Robust Design of Variation-Sensitive Digital Circuits

by

Hassan Mostafa

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, 2011

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

The nano-age has already begun, where typical feature dimensions are smaller than 100nm. The operating frequency is expected to increase up to 12 GHz, and a single chip will contain over 12 billion transistors in 2020, as given by the International Technology Roadmap for Semiconductors (ITRS) initiative. ITRS also predicts that the scaling of CMOS devices and process technology, as it is known today, will become much more difficult as the industry advances towards the 16nm technology node and further. This aggressive scaling of CMOS technology has pushed the devices to their physical limits. Design goals are governed by several factors other than power, performance and area such as process variations, radiation induced soft errors, and aging degradation mechanisms. These new design challenges have a strong impact on the parametric yield of nanometer digital circuits and also result in functional yield losses in variation-sensitive digital circuits such as Static Random Access Memory (SRAM) and flip-flops. Moreover, sub-threshold SRAM and flip-flops circuits, which are aggravated by the strong demand for lower power consumption, show larger sensitivity to these challenges which reduces their robustness and yield. Accordingly, it is not surprising that the ITRS considers variability and reliability as the most challenging obstacles for nanometer digital circuits robust design.

Soft errors are considered one of the main reliability and robustness concerns in SRAM arrays in sub-100nm technologies due to low operating voltage, small node capacitance, and high packing density. The SRAM arrays soft errors immunity is also affected by process variations. We develop statistical design-oriented soft errors immunity variations models for super-threshold and sub-threshold SRAM cells accounting for die-to-die variations and within-die variations. This work provides new design insights and highlights the important design knobs that can be used to reduce the SRAM cells soft errors immunity variations. The developed models are scalable, bias dependent, and only require the knowledge of easily measurable parameters. This makes them useful in early design exploration, circuit optimization as well as technology prediction. The derived models are verified using Monte Carlo SPICE simulations, referring to an industrial hardware-calibrated 65nm CMOS technology.

The demand for higher performance leads to very deep pipelining which means that hundreds of thousands of flip-flops are required to control the data flow under strict timing constraints. A violation of the timing constraints at a flip-flop can result in latching incorrect data causing the overall system to malfunction. In addition, the flip-flops power dissipation represents a considerable fraction of the total power dissipation. Sub-threshold flip-flops are considered the most energy efficient solution for low power applications in which, performance is of secondary importance. Accordingly, statistical gate sizing is conducted to different flip-flops topologies for timing yield improvement of super-threshold flip-flops and power yield improvement of sub-threshold flip-flops. Following that, a comparative analysis between these flip-flops topologies considering the required overhead for yield improvement is performed. This comparative analysis provides useful recommendations that help flip-flops designers on selecting the best flip-flops topology that satisfies their system specifications while taking the process variations impact and robustness requirements into account.

Adaptive Body Bias (ABB) allows the tuning of the transistor threshold voltage, V_t, by controlling the transistor body voltage. A forward body bias reduces V_t, increasing the device speed at the expense of increased leakage power. Alternatively, a reverse body bias increases V_t, reducing the leakage power but slowing the device. Therefore, the impact of process variations is mitigated by speeding up slow and less leaky devices or slowing down devices that are fast and highly leaky. Practically, the implementation of the ABB is desirable to bias each device in a design independently, to mitigate within-die variations. However, supplying so many separate voltages inside a die results in a large area overhead. On the other hand, using the same body bias for all devices on the same die limits its capability to compensate for within-die variations. Thus, the granularity level of the ABB scheme is a trade-off between the within-die variations compensation capability and the associated area overhead. This work introduces new ABB circuits that exhibit lower area overhead by a factor of 143X than that of previous ABB circuits. In addition, these ABB circuits are resolution free since no digital-to-analog converters or analog-to-digital converters are required on their implementations. These ABB circuits are adopted to high performance critical paths, emulating a real microprocessor architecture, for process variations compensation and also adopted to SRAM arrays, for Negative Bias Temperature Instability (NBTI) aging and process variations compensation. The effectiveness of the new ABB circuits is verified by post layout simulation results and test chip measurements using triple-well 65nm CMOS technology.

The highly capacitive nodes of wide fan-in dynamic circuits and SRAM bitlines limit the performance of these circuits. In addition, process variations mitigation by statistical gate sizing increases this capacitance further and fails in achieving the target yield improvement. We propose new negative capacitance circuits that reduce the overall parasitic capacitance of these highly capacitive nodes. These negative capacitance circuits are adopted to wide fan-in dynamic circuits for timing yield improvement up to 99.87% and to SRAM arrays for read access yield improvement up to 100%. The area and power overheads of these new negative capacitance circuits are amortized over the large die area of the microprocessor and the SRAM array. The effectiveness of the new negative capacitance circuits is verified by post layout simulation results and test chip measurements using 65nm CMOS technology.

Acknowledgements

First of all, I express my gratitude to Allah for providing me the blessings and strength to complete this work.

Then, I would like to give my sincere appreciation to my academic advisors Prof. Mohab Anis, Prof. Mohamed Elmasry, and Prof. Karim Karim, for their assistance, support, and strong encouragement throughout my research. I would also like to thank my committee members, Prof. Manoj Sachdev, Prof. Siddharth Garg, Prof. Eihab Abdel-Rahman, and Prof. Massimio Alioto. They have provided me with valuable comments and suggestions on my research work.

My stay at University of Waterloo was stimulating and entertaining, thanks to the friendship of Mohamed Abu-Rahma, Moataz Elayadi, Noman Hai, Tahseen Shakir, Hassan Hassan, Ahmed Youssef, Noman Hai, Aymen Ismail, Mohamed Amin, Mostafa Soliman, Hussien Attia, Yasser Atwa, Tarek El-Fouly, Osama Amin, Mohamed Feteeha, Hussien Attia, Hassan Elgohary, Mohamed Basha, Mohamed Sadek, Ahmed Mostafa, Mohamed Berisha, Mohab Elnashar, Mohab Elhakim, Adham Elmenoufey, and many other friends. I also thank all my lab mates in the VLSI lab for many fruitful discussions on different research topics. I would like to acknowledge Wendy Boles, Annette Dietrich, Karen Schooley, Susan King and Lisa Hendel from ECE graduate office for their help on administrative issues. I also thank Phil Regier for his great support for all IT related issues.

All of this was only made possible by the encouragement and love of my family. My adorable wife, Hoyda, shared with me every moment of my Ph.D. She supported me with her unconditional love and care during all the ups and downs of my research, and she was always there to motivate me. Despite her pain and suffering during the defense month, when our new baby, Leena, came to life, she kept me away from any stress focusing on preparing for my defense. Hoyda, I have really experienced that dreams come true, when I give my heart to you. My deepest gratitude goes to my mom and dad for their never ending support, and for remembering me in their prayers. No words of appreciation could ever reward them for all they have done for me. I am, and will ever be, indebted to them for all the achievements in my life. I could not forget the hard moments my kids have when we miss some trips and gatherings with our friends just because I have to meet some deadlines during this PhD journey. I am also thankful to my brother Hossam and my sister Donia, for their support and encouragement. Finally, I would like to express my warm thanks to Prof. Ahmed Soliman, from Cairo University, for his continued support and encouragement to me.

Dedication

To my beloved parents, my adorable wife (Hoyda), and my wonderful kids (Mostafa, Yassmina, and Leena).

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This chapter gives a short introduction on the importance of robust design of variation-sensitive digital circuits. Section 1.1 presents the motivation for this research. Section 1.2 provides the outline of the thesis.

Traditionally, performance and power estimations are based on the premise that the electrical characteristics and operating conditions of every device in the design matches the model specifications. However, with continued technology scaling of device dimensions to sub-100nm regime, it has become nearly impossible to maintain the same level of manufacturing control. This can cause devices to behave differently from model characteristics. Further, devices that are intended to be identical could differ vastly in their electrical characteristics [1]. The expected higher sensitivities to variations, radiation induced soft errors, aging degradations, and noise, make future CMOS devices prohibitively unreliable. Using unreliable devices to build a robust circuit is extremely challenging.

A design is robust when its performance is minimum sensitive to variations (i.e., process variations and environmental variations), reliability impacts (i.e., soft errors and aging), and noise, as shown in Figure 1.1. In other words, designing robust system means seeking win-win solutions for yield and reliability improvements. The global picture depicts variability and reliability impacts emerging as the major threats to the robust system design of future technology nodes [2]. These major threats can combine to make the actual design considerably different from the intended design. This translates into a reduced parametric yield, which is the number of manufactured chips that satisfy the required performance, power, and reliability specifications, and hence limits the number of shipped products. With the initial design cost being the same, the cost per good chip increases which has a direct impact on the bottom line dollar value of the design [1].

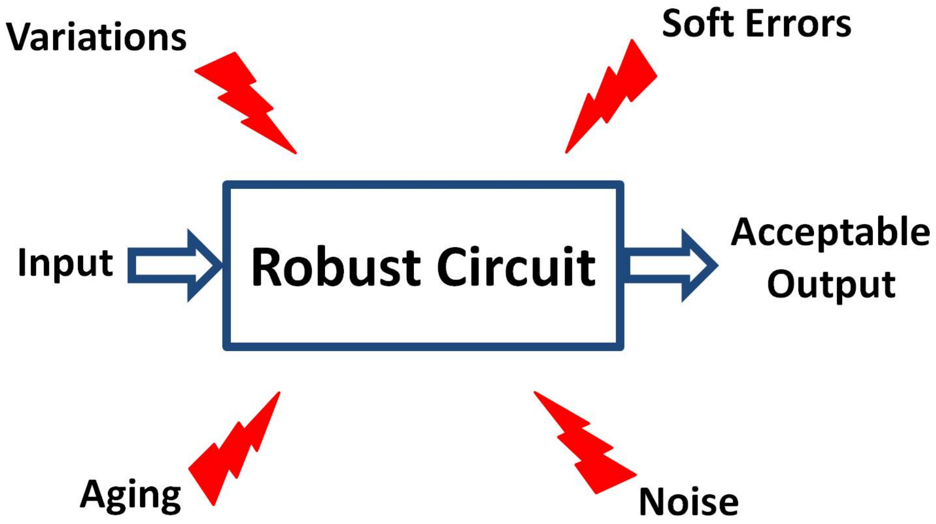


Figure 1.1: Robust circuit design

1.1 Motivation: Robust Circuit Design

The device parameters variation increases with technology scaling, and controlling it becomes extremely difficult [3–6]. These variations are caused by various sources such as limitations in getting precise control in chip manufacturing, especially, when standard lithography is used in printing nanometer scale technologies, and fundamental physical limits such as Random Dopants Fluctuations (RDF) and Line Edge Roughness (LER). This variability results in parametric yield loss, reduces the system robustness, and is currently one of the biggest challenges facing the semiconductor industry [6].

Process variations strongly impact different aspects of digital circuits operation. For example, in digital logic circuits, the overdrive voltage ($V_{DD} - V_t$) becomes unpredictable even for neighboring identically-sized transistors. As a result, the gate delay becomes a random variable. Moreover, traditional techniques that deal with die-to-die variability (such as slow/fast corner models and worst-case analysis) can not be used in dealing with the large increase in within-die variability. This is because these techniques tend to be inefficient and overly pessimistic in the presence of within-die variations. Therefore, statistical design techniques should be used instead of the worst-case techniques to deal with variations in nanometer technologies [7].

Not only does variability affect digital circuits, but it even has a much stronger impact on Static Random Access Memory (SRAM) and flip-flops [8, 9]. SRAM utilizes the most aggressive design rules to achieve the highest possible integration density, which makes SRAM the most sensitive circuit for process variations. With the exponential increase in SRAM density in microprocessors and System on Chips (SoCs), the SRAM yield loss, due to process variations, has strong impact on the overall product yield [6]. Moreover, the demand for higher performance has moved the clock frequencies up to multi-GHz in microprocessors and other advanced VLSI applications. These increased clock frequencies lead to very deep pipelining which means that hundreds of thousands of flip-flops are required to control the data flow under strict timing constraints. A violation of the timing constraints, due to process variations, at a flip-flop can result in latching incorrect data causing the overall system to malfunction [10]. In fact, the existence of several flip-flops topologies and circuits in the literature, makes it difficult for flip-flops designers to select the most robust flip-flop circuit to meet their design specifications while considering nanometer challenges such as variability and reliability. Thus, a comparative analysis between different flip-flops topologies considering these challenges helps in answering the question: *Given the design specifications, what is the flip-flop topology that achieves the highest robustness and yield?*.

Reliability degradation due to soft errors is higher than all other reliability mechanisms combined [11]. Soft errors are induced by energetic particle strikes from the chip package or the atmosphere. This particle strike induces a current pulse that can affect the potential of the struck node. Therefore, in memory elements such as SRAM cells and flip-flops, this may lead to a 1-to-0 flip or a 0-to-1 flip which corrupts the circuit logic state and destroys the stored data. However, in combinational logic circuits, it may cause a temporary change in the node voltage. This temporary change may be tolerated unless it is latched by a succeeding memory element. Thus, soft errors, especially in memory elements, result in reliability degradation and reduces the system robustness. In addition, reliability degradation occurs due to aging mechanisms, such as Negative Bias Temperature Instability (NBTI) and Hot Carrier Injection (HCI). Aging mechanisms are time-dependent degradation effects that cause a change of the transistor parameters such as the threshold voltage, as a function of the aging time. Therefore, these aging mechanisms might turn an initially fully functional circuit into a less or even non-functional circuit over time. This aging degradation depends on the stress applied to the device (i.e., the voltages applied to the transistor and the temperature).

Soft errors have a large impact on the memory elements such as SRAM and flip-flops circuits. In personal computers, soft errors are eclipsed by more common software bugs and may pass unnoticed in something like a graphical display. However, in some other applications such as networking and routing equipments, soft errors can cause hard network problems such as sending a packet to Toronto instead of Waterloo [12]. Soft errors are also affected by process variations, and accordingly, it will be beneficial to study the limitations imposed on the soft errors mitigation techniques by process variations (i.e., answering the question: *if a capacitor is added to enhance the soft errors immunity, what is the capacitor value that results in minimum soft errors immunity variations?*). On the other hand, the impact of process variations compensation methods on the soft errors immunity should be investigated (i.e., answering the question: *Does the timing yield improvement increase or decrease the soft errors immunity?*). Answering these questions helps in designing robust circuit and improving the parametric yield.

Unlike soft errors, which have the largest impact on memory elements, aging degradation mechanisms have large impact on both logic circuits and memory elements. Interestingly, most of the effects of device aging mechanisms can be understood by a change in the device threshold voltage (V_t) only [1]. For example, NBTI is modeled as an increase in the PMOS transistor V_t absolute value, and HCI is modeled as an increase in the NMOS transistor V_t value, with aging time. This V_t increase results in performance degradation in logic circuits, especially high performance circuits and may result in SRAM failures by reducing the noise margins. These aging mechanisms, combined with variations, result in reducing the circuit robustness and yield. Therefore, variations compensation techniques that take into account the aging mechanisms and circuit robustness are essential for parametric yield improvement.

1.2 Thesis Organization

In order to continue digital design success in the nanometer regime, it is critical to explore solutions to design a robust circuit by mitigating the impacts of the nanometer challenges (i.e., variations, soft errors, and aging mechanisms). This thesis focuses on dealing with the increase of these challenges in nanometer technologies and their impacts on SRAM, flip-flops, and high performance circuits. This research intends to fill the gap between different levels of abstraction by introducing models, methodologies, and circuits that help in designing a robust system by mitigating the impacts of process variations and aging degradation, as shown in Figure 1.2. In addition, this thesis discusses the limitations imposed by process variations on the soft error mitigation techniques, and the effect of process variations mitigation techniques on the soft errors immunity.

To set the stage for our discussion on robust design techniques, Chapter 2 reviews the CMOS technology scaling challenges such as process variations, soft errors, and aging mechanisms, and their impacts on digital circuits. Also, some background on the thesis digital circuits benchmarks such as SRAM, flip-flops, and high performance circuits, is presented in this chapter.

To study the impact of process variations on the soft errors immunity, in Chapter 3, new design-oriented statistical soft errors immunity variations models are presented that account for process variations for super-threshold and sub-threshold SRAM cells. These

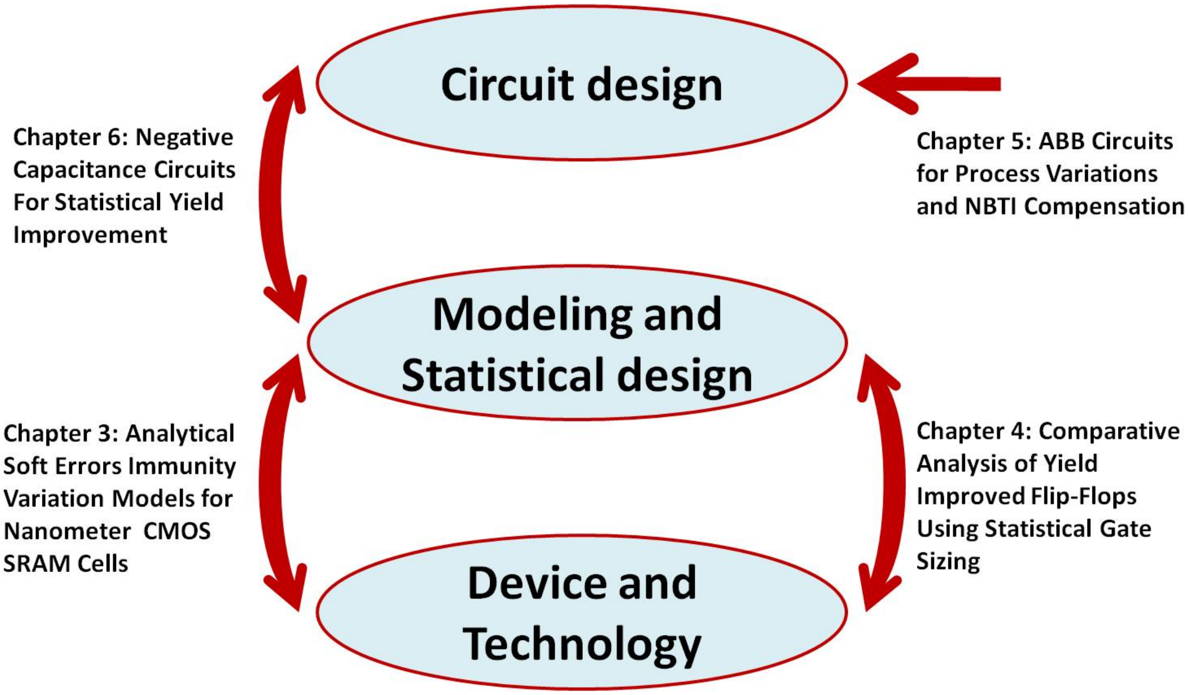

Circuit design
Chapter 6: Negative Capacitance Circuits For Statistical Yield Improvement
Chapter 5: ABB Circuits for Process Variations and NBTI Compensation
Modeling and Statistical design
Chapter 3: Analytical Soft Errors Immunity Variation Models for Nanometer CMOS SRAM Cells
Chapter 4: Comparative Analysis of Yield Improved Flip-Flops Using Statistical Gate Sizing
Device and Technology

Figure 1.2: Different levels of abstraction studied in this research.

models provide several design insights on how process variations and circuit level design decisions interact to affect the soft errors variations (i.e., how the soft errors immunity variations are affected when the SRAM cell design is modified to increase the nominal soft errors immunity). Also, these models can be extended to flip-flops circuits to explore similar design insights.

In Chapter 4, statistical gate sizing is used to improve the yield of flip-flops circuits. A comparative analysis between different flip-flops topologies considering the overhead required to achieve the yield improvement is presented. In addition, the effect of this yield improvement on the flip-flops soft errors immunity is discussed. This comparative analysis helps the flip-flops designers to decide which flip-flops topology is more robust for their system specifications while taking the process variations and soft errors impacts into account.

In Chapter 5, new ABB circuits are introduced and adopted to high performance circuits and SRAM arrays to reduce the impacts of process variations and NBTI. These ABB circuits exhibit low area overhead and resolution free operation compared to previously published ABB circuits. The low area overhead of these new ABB circuits makes it possible to use them at lower granularity levels, which results in more process variations compensation and higher robustness and yield. The effectiveness of these new ABB circuits is verified by using post layout simulation results and test chip measurements.

In Chapter 6, new negative capacitance circuits are proposed, for the first time, and adopted to high performance dynamic circuits for timing yield improvement and to SRAM arrays for read access yield improvement. These negative capacitance circuits show how the interaction between the statistical design and circuit design helps in achieving more circuit robustness and higher yield. The area and power overheads of the proposed negative capacitance circuits are amortized over the large die area. The effectiveness of these new negative capacitance circuits is verified by using post layout simulation results and test chip measurements.

Summary of this thesis work and suggestions for future work are discussed in Chapter 7.

Chapter 2

Background

As CMOS technology is scaled towards the deep sub-micron regime, digital circuits designers are facing increased variability, increased susceptibility to radiation induced soft errors, and more aging degradations. The design for robustness strategies are the key solutions to improve the overall system yield and reliability by mitigating these challenges.

This chapter is organized as follows. In Section 2.1, variability sources, impacts on the frequency and power, and the state-of-art variations mitigation techniques are presented. Sections 2.2 and 2.3 explain the soft errors and aging degradation mechanisms (i.e., the major reliability degradation mechanisms), respectively, and their impacts on the digital circuits robustness, followed by the state-of-art techniques for reliability improvement. Some background on the variation-sensitive digital circuits benchmarks (namely: SRAM, flip-flops, and high performance circuits), used in this thesis and how the nanometer design challenges reduce their robustness and yield, is displayed in Section 2.4. Finally, in Section 2.5, some conclusions are drawn

2.1 Variability

Variations in integrated circuits are the deviations from the desired or designed values for a structure or circuit parameter in concern [13]. The variations are usually caused by two different sources: physical factors and environmental factors. Physical factors cause a permanent variation in device parameters and are generally caused by the lack of exact controls and statistical variations during the fabrication process [13]. On the other hand, environmental factors cause variations in the operation of the circuit while the circuit is functioning, and include variations in the power supply and temperature.

2.1.1 Classification of Variations

For the purpose of variability analytical modeling, the variations can be separated into two classes [5, 7, 13, 14].

1. **Die-to-Die Variations (D2D)**

   The D2D variations are also called inter-die variations. They capture the variations from die to die and affect all devices on the same die in the same way (i.e., they may cause all the transistors threshold voltages on the same die to deviate from their nominal values by the same amount). These variations are independent and hence, they can be represented by a single value for each die. In addition, they represent a shift in the mean of the parameter from its nominal value. These variations are generally assumed to have a simple distribution, such as Gaussian, with a given variance [13]. D2D variations in a single process parameter are easily captured by corner-based models, which assume that all devices on a given design sample have a value that is shifted away from the mean by a fixed amount [13].

2. **Within-Die Variations (WID)**

   The WID variations are also called intra-die variations. These variations cause transistor parameters to vary across different locations within the same die. Thus, each device on a die requires a separate random variable to represent its WID variations. It can be classified into random variations and systematic variations [13].

   -i- **Random Variations**

   They are spatially uncorrelated variations which result from statistical quantization effects, such as Random Dopant Fluctuations (RDF) and Line Edge Roughness (LER). The impact of these random variations is expected to be worse as process parameters scale. The random variations impact can be reduced by increasing the logic depth due to the averaging effect. Unfortunately, the trend to increase the clock frequency of a design using aggressive pipelining has resulted in smaller logic depth, which increases the impact of this type of variations.

   -ii- **Systematic Variations**

   They are spatially correlated variations which are so difficult to be captured. This is because generating samples of correlated random variables of high dimensionality is a computationally expensive problem. All layout dependent variations such as channel length and width variations are treated as systematic variations.

2.1.2 Sources of Variations

1. **Process Variations (Static Variations)**

   The sources of process variations can be summarized as follows:

   -i- **Random Dopant Fluctuations (RDF)**

   As CMOS technology scales, the number of doping impurities in the channel depletion layer decreases. As can be seen in Figure 2.1, the atomicity of the dopants in the channel do not allow a constant concentration of dopants to appear across the channel. Thus, it is very unlikely to have two neighboring transistors with the same number and placement of dopants. This random number and placement of dopants bring uncertainty in the transistor threshold voltage, V_t. The statistical distribution of V_t due to RDF is found to follow a normal distribution [1, 15]. The standard deviation of V$_t$ distribution is calculated to be [16, 17]:

$$\sigma_{V_t} = \sqrt[4]{4q^3\epsilon_{Si}N_a\phi_F} \quad \frac{T_{ox}}{\epsilon_{ox}} \quad \frac{1}{\sqrt{3W * L}} \tag{2.1}$$

   where σ_{V_t} is the threshold voltage standard deviation due to RDF variations, q is the electron charge, ϵ_{Si} and ϵ_{ox} are the dielectric constants of the Silicon and gate oxide, respectively. N$_a$ is the channel dopant concentration, ϕ_F is the difference between Fermi level and intrinsic level, T$_{ox}$ is the gate oxide thickness, and W and L are the channel width and length, respectively. Equation (2.1) shows that σ_{V_t} is inversely proportional to the square root of the active device area. Hence, sizing up the transistors can be used to mitigate these variations, which is one of the most common techniques used in analog circuit design to reduce transistors mismatch [18]. Moreover, for SRAM cells, which typically have minimum size devices, σ_{V_t} is very large. Figure 2.2 shows that as the channel length is scaled for sub-100nm nodes, the RDF variations 3σ bounds become extremely large, especially when the number of dopant atoms is less than 100 atoms.

