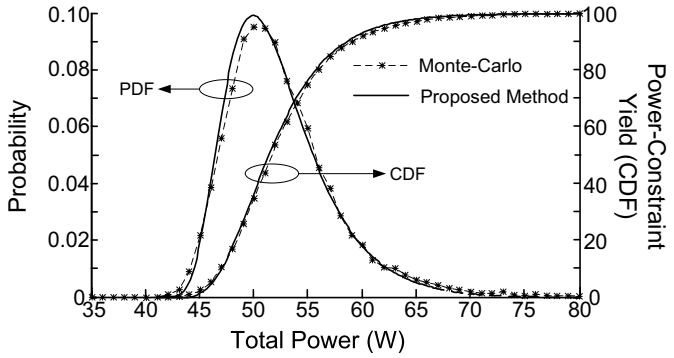


Figure 7.10: The obtained full-chip total power consumption PDF from our method compared with the Monte-Carlo simulations.

tistical estimations. The process parameters of all grids were set from $(-1.8\sigma_L,-1.8\sigma_{T_{ox}})$ to $(1.8\sigma_L,1.8\sigma_{T_{ox}})$ with $(0.1\sigma_L,0.1\sigma_{T_{ox}})$ steps, simultaneously, and the thermal profile was extracted for each case. As can be seen in Figure 7.9, the temperature of node B varies from 67-151°C. In fact, if all process parameters be less than $X_0 - 1.9\sigma_X$ the design will be too leaky and will experience thermal runaway ($T_i \rightarrow \infty$) [146]. However, the probability in which all (2×2500=5000) process parameters experience the worst case scenarios simultaneously is very low. The Monte-Carlo simulations showed 21 samples out of 10000 experiencing thermal runaway.

The power consumption uncertainty caused by the wide range of thermal variability due to process variations are quantified by using the analytical approach proposed in Section 7.4.1. The probability density function of the extracted full-chip total power is obtained in less than a second (after extraction of temperature moments) and are compared with the Monte-Carlo results in Figure 7.10. The total power consumption's expected value and standard deviation are 52.3 and 4.83 Watts. Moreover, based on a power consumption budget, one can find a probabilistic yield using Eq. (7.32) and the obtained total power PDF. This yield is also depicted and compared with the Monte-Carlo simulations in the figure.

In addition, for the hotspot movement evaluation, two grids of A and B are considered. It can be seen that the nominal temperature in grid A is 16.7°C higher than grid B when no variability is taken into account. However, to show the effect of the leakage variation on the hotspots formations, the $P(T_B > T_A)$ is calculated using the approach presented in Section IV. The analysis shows $P(T_B > T_A) = 4.28\%$ (Monte-Carlo result = 4.15%) which means 4.28% of fabricated dies experience higher temperature on grid B (FPAdd) rather than grid A (IntReg). This indicates that we should not only rely on the nominal location of a hotspot, but also examine other parts of a chip which has a high thermal variability. Therefore, quantative guidance can be provided to micro-architects regarding the hotspot locations to help them in devising thermal management solutions. The runtime of such evaluation was a fraction of a second given the previously estimated moments and covariances of temperatures.

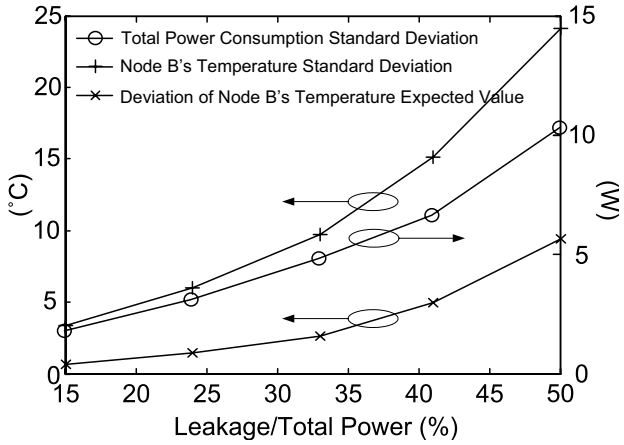
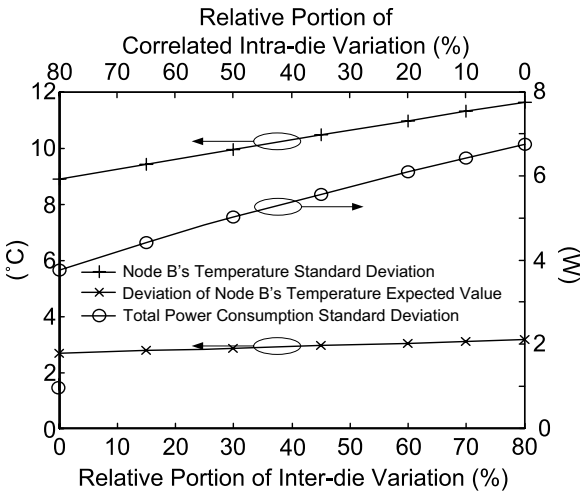


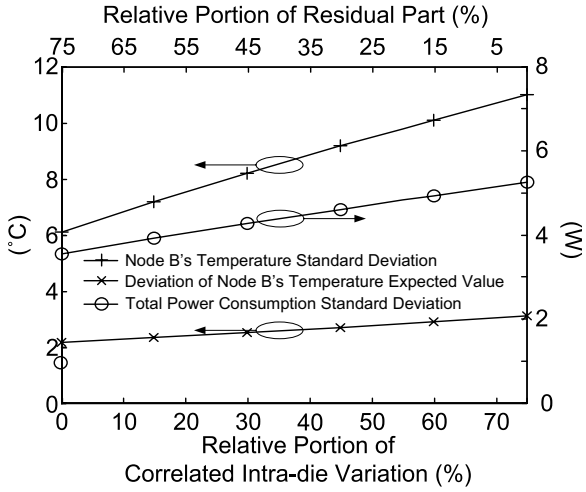
Figure 7.11: The total power consumption std and node B's temperature moments with respect to leakage/total power consumption ratio.

The behavior of temperature's moments and total power consumption are studied by varying the relative portion of the leakage/total power consumption. The nominal total power consumption was kept constant while the ratio of the leakage/total is changed from 15% to 50%. Figure 7.11 shows the standard deviation and deviation of the expected value of node B's temperature, the highest thermal variable point. It also shows the standard deviation of total power consumption when the ratio is varied. All values are obtained after the convergence of the step 4 calculations. It can be seen that, as expected, icreasing the leakage portion exacerbates the thermal uncertainty, and hence the total power consumption uncertainty, which emphasizes the importance of considering an statistical thermal analysis for scaled and leakier technologies.