   -ii- **Channel Length Variations**

   The patterning of features smaller than the wavelength of the light, used in optical lithography, results in distortions due to light diffraction, which is usually called Optical Proximity Effects (OPEs) [5, 13]. These effects are expected to be worse as technology scales since the light wavelength is not scaling at the same rate as the device feature size as shown in Figure 2.3. The OPEs will make it very difficult to print precise patterns on the Silicon wafer as technology scales

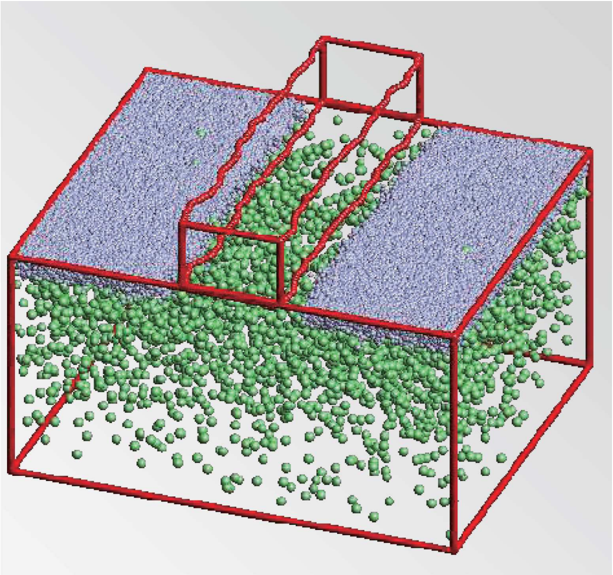


Figure 2.1: Atomistic process simulation incorporating Random Dopant Fluctuations (RDF) and Line Edge Roughness (LER) as the sources of intrinsic fluctuations [19].

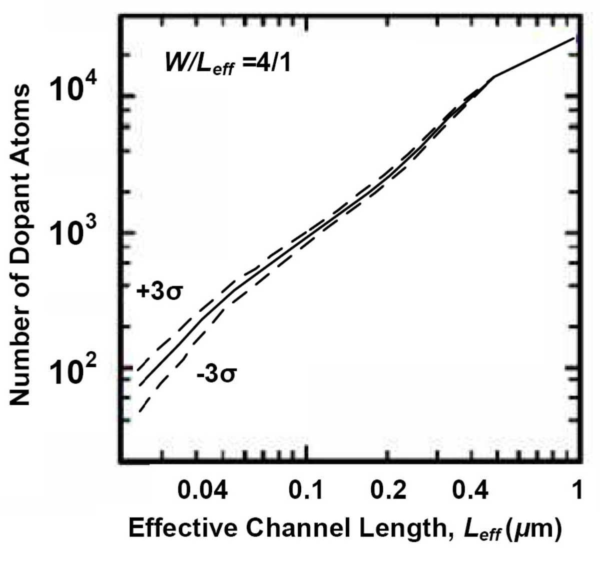


Figure 2.2: Number of dopant atoms in the depletion layer of a MOSFET versus channel length L_{eff} [20].

[21]. Feature sizes on Silicon wafer are now quarter the wavelength of the light source used in lithography, and will continue to get smaller in comparison to the wavelength of future light sources [21]. The OPEs cause large variations in defining the minimum feature sizes. OPEs are layout dependent; therefore result in different Critical Dimension (CD) variations depending on neighboring lines as well as orientation [14]. Controlling these variations has become very difficult in current technologies, and is expected to increase for future technology nodes [7].

The variation in the transistor channel length has direct impact on the transistor electrical parameters; however, the most affected parameter is the threshold voltage V$_t$ [15, 22, 23]. The reason for that is the exponential dependence of V$_t$ on channel length L for short channel devices, especially due to Drain Induced Barrier Lowering (DIBL) effect. DIBL causes V$_t$ to be strongly dependent on L as shown in Figure 2.4. This dependence can be modeled for zero body bias as [15, 22, 23]:

$$V_t \approx V_{to} - (\zeta + \eta V_{DS}) \exp(-L/L_{to}) \tag{2.2}$$

where V$_{to}$ is the long channel threshold voltage, ζ is the charge sharing coefficient, L_{to} is the characteristic length, and η is the DIBL coefficient. Hence, a slight variation in L will introduce large variation in V$_t$ due to the exponential dependence shown in (2.2) as shown in Figure 2.4.

-iii- **Line Edge Roughness (LER)**

LER refers to the roughness on the edge of the channel and contributes to the threshold voltage variations. Ideally, the edge of the channel should be a straight line, but as the edge of the channel is determined by a varying process, the edge will not be completely straight, as shown in Figure 2.1 [24]. Previously, the dimensions of the transistor channel were orders of magnitude larger than the roughness along the edge of the transistor channel (on the order of 5 nm), but as the transistor length is scaled down, the roughness does not scale correspondingly and can cause variations in transistor characteristics [24]. These random effects end up causing variations in the threshold voltage [24]. Figure 2.5 shows the predicted threshold voltage variations due to RDF and LER versus technology nodes [19, 25]. It is clear from this figure that the threshold voltage variations due to LER will be comparable to that due to RDF for sub-32nm technology nodes.

-iv- **Gate Oxide Thickness Variations**

A variation in the oxide thickness T$_{ox}$, will affect the transistor threshold voltage, V$_t$. Therefore, the T$_{ox}$ variations should be considered.

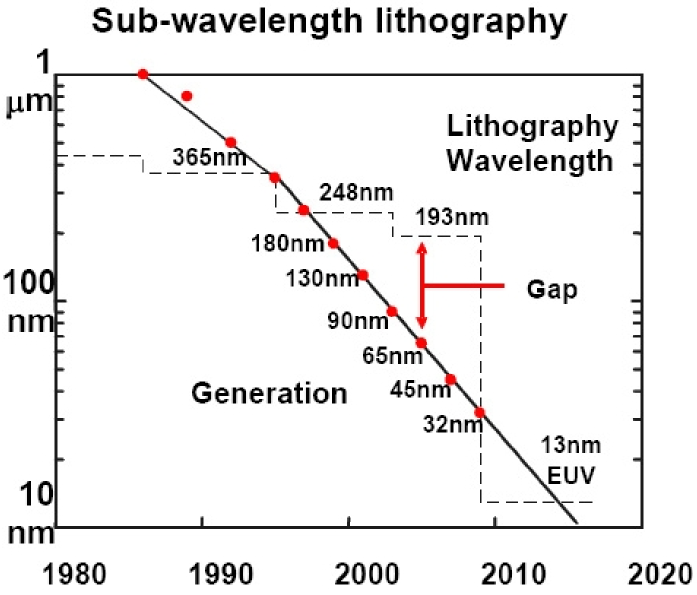


Figure 2.3: Lithography wavelength scaling for different technology nodes [3].

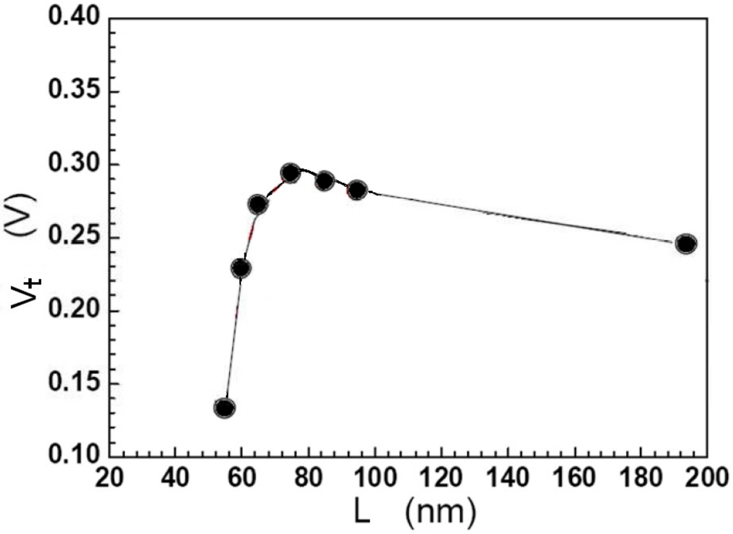
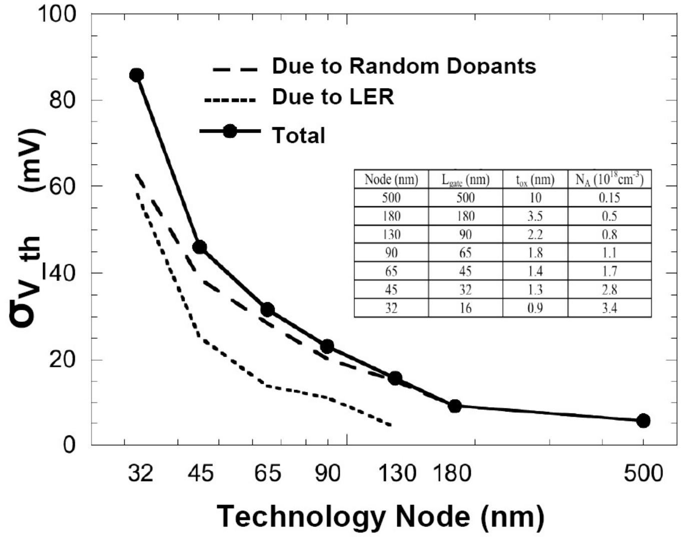


Figure 2.4: Measured V_t versus channel length L for a 90nm CMOS technology which shows strong short channel effects causing sharp roll-off for V_t for shorter L [15].



Node (nm)	L_{gate} (nm)	t_{ox} (nm)	N_A (10^{18}cm^{-3})
500	500	10	0.15
180	180	3.5	0.5
130	90	2.2	0.8
90	65	1.8	1.1
65	45	1.4	1.7
45	32	1.3	2.8
32	16	0.9	3.4

Figure 2.5: Predicted σ_{V_t} including RDF and LER versus technology nodes for the smallest transistor. The inset shows the technological parameters used [25].

-v- **Channel Width Variations**

Transistor channel width, W, will have variations as well because of the lithography limitations. These variations in W will contribute to V_t variations due to the Narrow-Width-Effects (NWEs), which cause V_t to be dependent on W. However, since W is typically 2-4 times larger than L, the impact of W variations on V_t can be considered much smaller than the impact due to L variations [15].

2. **Environmental Variations (Dynamic Sources)**

These variations affect the circuit operation while the circuit is functioning. They include variations in power supply voltage and temperature of the chip or across the chip [3, 7]. The power supply fluctuations are caused by the switching activity variations within the die that are mainly dependent on the input vectors. A reduced power supply lowers the drive strength of the transistors and, hence, degrades performance [14]. It should be noted that this power supply reduction will be problematic as technology scales since the headroom between the supply voltage and the device threshold voltage is consistently being reduced [26].

Within die temperature fluctuations are considered one of the major performance and packaging challenges. This is because both device and interconnect have temperature dependence that causes performance to degrade at higher temperature. Moreover, temperature variation across different communicating blocks on the same die may result in performance mismatches, which may lead to functional failures [5]. Figure 2.6 shows WID temperature fluctuations for a microprocessor unit, with the core exhibits a hot spot of 120^{o}C [27].

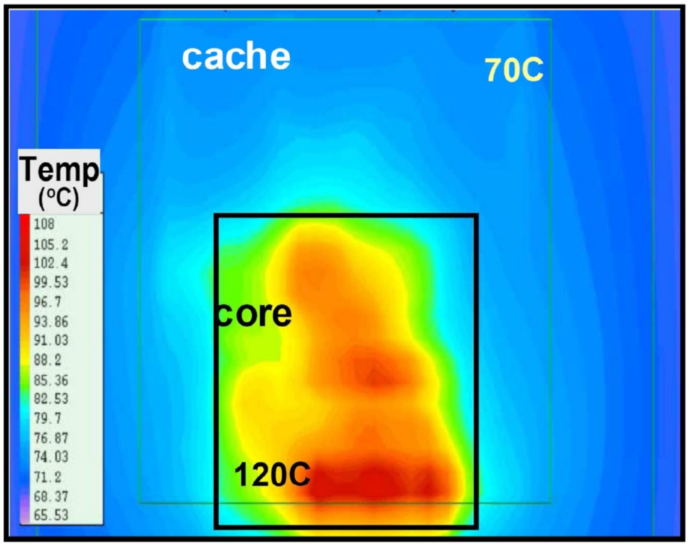


Figure 2.6: Thermal profile showing within die temperature variation for a microprocessor. Hot spots with temperatures as high as 120^{o}C are shown [27].

2.1.3 Impact of Variability on the Frequency and the Power

In the nanometer regime, the reduction of the threshold voltage results in a substantial increase in the device sub-threshold leakage current. Sub-threshold leakage current occurs between the drain and source of a transistor, when the gate voltage is less than the transistor threshold voltage, V$_t$ [20, 28]. The sub-threshold leakage current is exponentially dependent on the threshold voltage. Furthermore, sub-threshold leakage is also very sensitive to temperature, doubling for every 8^{o}K to 10^{o}K temperature increase [29]. In sub-100nm technology nodes, leakage power consumption is considered a significant part of the total power. It is expected that leakage power can reach more than 50% at 45nm technology as shown in Figure 2.7.

The large variability in advanced CMOS technologies is playing an essential role in determining the total leakage of a chip [30]. This has accentuated the need to account for statistical leakage variations during the design cycle [30, 31]. Figure 2.8 shows measured

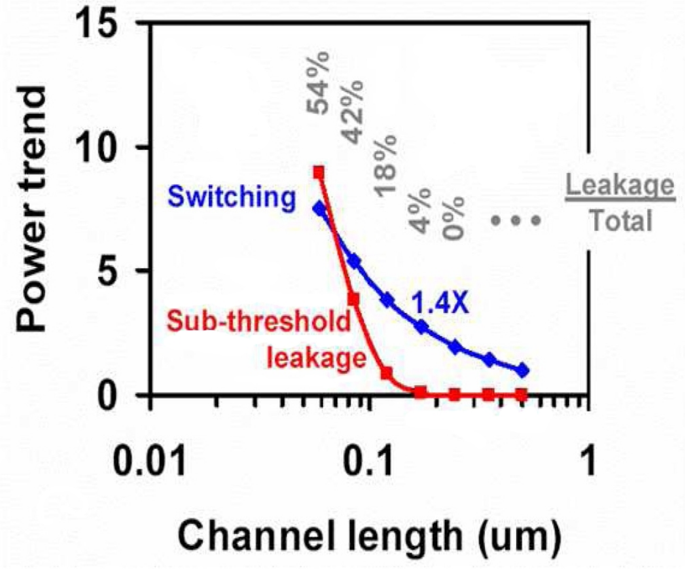


Figure 2.7: Dynamic (switching) and static (leakage) power versus technology scaling, showing the exponential increase in leakage power [28].

variations for frequency and leakage power, for 65nm technology [32]. As shown, there is a leakage variation of 10X for a 50% variation in chip frequency. The highest frequency chips have a wide spread of leakage and for a given leakage there is wide spread in the frequency of the chip. This is considered very large variations in leakage power, especially that leakage power is increasing exponentially with each successive technology node. This excessively large variation in the leakage power makes it very difficult to achieve the required speed while meeting the power constraints.

According to [27], a large fraction of the chips that meet the required operating frequency constraint, dissipate a large amount of leakage power. This makes them unsuitable for usage, and thus degrades the yield. This is due to the trade-off between leakage current and circuit performance. For devices with smaller V_t than nominal due to process variations, the sub-threshold leakage current increases exponentially. In the mean time, the circuit delay decreases due to the increase in the driving current, since the overdrive voltage (V_{DD} - V_t) is increased. Hence, these chips have higher operating frequency, but suffer from large leakage power which makes them unacceptable if their leakage power does not meet the power constraints [3, 27, 33].

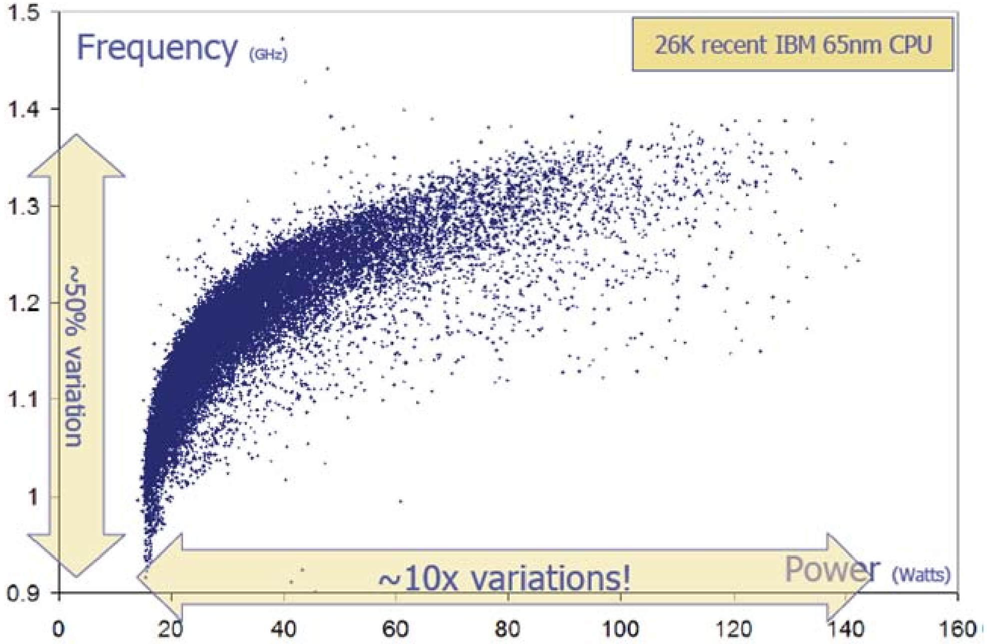


Figure 2.8: Leakage and frequency variations for IBM processor in 65nm technology [32].

2.1.4 State-of-Art Variations Mitigation Techniques

In this section, we review state-of-art related research work dealing with the increase in variability in nanometer technologies in order to improve the yield. The first method is the Computer Aided Design (CAD) tools and statistical design which attempt to model the variations and account for them in the design flow cycle. The second method tries to compensate for the variations with additional circuitry through adaptive techniques at the circuit level, and finally, the third method attempts to deal with variations at the architecture level.

1. **CAD Tools and Statistical Design**

   Recently, a large number of research work has been done in the area of CAD tools that attempt to account and model the random variations at the design flow level. One of the most researched topics in this area is Statistical Static Timing Analysis (SSTA) [7, 14, 23, 34, 35]. In SSTA, the circuit delay is considered a random variable and SSTA computes the probability density function (pdf) of the delay at a certain path [7]. It should be noted that the critical path is defined as the path that has the highest probability to provide a delay higher than a certain threshold delay. Similar to the SSTA, which is used to model the delay variations, few research work targets modeling the process variations impact on other metrics such as leakage power, noise margins, and soft errors [36–40].

Statistical design aims at changing the circuit parameters at the design phase statistically to reduce the impact of process variations and increase the circuit robustness and yield. One of the most common statistical design techniques is statistical gate sizing. In statistical gate sizing, either the length or width of the transistor is tweaked to modulate the current drive capability. For instance, process variations may increase the delay of the circuit, the statistical gate sizing algorithms are proposed to reduce the mean and standard deviation of the delay variations and improve the timing yield [34, 35].

2. **Adaptive Circuit Techniques**

   Adaptive circuit techniques try to reduce the effect of the variations in the device parameters and achieve acceptable chip robustness. This is achieved by monitoring various circuit characteristics that are sensitive to variations, and then changing circuit controls via feedback mechanism to reduce the effect of these variations [38, 41–45]. Moreover, they measure the variations by monitoring either the actual functional circuit, or additional replica circuitry which mimics the functional circuit. These techniques use control knobs such as supply voltage and body bias to achieve this control.

   In [46], an Adaptive Body Bias (ABB) circuit is proposed and fabricated, shown in Figure 2.9, where forward body bias and reverse body bias are used to improve performance and decrease leakage, respectively. This ABB allows each die on a wafer to have the optimum threshold voltage that maximizes the die frequency subject to leakage power constraint. A critical path mimic, which contains key circuit elements of a real microprocessor critical path, is used to model the effect of body bias on the frequency and leakage. In this ABB circuit, the phase detector compares the target frequency with the critical path mimic frequency. The counter and the bias selector provide a suitable body bias voltage according to the frequency difference.

   This ABB circuit is fabricated and tested in 150nm technology. With no body bias used, only 50% of the dies are acceptable, mainly in the lowest frequency bin. When global ABB is utilized using only one delay sensor, ABB reduces the relative frequency variation σ/μ from 4.1% to 1%, however a significant number of dies are still placed in the low frequency bin. In [46], the authors propose the WID-ABB (local ABB) technique, which provides each circuit block in the design with its unique bias combination which controls the frequency and leakage of that circuit block. This WID-ABB enables 99% of the dies to be accepted in the highest frequency bin. Figure 2.10 shows the results after using WID-ABB technique. The granularity of the WID-ABB has to be traded off as using WID-ABB for each gate will have a ter-

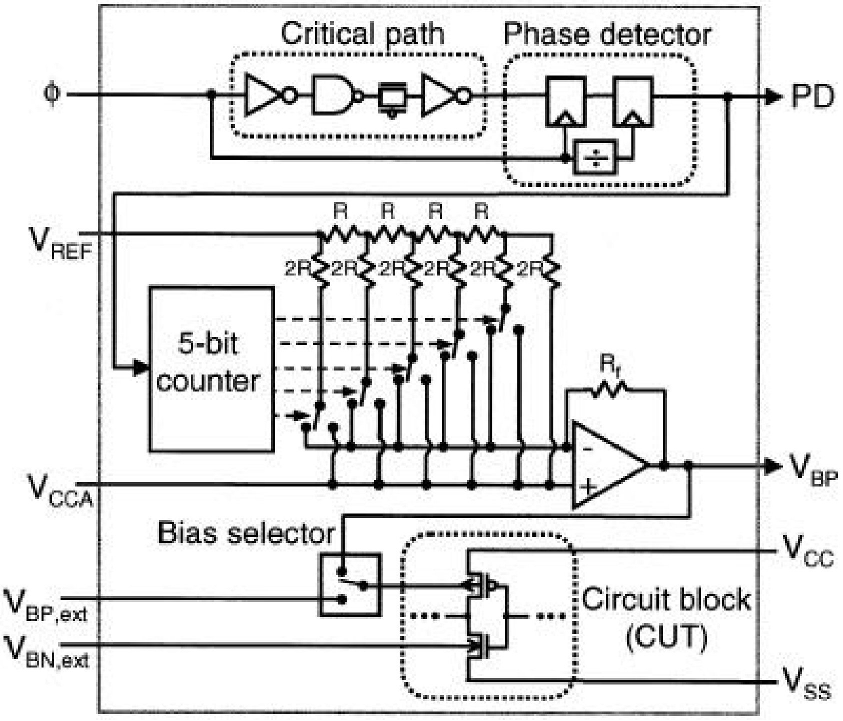


Figure 2.9: Block diagram of the fabricated ABB circuit used in [46].

rible power and area overheads while using it for very large circuit blocks will not help in reducing the WID variations impacts. Also, the resolution of the Digital-to-Analog Converter (DAC) used in this ABB circuit affects its capability in reducing the process variations.

In [47], the authors extend the ABB technique, and combine it with adaptive supply voltage V_{DD} to control the frequency and leakage distributions. It has been shown that using adaptive V_{DD} in conjunction with ABB is more effective than using either of them. Once again, ABB uses Forward Body Bias (FBB) to speed up dies that are too slow, and Reverse Body Bias (RBB) to reduce frequency and leakage power of dies that are too fast and leaky. Adaptive V_{DD} + ABB, on the other hand, recovers the dies that exceed the power limit by first lowering V_{DD}, and hence the operating frequency, thus bringing the total switching and leakage power below the power constraint, and then applying FBB to speed up and move them to the highest frequency bin allowed by the power limit. Using adaptive V_{DD} combined with WID-ABB, results in significant yield improvement. However, this yield improvement comes at the cost of additional area, design complexity and cost.

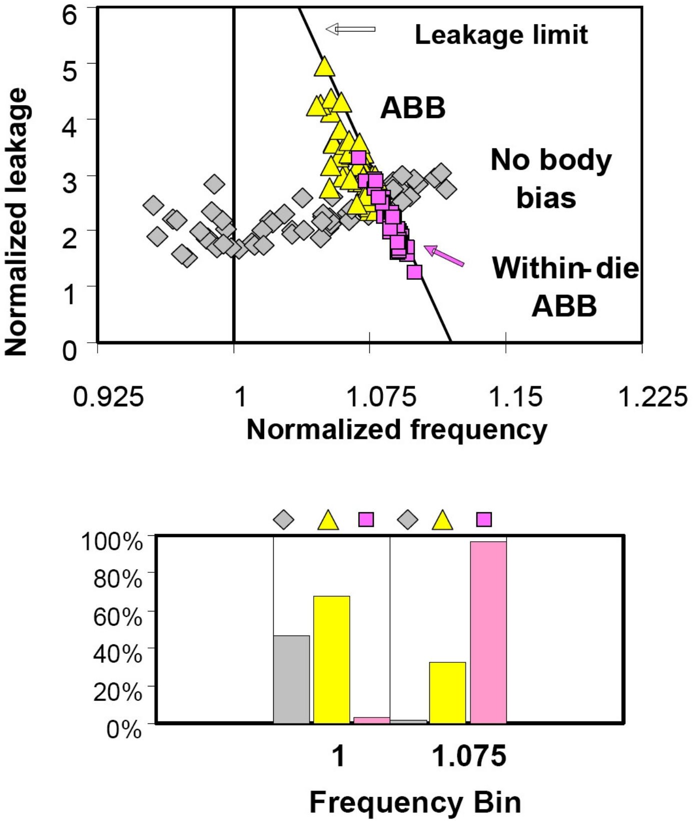


Figure 2.10: Measured leakage versus frequency distribution for 62 dies in a 150nm technology, showing the distributions after utilizing global ABB and WID-ABB (local ABB). In the lower Figure, the percentages of accepted dies at a certain frequency bin are shown [46].

3. **Coping with Variations at the Architecture Level**

   One of the primary research work that related variability to architecture is the work introduced in [48–50], which presents a statistical predictive model for the distribution of the maximum operating frequency, F_{MAX}, for a chip in the presence of process variations. This technique provides insight on the impact of different components of variations on the distribution of F_{MAX}. The delay distribution caused by the WID variations depends mainly on the total number of independent critical paths for the entire chip N_{cp}. For a larger number of critical paths, the mean value of the maximum critical path delay increases as shown in Figure 2.11, which is intuitive, since as the number of critical paths increases, the probability that one of them will be strongly affected by process variations is higher, and hence, the mean of the critical path delay is increased.

On the other hand, the standard deviation (called also delay spread) decreases with larger N_{cp}, thus making the spread of the overall critical path determined mainly by D2D variations. The model has been verified using measurements from Pentium processors. The results show that WID variations directly impact the mean of the maximum frequency, while D2D fluctuations impact the variance. Moreover, it is reported in [48] that when N_{cp} exceeds 14, there is no significant change in the frequency distribution.

Another factor that affects the delay distribution is the logic depth per critical path. The impact of logic depth on delay distribution is different when dealing with random or systematic WID variations. Random WID variations have an averaging effect on the overall critical path distribution, while systematic WID variations affect all the gates on the path, and hence increase the delay spread.

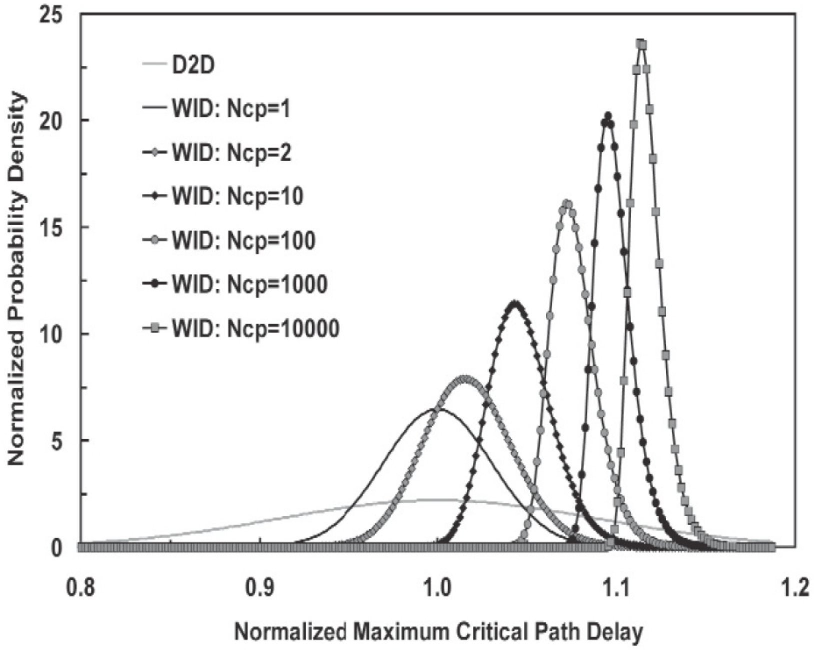


Figure 2.11: The WID maximum critical path distribution for different values of N_{cp} [49].

2.2 Soft Errors

Soft errors are due to external ionizing radiation events rather that a design or manufacturing defects [51]. Soft errors represent a considerable cost and reputation challenge for today's chip manufacturers. In safety critical applications, for example, unpredictable reliability can represent considerable risk, not only in terms of the potential human cost but also in terms of corporate liability, exposing manufacturers to potential litigation. In commercial consumer applications, there is again significant potential economic impact to consider. For high-volume and low-margin products, high levels of product failure may necessitate the costly management of warranty support or expensive field maintenance. Once again, the effect on brand reputation is considerable [52].

With the shrinking geometries and higher-density circuits in the nanometer regime, the issue of soft errors and reliability in complex SoCs design is set to become an increasingly challenging issue for the industry as a whole. In this section, the key radiation mechanisms that cause soft errors are presented, followed by how Soft Error Rate (SER) simulations are performed. Then, state-of-art soft error mitigation techniques are discussed.

2.2.1 Soft Errors Mechanisms

There are three key mechanisms that cause soft errors in Integrated Circuits (ICs) [53, 54]:

1. **Alpha Particles**

   Originally alpha particles were emitted by impurities in the packaging materials [53], but as better materials are now being used in the package, the soft errors that occur due to alpha particles are reduced [55]. As alpha particles strike the Silicon substrate, their positive charge induces the generation of electron-hole pairs. Alpha particles induce the creation of large quantity of electron-hole pairs in the path of their strike [54–56].

2. **High Energy Neutrons**

   High energy neutrons, resulting from the interaction of cosmic rays with the earth atmosphere, can also cause soft errors by producing secondary ions through collisions with Silicon nuclei [53]. The induced charge density of these interactions is considerably large compared to the charge generated due to alpha particle interactions [54–56].

3. **Low Energy Neutrons**

   A third significant soft-error mechanism is induced by low energy cosmic neutrons which cause Boron-10 (^{10}B) fission (since the ^{10}B is good at capturing neutrons from cosmic radiation) [55]. ^{10}B is used in IC materials such as BoroPhosphoSilicate Glass (BPSG) (which is used to form inter-metal layers insulators [53]). While neutrons with any energy level can cause ^{10}B fission, only low-energy neutrons need to be considered since the ^{10}B neutron capture cross-section rapidly decreases as the neutron energy increases [54, 56]. Over 90% of ^{10}B fission reactions are caused by neutrons with energies below 15 eV. In sub-250nm technology nodes, insulating materials free of ^{10}B have replaced BPSG, thus reducing the soft-errors that occur due to low-energy neutrons [53, 54, 56].

Figure 2.12 shows the different phases of charge generation and collection in a reverse-biased p-n junction [54, 56]. At the onset of an ionizing radiation event, such as an alpha particle or neutron collision, a track of electron-hole pairs is generated in the path of the ion's passage as shown in Figure 2.12(a) [54, 56]. When the path of the ionizing particles is in the proximity of a depletion region of a reverse-biased p-n junction, the generated electron-hole pairs are rapidly separated and collected by the depletion region built-in electric field creating a large current transient at the affected node as shown in Figure 2.12(b) [54, 56]. Note that the shape of the depletion region is modified by the ionizing particles into a funnel. The funnel effectively extends the area of the depletion region and greatly increases the drift collection efficiency by extending the high electric field of the depletion region deeper into the substrate. The current spike caused by the rapid drift charge collection phase is completed within tens of picoseconds. It is then followed by the diffusion collection phase that can last for hundreds of picoseconds until all the excessive charge generated by the ionizing particle has been collected, diffused and/or recombined as shown in Figure 2.12(c) [54, 56]. The resulting current pulse is shown in Figure 2.13 [54, 56].

$Q_{critical}$ is the critical amount of charge that has to be collected at a circuit node in order for a soft-error to occur [51]. $Q_{critical}$ depends on many factors including the node capacitance, the operating voltage, and the circuit topology [53]. Since the response of the circuit depends on the temporal characteristics of the charge injection, $Q_{critical}$ is not constant, but is a function of the characteristics of the radiation induced current pulse. For example, as this induced current pulse width increases, so does $Q_{critical}$ [58]. Intuitively this makes sense, since if the charge is spread over a larger period of time, the circuit can more easily recover against the injected charge. The shape of the current pulse that is injected onto a node after an ionizing radiation event can be approximated by a double exponential current pulse given by [59]:

$$i_{injected}(t) = \frac{Q}{\tau_f - \tau_r} \times [\exp(-t/\tau_f) - \exp(-t/\tau_r)] \tag{2.3}$$

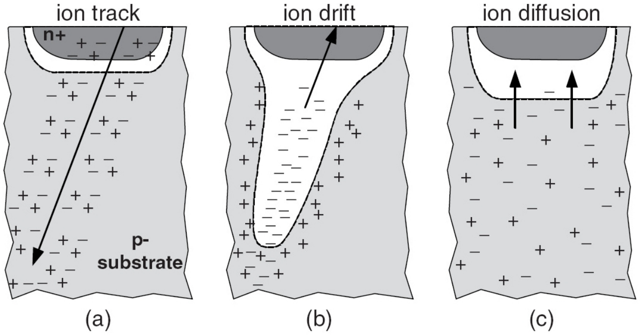


Figure 2.12: Generation of electron-hole pairs in a reverse-biased p-n junction: (a) Ionization of substrate atoms on the path of the striking particle; b) Formation of a funnel-shaped depletion region; (c) Drift charge collection is superseded by the diffusion charge collection [57].

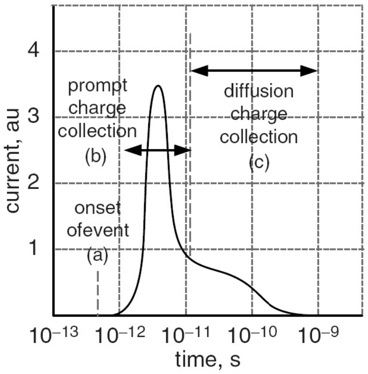


Figure 2.13: The time line and the current pulse shape for each phase [57].

where Q is the total charge deposited by this current pulse at the struck node, τ_f and τ_r are the falling time and the rising time constants, respectively [59]. This double exponential pulse exhibits a rapid rise time (∼10ps), but a more gradual fall time (∼200ps). Although there are several current pulse waveforms reported in [60], the current pulse model given in (2.3) has the advantage of being accurate as well as simple for further analytical modeling of soft errors. In memory elements such as SRAM cells and flip-flops, this particle strike induced current pulse may lead to a 1-to-0 flip or a 0-to-1 flip which corrupt the circuit logic state. However, in combinational circuits, it may cause a temporary change in the node voltage. This temporary change might be tolerated unless it is latched by a succeeding memory element.

2.2.2 Soft Error Rate (SER) Measurements

To directly measure the SER of an IC, one must have access to facilities where ionizing radiation events can be generated. SER testing can be performed at certain labs where accelerated neutron beam experiments can be performed [61]. This type of testing is costly and can only be done once the chip is fabricated. Another method for estimating the SER is to use simulations. Circuit simulators such as SPICE can be used to efficiently calculate the critical charge of a circuit, $Q_{critical}$ [62]. Then, with the help of empirical SER models, the SER is estimated. One empirical model has the form [58]:

$$SER \;\; \alpha \;\; N_{flux} \times CS \times \exp(-Q_{critical}/Q_s) \tag{2.4}$$

where N_{flux} refers to the intensity of the neutron flux, CS is the cross-section area of the struck node, and Q_s is the charge collection efficiency, which can be calculated by using process and device simulators. This empirical model can be used to determine the SER at a single node of a circuit. Calculations are made twice, once for when the node has a 1-to-0 flip error and once when the node has a 0-to-1 flip error, since $Q_{critical}$, Q_s and CS will be different for the two cases. For the 1-to-0 flip case, we consider only the Q_s and area of the NMOS drains that are attached to the node, since an upset can happen only when electrons are injected into the NMOS drain. Conversely, for the 0-to-1 flip case, we consider only the Q_s and area of the PMOS drains, since an upset can only happen when holes are injected into the PMOS drain [62]. The SER of the complete IC is determined by summing up the SER of every node (for both flips) [54].

2.2.3 State-of-Art Soft Error Mitigation Techniques

Electronic systems have different reliability requirements and, therefore, the same SER that is tolerated in some system applications may be unaccepted in other applications. For example, Central Processing Units (CPUs) with significant amounts of embedded SRAM can easily have the SER in excess of 50,000 FIT per chip, where 1 FIT = 1 error/10^9 device hours [57]. Assuming that a CPU with the SER = 50,000 FIT; a soft error is likely to occur once every 2.3 years. If such a CPU is used in a cell phone and the soft error causes bit flip in a critical memory element state, the customer would probably ignore the failure and, hence, such a SER amount is tolerated. However, when the same CPU is used in a life-support system, one failure per chip every 2 years may be unacceptable. Moreover, if a hundred of such CPUs are used in a mainframe computer, the resulting SER of the system has to be multiplied by their number. Simple calculations show that such a system will fail due to a soft error every week.

In the latter case, soft error mitigation techniques must be used to improve the SER and, hence, the electronic system robustness and reliability. State-of-art soft error mitigation techniques are divided into three levels. The first level targets at removing radiation sources or reducing their intensity, while the second level attempts to reduce the soft error at the circuit level by using additional circuitry. The last level is oriented at reducing the soft errors impact at the architecture level.

1. **Fabrication Level**

   Soft errors, caused by the alpha particles, can be mitigated by using pure materials in the chip packaging during the fabrication process. The most effective technique of mitigating the soft errors caused by the neutron-induced ^{10}B fission is to eliminate the BPSG from the semiconductor fabrication process flow, which has been already performed. Therefore, the high-energy cosmic neutron radiation has the main contribution to the system SER. Accordingly, soft error mitigation techniques at the circuit and architecture levels are used.

2. **Circuit Level**

   Increasing the capacitance of the storage node, that is susceptible to particle strike, is an obvious way of increasing $Q_{critical}$, and, therefore reducing the SER. Accordingly, upsetting a storage node in such a hardened circuit will require high energy particles. STMicroelectronics has reported using vertical Metal-Insulator-Metal (MIM) capacitors in SRAM cells for SER reduction [54]. The capacitors are located in the intermediate layers of interconnect between the first metal layer and the top Metal layer. Since there is no SRAM cell interconnects in those levels, the capacitors don't get in the way of the cell or increase its area. They just take up unused space above it. Unfortunately, that limits the ability to route over SRAM which results in about a 5% area overhead. A 120nm test chip was fabricated including conventional SRAM arrays and arrays of the new radiation hardened cells. The chip was bombarded with alpha particles and with neutrons. The hardened cells showed a large improvement in the SER over the standard ones [54].

   A few radiation hardening techniques are proposed in the literature for the SRAM. In [63], six extra PMOS transistors are added to the SRAM conventional six transistor (6T) cell targeting at holding the SRAM logic states when the SRAM cell is not accessed. This approach, despite its advantages in increasing the $Q_{critical}$, has a large area and power overhead (SRAM cell consists of 12 transistors (12T) instead of the conventional 6T). Some other hardened SRAM cells are proposed with less area and power overhead such as 8T SRAM and 10T SRAM.

   Moreover, system redundancy techniques are considered the most effective methods to reduce the SER. However, they have the largest overhead in terms of area and performance. Redundant systems are often used in highly-reliable real-time applications such as life-critical missions, aircraft and space apparatus control and transactional processing [57, 64]. Triple Modular Redundancy (TMR) allows a soft error in a single system to be ignored in favor of the majority data that is assumed to be correct. Thus, the correct data ”wins” the vote and appears at the output. The disadvantage of the TMR scheme is the extra area, power, and latency.

3. **Architecture Level**

   A number of techniques are developed for detection and correction of errors in SRAM cells at the architecture level using error detection and correction codes. All of the error detecting and correcting techniques used for SRAM arrays add a certain degree of redundancy into the system and therefore impact the system performance and occupy additional area. The choice of a detection/correction scheme is generally dictated by the tolerated SER level of the system [57]. The simplest error detection technique is the parity check. It works by adding an extra bit, the parity bit, to the data word so that the number of '1' data bits in the data word becomes even in case of even parity or odd in case of odd parity. If a single bit is flipped due to the particle strike soft error, this error will be detected by the parity check method. Unfortunately, when the number of corrupted data bits is even, the resulting error will not be detected by the parity check. In addition, the parity check scheme is incapable of identifying the corrupted data bits or correcting them. In contrast to the parity error detection, the Error Correction Codes (ECC) can identify the corrupted data bits locations and correct them by adding additional redundant bits.

2.2.4 Impact of Variability on Soft Errors

The impact of variability on $Q_{critical}$ and, accordingly, the SER has only gained attention recently. The primary research work focuses on the D2D variations using corner-based and worst-case based simulations such as [65]. Some other work tries to deal with WID variations using Monte Carlo simulations which are computationally expensive and time consuming [66]. Recently, the work in [37, 67] proposed an analytical model for the process variations impact on $Q_{critical}$ in the SRAM cells. However, the resulting model can account only for D2D variations. As a conclusion, The impact of variability on the SER using an analytical model that can account for both D2D and WID variations is still a missing work, which is developed in Chapter 3 of this thesis.

2.3 Aging Degradation Mechanisms

In Section 2.1, we have discussed how the transistor characteristics strongly depend on the fabrication processes and environmental factors. Once a chip is manufactured, packaged, tested (for correct functionality), and shipped to the customers, it is expected to function at the tested voltage, temperature, and frequency till the end of its usage life-time. However, physical changes can occur in the transistor due to movement of charges (i.e., electrons and holes) and breaking of atomic bonds with aging time. The key aging mechanisms that affect the transistor behavior with aging time are: (1) Hot Carrier Injection (HCI) that occurs due to defects in the gate stack by highly energized carriers under large electric fields causing shift in the threshold voltage, and (2) Bias-Temperature Instability (BTI) that arises due to the capturing of holes or electrons from the inverted channel in PMOS or NMOS transistors, respectively, by the broken Si-H bonds.

The BTI is called Negative BTI (NBTI) for the PMOS transistors and Positive BTI (PBTI) for the NMOS transistors. Among all the aging degradation mechanisms, NBTI is considered the major aging mechanism for advanced CMOS technologies and PBTI is also looming as a big concern [68]. Correspondingly, only the NBTI aging mechanism is considered in this thesis work.

2.3.1 NBTI Mechanism and Impact

NBTI is the generation of interface traps under negative bias conditions (i.e., $V_{GS} = -V_{DD}$) at elevated temperatures in PMOS transistors [69–74]. These interface traps are formed due to crystal mismatches at the Si-SiO$_2$ interface. During Si oxidation, the majority of the atoms are bonded to oxygen, whereas some of the atoms are bonded with hydrogen, leading to the formation of weak Si-H bonds. When a PMOS transistor is negatively biased, the holes in the channel dissociate these weak Si-H bonds and the interface traps are formed as shown in Figure 2.14. These interface traps can capture the holes in the inverted PMOS transistor channel. This makes the required V_{GS}, to reach the same strong inversion that occurs without these interface traps, more negative [1, 69, 75]. Hence, this results in an increase in the absolute PMOS transistor threshold voltage, $|V_{tp}|$. This $|V_{tp}|$ increase not only leads to reduced temporal performance but also causes reliability degradation and potential device failures [75]. Figure 2.15 shows the increase of $|V_{tp}|$ versus aging time indicating that there exist some partial recovery of the $|V_{tp}|$ shift when the stress voltage and temperature are removed.

In [22, 76, 77], it is stated that the PMOS transistor threshold voltage increase due to NBTI, $\Delta|V_{tp_{DC}}|$, under constant DC stress (i.e., the pMOS transistor gate voltage is grounded), follows a power law model with respect to the aging time as follows [22, 76]:

$$\Delta|V_{tp_{DC}}| = K_{DC} \times t^{\ell} \tag{2.5}$$

where K_{DC} is a technology dependent parameter (i.e., K_{DC} is a function of temperature, supply voltage, device geometry, and interfacial traps density), ℓ is an exponent depending on the NBTI mechanism ranging from 1/4 to 1/6, and 't' is the aging time in seconds. In real circuit operation, the effective ON time of the PMOS transistor is bounded by the operating frequency and the gate input probability. During the OFF time (i.e., the PMOS transistor gate voltage is connected to the supply voltage), the PMOS transistor experiences a partial recovery process, where $|V_{tp}|$ decreases back partially to its original value before stress [77] as shown in Figure 2.15. Accordingly, the PMOS transistor threshold voltage increase due to NBTI, $\Delta|V_{tp_{AC}}|$, under dynamic AC stress, is a scaled version of $\Delta|V_{tp_{DC}}|$ and given by [22, 76, 77]:

$$\Delta|V_{tp_{AC}}| \approx \rho \times \Delta|V_{tp_{DC}}| = \rho \times K_{DC} \times t^{\ell} \tag{2.6}$$

where ρ is a prefactor dependent on the operating frequency and the gate input probability. In [77], it is reported that the PMOS transistor life time is much longer under AC stress than DC stress by a factor of 4X.

2.3.2 NBTI State-of-Art Solutions

The NBTI degradation is characterized by using on-chip sensors and monitoring circuits. In [78], ring oscillators based on-chip NBTI degradation sensor is proposed. This NBTI sensor compares the frequencies of a stressed ring oscillator and a reference ring oscillator by utilizing a phase comparator. Another NBTI monitoring circuit is proposed in [79] that uses multiple Delay Locked Loops (DLLs) where the DLLs control voltage is modulated by the NBTI degradation. The main disadvantage of the aforementioned NBTI monitoring circuits is that there is no direct way to use them adaptively to compensate for NBTI degradation. In addition, the area overhead of these techniques is significant, especially, if additional circuitry is added to change them from NBTI monitoring circuits to NBTI compensation circuits. The requirements for a low area overhead NBTI compensation circuit is of paramount importance, especially, for NBTI-sensitive circuits. Thus, in Chapter 5, a new low area overhead ABB circuit is used for NBTI compensation by measuring the NBTI degradation and producing the appropriate body bias voltage that compensates for both NBTI degradation and process variations impacts.

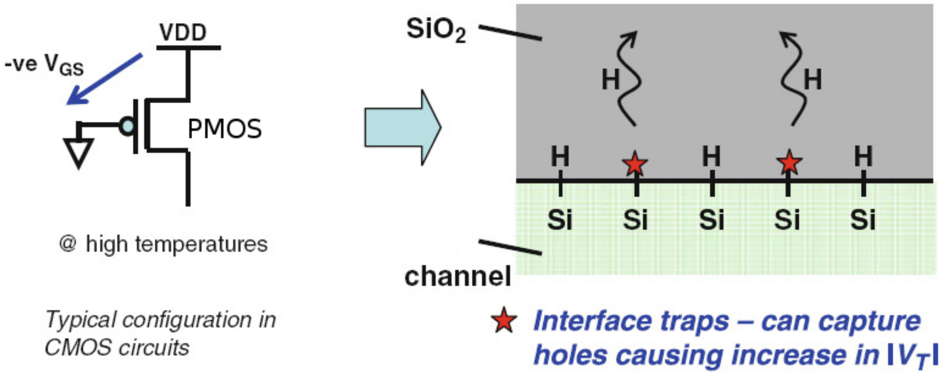

VDD
-ve V_{GS}
PMOS
@ high temperatures
Typical configuration in CMOS circuits
SiO_2
H
H
H
H
H
Si
Si
Si
Si
Si
channel
Interface traps – can capture holes causing increase in $|V_T|$

Figure 2.14: NBTI mechanism showing the interface traps [1].

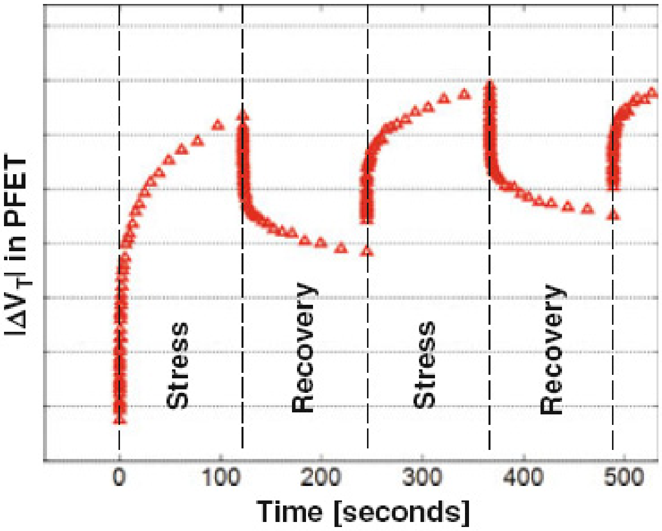


Figure 2.15: $|V_{tp}|$ stress-recovery-stress degradation behavior versus aging time [1].

2.4 Variation-Sensitive Digital Circuits Benchmarks

In this section, some background of the variation-sensitive digital circuits benchmarks, used in this thesis, is presented. These benchmarks include SRAM cells, flip-flops, and high performance circuits. In addition, the impacts of variability, soft errors, and NBTI on these circuits are discussed with the focus on the related research work that addresses these impacts and provides robustness solutions to mitigate them.

2.4.1 SRAM Cells

1. **Introduction**

   In today's SoCs, embedded SRAM dominates the chip area as shown in Figure 2.16. It is expected that SRAM area will exceed 90% of the overall chip area by 2014 as reported by International Technology Roadmap for Semiconductors (ITRS) [6]. This is driven by the demand of higher performance (multi-processing and multi-cores), lower power, and higher integration. To increase memory density, memory bitcells are pushed to achieve 50% area reduction each technology node as shown in Figure 2.17. This requires very aggressive design rules which makes SRAM cells more variation-sensitive due to their increased vulnerability to variations. For example, in state-of-art 45nm technology, an ultra high density bitcell area is approximately $0.25\mu\text{m}^2$.