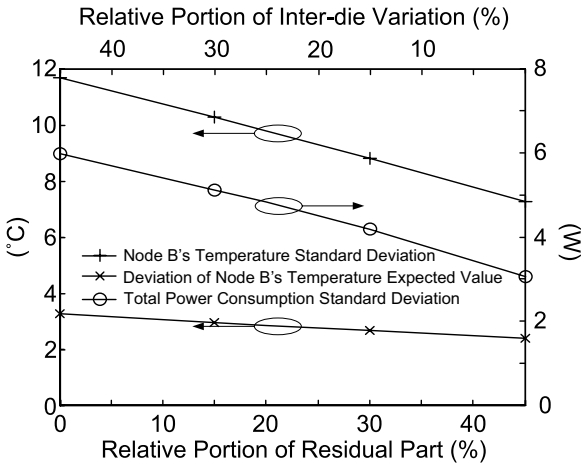
Moreover, to investigate how the relative magnitude of the inter, correlated intra die, and residual parts of variations affect the statistical behavior of temperature and total power consumption, the ratio of the inter-die variation is varied from 0% to 80% while the residual part is kept 20%. This means the correlated intra-die is varied from 80% to 0% (Fig. 7.12(a)). Also, in the second case the inter-die variation is kept constant to 25% while the correlated intra-die portion is varied from 0% to 75% which means the residual part is varied from 75% to 0% (Fig. 7.12(b)). Finally, in the last case the correlated intra-die part is kept constant 55% while the inter-die and residual parts are varied from 0% to 45% and 45% to 0%, respectively (Fig. 7.12(c)). The figures are obtained based on the leakage/total power ratio of 33%, after the iterations converged in step 4. As can be seen in figure 7.12(a) and 7.12(b), in constant residual part, more inter-die variation produces more thermal and power uncertainty. However, when the inter-die portion is constant more correlated intra-die variation brings more uncertainty. These are due to the fact that, in both cases, the physical parameters' covariances between grids were increased (please refer to Eq. (7.5)) which in fact increases the total uncertainty through the analysis. This is the same as the circuit delay uncertainty, in which the inter-die and spatial correlation (systematic



(a) Constant residual part (20%)



(b) Constant inter-die variation (25%)



(c) Constant correlated intra-die variation (55%)

Figure 7.12: The total power consumption standard deviation and node B's temperature moments with respect to relative portions of inter-die, correlated intra-die, and residual part variations.

intra-die) increase the overall variability [174, 175].

In closing, it can be concluded from the results that considering thermal uncertainty is a must for future VLSI design flow and should be considered in different applications. As shown in the plotted profiles, the magnitude of the temperature variation is not constant over a die suggesting to wisely investigate new placement techniques which not only target the minimization of the peak nominal temperature [160], but also optimize its variance. The leakage power reduction techniques (e.g., body biasing and supply gating) should also be more carefully utilized consid-

ering the bi-directional relation between leakage and temperature. In addition, the power rail analysis techniques have to be also modified in order to incorporate the thermal+leakage uncertainty information provided by the analyzer, to produce more reliable grid verification. Moreover, the electrothermal simulation techniques [156, 147] are other critical issues which have to be revisited since process-induced thermal uncertainty reveals more variations on delay and reliability (MTTF) than the past when process-independent temperatures were assumed. As a result, strong couplings between process variability, temperature, and leakage currents make the co-analyzing for (process variations/ leakage/ temperature) critical for future robust circuit designs.

7.6 Conclusions

In this chapter, a statistical temperature profile analyzer is proposed that estimates the probability density functions and covariances of temperatures over a die. The statistical behavior of temperature arises from the variable nature of leakage current due to physical parameter variabilities. The inter-die and spatially correlated intra-die gate length and oxide thickness variations are considered in this micro-architectural level analysis approach. The thermal dependence of leakage is also taken into account during estimations. Finally, as two applications of the extracted statistical moments, the migration of hotspots which appears when considering the variations, and the full chip total power consumption estimation are developed. Analysis done over the sample layout (Alpha 21364) showed that the temperature variances widely vary (1-9.6°C) over the blocks which can produce a temperature range of (65-140°C) on a location of the die and hence impact power/performance/reliability metrics.

Part IV

Thesis Closure

Chapter 8

Conclusions

In this thesis, computer-aided design methodologies are proposed to address some of the process variation concerns at the device, circuit, and micro-architectural levels.

At the device-level, it is shown that a methodology, enforcing variability minimization during the MOS device design process results in devices that are more immune to physical variations than traditionally designed devices. A theoretical study of various device parameters and their impacts on device characteristics are presented. An MOS device design approach is developed which finds appropriate values for oxide thickness, gate length, and channel doping profile characteristics (Halo and Retrograde Well) for a given MOS device structure and technology such that the extracted device parameters leads to a transistor which maximally satisfies three desired constraints on intrinsic delay, saturation, and total leakage currents, in the presence of variability. The algorithm is based on an optimization technique which places a maximized yield cube in the problem feasible space. The center of this cube is considered as the maximum yield design point. This method takes into account different possible variances on process parameters and desired performance-leakage metrics for a particular application. The designed devices are verified by comparing against some industrial devices and the the semiconductor roundmap. It is, therefore, concluded that the variability can be effectively considered from the device design.

At the circuit-level, advanced sampling and variance reduction-based methods (e.g., QMC, LHS, Control Variate, and Importance Sampling) are developed for the efficient yield estimation of digital, analog, and SRAM cells. The yield estimation of integrated circuits through the Monte-Carlo technique is inefficient. However, it is shown in this thesis that by proper engineering of the problems, the proposed MC-based methods are capable of providing an accurate yield estimation by using a low number of simulations, compared to that of the traditional-MC. Three types of VLSI circuits, the digital, analog, and SRAM cells are considered and different solutions are proposed for each. For the digital circuits statistical timing analysis problem, the fact that the timing yield problem contains some considerable higher that 1-D terms in its ANOVA decomposition is used toward improving the discrepancy of the Sobol's Quasi-MC sampling. This

problem is also shown to be a very suitable candidate for the application of the control variate by the extraction of the nominally critical path. However, for the analog circuits using a control variable requires extra overhead for the CV model training, lead us to use sampling-based methods such LHS for yield analysis. The linearity of the analog circuits are then used toward minimizing the inter-linear and quadratic correlation of the LHS samples, for a reduced variance estimation of the analog circuit yield. While the digital and analog circuit yield estimation problems suffer from the curse of dimensionality, the MC-based SRAM cell yield problem is challenging due to high variance of estimation as the failure rate is extremely low. As a result, an adaptive importance sampling is developed to provide an immune and accurate method of yield estimation with just a few thousand simulations.

Finally, a co-thermal-leakage analysis engine is developed at the micro-architectural level that accounts for an uncertain thermal profile due to process-induced leakage variations. The analysis is based on iterative calculation of the statistical thermal and leakage moments, and matching them into a shifted log-normal distribution. It is shown how this information can be used for the full-chip leakage power yield estimation, and investigation of the formation of the thermal hotspots.

Following is the list of related publications:

J1. J. Jaffari and M. Anis, "On Efficient LHS-Based Yield Analysis of Analog Circuits ", *Accepted by IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*.

J2. J. Jaffari and M. Anis, "Advanced Variance Reduction and Sampling Techniques for Efficient Statistical Timing Analysis", *Accepted by IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*.

J3. J. Jaffari and M. Anis, "Statistical Thermal Profile Considering Process Variations: Analysis and Applications", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 27, pp. 1027-1040, June 2008.

J4. J. Jaffari and M. Anis, "Variability-Aware Bulk-MOS Device Design", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 27, pp. 205-216, February 2008.

C1. J. Jaffari and M. Anis, "Correlation Controlled Sampling for Efficient Variability Analysis of Analog Circuits", *Proc. of IEEE/ACM Design Automation and Test in Europe* (DATE), pp. 1305-1308, 2010.

C2. J. Jaffari and M. Anis, "Practical Monte-Carlo Based Timing Yield Estimation of Digital Circuits", *Proc. of IEEE/ACM Design Automation and Test in Europe* (DATE), pp. 807-812, 2010.

C3. J. Jaffari and M. Anis, "Adaptive Sampling for Efficient Failure Probability Analysis of SRAM Cells", *Proc. of IEEE/ACM International Conference on Computer-Aided Design* (ICCAD), pp. 623-630, 2009.

C4. J. Jaffari and M. Anis, "Timing Yield Estimation of Digital Circuits using a Control Variate Technique", *Proc. of IEEE International Symposium on Quality Electronic Design* (ISQED), pp. 382-387, 2009.

C5. J. Jaffari and M. Anis, "On Efficient Monte Carlo-Based Statistical Static Timing Analysis of Digital Circuits", *Proc. of IEEE/ACM International Conference on Computer-Aided Design* (ICCAD), pp. 196-203, 2008.

C6. J. Jaffari and M. Anis, "Variability-Aware Device Optimization under I_{ON} and Leakage Current Constraints", *Proc. of IEEE/ACM International Symposium on Low Power Electronics and Design* (ISLPED), pp. 119-122,2006.

8.1 Future Works

Suggestions to extend the research of this thesis at different levels are listed as follows:

At the device-level, a methodology is developed to calculate the optimum doping and geometry parameters from the device yield point of view. However, to achieve certain physical characteristics at the device-level, the fabrication process should be tuned accordingly. A methodology that optimizes the fabrication process parameters (e.g. annealing, ionization, patterning and etching parameters) and targets the yield of the device, as defined in this thesis, is very helpful.

At the circuit-level, considering the promising advances in efficient MC-based yield estimation methods, an approach for future research should be on the design for maximizing the yield by using sampling-based methods as the core of the yield analysis. For example, in the 6T-SRAM cell design problem, a methodology that can progressively optimize the dimensions of the six transistors, during the progresses of the adaptive importance sampling yield estimation is very valuable. Such methods combine the design and the analysis of the yield in a unified flow rather than a very time-consuming iterative design-and-correct cyclic approach. A progressive update of the response surfaces during the sampling-based analysis, and a resultant design for the yield of analog and digital circuits should also be considered in the future.

Finally, at the micro-architectural-level, bringing the knowledge of the co-thermal-leakage statistics to the power grid verification, and statistically modeling the IR-drop profile might be a suitable direction to follow. The leakage and thermal variations introduce current and resistivity variations that are the two major contributors to total IR-drop variability on the power distribution network.

References

[1] L. R. Harriott, "Limits of lithography," *Proceedings of the IEEE*, vol. 89, pp. 366–374, Mar. 2001.

[2] P. Gupta, A. B. Kahng, C. H. Park, K. Samadi, and X. Xu, "Wafer topography-aware optical proximity correction," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 25, pp. 2747–2756, Dec. 2006.

[3] K. Bernstein, D. J. Frank, A. E. Gattiker, W. Haensch, B. L. Ji, S. R. Nassif, E. J. Nowak, D. J. Pearson, and N. J. Rohrer, "High-performance cmos variability in the 65-nm regime and beyond," *IBM Journal of Research and Development*, vol. 50, pp. 433–449, July-September 2006.

[4] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De., "Parameter variations and impact on circuits and microarchitecture," in *Proc. of IEEE/ACM Design Automation Conference*, 2003, pp. 338–342.

[5] W. Huang, S. Ghosh, S. Velusamy, K. Sankaranarayanan, K. Skadron, and M. R. Stan, "HotSpot: A compact thermal modeling methodology for early-stage VLSI design," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 14, pp. 501–513, May 2006.

[6] S. R. Nassif, "Delay variability: sources, impacts and trends," in *Proc. of IEEE International Solid State Circuits Conference*, 2000, pp. 368–369.

[7] T. A. Tugbawa, "Chip-scale modeling of pattern dependencies in copper chemical mechanical polishing processes," Ph.D. dissertation, Massachusetts Institute Of Technology, 2002.

[8] D. Ouma, D. Boning, J. Chung, W. Easter, V. Saxena, S. Misra, and A. Crevasse, "Characterization and modeling of oxide chemical-mechanical polishing using planarization length and pattern density concepts," *IEEE Transactions on Semiconductor Manufacturing*, vol. 15, pp. 232–244, May 2002.

[9] X. Tang, V. K. De, and J. D. Meindl, "Intrinsic mosfet parameter fluctuations due to random dopant placement," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 5, pp. 369–376, Dec. 1997.

[10] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*. Cambridge University Press, 1998.

[11] International Technology Roadmap for Semiconductors, 2005.

[12] Y. C. P. A. B. Kahng, "Subwavelength lithography and its potential impact on design and eda," in *Proc. of IEEE/ACM Design Automation Conference*, 1999, pp. 799–804.

[13] K. A. Bowman, S. G. Duvall, and J. D. Meindl, "Impact of die-to-die and within-die parameter fluctuations on the maximum clock frequency distribution for gigascale integration," *IEEE Journal of Solid-State Circuits*, vol. 37, pp. 183–190, Feb. 2002.

[14] R. Rao, A. Devgan, D. Blaauw, and D. Sylvester, "Parametric yield estimation considering leakage variability," in *Proc. of IEEE/ACM Design Automation Conference*, 2004, pp. 442–447.