   SRAM cells are usually used to implement memories that require short access times, low power, and robustness to environmental conditions [80]. SRAM cells, conventionally, have 6-transistors (6T) as seen in Figure 2.18. Transistors M_{pL}, M_{nL}, M_{pR} and, M_{nR} comprise a pair of cross-coupled inverters that use positive feedback to store a value. Transistors M_{aL} and M_{aR} are two pass transistors that allow access to the storage nodes for reading and writing. To write a value into an SRAM cell, the new value and its complement are driven on the bitlines (BL and BLB), and then the wordline (WL) is raised. The new value will overwrite the old value since the bitlines are actively driven by the write circuitry. To read a value from an SRAM cell, the bitlines are precharged high to the supply voltage, and then the WL is raised turning ON the pass transistors. Because one of the internal storage nodes is low, one of the bitlines starts discharging and a sense amplifier, which is connected to the bitlines, senses which of the bitlines is discharging and reads the stored value.

2. **Variability in SRAM Cells**

   While process variation affects performance and leakage of digital logic circuits, its impact on SRAM cells is much stronger. In advanced CMOS technology nodes, the predominant yield loss comes from the increase of process variations, which strongly impacts SRAM functionality as the supply voltage is reduced and higher density is packed. In particular, WID variations due to RDF and LER strongly impact SRAM operation. Figure 2.19 shows that V_t variation for SRAM devices increases significantly with scaling, which poses a major challenge for SRAM design. As an example, due to WID variations, each transistor in the bitcell experiences different type of variation, hence, the symmetry of the bitcell is lost. There are three main parametric failure mechanisms (also known as SRAM robustness failures) [8, 81, 82]: (1) Read access failure, (2) Read stability failure, and (3) Write stability failure.

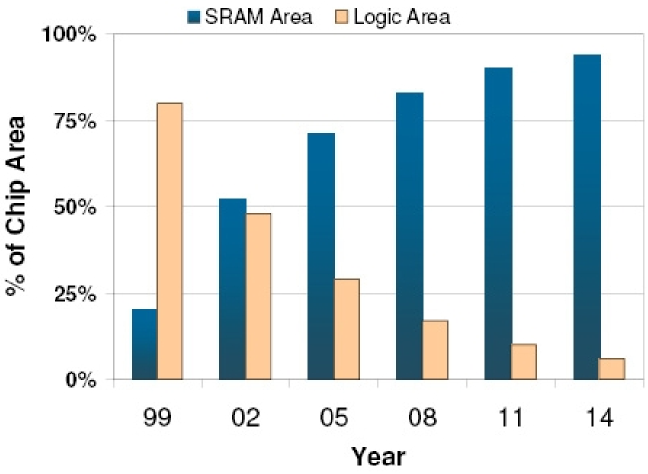


Figure 2.16: SRAM and logic area versus technology scaling. SRAM dominates chip area in modern SoCs and microprocessors [6, 54].

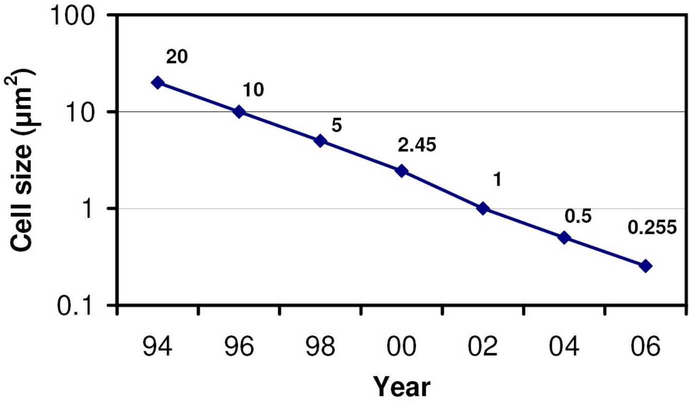


Figure 2.17: SRAM bitcell area scaling from 350nm down to 45nm technology nodes [6, 54].

(1) **Read Access Failure**

During the read operation, the WL is activated for a small period of time determined by the cell read current and the bitline capacitance as shown in Figure 2.20. The content of the cell is read by sensing the differential voltage between the bitlines. For successful read operation, the precharged to V_{DD} bitlines should discharge to a sufficient value which can trigger the sense amplifier correctly.

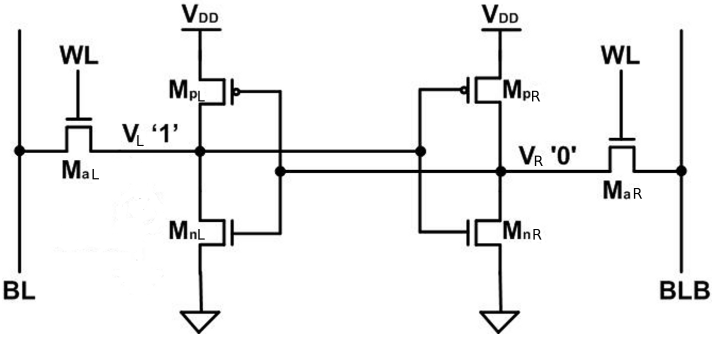


Figure 2.18: The 6T SRAM cell with Node V$_L$ is assumed to be at logic '1' and node V$_R$ is assumed to be at logic '0'.

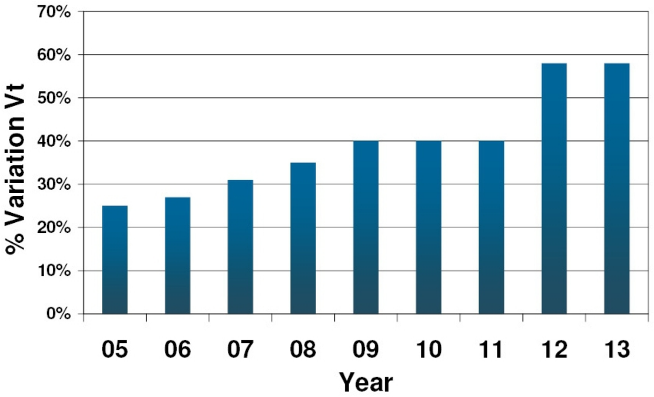


Figure 2.19: SRAM devices V_t variation scaling trend [6]

A failure happens if the bitcell read current (I_{read}) decreases below a certain limit. This may occur due to the increase in V_t for transistor M_{aR}, transistor M_{nR}, or both. This decrease in I_{read} reduces the bitline differential voltage sensed using the sense amplifier. This may result in wrong evaluation using the sense amplifier. This type of failure shows a strong impact on memory speed [8, 81, 82]. This is because the WL activation period is about 30% of the memory access time, and it is always desirable to reduce it to achieve higher speed operation [83].

(2) **Read Stability Failure**

SRAM cells are designed to ensure that the contents of the cell do not get altered during the read operation while the cell should be able to quickly change its state during the write operation. These conflicting requirements for read and write

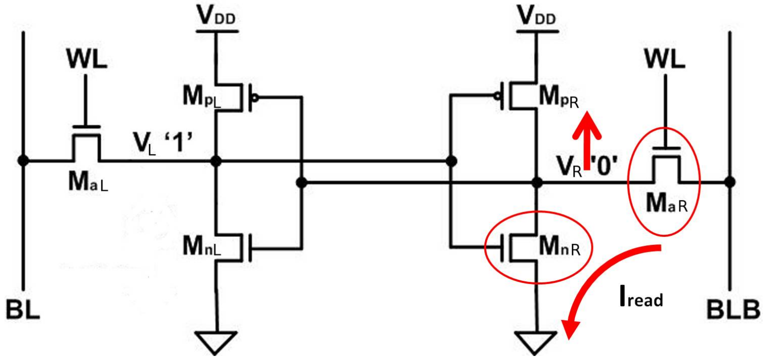
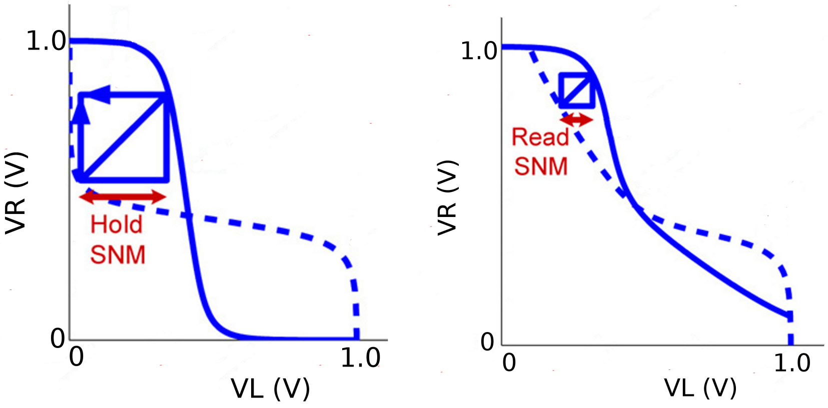


Figure 2.20: SRAM cell read operation.

operations are satisfied by sizing the bitcell transistors to provide stable read and write operations [8, 81, 82]. In read operation, the SRAM bitcell is most prone to failures. After the WL is enabled, the internal storage node storing a zero, V_R in Figure 2.20, slightly rises due to the voltage divider between transistors M_{aR} and M_{nR}, as displayed in Figure 2.20. If the voltage at V_R exceeds to the trip voltage of inverter (M_{nL}-M_{pL}), the cell may flip its state. In this case, stable read operation requires that transistor M_{nR} should be stronger than M_{aR}. Read stability is exacerbated by process variations which affect all the transistors in the bitcell [8, 81, 82]. To quantify the SRAM robustness against this type of failure, Static Noise Margin (SNM) is the most commonly used metric [83].

SNM is the noise stability measure of the SRAM cells and is defined as the minimum DC noise voltage necessary to change the state of the SRAM cell [84]. SNM is computed as the side length of a maximum square nested between the two Voltage Transfer Characteristic (VTC) curves of the SRAM cell (i.e., one VTC for inverter M_{nL}-M_{pL} and the other VTC for inverter M_{nR}-M_{pR} as shown in Figure 2.21). Depending on the SRAM operation, the SNM is classified as HOLD SNM (when the wordline is '0' and the cell is holding the data) or READ SNM (when the wordline is '1' and the data is read from the cell) as shown in Figure 2.21 [77, 84].

The READ SNM is more sensitive to threshold voltage deviations than the HOLD SNM. This is because in the HOLD mode, the nodes V_L and V_R are strongly coupled to each other making the cell less sensitive to threshold voltage deviations. However, in the READ mode, the connection of the bitlines to nodes V_L and V_R through the access transistors increases the cell sensitivity to the threshold voltage deviations [77]. Process variations cause large spread in SNM as shown in the measured SNM curves in Figure 2.22 [85].



(a)

(b)

Figure 2.21: The 6T SRAM cell SNM computation through the VTC curves for (a) HOLD mode when the wordline is '0' and (b) READ mode when the wordline is '1'.

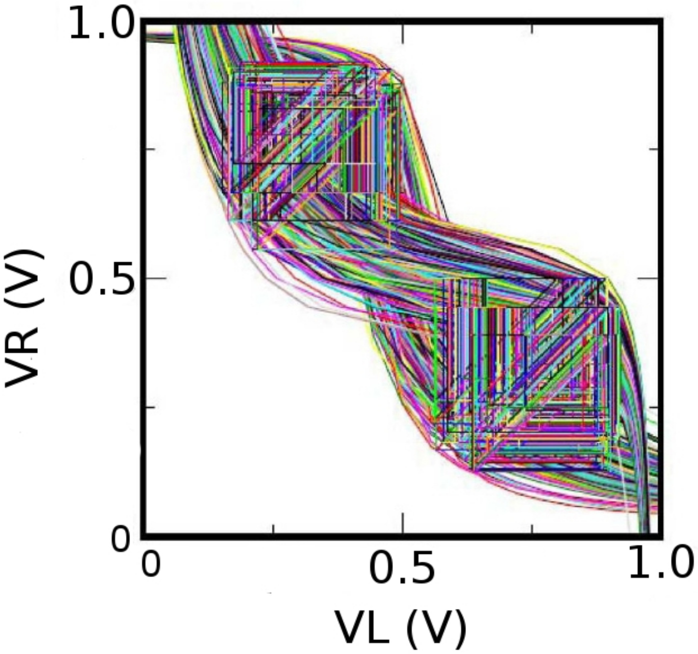


Figure 2.22: Measured SNM curves for 512 bitcells in 65nm technology node showing the strong impact of WID variations on SNM [85].

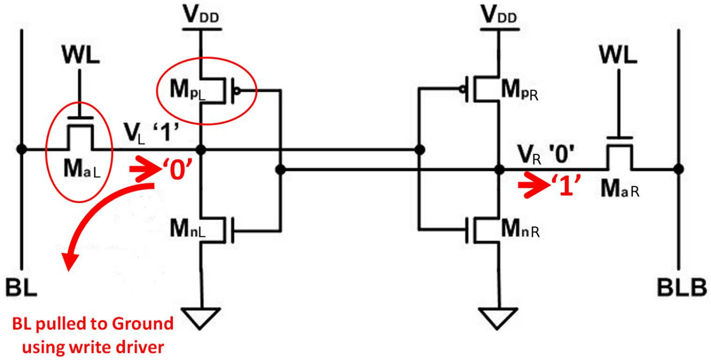


Figure 2.23: SRAM cell write operation.

(3) **Write Stability Failure**

In write operation, BL is pulled to zero using the write driver as shown in Figure 2.23. When transistor M_{aL} is turned ON, a voltage drop in the storage node, V_L, holding data '1' occurs, until it reaches below the trip voltage of inverter (M_{nR}-M_{pR}), where the positive feedback action begins. For stable write operation, transistor M_{aL} should be stronger than M_{pL}. The cell write stability is quantified by the Write Margin (WM). WM is the measure of the SRAM cell write stability and is defined as the maximum BL voltage that cause the cell to flip when the BLB is kept at V_{DD} (assuming V_L = '1' and V_R = '0', as shown in Figure 2.23) [86]. Due to WID variations, write stability failure happens when an SRAM cell fails to write a desired state during the write operation [82].

3. **Soft Errors in SRAM Cells**

As technology scales, SRAM junction capacitance, cell area and supply voltage are all scaled down. These reductions have opposing effects on $Q_{critical}$ (the critical amount of charge that has to be collected at a circuit node in order for a soft-error to occur). However, it has been shown that the combined effect causes SRAM single bit SER to saturate or slightly decrease with technology scaling [54], as shown in Figure 2.24. It is important to note that this trend in single-bit SER does not translate to reduction in the overall system failure rate due to the rapid growth in SRAM density as displayed in the same figure. In fact, SRAM failure rates are increasing significantly with scaling and have now become a major reliability concern for many applications. Therefore, SRAM memories must be protected by using ECC design. However, the probability of Multi-Cell Upsets (MCUs) is increasing with technology scaling [87], which limits the capability of the ECC codes. Hence, soft error mitigation techniques at the circuit level are required even if ECC codes are utilized.

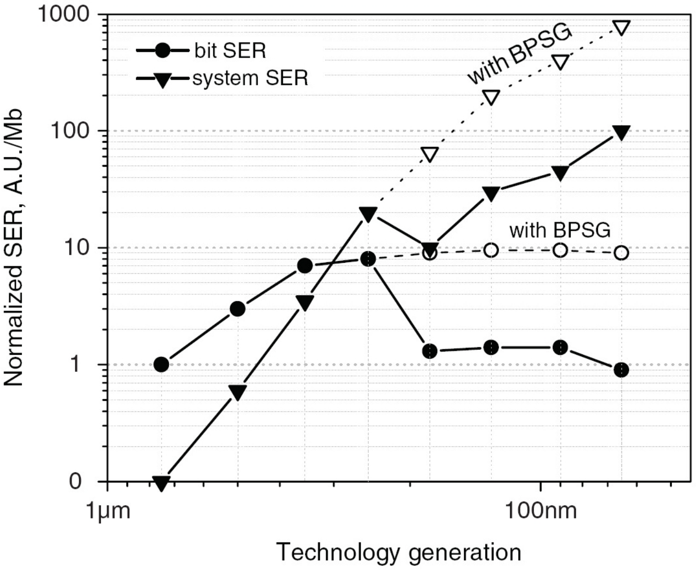


Figure 2.24: SRAM single bit and SRAM system SER as a function of the technology generation. Note a drop in the SER levels beyond 250nm technology generation due to the elimination of BPSG dielectric. Dotted lines show the simulated SER levels in case BPSG would have been used in the corresponding technology generations [54, 57]

Moreover, process variations lead to variations in $Q_{critical}$ which also impacts the SER [54]. Therefore, the techniques used to mitigate the soft errors of the SRAM array such as adding a coupling capacitor, as mentioned in Section 2.2.3, may result in increasing the SER variations and correspondingly, the value of this coupling capacitor is limited by the process variations. Some research work attempts to deal with WID variations using Monte Carlo simulations which are computationally expensive and time consuming [66]. Recently, the work in [37, 67] introduces an analytical model for the process variations impact on $Q_{critical}$ in the SRAM cells. However, this model can account only for D2D variations. As a conclusion, The impact of variability on the SER using an analytical model that can account for both D2D and WID variations is still a missing work, which is developed in Chapter 3 of this thesis.

4. **NBTI in SRAM Cells**

Under NBTI degradation, the $|V_{tp}|$ of the two PMOS transistors M_{pL} and M_{pR}, shown in Figure 2.18, increases with aging time according to (2.6). The gate input probabilities (the probability that the PMOS transistor is ON) at nodes V_L and V_R are denoted by p_L and p_R, respectively. Due to the symmetric structure of the SRAM cell, p_L and p_R add up to 1.0 (i.e., $p_L + p_R = 1.0$) [77]. Therefore, the

$|V_{tp}|$ degradation of the two PMOS transistors is not equal. In the following, the impacts of the NBTI $|V_{tp}|$ increase on the SRAM cell read robustness (i.e., SNM and read failure probability), write robustness (i.e., WM and write failure probability), sub-threshold leakage, and soft errors immunity, are discussed.

-i- **Read Robustness**

The SNM degrades over aging time under stressed conditions because the trip point of the left inverter (inverter M_{nL}-M_{pL} in Figure 2.18) is reduced due to the $|V_{tp}|$ increase of transistor M_{pL}. Accordingly, the cell becomes more vulnerable to flipping compared to the unstressed conditions [88]. In [77], it is shown that the HOLD SNM degrades by less than 3% whereas the READ SNM degrades by more than 10% at a temperature of 125^oC over 3 years aging time (t = 10^8 seconds) with $p_L = p_R = 0.5$.

Read failure probability is defined as the probability of a destructive read operation. The destructive read operation occurs when a voltage rise at the node storing '0' (i.e., V_R in Figure 2.18) exceeds the trip point of the load inverter (i.e., M_{nL}-M_{pL} inverter in Figure 2.18) and flips the original data during a read operation. This destructive read does not occur at the node storing '1' because the bitlines are precharged to V_{DD} before the read operation. Correspondingly, the node storing '1' is not affected during the read operation. The aging NBTI effect results in increasing the read failure probability further by reducing the inverter trip point. In [77], the read failure probability increases under stressed conditions by a factor of 2.9X compared to the unstressed case at a temperature of 125^oC over 3 years aging time (t = 10^8 seconds) with $p_L = p_R = 0.5$. The read access time is not impacted by the NBTI because this time is determined by discharging the bitline through the NMOS transistors M_{nR} and M_{aR}, which are not affected by the NBTI (assuming V_R = '0' as shown in Figure 2.18) [88].

-ii- **Write Robustness**

As the PMOS transistor threshold voltage, $|V_{tp}|$, increases with aging time due to NBTI, the node storing '1' (i.e., V_L in Figure 2.18) gets weaker and writing a '0' to this node becomes easier [77, 88]. Accordingly, the WM improves (i.e., increases) and the write failure probability is reduced over aging time. In [77], the WM is increased by 1.4% and the write failure probability is reduced by a factor of 2.4X at a temperature of 125^oC over 3 years aging time (t = 10^8 seconds) with $p_L = p_R = 0.5$.

-iii- **Sub-threshold Leakage**

As the PMOS transistor threshold voltage increases with aging time, the sub-threshold leakage current decreases exponentially, and accordingly, the SRAM total leakage is reduced with aging. In [77], the leakage current of the SRAM cell is reduced by 13% at a temperature of 125^oC over 3 years aging time (t = 10^8 seconds) with $p_L = p_R = 0.5$. It should be noted that the leakage reduction is maximized when $p_L = p_R = 0.5$. However, if the gate input probabilities are not equal (i.e., $p_L \neq p_R$), one of the PMOS transistors exhibits more $|V_{tp}|$ increase compared to the other. Unfortunately, the reduced leakage current through this higher $|V_{tp}|$ transistor is compensated by less probabilities of OFF time. For example, if $p_L = 1.0$ and $p_R = 0$, there is no leakage reduction because the higher $|V_{tp}|$ transistor (i.e., M_{pL} in this example) is always ON and accordingly, the leakage current is only determined through M_{pR}, which is not impacted by NBTI.

-iv- **Soft Errors Immunity**

For the proper operation of the SRAM cell, the PMOS pull-up transistors are sized to be weaker than the NMOS pull-down transistors. Consequently, the data node storing logic '1' (i.e., V_L in Figure 2.18) is the most susceptible to particle strikes [55, 89]. When a particle strike occurs at node V_L, the injected current pulls this node voltage down to '0' against the PMOS transistor M_{pL} current, which tries to recover the node voltage. Due to NBTI, M_{pL} current is reduced due to the increased $|V_{tp}|$ which reduces Q$_{critical}$ and increases the SER. In [89], the sensitivity of Q$_{critical}$ to transistor M_{pL} threshold voltage, $|V_{tpL}|$, is given by:

$$\frac{\Delta Q_{critical}}{\Delta|V_{tpL}|} = -\frac{Q_{critical}}{V_{DD} - |V_{tpL}|} \tag{2.7}$$

where V_{DD} is the supply voltage. According to (2.7), Q$_{critical}$ is reduced by 6.25% for $V_{DD} = 1.0$V, $|V_{tpL}| = 0.204$V, and $\Delta|V_{tpL}| = 50$mV. This Q$_{critical}$ reduction results in a large increase in the SER due to the exponential relationship expressed in (2.4), especially, in large size SRAM modules.

2.4.2 Flip-Flops

1. **Introduction**

The absolute majority of high-performance digital designs today utilize a synchronous clock to order events [90]. Although the principle of synchronization is easy in the system design perspective, ordering all events in a high performance design in a synchronous fashion requires generation and distribution of clock signals at multi-GHz clock frequencies, which is extremely challenging. It is therefore of highest interest to design the flip-flops such that they are optimized for their desired task while taking into account the yield and the reliability constraints.

Synchronization circuits such as latches and flip-flops constitute the clocked registers that synchronize the data flow in a VLSI circuit. Hence, flip-flops and latches are among the most important circuit blocks in a digital synchronous chip design. Ideally, timing circuits like flip-flops and latches should add as little latency as possible, and have low power dissipation. In practice, however, clocked registers can actually consume a substantial fraction of the clock-cycle period, and dissipate a considerable portion of the total power.

Latches are the simplest kind of synchronizing circuits in a sequential design. A latch is a level sensitive device that is either transparent or opaque, depending on the signal level of the clock input. A simple schematic of a transmission gate latch is shown in Figure 2.25. When the clock signal (Clk) is high, the latch lets the input (D) pass to the output (Q), while if the clock is low, the output (Q) will hold the previous input data on the output.

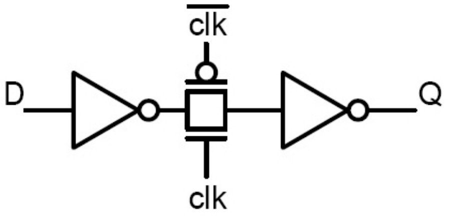


Figure 2.25: Schematic example of a simple level-sensitive latch [91].

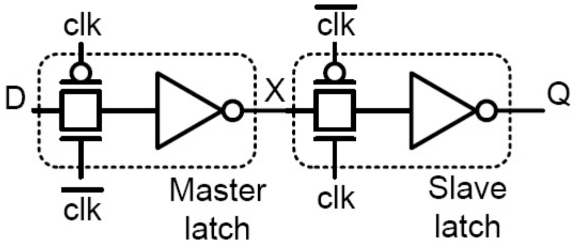


Figure 2.26: Schematic example of a positive edge-triggered flip-flop [91].

An edge-triggered flip-flop samples the data input on one edge of the clock, but in contrast to a level-sensitive latch, keeps the sampled data on the output during the remainder of the clock period. A simple master-slave flip-flop can be constructed from two cascaded level-sensitive latches, as shown in Figure 2.26. When the clock signal (Clk) is low, the first latch, called master latch, is transparent and the input is transferred to the intermediate node (X). The second latch, called the slave latch, is opaque so the output (Q) is held at its previous state. When the clock signal (Clk) makes a low-to-high transition, the master latch becomes opaque, and the slave latch becomes transparent, and the intermediate data at (X) is transferred to the output (Q). The data on the output is valid for the remainder of the clock period [90].

In the following, the timing terminologies of the flip-flops are defined, the Power-Delay Product (PDP) design space is explained, and a brief introduction about the main flip-flops topologies is presented.