[15] B. D. Cory, R. Kapur, and B. Underwood, "Speed binning with path delay test in 150-nm technology," *Proceedings of the IEEE*, vol. 20, pp. 41–45, Sep-Oct 2003.

[16] D. Schroder and J. Babcock, "Negative bias temperature instability: Road to cross in deep submicron silicon semiconductor manufacturing," *Journal of Applied Physics*, vol. 94, pp. 1–18, July 2003.

[17] K. Roy, S. Mukhopadhyay, and H. M. Meymand, "Leakage current mechanisms and leakage reduction technique in deep-submicrometer cmos circuits," *Proceedings of the IEEE*, vol. 91, pp. 305–327, Feb. 2003.

[18] C. S. Murthy and M. Gall, "Process variation effects on circuit performance: TCAD simulation of 256-MBit technology," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 16, pp. 1383–1389, Nov. 1997.

[19] S. R. Nassif, "Design for variability in DSM technologies," in *Proc. of IEEE International Symposium on Quality Electronic Design*, 2000, pp. 451–454.

[20] A. Srivastava, R. Bai, D. Blaauw, and D. Sylvester, "Modeling and analysis of leakage power considering within-die process variations," in *Proc. of IEEE/ACM International Symposium on Low-Power Electronics*, 2002, pp. 64–67.

[21] T. Chen, "Where CMOS is going: Trendy hype vs. real technology," in *Proc. of IEEE International Solid State Circuits Conference*, 2006, pp. 1–18.

[22] M. Anis and M. H. Aburahma, “Leakage current variability in nanometer technologies,” in *Proc. of IEEE International Workshop System-on-Chip for Real-Time Applications*, 2005, pp. 60–63.

[23] S. Mukhopadhyay and K. Roy, “Modeling and estimation of total leakage current in nano-scaled cmos devices considering the effect of parameter variation,” in *Proc. of IEEE/ACM International Symposium on Low-Power Electronics*, 2003, pp. 172–175.

[24] D. J. Frank, W. Haensch, G. Shahidi, and O. H. Dokumaci, “Optimizing CMOS technology for maximum performance,” *IBM Journal of Research and Development*, vol. 50, pp. 419–431, July 2006.

[25] S. Thompson, P. Packan, and M. Bohr, “MOS scaling: Transistor challenges for the 21st century,” *Intel Technology Journal*, Q3 1998.

[26] S. Chung and C. Li, “An analytical threshold-voltage model of trench-isolated MOS devices with nonuniformly doped substrates,” *IEEE Transactions on Electron Devices*, vol. 39, pp. 614–622, Mar. 1992.

[27] Z. Liu, C. Hu, J. Huang, T. Chan, M. Jeng, P. KO, and Y. Cheng, “Threshold voltage model for deep-submicrometer MOSFET’s,” *IEEE Transactions on Electron Devices*, vol. 40, pp. 86–95, Jan. 1993.

[28] B. Sheu, D. Scharfetter, P. Ko, and M. Jeng, “BSIM: Berkeley short-channel IGFET model for MOS transistors,” *IEEE Journal of Solid-State Circuits*, vol. 22, pp. 558–566, Aug. 1987.

[29] M. Rodder, M. Hanratty, D. Rogers, T. Laaksonen, J. Hu, S. Murtaza, and C. Chao, “A 0.1μm gate length CMOS technology with 30A gate dielectric for 1.0V–1.4V applications,” in *Proc. of IEEE International Electron Devices Meeting*, 1997, pp. 223–226.

[30] R. Gwoziecki and T. Skotnicki, “Smart pockets - total suppression of roll-off and roll-up,” in *Proc. of IEEE Symposium on VLSl Technology, Digest of Technical Papers*, 1999, pp. 91–92.

[31] C. Wu and et al., “A 90-nm CMOS device technology with high-speed, general-purpose, and low-leakage transistors for system on chip applications,” in *Proc. of IEEE International Electron Devices Meeting*, 2002, pp. 65–68.

[32] K. Schuegraf and C. Hu, “Hole injection SiO2 breakdown model for very low voltage lifetime extrapolation,” *IEEE Transactions on Electron Devices*, vol. 41, pp. 761–767, May 1994.

[33] A. Srivastava and C.M.Osburn, "Response surface based optimization of 0.1 μm PMOSFETs with ultra-thin oxide dielectrics," in *Proc. of SPIE Conference on Microelectronic Device Technology*, 1998, pp. 253–264.

[34] A. Pouydebasque, M. Muller, F. Boeuf, D. Lenoble, F. Lallement, A. Grouillet, A. Halimaoui, R. E. Farhane, D. Delille, and T. Skotnicki, "Improved Vt and Ioff characteristics of NMOS transistors featuring ultra-shallow junctions obtained by plasma doping (PLAD)," in *Proc. of IEEE European Solid-State Device Research Conference*, 2003, pp. 35–38.

[35] C. M. Osburn, I. De, K. F. Yee, and A. Srivastava, "Design and integration considerations for end-of-the roadmap ultrashallow junctions," *Journal of Vacuum Science and Technology B - Microelectronics and Nanometer Structures*, vol. 18, pp. 338–345, Jan. 2000.

[36] S. Sahaa, "Scaling considerations for high performance 25 nm metal–oxide–semiconductor field effect transistors," *Journal of Vacuum Science and Technology B - Microelectronics and Nanometer Structures*, vol. 19, pp. 2240–2246, Nov. 2001.

[37] M. Ono, M. Saito, T. Yoshitomi, and C. Fiegna, "Sub-50 nm gate length N-MOSFETS with 10 nm phosphorus source and drain junctions," in *Proc. of IEEE International Electron Devices Meeting*, 1993, pp. 119–122.

[38] M. Müller and M. Bidaud, "Advanced junction engineering for 60nm-CMOS transistors," in *Proc. of IEEE European Solid-State Device Research Conference*, 2002, pp. 315–318.

[39] K. Lee, J. Si, Y. Li, W. Kang, R. Malik, R. Rengarajan, S. Chaloux, J. Bernstein, and P. Kellerman, "Shallow n^+/p^+ junction formation using plasma immersion ion implantation for CMOS technology," in *Proc. of IEEE Symposium on VLSl Technology, Digest of Technical Papers*, 2001, pp. 21–22.

[40] M. Y. Kwong, R. Kasnavi, P. Griffin, J. D. Plummer, and R. W. Dutton, "Impact of lateral Source/Drain abruptness on device performance," *IEEE Transactions on Electron Devices*, vol. 49, pp. 1882–1890, Nov. 2002.

[41] T. Ghani, K. Mistry, P. Packan, S. Thompson, M. Stettler, S. Tyagi, and M. Bohr, "Scaling challenges and device design requirements for high performance sub-50 nm gate length planar CMOS transistors," in *Proc. of IEEE Symposium on VLSl Technology, Digest of Technical Papers*, 2000, pp. 174–175.

[42] R. Gwoziecki, T. Skotnicki, P. Bouilon, and A. Poncet, "Junction design guideline for 0.18μm CMOS," in *Proc. of IEEE European Solid-State Device Research Conference*, 1997, pp. 388–391.

[43] S. Mukhopadhyay, A. Raychowdhury, and K. Roy, "Accurate estimation of total leakage in nanometer-scale bulk cmos circuits based on device geometry and doping profile," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 24, pp. 363–381, Mar. 2005.

[44] X. Xi, M. Dunga, J. He, W. Liu, . K. M. Cao, X. Jin, J. J. Ou, M. Chan, A. M. Niknejad, and C. Hu, "BSIM4.5 MOSFET model," [Online]. Available: http://www-device.eecs.berkeley.edu/ bsim3/bsim4.html.

[45] S. Ogura, C. F. Codella, N. Rovedo, J. F. Shepard, and J. Riseman, "A half micron MOSFET using double implanted LDD," in *Proc. of IEEE International Electron Devices Meeting*, 1982, pp. 718–722.

[46] T. Hori, "A 0.1–μm CMOS technology with Tilt–Implanted Punchthrough Stopper (TIPS)," in *Proc. of IEEE International Electron Devices Meeting*, 1994, pp. 75–78.

[47] Y. Taur and E. Nowak, "CMOS devices below 0.1 pm: How high will performance go?" in *Proc. of IEEE International Electron Devices Meeting*, 1997, pp. 215–218.

[48] S. Venkatesan, J. W. Lutze, C. Lage, and W. J. Taylor, "Device drive current degradation observed with retrograde channel profiles," in *Proc. of IEEE International Electron Devices Meeting*, 1995, pp. 419–422.

[49] S. E. Thompson, P. A. Packan, and M. T. Bohr, "Linear versus saturated drive current: Tradeoffs in super steep retrograde well engineering," in *Proc. of IEEE Symposium on VLSI Circuits, Digest of Technical Papers*, 1996, pp. 154–155.

[50] D. A. Antoniadis, I. J. Djomehri, K. M. Jackson, and S. Miller, "Well-Tempered Bulk-Si NMOSFET device home page," [Online]. Available: http://www-mtl.mit.edu/researchgroups/Well/.

[51] P. Kumaraswamy, "A generalized probability density function for double-bounded random processes," *Elsevier Journal of Hydrology*, vol. 46, pp. 79–88, Mar. 1980.

[52] K. Ponnambalam, A. Seifi, and J. Vlach, "Probabilistic design of systems with general distributions of parameters," *International Journal of Circuit Theory and Applications*, vol. 29, pp. 527–536, Aug. 2001.

[53] T. F. Colemen and Y. Zhang, *Optimization Toolbox for use with Matlab*. Natick MA: The Mathworks Inc., 2005.

[54] G. S. Samudra, H. Chen, D. Chan, and Y. Ibrahim, "Yield optimization by design centering and worst-case distance analysis," in *Proc. of IEEE International Conference on Computer Design*, 1999, pp. 289–290.

[55] J. Jaffari and M. Anis, "Variability-aware device optimization under I_{ON} and leakage current constraints," in *Proc. of IEEE/ACM International Symposium on Low-Power Electronics*, 2006, pp. 119–122.

[56] Synopsys Taurus MEDICI Version V-2003.12, 2003.

[57] M. Darwish, J. Lentz, M. Pinto, P. Zeitzoff, T. Krutsick, and H. Vuong, "An improved electron and hole mobility model for general purpose device simulation," *IEEE Transactions on Electron Devices*, vol. 44, pp. 1529–1538, Sept. 1997.

[58] E. Kane, "Zener tunneling in semiconductors," *Journal of Physics and Chemistry of Solids*, vol. 12, pp. 181–188, Jan. 1960.

[59] W. Harrison, "Tunneling from an independent-particle point of view," *APS Physical Review*, vol. 123, pp. 85–89, July 1961.

[60] P. M. Zeitzoff, A. F. Tasch, W. E. Moore, S. A. Khan, and D. Angelo, "Modeling of manufacturing sensitivity and of statistically based process control requirements for a $0.18\mu m$ NMOS device," in *Proc. of International Conference on Characterization and Metrology for ULSI Technology*, 1998, pp. 73–81.

[61] M. D. McKay, W. J. Conover, and R. J. Beckman, "A comparison of three methods for selecting values of input variables in the analysis of output from a computer code," *Technometrics*, vol. 42, pp. 55–61, Feb. 2000.

[62] M. Stein, "Large sample properties of simulations using latin hypercube sampling," *American Statistical Association*, vol. 29, pp. 143–151, May 1987.

[63] M. Keramat and R. Kielbasa, "Latin hypercube sampling monte carlo estimation of average quality index for integrated circuits," *Analog Integrated Circuits and Signal Processing*, vol. 14, pp. 131–142, Sept. 1997.

[64] S. Ogura, C. F. Codella, N. Rovedo, J. F. Shepard, and J. Riseman, "Random number generation and quasi-monte carlo methods," in *SIAM CBMS-NSF Regional Conference Series in Applied Math*, 1992.

[65] R. Caflisch, W. Morokoff, and A. Owen, "Valuation of mortgage backed securities using brownian bridges to reduce effective dimension," *Journal of Computational Finance*, vol. 1, pp. 27–46, Aug. 1997.

[66] X. Wang and K. T. Fang, "The effective dimension and quasi-monte carlo integration," *Journal of Complexity*, vol. 19, pp. 101Ű–124, Apr. 2003.

[67] X. Wang and I. H. Sloan, "Why are high-dimensional finance problems often of low effective dimension?" *SIAM Journal on Scientific Computing*, vol. 27, pp. 159–183, 2005.

[68] W. J. Morokoff and R. E. Caflisch, "Quasi-random sequences and their discrepancies," *SIAM Journal on Scientific Computing*, vol. 15, pp. 1251–1279, July 1994.

[69] S. S. Lavenberg and P. D. Welch, "A perspective on the use of control variables to increase the efficiency of monte carlo simulations," *Management Science*, vol. 27, pp. 322–335, Mar. 1981.

[70] D. E. Hocevar, M. R. Lightner, and T. N. Trick, "A study of variance reduction techniques for estimating circuit yields," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. CAD-2, pp. 180–192, July 1983.

[71] J. C. Hsu and B. L. Nelson, "Control variates for quantile estimation," *Management Science*, vol. 36, pp. 835–851, Mar. 1990.