A) **Flip-Flops Timing Characteristics**

A timing diagram of a positive edge-triggered flip-flop is shown in Figure 2.27. All timing relations for the edge-triggered flip-flop are referred only to the sampling clock edge. The timing relations for an edge-triggered flip-flop are defined by essentially four different delays, which are [90]:

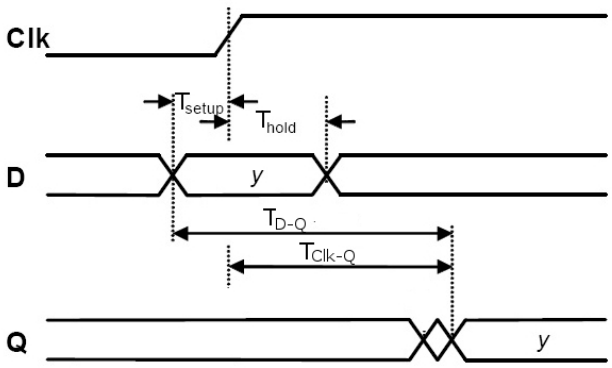


Figure 2.27: Timing characteristics for a positive edge-triggered flip-flop [91].

i. **Setup time (T$_{setup}$)**
It is defined as the minimum time that the input data should be available before the clock sampling edge arrival.

ii. **Hold time (T$_{hold}$)**
It is defined as the minimum time that the input data should be available after the clock sampling edge arrival.

iii. **Clock-to-output delay (T$_{Clk-Q}$)**
It represents the delay from the sampling clock edge (Clk) to the time at which the latched data is valid at the output (Q).

iv. **Data-to-output delay (T$_{D-Q}$)**
It represents the delay from a transition of the input data (D) to the time at which the latched data is valid at the output (Q). This delay can be determined as the sum of the setup time and the clock-to-output delay.

B) **Power-Delay Product (PDP) Design Space**

When optimizing flip-flop circuits, trade-offs between power and delay can be made as for all logic design. A power-efficient flip-flop is one that for a certain delay has the minimal power dissipation and vice versa. This can be illustrated in a design-space graph shown in Figure 2.28, which shows the total power for two flip-flops plotted versus the total minimum latency (T_{D-Q}). If a fair and accurate comparison between different flip-flops topologies should be done, a power-delay plot like Figure 2.28 is needed. As an example comparing flip-flop FF-1 with FF-2 only at one point will yield that one of the topologies is better than the other in general. However, the truth might be that they are the better choice in different parts of the design space. For instance, a low-latency flip-flop that dissipates more power (FF-1 at τ_1) could be used in critical parts of a design, while using a slower less power-consuming flip-flop (FF-2 at τ_2) in noncritical parts. Therefore, using an optimization program to find the optimal sizing of the flip-flop that will exhibit the minimum PDP is the best method to trade-off the flip-flops power consumption and latency delay.

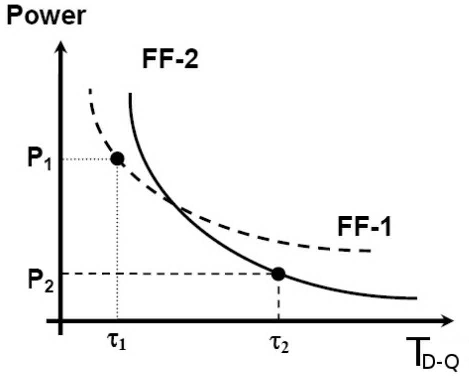


Figure 2.28: Power-delay design space for two different flip-flops topologies.

C) **Common Flip-Flops Topologies**

In the literature, there exists a large number of flip-flop circuits proposed, which can be classified mainly into three categories. These are master-slave latch pairs, pulsed latches, and sense-amplifier based flip-flops [90, 92, 93].

-i- **Master-Slave Flip-Flop (MSFF)**

The most common approach to build an edge-triggered flip-flop is to combine two level-sensitive latches, which are clocked on opposite clock phases. An example of a common static Transmission Gate Master-Slave Flip-Flop (TG-MSFF) is shown in Figure 2.29. The setup time of this flip-flop is mainly determined by the propagation delay of the master latch, and the output latency is determined by the propagation delay through the slave latch, resulting in a quite large latency delay. Therefore, this flip-flop is frequently used in non-critical data paths, where large latency delay is not impacting the performance of the system.

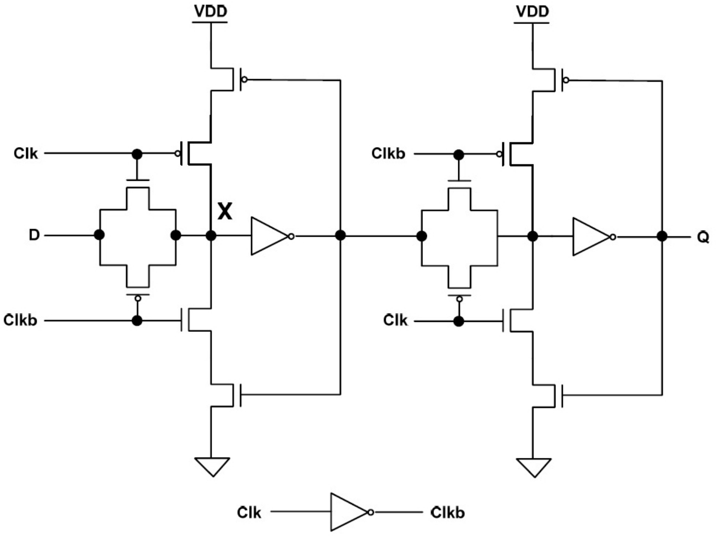


Figure 2.29: Transmission Gate Master-Slave Flip-Flop (TG-MSFF) [91–93]

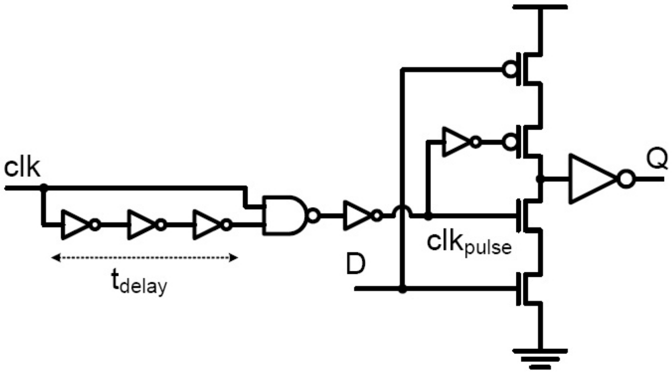


Figure 2.30: Pulsed C^{2}MOS latch with external pulse-generator [91–93].

-ii- **Pulsed Latches**

The principle of a pulsed latch is to create a short pulse on the latching edge of the clock, and then clock the latch with that pulse, thereby obtaining an edge-triggering behavior. A simple example of a pulsed-latch using a clocked-CMOS (C^{2}MOS) latch is shown in Figure 2.30. The pulse generator could be an external circuit or integrated in the latch design [90]. However, an external clock-pulse generator could be shared with a number of other latches in order to reduce the total clock power. For a rising edge of the clock (Clk), the output of the pulse-generator (Clk$_{pulse}$) will go high, making the

latch transparent. After a delay (t_{delay}) the output of the pulse-generator will go low, thus making the latch opaque. During the high pulse the latch transfers any change of data on the input. This property is referred to as negative setup time, because data is correctly latched even though arriving after the rising edge of the main clock signal. However, this negative setup time property exists at the expense of poor hold time behavior, because during the duration of the latching pulse, the input is not allowed to change data erroneously in order not to corrupt the output value. Hence, pulsed flip-flops with negative setup time usually have large positive hold times.
Several pulsed latches have been described in the literature, and some of them are utilized as low-latency flip-flops in critical pipeline stages in high-performance microprocessors. One of the most popular type of this topology is the Semi-Dynamic pulsed-latch Flip-Flop (SD-FF), presented in [92–94]. This flip-flop is used frequently in high performance circuits and will be considered as a representative of the pulsed flip-flops topology.

-iii- **Sense-Amplifier Based Flip-Flops**
A third technique to implement an edge-triggered flip-flop is to utilize a sense-amplifier to sample the data [90–93]. A typical sense-amplifier based flip-flop is shown in Figure 2.31, where a pre-charged sense-amplifier front-end is used to sample the complementary data inputs when the clock makes a rising transition. A NAND-based SR-latch captures the sampled data and holds it until next rising clock-edge. Due to the amplification provided by the feedback in the cross-coupled inverters, the flip-flop can sample input signals with small amplitude differences. Therefore, sense-amplifier flip-flops could be utilized as synchronous level-converters between different power-supply regions [95]. Another advantage with the sense-amplifier flip-flop is the low number of clocked transistors, which gives low clock load. One of the largest drawbacks with the sense-amplifier flip-flops is the precharged behavior of the sample-stage, which is power-consuming, especially when the data activity on the inputs is low.

2. **Variability in Flip-Flops**

The increased clock frequencies, in microprocessors and high performance VLSI applications, lead to very deep pipelining which means that hundreds of thousands of flip-flops are required to control the data flow under strict timing constraints. A violation of the timing constraints at a flip-flop can result in latching incorrect data causing the overall system to malfunction [10]. Deterministic gate sizing tools size the flip-flops circuits to optimize the PDP, as shown in Figure 2.28. However, due to process variations, a large number of circuits might not meet the target delay.

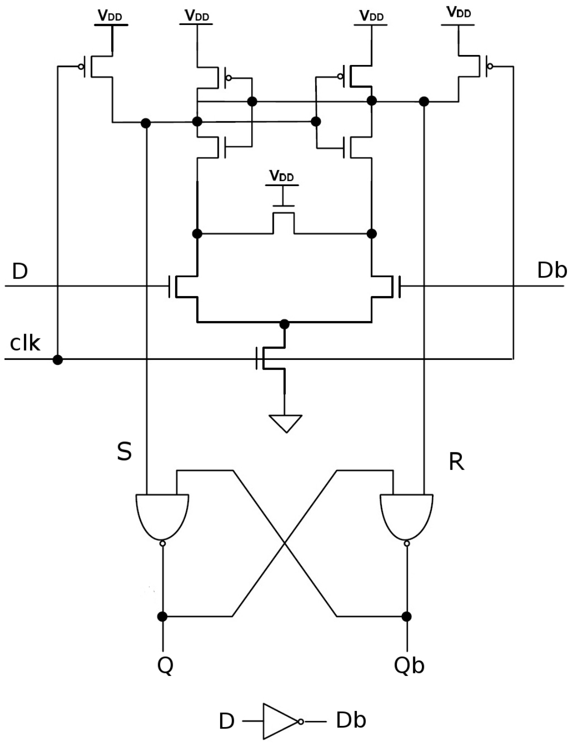


Figure 2.31: Sense Amplifier based Flip-Flop (SA-FF) [90–93]

Consider as an intuitive example a flip-flop that is designed for optimum PDP, which exhibits a specific target delay. Due to process variations, the delay is normally distributed with the probability density function (pdf) shown in Figure 2.32. This figure shows that 50% of the total number of flip-flops will not meet the desired target delay constraint. Therefore, the flip-flops have to be designed using statistical gate sizing tools to improve the timing yield [34, 35].

The research work in [96] presents a comparative analysis between different flip-flops topologies considering process variations (i.e., which flip-flop topology exhibits the highest delay variations). However, this work in [96] performs this comparative analysis by using deterministic flip-flops sizing to achieve minimum PDP. This makes this comparative analysis, in [96], impractical because these flip-flops will not be used in actual circuits with that minimum PDP sizing that results in 50% timing yield as shown in Figure 2.32. Therefore, a comparative analysis between different flip-flops topologies, by using statistical flip-flops sizing to achieve 99.87% timing yield, considering the required power and energy overheads to achieve this timing yield improvement, is more practical and fair. This comparative analysis is conducted in chapter 4.

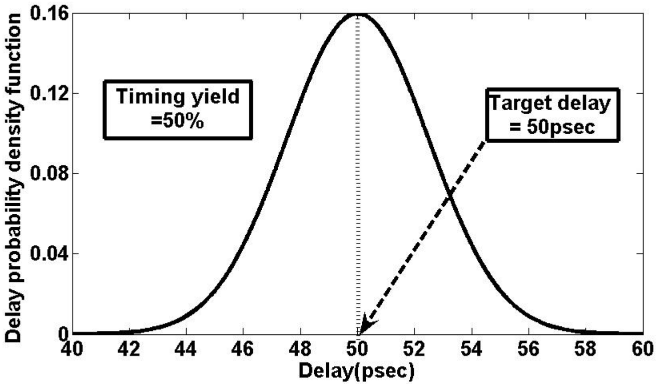


Figure 2.32: The delay pdf due to process variations using deterministic sizing algorithms illustrating that up to 50% of flip-flops will not meet the target delay.

3. **Soft Errors in Flip-Flops**

At the chip level, the contribution to the SER from flip-flops is growing due to feature size scaling and supply voltage reduction [56, 97]. In the mean time, it is more common to protect SRAM memories with ECC circuitry, thus reducing their SER. As a consequence, the relative SER contribution from the flip-flops is increasing since flip-flops can not be protected by ECC. Combinational logic circuits currently have a minor impact on the chip SER, particularly at moderate operating frequencies. However, their contribution to chip SER is also growing with technology scaling as shown in Figure 2.33 [97, 98].

An IC designer utilizing standard-cell libraries generally can choose from a large variety of flip-flops and latches. The choice depends on the desired performance and power dissipation. When moving into the technology nodes beyond 100nm, the process variations and the SER of the flip-flops are new design metrics that have to be taken into account. Therefore, it is important that accurate data are available about the SER of flip-flops that are used in production designs.

Several studies on the SER of flip-flops have been reported. SER measurements are published for a 90nm test chip [97]. Then, a study is presented on dedicated latch designs processed in a 65nm technology [36]. A comparative analysis between different flip-flops considering soft errors immunity is presented in [99, 100]. However, this comparative analysis is conducted on minimum PDP flip-flops sizing scenario. In the presence of process variations, it is more relevant to conduct this comparative analysis while the flip-flops are sized to achieve timing yield improvement. This comparative analysis is performed in Chapter 4.

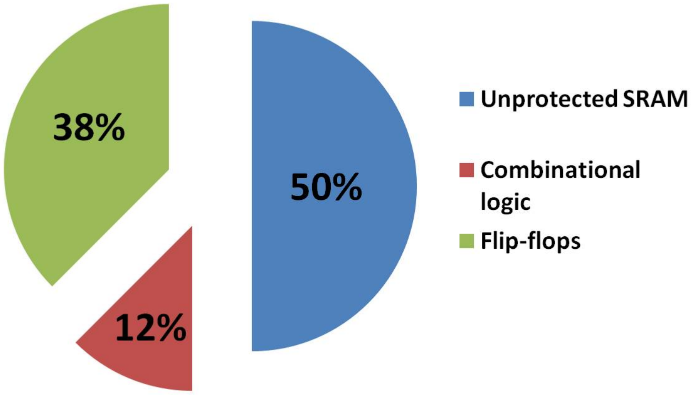


Figure 2.33: The contribution of unprotected SRAM memories, flip-flops, and combinational logic to the chip SER [101].

4. **NBTI in Flip-Flops**

   The degradation in the PMOS transistor V_t, due to NBTI, results in increasing T_{setup}, T_{hold}, and T_{D-Q} of the flip-flops. This increase in the flip-flops delay metrics results in timing violation and reduces the flip-flops robustness [102–104]. The effect of NBTI degradation on the setup and hold times of the flip-flops is discussed in [104] and it is shown that NBTI tightens the setup and hold timing constraints imposed on the flip-flops. Moreover, it is found in [104] that different topologies of flip-flops exhibit different levels of susceptibility to NBTI induced degradation in their setup and hold time values.

2.4.3 High Performance Circuits

In this thesis, the high performance circuits refer to these logic circuits that are used in the critical path design of microprocessors and high performance VLSI applications. These circuits consist of static logic CMOS gates (i.e., inverters, NAND, NOR, and transmission gates) and dynamic logic gates (i.e., Domino logic gates).

1. **Variability in High Performance Circuits**

   The V_t variations of the transistors result in variations in the speed between different chips. Thus, if the chips are designed using nominal V_t of the transistors to run at a particular speed, some of them will fail to meet the desired frequency constraint, which leads to parametric timing yield loss [1, 105]. Figure 2.8 portrays how the variations in the threshold voltage (under the influence of both D2D and WID variations) translate into frequency distribution for 65nm CMOS technology, which is expected to be worse with the continued technology scaling.

Moreover, the V_t variations have different impacts on dynamic logic circuits. These circuits operate on the principle of precharge and evaluate. The output is precharged to logic '1' in the negative phase of the clock (i.e., CLK = '0'). The positive phase of the clock (i.e., CLK= '1') allows the inputs to decide if the output will be kept precharged to logic '1' (through the weak PMOS keeper transistor) or will be discharged to ground through the Pull Down Network (PDN), which consists of NMOS transistors. The keeper transistor is used to keep the output node at logic '1' when the PDN network NMOS transistors are OFF, and designed weak (i.e., smaller W/L) to allow the output node to discharge to ground when the PDN is ON.

Since the information is saved as charge at the output node capacitor, dynamic logic is highly susceptible to noise and timings of input signals. Due to the inherent nature of the circuit, a slight variation in transistor threshold voltage can kill the logic functionality. For example, consider the domino logic shown in Figure 2.34, If the V_t of the NMOS transistors in the second stage is low due to process variation, then a small change in Out1, IN4, or IN5 can turn the PDN path ON and result in wrong evaluation of Out2 [1, 105]. In register files, increased leakage due to lower V_t dies has forced the circuit designers to upsize the keeper to obtain an acceptable robustness under worst-case V_t conditions. Large variation in V_t indicates that a large number of low leakage dies suffer from the performance loss due to an unnecessarily strong keeper [1, 105].

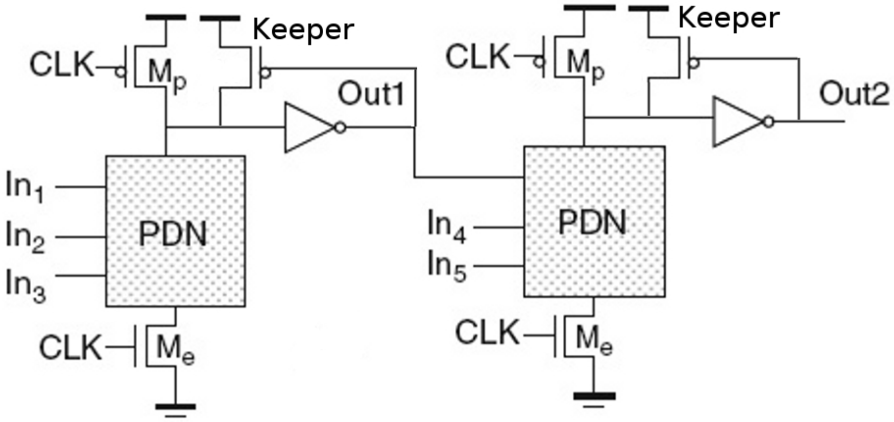


Figure 2.34: Example of a dynamic logic circuit [1, 105].

2. **Soft Errors in High Performance Circuits**

   Single Event Upset (SEU) is a voltage transient caused by neutron or alpha particles from cosmic ray or package materials, respectively. These voltage transients may flip bits in SRAM and flip-flops, causing soft errors. However, these voltage transients can happen on any node in combinational logic producing a transient pulse. This transient pulse can propagate through logic gates and finally be latched by a memory element, resulting in a soft error [106]. Figure 2.33 shows that the contribution of combinational logic to the SER is 12%. However, this contribution is growing with technology scaling.

3. **NBTI in High Performance Circuits**

   The PMOS transistors V_t shift due to NBTI rises up to 50mV through the life time in 90nm technology. This shift is translated to more than 20% degradation in circuit speed or in extreme cases to a functional failure [107]. Experimental data further indicates that NBTI worsens exponentially with thinner gate oxide and higher operating temperature. In fact, as gate oxide scales thinner than 4nm, NBTI has gradually become the dominant factor to limit circuit life time [1]. Even though tremendous efforts have been spent to improve the fabrication process, the impact of NBTI on circuit performance becomes so severe that technology improvement alone is not sufficient. In the nanometer regime, it is essential to develop design methods to understand, simulate, and minimize the degradation of circuit performance in the presence of NBTI, in order to ensure robust circuit operation over a desired period of time [107].

2.5 Summary

In this chapter, we have presented a survey on the sources and impacts of variations, soft errors, and NBTI aging degradation that affect on the circuit design robustness in the nanometer regime. It is shown that the impacts of these challenges are getting worse with technology scaling, especially, on variation-sensitive digital circuits such as SRAM, flip-flops, and high performance circuits. We also presented an overview on different research works in the area of mitigating these challenges to increase the circuit robustness and yield. Moreover, in this chapter, some background on the variation-sensitive circuits benchmarks, used throughout this thesis work, is introduced. In the following chapters, we will target:

-i- Exploring the impact of process variations on the soft errors immunity of SRAM cells and flip-flops. This will help in finding the limitations imposed by the process variations on the soft errors mitigation techniques. (Chapter 3)

-ii- Conducting a comparative analysis between different flip-flops topologies considering the power and energy overheads required for yield improvement. The effect of this yield improvement, which is performed by using statistical gate sizing, on the flip-flops soft errors immunity is also discussed. This comparative analysis will help flip-flops designers to select the best flip-flop topology that satisfy their system delay, power, and robustness requirements. (Chapter 4)

-iii- Introducing new low area overhead ABB circuits for process variations compensation of high performance circuits and for NBTI compensation of SRAM arrays. The effectiveness of these ABB circuits is verified by using post layout simulation results and test chip measurements. (Chapter 5)

-iv- Proposing new negative capacitance circuits, for the first time, for timing yield improvement of dynamic logic circuits and for read access yield improvement of SRAM arrays. These negative capacitance circuits are connected to the highly capacitive circuit nodes to reduce the parasitic capacitance at these nodes. The effectiveness of these negative capacitance circuits is verified by using post layout simulation results and test chip measurements. (Chapter 6)

Chapter 3

Analytical Soft Errors Immunity Variation Models for Nanometer CMOS SRAM Cells

In this chapter, design-oriented analytical models, for the critical charge variability, accounting for both D2D and WID variations, are proposed. These models are derived for the super-threshold SRAM cells and the sub-threshold SRAM cells. Several design insights, showing the design knobs that can be used to reduce the SER and its variability, are presented in this chapter. These design insights help in understanding the limitations imposed by the process variations on the soft errors mitigation techniques.

This chapter is organized as follows. In Section 3.1, a brief introduction is presented. The analytical critical charge variability models and the corresponding design insights are presented in Sections 3.2 and 3.3 for the super-threshold SRAM cells and the sub-threshold SRAM cells, respectively. Finally, in Section 3.4, some conclusions are drawn.

3.1 Introduction

SRAM occupies the majority of the die area in SoCs and microprocessors. Accordingly, several leakage reduction techniques such as supply voltage reduction and dynamic voltage scaling, are applied to SRAMs to limit the overall chip leakage. These leakage reduction techniques in conjunction with the SRAM lower nodes capacitances increase the SRAM soft errors vulnerability. In addition, process variations are expected to worsen in future technologies. Therefore, nanometer SRAM cells are more susceptible to the particle strike soft errors and the increased statistical process variations. Due to the existence of process

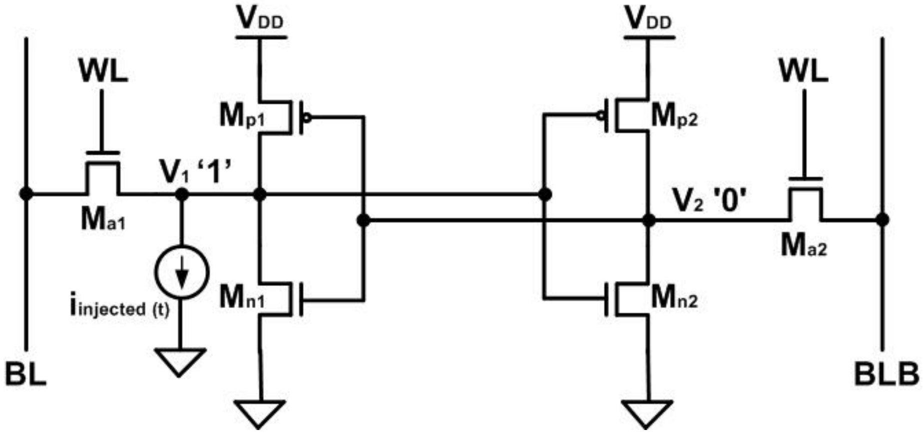


Figure 3.1: The SRAM cell with the particle strike induced current pulse ($i_{injected}(t)$). Node V_1 is assumed to be at logic '1' and node V_2 is assumed to be at logic '0'.

variations, the Soft Error Rate (SER) has variations around its nominal value which can result in SRAM failure to meet robustness constraints.

Figure 3.1 shows a typical six transistor (6T) SRAM cell. It consists of two cross-coupled inverters, that store two complementary logic values ('1' and '0') at their output nodes. These output nodes are denoted by V_1 and V_2. As discussed in Chapter 2, if the charge, collected by the particle strike at the storage nodes of the SRAM cell (i.e., nodes V_1 and V_2 in Figure 3.1), is more than a minimum value, the node is flipped and a soft error occurs. This minimum value is called a critical charge ($Q_{critical}$), which can be used as a measure of the SRAM cell vulnerability to soft errors [51, 56, 58, 60, 99, 100]. This critical charge, $Q_{critical}$, exhibits an exponential relationship with the SER as expressed in (2.4) [60], and consequently, $Q_{critical}$ should be designed high enough to limit the SER.