[72] R. Kanj, R. Joshi, and S. Nassif, "Mixture importance sampling and its application to the analysis of sram designs in the presence of rare failure events," in *Proc. of IEEE/ACM Design Automation Conference*, 2006, pp. 69–72.

[73] S. M. Ross, *Simulation*. Academic Press, 2006.

[74] V. Veetil, D. Sylvester, and D. Blaauw, "Efficient monte carlo based incremental statistical timing analysis," in *Proc. of IEEE/ACM Design Automation Conference*, 2008, pp. 676–681.

[75] M. Keramat and R. Kielbasa, "A study of stratified sampling in variance reduction techniques for parametric yield estimation," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 45, pp. 575–583, May 1998.

[76] D. Blaauw, K. Chopra, A. Srivastava, and L. Scheffer, "Statistical timing analysis: from basic principles to state-of-the-art," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 27, pp. 589–607, Apr. 2008.

[77] L. Scheffer, "The count of monte carlo," in *IEEE TAU*, 2004.

[78] S. Tasiran and A. Demir, "Smart monte carlo for yield estimation," in *IEEE TAU*, 2006.

[79] A. Singhee, S. Singhal, and R. A. Rutenbar, "Practical, fast monte carlo statistical static timing analysis: Why and how?" in *Proc. of IEEE/ACM International Conference on Computer-Aided Design*, 2008, pp. 190–195.

[80] J. F. Traub and A. G. Werschulz, *Complexity and information*. Cambridge University Press, 1998.

[81] I. M. Sobol, "The distribution of points in a cube and the approximate evaluation of integrals," *USSR Computational Mathematics and Mathematical Physics*, vol. 7, pp. 86–112, 1967.

[82] P. Bratley and B. L. Fox, "Algorithm 659: Implementing sobol's quasirandom sequence generator," *ACM Transactions on Mathematical Software*, vol. 14, pp. 88–100, Mar. 1988.

[83] F. Brglez and H. Fujiwara, "A neutral netlist of combinational benchmark circuits and a target translator in fortran," in *Proc. of IEEE International Symposium on Circuits and Systems*, 1985.

[84] F. Brglez, D. Bryan, and K. Koiminski, "Combinational profiles of sequential benchmark circuits," in *Proc. of IEEE International Symposium on Circuits and Systems*, 1989, pp. 1929–1934.

[85] J. Xiong, V. Zolotov, and L. He, "Robust extraction of spatial correlation," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 26, pp. 619–631, Apr. 2007.

[86] P. Friedberg, Y. Cao, J. Cain, R. Wang, J. Rabaey, and C. Spanos, "Modeling within-field gate length spatial variation for process-design co-optimization," in *Proc. of SPIE*, 2005, pp. 178–188.

[87] A. Agarwal, D. Blaauw, V. Zolotov, S. Sundareswaran, M. Zhao, K. Gala, and R. Panda, "Statistical delay computation considering spatial correlations," in *Proc. of IEEE/ACM Asia and South Pacific Design Automation Conference*, 2003, pp. 271–276.

[88] "Capo: A large-scale fixed-die placer from ucla," [Online]. Available: http://vlsicad.ucsd.edu/GSRC/bookshelf/Slots/Placement.

[89] J. H. Halton, "On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals," *Numerische Mathematik*, vol. 2, pp. 84–90, 1960.

[90] B. L. Fox, "Algorithm 647: Implementation and relative efficiency of quasirandom sequence generators," *ACM Transactions on Mathematical Software*, vol. 12, pp. 362–376, Dec. 1986.

[91] H. Niederreiter and C. Xing, "Low-discrepancy sequences and global function fields with many rational places," *Finite Fields and Their Applications*, vol. 2, pp. 241–273, July 1996.

[92] I. H. Sloan and H. Wozniakowski, "When are quasi-monte carlo algorithms efficient for high dimensional integrals?" *Journal of Complexity*, vol. 14, pp. 1–33, 1998.

[93] A. Papageorgiou, "Sufficient conditions for fast quasi-monte carlo convergence," *Journal of Complexity*, vol. 19, pp. 332–351, June 2003.

[94] C. LeMieux and A. Owen, "Quasi-regression and the relative importance of the anova components of a function," in *Monte Carlo and Quasi-Monte Carlo Methods*, 2002, pp. 331–344.

[95] A. Jian and A. Owen, "Quasi-regression," *Journal of Complexity*, vol. 17, pp. 588Ű–607, Dec. 2001.

[96] H. Chang and S. Sapatnekar, "Statistical timing analysis under spatial correlations," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 24, pp. 1467–1482, Sept. 2005.

[97] H. Niederreiter, "Point sets and sequences with small discrepancy," *Monatshefte für Mathematik*, vol. 104, pp. 273–337, Dec. 1987.

[98] D. E. Knuth, *The art of computer programming, vol. 2: Seminumerical Algorithms*. Addison-Wesley, 1981.

[99] P. Jaeckel, *Monte Carlo methods in finance*. Wiley, 2002.

[100] I. M. Sobol, "Uniformly distributed sequences with an additional uniform property," *USSR Computational Mathematics and Mathematical Physics*, vol. 16, pp. 236–242, 1976.

[101] J. Cheng and M. J. Druzdzel, "Computational investigation of low-discrepancy sequences in simulation algorithms for bayesian networks," in *Proc. of Annual Conference on Uncertainty in Artificial Intelligence*, 2000, pp. 72Ű–81.

[102] H. Niederreiter, "Constructions of (t,m,s)-nets and (t,s)-sequences," *Finite Fields and Their Applications*, vol. 11, pp. 578–Ű600, Jan. 2005.

[103] H. S. Hong and F. J. Hickernell, "Algorithm 823: Implementing scrambled digital sequences," *ACM Transactions on Mathematical Software*, vol. 29, pp. 95–109, June 2003.

[104] H. A. David, *Order statistics*. John Wiley, 1981.

[105] A. Azzalini, "A class of distributions which includes the normal ones," *Scandinavian Journal of Statistics*, vol. 12, no. 2, pp. 171–178, 1985.

[106] X. Li, J. Le, P. Gopalakrishnan, and L. T. Pileggi, "Asymptotic probability extraction for nonnormal performance distributions," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 26, pp. 16–37, Jan. 2007.

[107] P. R. Kinget, "Device mismatch and tradeoffs in the design of analog circuits," *IEEE Journal of Solid-State Circuits*, vol. 40, pp. 1212–1224, June 2005.

[108] M. J. M. Pelgrom, A. C. J. Rens, M. Vertregt, and M. B. Dijkstra, "A 25-ms/s 8-bit cmos ad converter for embedded application," *IEEE Journal of Solid-State Circuits*, vol. 29, pp. 879–886, Aug. 1994.

[109] J. Bastos, A. M. Marques, M. S. J. Steyaert, and W. Sansen, "A 12-bit intrinsic accuracy high-speed cmos dac," *IEEE Journal of Solid-State Circuits*, vol. 33, pp. 1959–1969, Dec. 1998.

[110] S. J. Lovett, G. A. Gibbs, and A. Pancholy, "Yield and matching implications for static ram memory array sense amplifier design," *IEEE Journal of Solid-State Circuits*, vol. 35, pp. 1200–1204, May 2000.

[111] A. Balankutty, T. C. Chih, C. Y. Chen, and P. R. Kinget, "Mismatch characterization of ring oscillators," in *Proc. of IEEE Custom Integrated Circuits Conference*, 2007, pp. 515–518.

[112] J. Oehm and U. Gruenebaum, "Statistical analysis and optimization of a bandgap reference for vlsi applications," *Analog Integrated Circuits and Signal Processing*, vol. 29, pp. 213–220, Dec. 2001.

[113] M. Conti, P. Crippa, S. Orcioni, and C. Turchetti, "Parametric yield formulation of mos ic's affected by mismatch effect," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 18, pp. 582–596, May 1999.

[114] J. Oehm and K. Schumacher, "Quality assurance and upgrade of analog characteristics by fast mismatch analysis option in network analysis environment," *IEEE Journal of Solid-State Circuits*, vol. 28, pp. 865–871, July 1993.

[115] P. Bratley, B. J. Fox, and L. E. Schrage, *A guide to simulation*. Springer, 1987.

[116] J. F. Kenney and E. S. Keeping, *Mathematics of statistics*. Princeton, NJ: Van Nostrand, 1951.

[117] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. New York: Dover, 1972.

[118] C. Rose and M. D. Smith, *Mathematical statistics with mathematica*. New York: Springer-Verlag, 2002.

[119] H. Masuda, S. Ohkawa, A. Kurokawa, and M. Aoki, "Challenge: variability characterization and modeling for 65 to 90-nm processes," in *Proc. of IEEE Custom Integrated Circuits Conference*, 2005, pp. 593–599.

[120] B. Razavi, *Design of analog CMOS integrated circuits*. McGraw Hill, 2000.

[121] A. B. Owen, "Controlling correlations in latin hypercube samples," *Journal of the American Statistical Association*, vol. 89, pp. 1517–1522, Dec. 1994.

[122] K. Iwase and K. Kanefuji, "Estimation for 3-parameter lognormal distribution with unknown shifted origin," *Statistical Papers*, vol. 35, pp. 81–90, Dec. 1994.

[123] A. Genz, "Numerical computation of multivariate normal probabilities," *Journal of Computational and Graphical Statistics*, vol. 1, pp. 141–149, June 1992.

[124] R. V. Joshi, S. Mukhopadhyay, D. W. Plass, Y. H. Chan, , and C. C.-T. A. Devgan, "Variability analysis for sub-100 nm pd/soi cmos sram cell," in *Proc. of IEEE European Solid-State Circuits Conference*, 2004, pp. 211–214.

[125] R. Heald and P. Wang, "Variability in sub-100nm sram designs," in *Proc. of IEEE/ACM International Conference on Computer Aided Design*, 2004, pp. 347–352.

[126] A. J. Bhavnagarwala, X. Tang, and J. D. Meindl, "The impact of intrinsic device fluctuations on cmos sram cell stability," *IEEE Journal of Solid-State Circuits*, vol. 36, pp. 658–665, Apr. 2001.

[127] S. Mukhopadhyay, H. Mahmoodi, and K. Roy, "Statistical design and optimization of sram cell for yield enhancement," in *Proc. of IEEE/ACM International Conference on Computer Aided Design*, 2004, pp. 10–13.

[128] K. Agarwal and S. Nassif, "Statistical analysis of sram cell stability," in *Proc. of IEEE/ACM Design Automation Conference*, 2006, pp. 57–62.

[129] A. Owen and Y. Zhou, "Safe and effective importance sampling," *Journal of the American Statistical Association*, vol. 95, pp. 135–143, Mar. 2000.

[130] T. C. Hesterberg, *Advances in importance sampling*. Ph.D. Dissertation, Statistics Department, Stanford University, 1988.

[131] A. S. Householder, *The numerical treatment of a single nonlinear equation*. McGraw-Hill, 1970.

[132] P. J. Smith, M. Shafi, and H. Gao, "Quick simulation: A review of importance sampling techniques in communications systems," *IEEE Journal on Selected Areas in Communications*, vol. 15, pp. 597–613, May 1997.

[133] J. S. Sadowsky and J. A. Bucklew, "On large deviations theory and asymptotically efficient monte carlo estimation," *IEEE Transactions on Information Theory*, vol. 36, pp. 579–588, Mar. 1990.

[134] P. Glasserman, P. Heidelberger, and P. Shahabuddin, "Asymptotic optimal importance sampling and stratification for pricing path-dependent options," *Mathematical Finance*, vol. 9, pp. 117–152, Apr. 1999.

[135] R. E. Melchers, "Search-based importance sampling," *Structural Safety*, vol. 9, pp. 117–128, 1990.

[136] J. S. Stadler and S. Roy, "Adaptive importance sampling," *IEEE Journal on Selected Areas in Communications*, vol. 11, pp. 309–316, Apr. 1993.

[137] I. Dimov, A. Karaivanova, R. Georgieva, and S. Ivanovska, "Parallel importance separation and adaptive monte carlo algorithms for multiple integrals," in *LNCS: Numerical Methods and Applications*, 2003, pp. 99–107.

[138] H. J. Kushner and G. G. Yin, *Stochastic approximation algorithms and applications*. Springer, 1997.

[139] B. Arouna, "Adaptative monte carlo method, a variance reduction technique," in *Monte Carlo Methods and Applications*, vol. 10, 2004, pp. 1–24.

[140] H. Robbins and S. Monro, "A stochastic approximation method," *Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, 1951.

[141] R. Kumar and V. Kursun, "Reversed temperature-dependent propagation delay characteristics in nanometer cmos circuits," *IEEE Transactions on Circuits and Systems—Part II: Analog and Digital Signal Processing*, vol. 53, pp. 1078–1082, Oct. 2006.

[142] A. H. Ajami, K. Banerjee, and M. Pedram, "Modeling and analysis of nonuniform substrate temperature effects on global ULSI interconnects," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 24, pp. 849–861, June 2005.

[143] S. Bota, M. Rosales, J. Rosello, A. Keshavarzi, and J. Segura, "Within die thermal gradient impact on clock-skew: a new type of delay-fault mechanism," in *Proc. of IEEE International Test Conference*, 2004, pp. 1276–1283.