In the following, design-oriented analytical models, for the critical charge variability, accounting for both D2D and WID variations, are proposed. The derived models are verified and compared to Monte Carlo simulations by using industrial hardware-calibrated 65nm CMOS technology transistor model, reported in Appendix D. These models are derived for the super-threshold SRAM cells (used in high performance applications) and for the sub-threshold SRAM cells (used in low power systems). In addition, it is explained how these models can be extended to account for both super-threshold flip-flops and sub-threshold flip-flops as well.

3.2 Super-Threshold SRAM Cells

Recently, researchers have attempted to calculate the critical charge nominal value as well as addressing the impact of process variations on the critical charge in super-threshold SRAM cells. However, most of this research is conducted by using Monte Carlo analysis tools [65, 66, 108, 109], which are time consuming and provide little design insights. Moreover, these Monte Carlo analysis tools are not scalable with technology. From a design perspective, few articles have been published on modeling the critical charge and its variations. In [59, 110, 111], different models for the critical charge are proposed, however, these models overestimate the critical charge value and provide little insights to circuit designers. In [37, 67], an analytical model to estimate the critical charge is presented. Despite its accuracy in modeling the critical charge, this model depends mainly on SPICE simulations. Thus, this model can be used only when dealing with D2D variations. These D2D variations are estimated by applying corner-based analysis that have been already performed in [37, 67]. These techniques tend to be inefficient, and completely pessimistic in the presence of relatively large variations. Therefore, statistical design-oriented techniques are required, especially, when dealing with the WID variations [23].

In this section, an accurate analytical model of the critical charge, accounting for both D2D and WID variations, is proposed. This model is further simplified to provide more design insights on the impact of process variations on the critical charge. The derived model is simple, scalable in terms of technology scaling. Moreover, it shows explicit dependence on design parameters such as node capacitance, transistors sizing, transistor parameters, and supply voltage. The results are verified by using SPICE transient and Monte Carlo simulations and an industrial hardware-calibrated 65nm CMOS technology transistor model reported in Appendix D. These results are particularly important for the design of nanometer technology, when WID variations dominate the process variations [4].

3.2.1 Review of the Previous Critical Charge Models

The previous critical charge models, introduced in [37, 59, 67, 110, 111], exhibit some limitations, that make them incapable of modeling the WID variations. For example, the model introduced in [59] modeled $Q_{critical}$ as follows:

$$Q_{critical} = C_1 V_{DD} + i_{p1max}\ t_f \tag{3.1}$$

where i_{p1max} is the maximum restoring current of the transistor M_{p1}, as shown in Figure 3.1. The critical charge obtained from this model is overestimated, because of the following two reasons: (1) The flipping threshold voltage of an inverter is less than V_{DD} (around

$V_{DD}/2$) and (2) The restoring current term ($i_{p1max}\,t_f$) considers only the maximum current value which is not a valid assumption for the time varying restoring current. These issues have been refined to some extent in [110], by defining the critical charge as:

$$Q_{critical} = \int_0^{V_{trip}} C_1 \, dV + \rho \; i_{p1} \; t_{pulse} \tag{3.2}$$

where V_{trip} is the tripping point of the SRAM cell, ρ is a correction factor, and t_{pulse} is the duration of the particle induced current pulse. This model provides a better estimation of $Q_{critical}$. However, both models in [59] and [110] can not be used to model the variations (D2D or WID variations), since they account only for M_{p1} current and ignore the currents of M_{n2} and M_{p2} which can have a significant contribution to the critical charge variability.

The work in [111] presents an analytical method to calculate $Q_{critical}$ in terms of the transistor parameters and the injected current pulse magnitude and duration. This model utilizes a rectangular current pulse, instead of using an exponential current pulse, to model the particle strike induced current pulse, which makes its accuracy in calculating $Q_{critical}$ very poor. If an exponential current pulse is to be used, the model becomes complex and provides little insights. In addition, the model ignores the NMOS transistors current (i.e., M_{n2}), and does not show its effectiveness in calculating $Q_{critical}$, when different transistor parameters vary.

Finally, the work in [37, 67] introduces a very accurate model in calculating $Q_{critical}$. However, the value of the injected current pulse charge, Q, is obtained via iterative transient simulations by increasing Q by a small amount (∼0.001fC) in SPICE till flipping occurs. Although this method can be used in calculating D2D variations by using corner-based or worst-case methods, in which the value of Q can be obtained by using SPICE simulations. This technique can not be used for the WID statistical variations, since Q must be calculated for each statistical run. Consequently, this model accounts only for D2D variations, which have been already performed in [37, 67].

The proposed accurate model overcomes all the previous limitations, and introduces analytical formulas for $Q_{critical}$ which can be employed without SPICE simulations (assuming that the transistor parameters are known). Moreover, the developed accurate model accounts for both D2D and WID variations. The disadvantage of this accurate model is its complexity in the WID variations modeling, which is refined by using the simplified model. The simplified model introduces only three equations that provide useful design insights. Based on these design insights, the design knobs that can be used to reduce the SER and its variability are extracted.

3.2.2 Accurate Model Assumptions and Derivations

The SRAM cell has its highest susceptibility to particle strikes in the standby mode (i.e., when WL = '0'), since, in the standby mode, the storage nodes are disconnected from the highly capacitive bitlines. Therefore, their critical charge is smaller than that when the SRAM cell is operating in the read mode. In addition, the SRAM cell is most likely to be in the standby mode during its operating time. Thus, the access transistors M_{a1} and M_{a2} are excluded from the analysis.

For the proper operation of the SRAM cell, the PMOS pull-up transistors are sized to be weaker than the NMOS pull-down transistors, as discussed in Chapter 2. Consequently, the data node storing logic '1' is the most susceptible to particle strikes. It has been reported that $Q_{critical}$ of a 0-to-1 flip in SRAM is about 22X larger than that for a 1-to-0 flip [55]. Therefore, the proposed critical charge models account for the 1-to-0 flip case only. Assuming that node V_1 stores logic '1' and accordingly node V_2 stores logic '0', as shown in Figure 3.1, only transistors M_{p1} and M_{n2} are ON.

1. **Critical Charge Model**

   In order to determine the critical charge model at node V_1, which is more susceptible to soft errors, the particle strike is modeled by a double exponential current pulse given by (2.3) [112]. Typically, for a particle induced current pulse, τ_f is much larger than τ_r as discussed in Chapter 2 [37, 60, 67]. Based on this fact, and for model simplicity, we further approximate (2.3) as a single exponential current pulse, as given in the following equation:

   $$i_{injected}(t) \approx \frac{Q}{\tau} \times \exp(-t/\tau) \tag{3.3}$$

   where τ is equal to τ_f in (2.3). The nodal current equation at node V_1 is written as:

   $$C_1 \frac{dV_1}{dt} = i_{p1}(t) - i_{injected}(t) \tag{3.4}$$

   where C_1 is node V_1 capacitance; $i_{p1}(t)$ is the PMOS transistor, M_{p1}, restoring current, which tries to pull-up node V_1; and $i_{injected}(t)$ is the injected current pulse given in (3.3). It should be noted that transistor M_{n1} sub-threshold current is ignored in this analysis because it is very small with respect to $i_{p1}(t)$ [37, 67].

   From (3.4), the values of Q and τ, that equalize $i_{p1}(t)$ and $i_{injected}(t)$ currents, can be obtained. Hence, node V_1 voltage attains a certain minimum value, V_{min}, which can be obtained by equating these two currents. Since transistor M_{p1} is in the linear region, M_{p1} can be modeled by a resistor R_{p1}. As a result, (3.4) is rewritten as follows:

$$C_1\frac{dV_1}{dt} = \frac{V_{DD} - V_1}{R_{p1}} - \frac{Q}{\tau} \times \exp(-t/\tau) \tag{3.5}$$

where V_{DD} is the supply voltage. The minimum voltage, V_{min}, is computed by equating the two currents and the time at which this V_{min} occurs, t_{min}, is obtained by solving the differential equation in (3.5) and finding the time at which $V_1 = V_{min}$. t_{min} and V_{min} are expressed as [37, 67]:

$$t_{min} = \frac{\tau R_{p1} C_1}{\tau - R_{p1} C_1} \times \ln(\frac{\tau}{R_{p1} C_1}) \tag{3.6}$$

$$V_{min} = V_{DD} - \frac{Q R_{p1}}{\tau} \times (\frac{R_{p1} C_1}{\tau})^{\frac{R_{p1} C_1}{\tau - R_{p1} C_1}} \tag{3.7}$$

The work in [37, 67] finds Q by using transient SPICE simulations. Therefore, if the model in [37, 67] is to be used for statistical WID variations modeling, this value of Q must be found for each statistical run, which turns out to be completely inefficient. This is the reason why this model in [37, 67] can only be used for the D2D variations modeling, which has been already performed in [37, 67].

In the proposed model, we assume that once node V_1 voltage hits its minimum value, V_{min}, the PMOS transistor, M_{p1}, restoring current causes V_1 voltage to either recover to logic '1' and no flipping occurs, or flip to logic '0' and flipping occurs. This assumption is justified by noting that after the time t_{min}, the injected current $i_{injected}(t)$ continues decaying exponentially according to (3.3). Therefore, the goal is to find the condition on the restoring current, $i_{p1}(t)$, that causes node V_1 to flip. This restoring current is controlled by its gate voltage, V_2. Accordingly, if V_2 is rising, the source to gate voltage of M_{p1} decreases, and correspondingly, the restoring current decreases resulting in a soft error. On the other hand, if V_2 is falling, the restoring current increases, and correspondingly, node V_1 voltage recovers and no flipping occurs.

Due to the fact that the inverter switching voltage, V_M, is defined as, the threshold between logic '1' and logic '0' (i.e., when the inverter input slightly exceeds V_M, the inverter output is assumed to be at logic '0', and vice versa). If V_{min} is slightly below the switching voltage of the second inverter, V_{M2}, V_2 rises to logic '1' decreasing the restoring current, and resulting in a soft error.

Consider the flipping case (i.e., $V_{min} < V_{M2}$), node V_2 voltage stays around 0V, for the time interval over which V_1 is approaching V_{min} (i.e., t_{min}), and then starts to rise. Furthermore, V_1 is assumed to remain constant at V_{min}, until V_2 rises and exceeds the switching threshold of the first inverter, V_{M1}. The time at which V_2

hits (V_{M1}) is denoted by t_f, which refers to the SRAM cell flipping time. These assumptions are validated by noticing that once V_2 hits V_{M1}, the positive feedback of the cell becomes strong enough to continue flipping the cell state. Moreover, these assumptions allow us to decouple the cross-coupled inverters of the SRAM cell, as proposed in [37, 67]. From (3.7), and for a given τ, the value of Q, that just cause V_1 to flip, is obtained by equating V_{min} to V_{M2}. Correspondingly, Q is determined by:

$$Q = \frac{(V_{DD} - V_{M2})\tau}{R_{p1}\beta} \text{ where } \beta = (\frac{R_{p1}C_1}{\tau})^{\frac{R_{p1}C_1}{\tau - R_{p1}C_1}} \tag{3.8}$$

From (3.8), Q is obtained without SPICE simulations. Therefore, the main limitation in [37, 67] for WID variations modeling is refined.

Now, the objective is to find the flipping time, t_f. The flipping time, t_f, is the sum of t_{min}, and the time delay that V_2 takes to rise from 0V to V_{M1} (this time is denoted by t_{up}). This t_{up} delay is driven by transistors M_{p2} and M_{n2}, where their gate voltage V_1 is constant at V_{M2}. Transistor M_{p2} is in the saturation region. However, transistor M_{n2} is in the linear region, when V_2 rises from 0V to (V_{M2}-V_{tn2}), where V_{tn2} is the threshold voltage of M_{n2}. When V_2 exceeds (V_{M2}-V_{tn2}), transistor M_{n2} is in the saturation region. The currents of these two transistors are given by:

$$\begin{aligned} i_{n2} &= \begin{cases} \frac{V_2}{R_{n2}} & 0 \le V_2 \le (V_{M2} - V_{tn2}) \\ i_{n2sat} & (V_{M2} - V_{tn2}) \le V_2 \le V_{M1} \end{cases} \\ i_{p2} &= \left\{ i_{p2sat} \quad 0 \le V_2 \le V_{M1} \right. \end{aligned} \tag{3.9}$$

where i_{p2} and i_{n2} are the currents of transistors, M_{p2} and M_{n2}, respectively, i_{p2sat} and i_{n2sat} are the saturation currents of transistors M_{p2} and M_{n2}, respectively, and R_{n2} is the linear region equivalent resistance of transistor M_{n2}. The nodal current equation at node V_2 is given by:

$$C_2\frac{dV_2}{dt} = i_{p2} - i_{n2} \tag{3.10}$$

where C_2 is the node capacitance of node V_2. From (3.9) and (3.10), it is obvious that t_{up} can be divided into two time delays. The first time delay, t_{up1}, is the time delay taken when V_2 rises from 0V to (V_{M2}-V_{tn2}), while transistor M_{n2} is in the linear region. The other time delay, t_{up2}, is the time elapsed when V_2 rises from (V_{M2}-V_{tn2}) to V_{M1}, while M_{n2} is in the saturation region. Following that, the differential equation in (3.10) is solved in two time intervals with the following boundary conditions, V$_2$(t$_{min}$) = 0V, V$_2$(t$_{min}$+t$_{up1}$) = (V$_{M2}$-V$_{tn2}$), and V$_2$(t$_f$) = V$_{M1}$, yielding:

$$t_{up1} = C_2 R_{n2} \ln(\frac{i_{p2sat} R_{n2}}{i_{p2sat} R_{n2} - (V_{M2} - V_{tn2})}) \text{ and } t_{up2} = C_2 \frac{V_{M1} - (V_{M2} - V_{tn2})}{i_{p2sat} - i_{n2sat}} \quad (3.11)$$

By using (3.6), (3.7), and (3.11), the flipping time t$_f$ is expressed as:

$$t_f = \tau \ln(\frac{1}{\beta}) + t_{up1} + t_{up2} \quad (3.12)$$

Thus, the critical charge, Q$_{critical}$, is obtained as follows [37, 59, 67, 110, 111, 113]:

$$Q_{critical} = \int_0^{t_f} i_{injected}(t) dt = Q\ (1 - \exp(-t_f/\tau)) \quad (3.13)$$

In this derivation, the focus is on the supply voltage range covering the super-threshold region, without accounting for the sub-threshold operation. To simplify the analysis, the well-known alpha-power law model for the transistor current [114], is adopted. In [114], the transistor current in the saturation region is modeled by:

$$i_n = K_{n'}(W/L)(V_{GS} - V_{tn})^{\alpha_n} \quad (3.14)$$

where V$_{tn}$ is the threshold voltage, K$_{n'}$ is a technological parameter, α_n is the velocity saturation exponent ranging from 1 to 2, depending on whether the transistor is in deep velocity or pinch-off saturation, and W and L are the width and length of the transistor channel, respectively.

According to this model, the inverter switching voltage, V$_M$, is given by [114]:

$$\begin{aligned} V_M &= \frac{r(V_{DD} - |V_{tp}|) + V_{tn}}{1 + r} \\ \text{where } r &= (\frac{K_{p'}(W/L)_p}{K_{n'}(W/L)_n})^{1/\alpha} \text{ and } \alpha = \alpha_n = \alpha_p \end{aligned} \quad (3.15)$$

where V$_{tn}$ and V_{tp} are the threshold voltages, α_n and α_p are the velocity saturation exponents, K$_{n'}$ and K$_{p'}$ are the technology parameters, and (W/L)$_n$ and (W/L)$_p$ are the aspect ratios of the NMOS and PMOS transistors, respectively.

In addition, the currents i$_{p2sat}$ and i$_{n2sat}$ are given by:

$$\begin{aligned} i_{p2sat} &= K_{p'}(W/L)_{p2}(V_{DD} - V_{M2} - |V_{tp2}|)^{\alpha} \\ i_{n2sat} &= K_{n'}(W/L)_{n2}(V_{M2} - V_{tn2})^{\alpha} \end{aligned} \quad (3.16)$$

and the resistances R$_{p1}$ and R$_{n2}$ are computed by:

$$\begin{aligned} R_{p1} &= \frac{1}{K_{p'}(W/L)_{p1}(V_{DD} - |V_{tp1}|)} \\ R_{n2} &= \frac{1}{K_{n'}(W/L)_{n2}(V_{M2} - V_{tn2})} \end{aligned} \quad (3.17)$$

Using (3.6)-(3.8), (3.11)-(3.13), and (3.15)-(3.17), the critical charge, Q$_{critical}$, can be obtained without doing any SPICE simulations.

2. **Statistical Critical Charge Variation Model**

   Process variations affect device parameters, resulting in fluctuations in the critical charge. The primary sources of process variations, that affect the device parameters, are RDF, LER, and channel length variations as discussed in Chapter 2.

   From the above derivations, it is evident that the critical charge, Q$_{critical}$, is dependent on the threshold voltages of transistors M$_{p1}$, M$_{p2}$, M$_{n1}$, and M$_{n2}$, which are denoted by V$_{tp1}$, V$_{tp2}$, V$_{tn1}$, and V$_{tn2}$, respectively. A small change in these threshold voltages results in an incremental change in the critical charge, $\Delta Q_{critical}$, that is calculated by using Taylor expansion around the nominal value as follows:

$$\begin{aligned} \Delta Q_{critical} &= \frac{\partial Q_{critical}}{\partial V_{tp1}}\Delta V_{tp1} + \frac{\partial Q_{critical}}{\partial V_{tp2}}\Delta V_{tp2} \\ &+ \frac{\partial Q_{critical}}{\partial V_{tn1}}\Delta V_{tn1} + \frac{\partial Q_{critical}}{\partial V_{tn2}}\Delta V_{tn2} \end{aligned} \quad (3.18)$$

   where $\Delta V_{tp1}, \Delta V_{tp2}, \Delta V_{tn1}$, and ΔV_{tn2} are the variations of the threshold voltages. The partial derivative terms in (3.18) can be computed numerically at the mean threshold voltages. Therefore, the standard deviation of the critical charge variations is calculated as follows:

$$\begin{aligned}\sigma_{Q_{critical}} &= \{(\frac{\partial Q_{critical}}{\partial V_{tp1}})^2\sigma^2_{V_{tp1}} + (\frac{\partial Q_{critical}}{\partial V_{tp2}})^2\sigma^2_{V_{tp2}} \\ &+ (\frac{\partial Q_{critical}}{\partial V_{tn1}})^2\sigma^2_{V_{tn1}} + (\frac{\partial Q_{critical}}{\partial V_{tn2}})^2\sigma^2_{V_{tn2}}\}^{0.5} \qquad (3.19)\end{aligned}$$

where $\sigma_{Vtp1}, \sigma_{Vtp2}, \sigma_{Vtn1}$, and σ_{Vtn2} are the standard deviations of the threshold voltages $V_{tp1}, V_{tp2}, V_{tn1}$, and V_{tn2}, respectively.

This model is valid under the following assumptions:

-i- The dominant source of variations is the transistor V_t variations. The channel length variations are assumed to affect only V_t through DIBL effect as explained in Chapter 2. While the variations in the channel length introduce also fluctuations in the input gate capacitance, nevertheless, this contribution is much smaller than that in the threshold voltage variations [23, 115].

-ii- The impact of process variations on the critical charge variations is computed by using a linear approximation. This assumption is accurate, since, WID variations can be linearized around the nominal value [115–119]. Under this linear approximation, the critical charge mean value is assumed to be equal to its deterministic value, when no variations are introduced. Therefore, process variations affect only the variance of the critical charge (i.e., the critical charge spread around its nominal value).

-iii- According to [120], the correlation between the different transistors threshold voltages can be neglected for random WID variations. This is due to the fact that the RDF is random, and therefore, V_t of the four transistors, in consideration, are identified as four independent and uncorrelated gaussian random variables [9]. This assumption simplifies the derivation of (3.19).

3.2.3 Simplified Model for Statistical Design-Oriented Critical Charge Variation

1. **Simplified Model Assumptions and Derivations**

   The model, which is introduced in Section 3.2.2, for the critical charge variations, is calculated numerically. Therefore, it does not present obvious design insights for WID variations. In this section, this accurate model is simplified for the case of a symmetric 6T SRAM, to account for the critical charge variations from a design perspective. The following assumptions are made to derive this simplified model:

-i- The inverters switching voltages are equal to half the supply voltage (i.e., $V_{M1} = V_{M2} = 0.5V_{DD}$). Thus, the variations in V_{M1} and V_{M2} are ignored. It should be noted that the inverters threshold voltage can be assumed of any other value depending on the SRAM sizing such as $V_{DD}/3$ or $V_{DD}/4$.

-ii- The variation of the factor β, expressed in (3.8), which is dependent only on V_{tp1} through R_{p1} is calculated to be less than 0.8%, relative to its mean value. As a result, the variations in this factor are ignored, and this factor is assumed constant from the variability perspective.

-iii- The time delay t_{up} is obtained simply by using a first order approximation of the low to high propagation delay of an inverter, which can be modeled as follows:

$$t_{up} = \frac{C_2 \Delta V}{i_{average}} \tag{3.20}$$

where ΔV is the output voltage swing, that is usually assumed to be $0.5V_{DD}$, and $i_{average}$ is the average charging current, that is the difference between transistor M_{p2} current and transistor M_{n2} current. From (3.9), M_{p2} is in the saturation region during the entire charging process time, hence, its average current is i_{p2sat}. While, transistor M_{n2} current rises from 0A, when the output voltage V_2 (V_{DS} of transistor M_{n2}) equals 0V, up to i_{n2sat}, when the transistor enters the saturation region. This current is assumed linear with V_2 in the linear region, as depicted in Figure 3.2. The average of this current is obtained from Figure 3.2 as follows:

$$i_{n2average} = i_{n2sat}(0.5 + V_{tn2}/V_{DD}) \tag{3.21}$$

The relative variations of this current are given by:

$$\frac{\Delta i_{n2average}}{i_{n2average}} = [\frac{-\alpha}{(V_{DD}/2) - V_{tn2}} + \frac{1}{V_{DD}(0.5 + V_{tn2}/V_{DD})}]\Delta V_{tn2} \tag{3.22}$$

The variations due to the first term in (3.22) dominate the second term (as a numeric example, when $V_{DD} = 1$V, $\alpha = 1.25$, and $V_{tn2} = 0.342$V, the first term is 7X higher than the second term). Therefore, in the following derivations, $i_{n2average}$ is assumed to be equal $i_{n2sat}(0.5 + V_{tn2}/V_{DD})$, while the variations of the term $(0.5 + V_{tn2}/V_{DD})$ are not considered, and this factor is assumed constant, from the variability perspective.

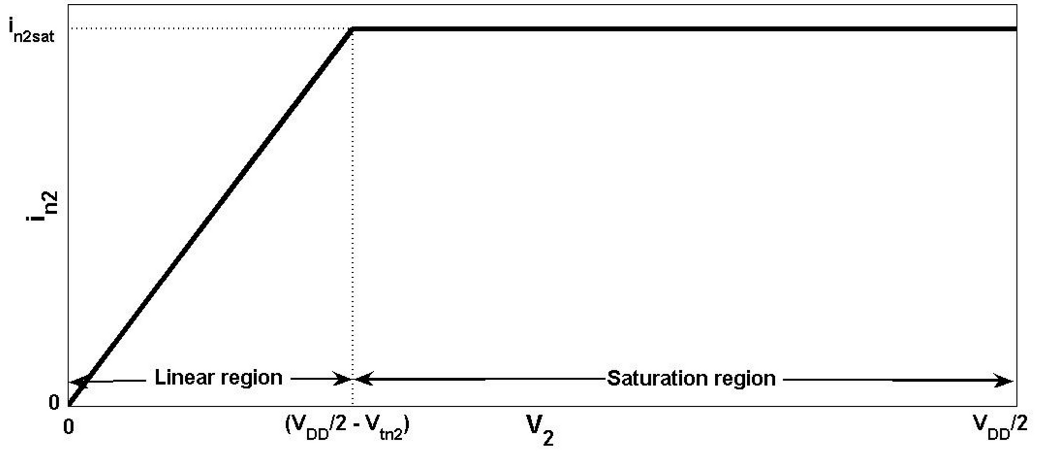


Figure 3.2: Transistor M$_{n2}$ current approximation. This current is assumed linear as V$_2$ changes from 0 to $(V_{DD}/2 - V_{tn2})$ then it saturates at i$_{n2sat}$ when V$_2$ changes from $(V_{DD}/2 - V_{tn2})$ to $V_{DD}/2$.