[144] H. Su, F. Liu, A. Devgan, E. Acar, and S. Nassif, "Full chip leakage estimation considering power supply and temperature variations," in *Proc. of IEEE/ACM International Symposium on Low-Power Electronics*, 2003, pp. 78–83.

[145] T. Wang, J. Tsai, and C. Chen, "Thermal and power integrity based power/ground networks optimization," in *Proc. of IEEE/ACM Design, Automation and Test in Europe*, 2004, pp. 830–835.

[146] A. Vassighi and M. Sachdev, "Thermal runaway in integrated circuits," *IEEE Transactions on Device and Materials Reliability*, vol. 6, pp. 300–305, June 2006.

[147] Y. Cheng, P. Raha, C. Teng, E. Rosenbaum, and S. Kang, "ILLIADS-T: An electrothermal timing simulator for temperature-sensitive reliability diagnosis of CMOS VLSI chips," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 17, pp. 668–680, Aug. 1998.

[148] P. Li, L. Pileggi, M. Asheghi, and R. Chandra, "IC thermal simulation and modeling via efficient multigrid-based approaches," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 25, pp. 1763–1776, Sept. 2006.

[149] Y. Zhan and S. Sapatnekar, "Fast computation of the temperature distribution in VLSI chips using the discrete cosine transform and table look-up," in *Proc. of IEEE/ACM Asia and South Pacific Design Automation Conference*, 2005, pp. 87–92.

[150] "Hotspot 3.1 temperature modeling tool," [Online]. Available: http://lava.cs.virginia.edu/HotSpot/.

[151] J. Kao, S. Narendra, and A. Chandrakasan, "Subthreshold leakage modeling and reduction techniques," in *Proc. of IEEE/ACM International Conference on Computer Aided Design*, 2002, pp. 141–148.

[152] R. Rao, A. Srivastava, D. Blaauw, and D. Sylvester, "Statistical estimation of leakage current considering inter- and intera-die process variation," in *Proc. of IEEE/ACM International Symposium on Low-Power Electronics*, 2003, pp. 84–89.

[153] H. Chang and S. Sapatnekar, "Prediction of leakage power under process uncertainties," *ACM Transactions on Design Automation of Electronic Systems*, vol. 12, pp. 1–27, Apr. 2007.

[154] S. Borkar, "Design challenges of technology scaling," *IEEE Micro*, vol. 19, pp. 23–29, July-August 1999.

[155] W. Huang, E. Humenay, K. Skadron, and M. R. Stan, "The need for a fullchip and package thermal model for thermally optimized IC designs," in *Proc. of IEEE/ACM International Symposium on Low-Power Electronics*, 2005, pp. 245–250.

[156] Y. Zhan, B. Goplen, and S. Sapatnekar, "Electrothermal analysis and optimization techniques for nanoscale integrated circuits," in *Proc. of IEEE/ACM Asia and South Pacific Design Automation Conference*, 2006, pp. 219–222.

[157] G. M. Link and N. Vijaykrishnan, "Thermal trends in emerging technologies," in *Proc. of IEEE International Symposium on Quality Electronic Design*, 2006.

[158] S. Zhang, V. Wason, and K. Banerjee, "A probabilistic framework to estimate full-chip subthreshold leakage power distribution considering within-die and die-to-die p-t-v variations," in *Proc. of IEEE/ACM International Symposium on Low-Power Electronics*, 2004, pp. 156–161.

[159] K. Cheng, C. Tsai, C. C. Teng, and S. M. Kang, *Electrothermal analysis of VLSI systems*. Kluwer Academic Publishers, 2000.

[160] C. Tsai and S. Kang, "Cell-level placement for improving substrate thermal distribution," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 19, pp. 253–266, Feb. 2000.

[161] Y. Saad, *Iterative Methods for Sparse Linear Systems, 2nd Edition*. Society for Industrial and Applied Mathematics, 2003.

[162] P. Teck and B. Nikolic, "Impact of layout on 90nm cmos process parameter fluctuations," in *Proc. of IEEE Symposium on VLSI Circuits, Digest of Technical Papers*, 2006, pp. 69–70.

[163] [Online]. Available: http://www.eas.asu.edu/p̃tm/.

[164] S. Chandra, K. Lahiri, A. Raghunathan, and S. Dey, "Considering process variations during system-level power analysis," in *Proc. of IEEE/ACM International Symposium on Low Power Electronics and Design*, 2006, pp. 342–345.

[165] S. M. Ross, *Introduction to Probability Models*. Academic Press, 2003.

[166] R. C. Whaley and A. Petitet, "Minimizing development and maintenance costs in supporting persistently optimized BLAS," *Software: Practice and Experience*, vol. 35, no. 2, pp. 101–121, Feb. 2005.

[167] D. Coppersmith and S. Winograd, "Matrix multiplication via arithmetic progressions," *Elsevier Journal of Symbolic Computation*, vol. 9, no. 3, pp. 251–280, 1990.

[168] N. C. Beaulieu, A. A. Abu-Dayya, and P. J. McLane, "Comparison of methods of computing lognormal sum distributions and outages for digital wireless applications," in *Proc. of IEEE International Conference on Communications*, 1994, pp. 1270–1275.

[169] S. Schwartz and Y. Yeh, "On the distribution function and moments of power sums with lognormal components," *Bell System Technical Journal*, vol. 61, pp. 1441–1462, Sept. 1982.

[170] M. Pedram and S. Nazarian, "Thermal modeling, analysis, and management in vlsi circuits: Principles and methods," *Proceedings of the IEEE*, vol. 94, pp. 1487–1501, Aug. 2006.

[171] L. S. Chih, R. Mahajan, D. Vivek, and K. Banerjee, “Analysis and implications of IC cooling for deep nanometer scale CMOS technologies,” in *Proc. of IEEE International Electron Devices Meeting*, 2005, pp. 1018–1021.

[172] J. Aitchison and J. Brown, *The Lognormal Distribution*. Cambridge University Press, 1957.

[173] A. Jain and et al., “A 1.2ghz alpha microprocessor with 44.8gb/s chip pin bandwidth,” in *Proc. of IEEE International Solid State Circuits Conference*, 2001, pp. 240–241.

[174] K. A. Bowman, S. G. Duvall, and J. D. Meindl, “Impact of die-to-die and within-die parameter fluctuations on the maximum clock frequency distribution,” in *Proc. of IEEE International Solid State Circuits Conference*, 2001, pp. 278–288.

[175] K. A. Bowman and J. D. Meindl, “Impact of within-die parameter fluctuations on future maximum clock frequency distribution,” in *Proc. of IEEE Custom Integrated Circuits Conference*, 2001, pp. 229–232.