2. **Statistical Design-Oriented Critical Charge Variation Model Accounting for WID Variation**

   By using the simplified model formulas, the partial derivatives, defined in (3.19), are calculated analytically and normalized to the mean value of Q$_{critical}$ as follows:

$$\frac{\frac{\partial Q_{critical}}{\partial |V_{tp1}|}}{Q_{critical}} = \frac{-1}{(V_{DD} - |V_{tp1}|)} \tag{3.23}$$

$$\frac{\frac{\partial Q_{critical}}{\partial |V_{tp2}|}}{Q_{critical}} = \frac{(\alpha/\tau) i_{p2avergae}}{C_2(\frac{V_{DD}}{2})(\frac{V_{DD}}{2} - |V_{tp2}|)} \frac{\beta t_{up}^2 \exp(\frac{-t_{up}}{\tau})}{1 - \beta \exp(\frac{-t_{up}}{\tau})} \tag{3.24}$$

$$\frac{\frac{\partial Q_{critical}}{\partial V_{tn2}}}{Q_{critical}} = \frac{-(\alpha/\tau) i_{n2avergae}}{C_2(\frac{V_{DD}}{2})(\frac{V_{DD}}{2} - V_{tn2})} \frac{\beta t_{up}^2 \exp(\frac{-t_{up}}{\tau})}{1 - \beta \exp(\frac{-t_{up}}{\tau})} \tag{3.25}$$

   From (3.23), it is clear that reducing |V$_{tp1}$| results in reducing the relative variations. Accordingly, it is recommended that transistor M$_{p1}$ is used as a low-V$_t$ device, if the dual-V$_t$ technique is to be used (The same for M$_{p2}$, when the hit occurs at the other node). Moreover, as the supply voltage V$_{DD}$ is reduced, the variations due to V$_{tp1}$ are increased.

   Since increasing the node capacitance is one of the most common techniques to mitigate soft errors in SRAM cells, it is important to see the impact of increasing the node capacitance on the relative critical charge variations. Usually, a coupling capacitor, C$_c$, is employed between the storage nodes (V$_1$ and V$_2$) as shown in Figure 3.3. This coupling capacitor, C$_c$, increases the nodal capacitances of the SRAM cell storage nodes, and therefore, their critical charge is increased significantly. This C$_c$ is stacked on top of the SRAM cell (Metal-Insulator-Metal (MIM) capacitor) to minimize the required area overhead. The model capacitances C$_1$ and C$_2$, have to be modified to account for C$_c$, by applying the Miller effect as follows [37, 67]:

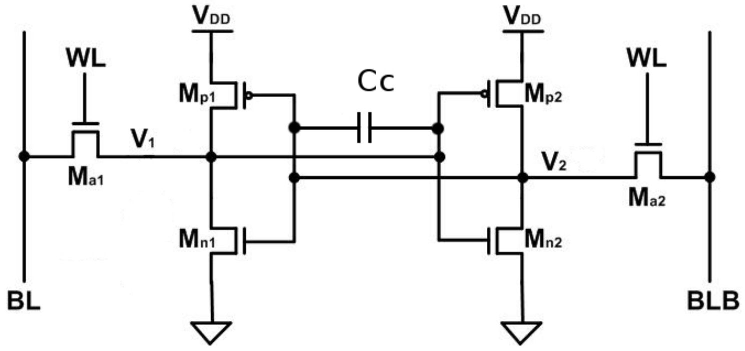


Figure 3.3: The SRAM cell with the coupling capacitor, C_c, which increases the critical charge value of its storage nodes (V$_1$ and V$_2$).

$$C_1^{'} = C_1 + 2C_c \quad \text{and} \quad C_2^{'} = C_2 + 2C_c \tag{3.26}$$

From (3.24) and (3.25), and by using t$_{up}$ and β formulas, the relative critical charge variations ($\frac{\frac{\partial Q_{critical}}{\partial |V_{tp2}|}}{Q_{critical}}$ and $\frac{\frac{\partial Q_{critical}}{\partial V_{tn2}}}{Q_{critical}}$) have the same dependence on the node capacitance, C$^{'}$ (assuming $C_1^{'} = C_2^{'} = C^{'}$ for a symmetric SRAM cell). This dependence is in the form $\frac{\beta C^{'} \exp(-\gamma C^{'})}{(1-\beta \exp(-\gamma C^{'}))}$, where $\gamma = \frac{(\frac{V_{DD}}{2})}{\tau(i_{p2avergae} - i_{n2average})}$. Therefore, it is possible to obtain the value of the node capacitance, C$^{'}$, that maximizes these relative variations, by differentiating with respect to C$^{'}$, and equating the result to zero. After some simplifications, the condition on C$^{'}$ for the maximum possible relative variations is given by:

$$\begin{aligned} 1-\beta\exp(-\gamma C^{'}) &= C^{'}(\gamma-\theta) \\ \text{where} \quad \gamma &= \frac{(\frac{V_{DD}}{2})}{\tau(i_{p2avergae}-i_{n2average})} \\ \text{and} \quad \theta &= (\frac{\frac{\partial\beta}{\partial C^{'}}}{\beta}) = \frac{\frac{R_{p1}}{\tau}}{1-\frac{R_{p1}C^{'}}{\tau}}(1+\frac{\ln(\frac{R_{p1}C^{'}}{\tau})}{1-\frac{R_{p1}C^{'}}{\tau}}) \end{aligned} \tag{3.27}$$

From (3.27), the value of $C^{'}$ that maximizes the relative variations, denoted by $C^{'}_{m}$, is obtained for a given value of V_{DD}, τ, and average currents ($i_{p2averge}$ and $i_{n2average}$). These average currents are dependent on transistors M_{p2} and M_{n2} parameters (W/L and V_t). Since $C^{'}_{m}$ results in the maximum relative variations, it is essential at the design level to avoid the satisfaction of this condition reported in (3.27). Otherwise, the SRAM cell will exhibit the maximum possible relative critical charge variations. These maximum variations are calculated by substituting the condition in (3.27) in (3.24) and (3.25) and are given by:

$$(\frac{\frac{\partial Q_{critical}}{\partial|V_{tp2}|}}{Q_{critical}})|_{max} = \frac{\alpha\beta\tau(\frac{\gamma^2}{(\gamma-\theta)})i_{p2average}}{(\frac{V_{DD}}{2})(\frac{V_{DD}}{2}-|V_{tp2}|)}\exp(-\gamma C^{'}_{m}) \tag{3.28}$$

$$(\frac{\frac{\partial Q_{critical}}{\partial V_{tp2}}}{Q_{critical}})|_{max} = \frac{\alpha\beta\tau(\frac{\gamma^2}{(\gamma-\theta)})i_{n2average}}{(\frac{V_{DD}}{2})(\frac{V_{DD}}{2}-V_{tn2})}\exp(-\gamma C^{'}_{m}) \tag{3.29}$$

By using (3.23) with (3.28) and (3.29), the maximum possible relative critical charge variations, for a give SRAM cell design with respect to $C^{'}$, are estimated.

In addition, (3.24), and (3.25) indicate that the relative variations, due to V_{tn2} and V_{tp2}, are decaying exponentially with (t_{up}/τ). From (3.20), t_{up} is dependent on $C^{'}$, therefore, there exists a certain value of $C^{'}$ for a given τ that makes the relative variations contributions of V_{tn2} and V_{tp2} smaller than that of V_{tp1}. In this situation, the variations of V_{tp1} dominate, and further increasing $C^{'}$ does not reduce the overall relative variations which are at a minimum value. The knowledge of $C^{'}$, which results in maximum and minimum relative variations, provides a vital design insight for circuit designers, who target at mitigating the soft errors, while keeping the variability at a certain level.

Finally, the proposed models can be used for future CMOS technology nodes (i.e., 45-nm, 32-nm, and 22-nm), since, the transistor model parameters such as the technology parameters and the threshold voltage standard deviation, σ_{V_t}, can be easily obtained. Therefore, the proposed models are scalable in terms of technology scaling and can be used to predict the critical charge variability for future technology nodes as long as the models assumptions are satisfied.

3.2.4 Results and Discussion

In all the following simulations, an industrial hardware-calibrated 65nm CMOS technology transistor model, with technological parameters shown in Table 3.1, is employed.

Table 3.1: 65-*nm* Technology information and SRAM sizing

	NMOS	PMOS
Nominal V_{DD}	1.0-1.2 V	
W/L ($\mu m/\mu m$)	0.195/0.06	0.12/0.06
V_{to} (mV)	342	-204
$\sigma_{V_{to}}$ (mV)	25.8	34.3

1. **Verification of the Models Assumptions**

   First, the assumptions, used in deriving (3.6) and (3.7), are verified. Figure 3.4 illustrates the non-flipping case, where the SRAM cell recovers for different values of V_{DD}. Node V_1 voltage falls down till it hits a minimum voltage (which is called V_{min}, and given in (3.7)) then recovers back to V_{DD}. From Figure 3.4, this minimum voltage V_{min} is close to $V_{DD}/2$ justifying the assumptions used in the simplified model. Figure 3.5.a shows the two nodes V_1 and V_2 voltages in the non-flipping case. It is clear that, since V_2 voltage can not hit V_{M1}, the SRAM cell is recovered. However, in Figure 3.5.b, the V_2 node voltage hits V_{M1}, and hence, the SRAM cell exhibits a soft error.

   Moreover, Figure 3.5.b shows that node V_2 voltage is around 0V as long as node V_1 voltage is falling. Once node V_1 voltage hits V_{min}, V_1 stays constant at V_{min}, whereas, node V_2 voltage rises to V_{M1}. It should be mentioned that the minimum

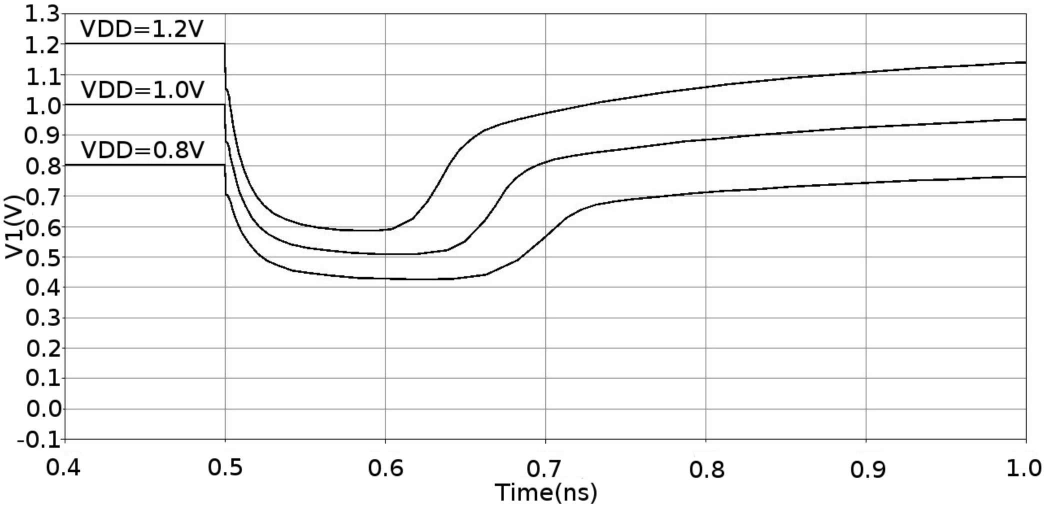
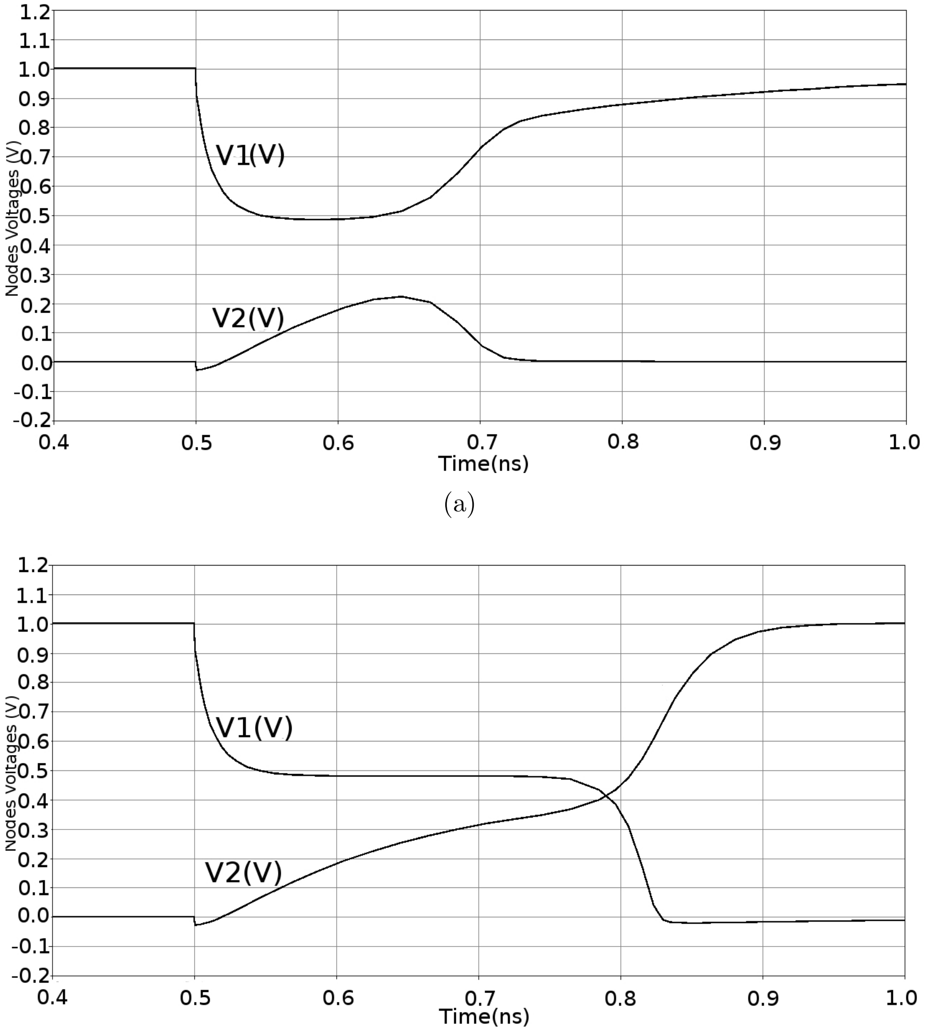


Figure 3.4: The non-flipping case when the SRAM cell recovers for different values of V_{DD}.

voltage, V_{min}, shown in Figure 3.5.b, at which V_1 stays constant before flipping to 0V, is slightly less than V_{min} shown in Figure 3.5.a for the non-flipping case. The difference between these two minima is approximately 15mV, which demonstrates that the flipping occurs, when V_{min} is less than V_{M2}.

2. **Verification of the Models Estimated Critical Charge**

   To verify the critical charge nominal value, and the critical charge variations models, the analytical models are compared to the simulation results using SPICE transient and Monte Carlo simulations. These simulations are performed to validate the nominal critical charge, and the critical charge variability models, respectively, for both the accurate and the simplified models. In the following, the validation results for these models are presented. 5,000 Monte Carlo runs are used to provide a good accuracy in determining the critical charge mean and standard deviation. For each Monte Carlo run, the value of the current pulse charge, Q, that causes the cell to flip is determined. Then, the simulations are repeated for different V_{DD} (from 0.7V to 1.2V), to find the effect of reducing V_{DD} on the critical charge mean and standard deviation. The SRAM sizing, shown in Table 3.1, is used in the simulation setups. Hardware-calibrated statistical transistor models, reported in Appendix D, are used to account for V_t variations. The PMOS transistors have higher V_t variations than the NMOS transistors, since the PMOS transistors exhibit lower driving strength (weaker) than the NMOS transistors in the SRAM cell.



(a)

(b)

Figure 3.5: The two nodes V_1 and V_2 in (a) The non-flipping case and (b) The flipping case.

-i- **Nominal Critical Charge:**

Figure 3.6 displays the nominal critical charge, which is obtained by using the transient simulations ($Q_{critical}$) and Monte Carlo simulations ($\mu_{Q_{critical}}$). Clear agreement between $Q_{critical}$ and $\mu_{Q_{critical}}$ justifies the linearity approximation assumption, down to $V_{DD} = 0.7$V (i.e., WID variations affect only on the critical charge variance (spread) and have no effect on its mean).

Figure 3.7 shows the nominal critical charge value calculated from the proposed

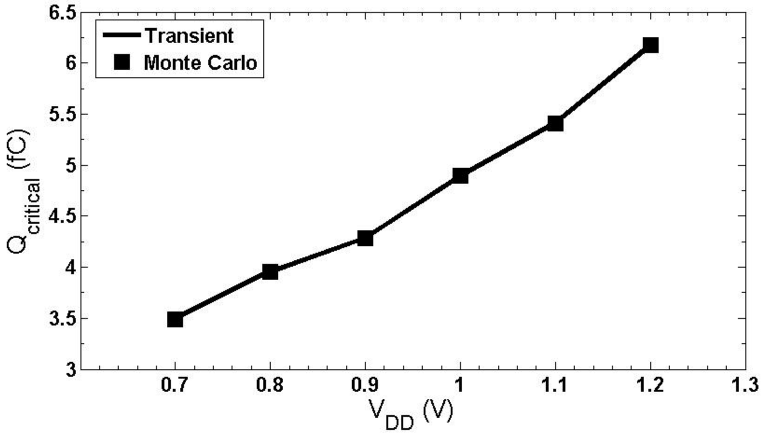


Figure 3.6: $Q_{critical}$ versus V$_{DD}$ from the transient simulations (when no variations are introduced) and from Monte Carlo simulations.

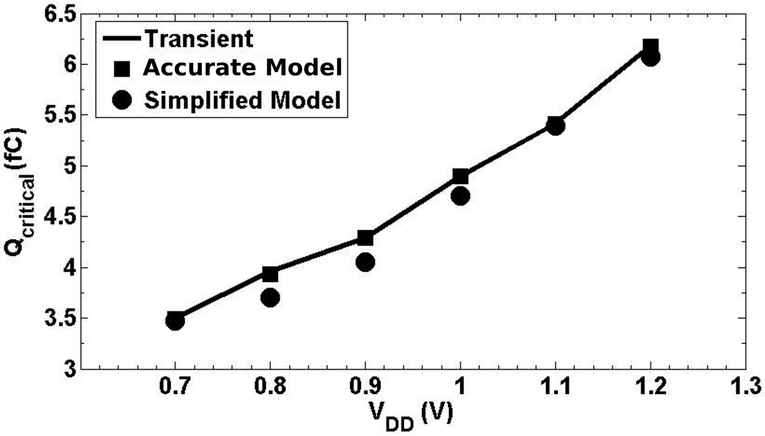


Figure 3.7: $Q_{critical}$ versus V$_{DD}$ from Monte Carlo simulations. Also shown the results from the proposed accurate and simplified models.

accurate and simplified model versions, and compared to the transient simulations results for different supply voltage values. It should be highlighted that the simplified model is proposed only for the WID variations estimation, although it still shows an acceptable match for the nominal critical charge value. These results are obtained by using τ = 250ps. It is obvious from Figures 3.6 and 3.7 that reducing the supply voltage decreases the critical charge, which is expected.

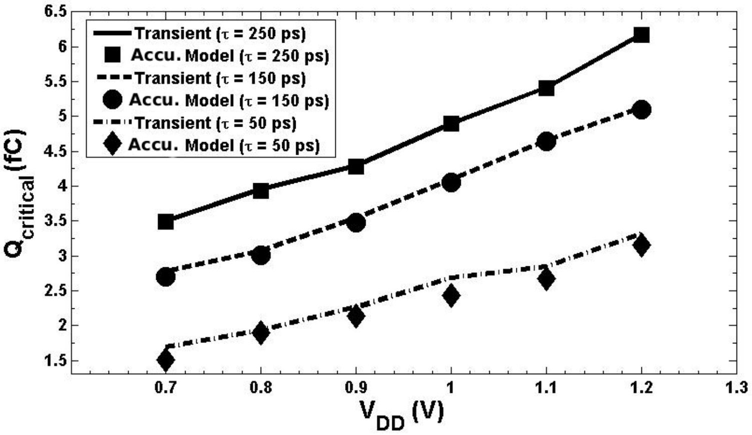


Figure 3.8: $Q_{critical}$ versus V$_{DD}$ for different values of τ (50ps to 250ps)from the transient simulations and from the proposed accurate model.

According to [60], the current pulse, used in circuit level modeling of soft errors, might have a varying width from a few picoseconds to hundreds of picoseconds. The narrow current pulse represents the worst-case situation, because the critical charge, Q$_{critical}$, is minimal. This narrow current pulse corresponds to an event, in which the track of an ionized particle intersects the drain of the NMOS transistor in the OFF-state (like M$_{n1}$ in the analyzed case). This means that the charge collection mechanism is dominated by the drift current (due to local electric fields) in a very short time. On the other hand, the charge collection mechanism is dominated by diffusion current in the events in which the ion track does not intersect the drain [60]. Theoretical studies showed that, typically, 80-90% of the neutron induced SER is represented by the latter events in which the current pulse is relatively wide [121]. Such a discussion demonstrates that both narrow and wide current pulses must be considered in Q$_{critical}$ calculations.

Therefore, the values of Q$_{critical}$, calculated from the proposed accurate model and from SPICE transient simulations for different current pulse widths (by varying τ from 50ps to 250ps), are shown in Figure 3.8. The simplified model results are not shown in this figure as the simplified model is mainly introduced for WID variations estimation. In Figure 3.8, it is shown that as the current pulse width increases (i.e., diffusion current dominates), the critical charge increases.

-ii- **Critical Charge Variations:**

Figure 3.9 shows the simulation result for $\sigma_{Q_{critical}}$ for different V$_{DD}$ values. Note that each data point represents $\sigma_{Q_{critical}}$ calculated from 5,000 Monte Carlo runs.

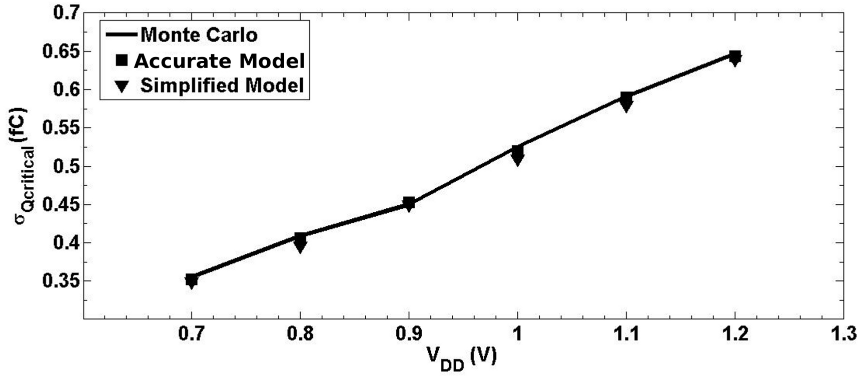


Figure 3.9: Critical charge variations $\sigma_{Q_{critical}}$ versus V_{DD} from Monte Carlo simulation and from the proposed accurate and simplified model.

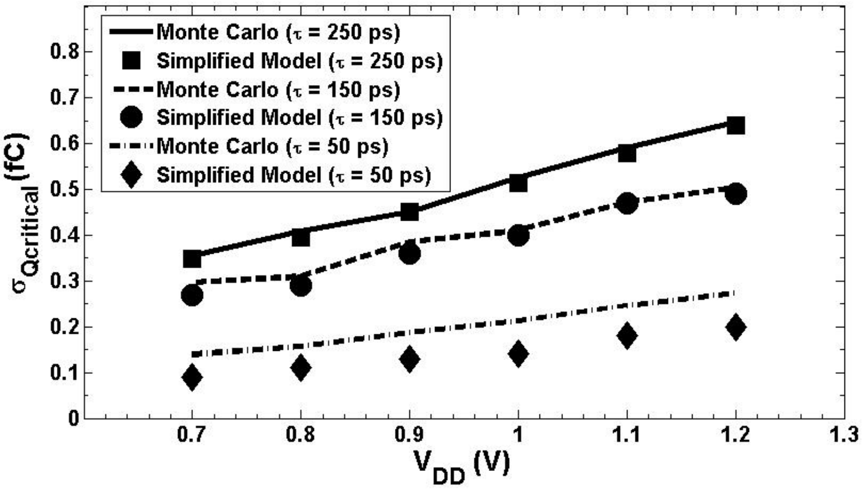


Figure 3.10: Critical charge variations $\sigma_{Q_{critical}}$ versus V_{DD} for different values of τ (50ps to 250ps) from Monte Carlo simulation and from the proposed simplified model.

Also, Figure 3.9 shows the results from the proposed models. Both models results exhibit a good match with the simulation results. Figure 3.10 shows $\sigma_{Q_{critical}}$ obtained from Monte Carlo simulations and from the simplified model for different values of τ which demonstrates that, as τ is reduced, the critical charge variations are reduced as well. It is important to show that as V_{DD} is reduced, $\sigma_{Q_{critical}}$ is decreased, which is counter-intuitive since increasing V_{DD} normally means lower variations.

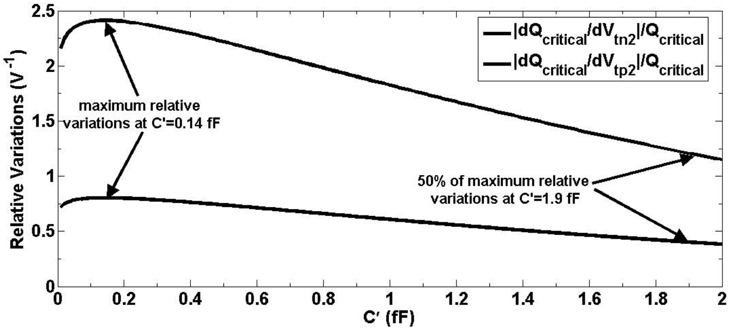


Figure 3.11: The relative variations $(\partial Q_{critical}/\partial|V_{tp2}|)/Q_{critical}$ and $(\partial Q_{critical}/\partial V_{tn2})/Q_{critical}$ versus C' showing that the maximum relative variation occurs at C' = 0.14fF and 50% of the maximum variation occurs at C' = 1.9fF.

3. **The Effect of the Coupling Capacitor on the Critical Charge Relative Variability**

In Section 3.2.3, it has been shown that the capacitance $C_m^{'}$, which results in the maximum relative variations, can be obtained from the condition given in (3.27). For V_{DD} = 1 V, α= 1.25 (extracted from fitting Log (I_D)-Log (V_{GS}) characteristics to the alpha-power model), i_{p2sat} = 12.9 μA, i_{2nsat} = 11.2 μA, and τ = 250ps. By using (3.27), γ= 0.6*10^{15}F^{-1}, and solving this equation yields that $C_m^{'}$ = 0.143fF. The node capacitance C equals 0.93fF, therefore, the condition for the maximum relative variations is not met in this case, since C is already larger than $C_m^{'}$. Figure 3.11 shows how the relative variations in (3.24) and (3.25) vary with the capacitance $C^{'}$. For a given relative variations specifications, the value of the capacitance $C^{'}$, that results in these relative variations, can be obtained from this figure. For example, the value of $C^{'}$, that results in 50% of the maximum relative variations value, equals 1.9fF. Consequently, the coupling capacitor, that results in half maximum relative variations, is C_c = (1.9-0.93)/2 = 0.485fF by using (3.26).

Figures 3.12 and 3.13 portray the overall relative variations ($\sigma_{Q_{critical}}/\mu_{Q_{critical}}$) versus C_c obtained from Monte Carlo simulations, and from the proposed model, for different values of V_{DD}, when τ = 250ps and τ = 50ps. The proposed model is in good agreement with the simulation results.

It is obvious from Figures 3.12 and 3.13 that, as C_c increased, $\sigma_{Q_{critical}}/\mu_{Q_{critical}}$ decreases, till reaching a minimum value at which increasing C_c has no effect on $\sigma_{Q_{critical}}/\mu_{Q_{critical}}$. The reason for this is readily explained by recalling (3.24) and (3.25), which show that for large values of C_c, the variations from V_{tn2} and V_{tp2} are vanished (since increasing C_c increases t_{up}) and, hence, the variations from V_{tp1} dominate the overall variations. Therefore, $\sigma_{Q_{critical}}/\mu_{Q_{critical}}$ is proportional to (1/(V$_{DD}$-|V$_{tp1}$|). This latter observation explains why $\sigma_{Q_{critical}}/\mu_{Q_{critical}}$ saturates at the highest value for the case when V$_{DD}$ = 0.8V.

Also, Figures 3.12 and 3.13 show also that $\sigma_{Q_{critical}}/\mu_{Q_{critical}}$ decreases, as V$_{DD}$ is reduced, before reaching its minimum level. However, $\sigma_{Q_{critical}}/\mu_{Q_{critical}}$ decreases, as V$_{DD}$ increases, when V$_{tp1}$ variations dominate (at large values of C$_c$).

Finally, as shown in these two figures, $\sigma_{Q_{critical}}/\mu_{Q_{critical}}$ reaches a minimum value, at smaller values of C$_c$, for smaller τ values. Hence, for small τ values, C$_c$, that results in the minimum $\sigma_{Q_{critical}}/\mu_{Q_{critical}}$, is smaller than that for large τ values. The value of C$_c$, that causes $\sigma_{Q_{critical}}/\mu_{Q_{critical}}$ to reach its minimum value, is denoted by C$_{cmin}$, and is obtained from $\sigma_{Q_{critical}}/\mu_{Q_{critical}}$ plots. It might be beneficial for designers to know, in advance, the value of C$_{cmin}$, and the impact of V$_{DD}$ and τ on it.

Figure 3.14 shows how V$_{DD}$ and τ affect on C$_{cmin}$, as obtained from the proposed simplified model and from Monte Carlo simulations. According to Figure 3.14, it is clear that C$_{cmin}$ increases when V$_{DD}$ increases, and also when τ increases. This result is promising for low power SRAM cells, since a smaller coupling capacitor is required to have the minimum relative critical charge variations.

Now, the values of C_c, that result in maximum and minimum $\sigma_{Q_{critical}}/\mu_{Q_{critical}}$, are calculated. Thus, a good design insight is to use a coupling capacitor between these two extremes, to enhance the critical charge mean, and minimize the relative critical charge variations, under certain power and performance constraints.

4. **SRAM Cell Transistors Contribution to the Overall Critical Charge Variability**

   The overall critical charge standard deviation ($\sigma_{Q_{critical}}$) has contributions from different transistors threshold voltages variations (i.e., V$_{tn1}$, V_{tn2}, V_{tp1}, and V$_{tp2}$). Figure 3.15 shows the percentage contribution of each transistor threshold voltage variations for different values of V$_{DD}$, when τ = 50ps and 250ps, obtained from the two proposed models. It is evident that the contribution of V$_{tn1}$ in the accurate model is small (less than 6%). This justifies the assumptions used in deriving the simplified model, which ignores its variations contribution (when we assume that V$_{M1}$ = V_{DD}/2). According to Figure 3.15, the contribution of V$_{tp1}$ increases, as the supply voltage is reduced which is well explained by (3.23) (inversely proportional to (V$_{DD}$-|V$_{tp1}$|)). At V$_{DD}$ = 0.7V, the transistor M$_{p1}$ dominates the variations (62%) for the case τ = 50ps.

   Moreover, when τ increases, V$_{tn2}$ and V$_{tp2}$ contributions to the critical charge variance are increased, and V$_{tp1}$ contribution is decreased. These results agree with (3.24) and (3.25). In addition, Figure 3.15 shows that the contributions of the PMOS transistors, M$_{p1}$ and M$_{p2}$, dominate the variations, because their percentage contributions is

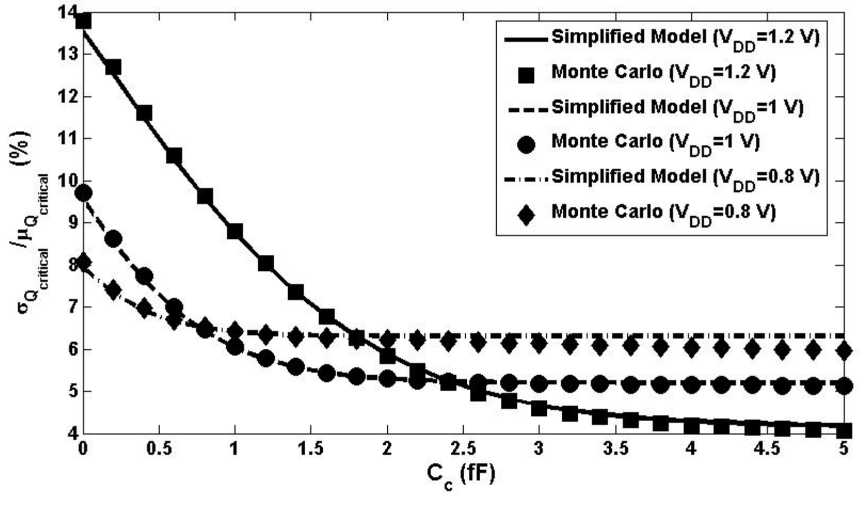


Figure 3.12: The overall relative variations ($\sigma_{Q_{critical}}/\mu_{Q_{critical}}$) versus C$_c$ obtained from Monte Carlo simulations and from the proposed simplified model for different values of V$_{DD}$ when τ = 250ps.

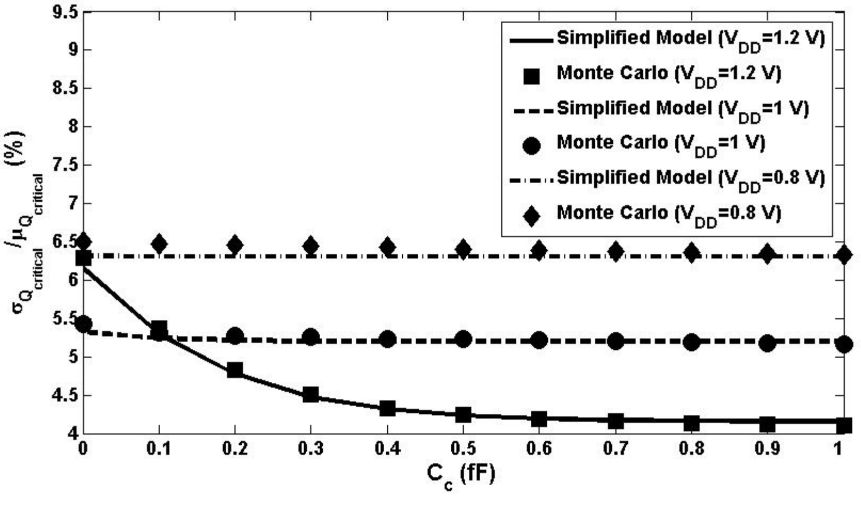


Figure 3.13: The overall relative variations ($\sigma_{Q_{critical}}/\mu_{Q_{critical}}$) versus C$_c$ obtained from Monte Carlo simulations and from the proposed simplified model for different values of V$_{DD}$ when τ = 50ps.

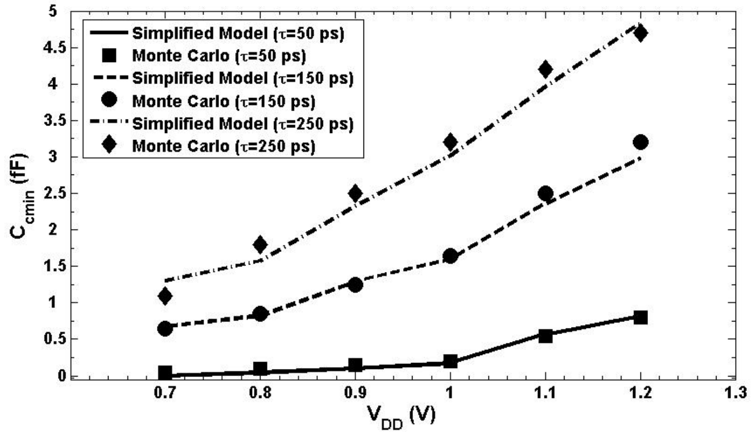


Figure 3.14: The coupling capacitor that results in minimum relative critical charge variations C$_{cmin}$ versus V$_{DD}$ for different values of τ which shows that when V$_{DD}$ is reduced, the value of C$_{cmin}$ that results in minimum relative variations is decreased. These results are obtained from the proposed simplified model and from Monte Carlo simulations.

larger than 84% in all cases. This fact can be justified by noting that the gate area of the PMOS transistors is smaller than that of the NMOS transistors (as reported in Table 3.1). Since the threshold voltage variations are inversely proportional to the square root of the gate area (W*L), the PMOS transistors dominate the variations.

5. **Accuracy of the Proposed Models**

In Figure 3.16, Q$_{critical}$ from the proposed accurate model is plotted versus the transient simulations results for different values of τ, V$_{DD}$, and C$_c$. The maximum error is 6.2%, and the average error is 1.8%. Figure 3.17 shows $\sigma_{Qcritical}$ from the simplified model plotted versus Monte Carlo simulation results for different values of τ, V$_{DD}$, and C$_c$. The maximum error is 9.2%, and the average error is 4%. Good agreement between the proposed models and the simulation results justifies all the assumptions used to derive the models, as explained in Sections 3.2.2, and 3.2.3.

As shown in the previous discussions, the proposed models are based on easily measurable parameters, which can be directly extracted from the measurements or technology information (i.e., C, σ_{V_t}, V$_{to}$, and α). In addition, the proposed models are very efficient when compared to the computationally expensive, and time consuming Monte Carlo simulations. The models can be used to explore design trade-offs to increase the critical charge or control its variability. The proposed models show how the coupling capacitor, one of the most common soft error mitigation techniques in

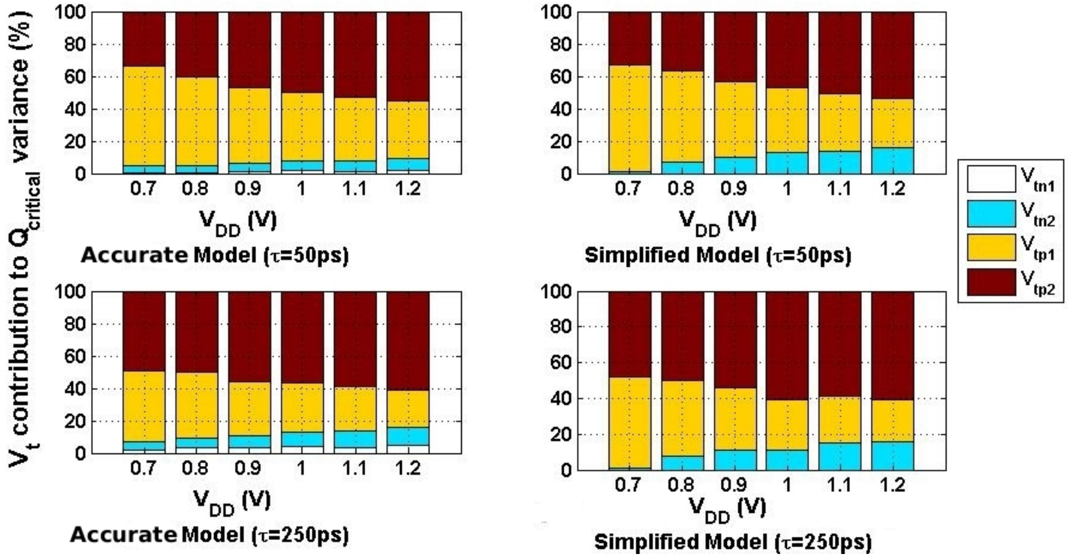


Figure 3.15: The percentage contribution of each transistor threshold voltage variations for different values of V$_{DD}$ when τ = 50ps and 250ps obtained from the two proposed models. The contribution of V$_{tp1}$ increases as the supply voltage is reduced which is well explained from (3.23) (inversely proportional to (V$_{DD}$-|V$_{tp1}$|)).

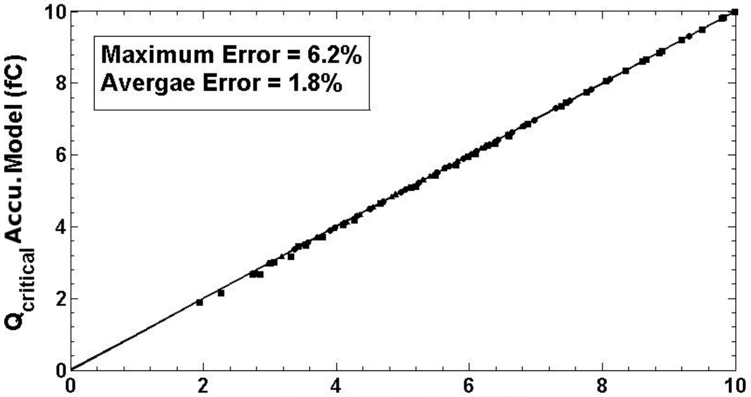


Figure 3.16: Q$_{critical}$ from the proposed accurate model is plotted versus the transient simulations results for different values of τ, V$_{DD}$, and C$_c$.

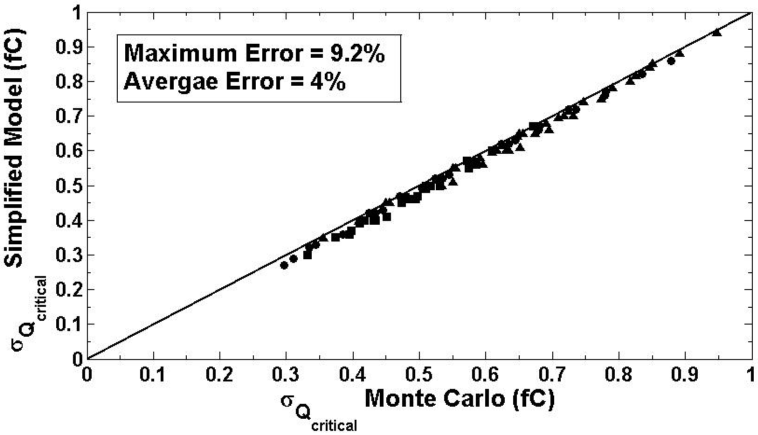


Figure 3.17: $\sigma_{Qcritical}$ from our simplified model plotted versus Monte Carlo simulation results for the same ranges of τ, V$_{DD}$, and C$_c$.

SRAM cells, affects on the critical charge relative variability. Moreover, the proposed models provide a certain range for this coupling capacitor, C$_c$, to keep the variability within an acceptable limit.

3.2.5 Design Insights

In this section, some design insights, extracted from the proposed models, are reported. The proposed models provide the following design insights:

-i- Increasing the supply voltage, V$_{DD}$, results in increasing both Q$_{critical}$ and $\sigma_{Q_{critical}}$. Therefore, the choice of V$_{DD}$, that yields acceptable values of $\sigma_{Q_{critical}}/\mu_{Q_{critical}}$, is essential as explained in the proposed models.

-ii- From the formulas derived in Section 3.2.2, the critical charge nominal value for the SRAM cell is estimated accurately without time consuming transient simulations. For a target SER, the critical charge value can be calculated by using (2.4). Once the required critical charge is known, the circuit parameters are designed to achieve it without doing any SPICE simulations. For example, if the target critical charge, Q_o, the SRAM cell power supply, V_{DD}, should be selected such that $\mu_{Q_{critical}} - 3 \times \sigma_{Q_{critical}}$ is larger than Q_o to ensure that 99.87% of the SRAM samples will achieve the target SER, as portrayed in Figure 3.18.

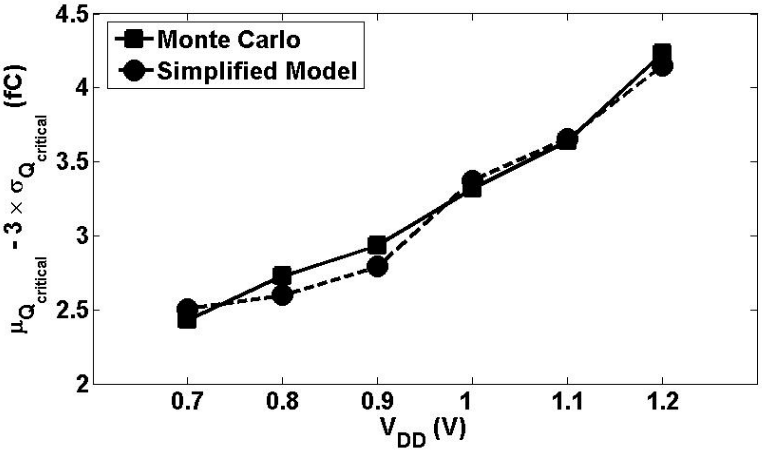


Figure 3.18: The $\mu_{Q_{critical}} - 3 \times \sigma_{Q_{critical}}$ curves obtained from Monte Carlo simulations and from the simplified model

-iii- The coupling capacitor, C$_c$, can result in a maximum $\sigma_{Q_{critical}}/\mu_{Q_{critical}}$, as depicted in (3.27). Although, this occurs in the designed SRAM, proposed in this work, only when τ exceeds 1000ps, it can occur at lower τ values for a different SRAM design, when the condition, in (3.27), is satisfied. Therefore, the circuit designer must be aware, at the design level, of this condition and avoid it.

-iv- For C$_c$ $= C_{cmin}$, V$_{tp1}$ variations dominate the overall critical charge variations. Thus, $\sigma_{Q_{critical}}/\mu_{Q_{critical}}$ is at its minimum value and inversely proportional to (V$_{DD}$ -|V$_{tp1}$|). Therefore, a further increase in C$_c$ results in increasing Q$_{critical}$, while keeping $\sigma_{Q_{critical}}/\mu_{Q_{critical}}$ constant.

-v- For C$_c$ $< C_{cmin}$, the variations of both V$_{tn2}$ and V$_{tp2}$ dominate $\sigma_{Q_{critical}}/\mu_{Q_{critical}}$. These variations decay exponentially with (t$_{up}$/τ). Therefore, to reduce $\sigma_{Q_{critical}}/\mu_{Q_{critical}}$ in this case, either increasing t$_{up}$ or reducing the average charging current.

-vi- Since the two extremes of C$_c$, that result in maximum and minimum $\sigma_{Q_{critical}}/\mu_{Q_{critical}}$, can be obtained from the proposed models, the circuit designer can determine C$_c$ that results in a certain Q$_{critical}$ and $\sigma_{Q_{critical}}/\mu_{Q_{critical}}$, while satisfying the power and performance constraints at the design level.

3.2.6 Models Extension to Super-Threshold Flip-Flops Circuits

Although this section has focused on the critical charge and its variability modeling for the SRAM cell, it can be extended to model them in flip-flops circuits. This is possible, because all the flip-flops topologies consist of an embedded cross-coupled inverters as those in the SRAM cell. However, these inverters are not symmetric like those in the SRAM cell. The proposed models can be extended to account for asymmetrical inverters by simply assuming that $V_{M1} \neq V_{M2}$.

3.3 Sub-Threshold SRAM Cells

Sub-threshold digital circuit design is one of the best energy saving techniques for applications with strict energy constraints [39, 122–127]. SRAM cells comprise a significant percentage of the total area of many digital chips [128, 129]. For this reason, SRAM cells leakage power dominates the total leakage power of the chip. Moreover, the large switched capacitances in the SRAM cells bitlines and wordlines increase the SRAM cells access energy [128, 129]. The design of sub-threshold SRAM cells reduces both leakage power and access energy. However, robustness and process variations are the main design challenges for sub-threshold SRAM cells [122]. In the following, analytical models for the soft errors immunity variations for sub-threshold SRAM cells are presented.

3.3.1 Accurate Model Assumptions and Derivations

The 6T SRAM cell shown in Figure 3.1 can not be used in the sub-threshold operation due to poor noise margins and is limited to the super-threshold operation [123, 124]. Therefore, several SRAM cells are reported in the literature to overcome the poor noise margins problem by adding extra transistors. For example, the 10T SRAM cell shown in Figure 3.19 has four extra transistors implementing a read buffer that isolates the reading and writing ports [124]. Thus, the WL, BL, and BLB lines are used for the writing operation whereas the RWL and RBL are used for the reading operation. Most of the sub-threshold SRAM cells reported in the literature utilize the same two port cell topology [86, 123, 130, 131]. The core of these sub-threshold SRAM implementations is the two cross-coupled inverters. These two cross-coupled inverters used in the sub-threshold SRAM cells might be designed asymmetrically to improve the noise margins.

The conventional SRAM cell shown in Figure 3.1 has its highest susceptibility to particle strikes in the standby mode because the storage nodes are disconnected from the highly capacitive bitlines [37, 67]. However, the sub-threshold SRAM cell shown in Figure 3.19 has its highest susceptibility to particle strikes when in the reading mode or the standby mode. This is because in both reading and standby modes, the storage nodes are disconnected from the highly capacitive bitlines. The buffer node V_C is driven by V_B and is connected to the highly capacitive RBL during the reading operation. Thus, V_C is less susceptible to particle strikes than V_A and V_B.

Accordingly, only the storage nodes V_A and V_B are considered in the following analysis. Thus, the access transistors, M_{a1} and M_{a2}, and the read buffer extra transistors, M_{bn1}, M_{bn2}, M_{bn3}, and M_{bp1} are excluded from the analysis. Assume that V_A stores logic “1” and accordingly V_B stores logic “0”. Thus, transistors, M_{p1} and M_{n2}, are conducting more sub-threshold currents than transistors, M_{n1} and M_{p2}, to maintain V_A and V_B voltages, respectively.

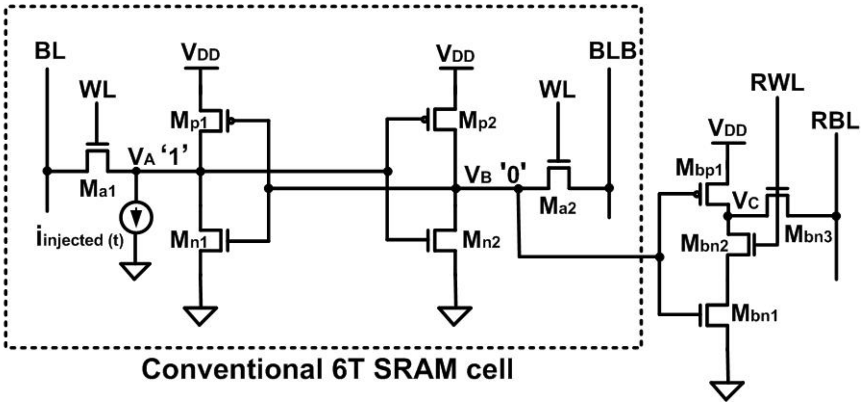


Figure 3.19: The sub-threshold 10T SRAM cell which consists of a conventional 6T SRAM cell and an extra read buffer to improve the noise margins [124]. In this 10T sub-threshold SRAM cell, the particle strike is modeled by a current pulse source ($i_{injected}(t)$). V_A is assumed to be at logic "1" and V_B is assumed to be at logic "0".

In the super-threshold 6T SRAM cell design, the data node storing logic "1" (V_1 in Figure 3.1) is the most susceptible to particle strikes [55]. Therefore, super-threshold SRAM critical charge modeling considers only the 1-to-0 flipping case and ignores the 0-to-1 flipping case [37, 67] as performed in Section 3.2. However, this assumption cannot be applied in the sub-threshold SRAM cell, because the NMOS transistor might become the weaker transistor, depending on the sizing and the technology parameters. In the sub-threshold region, the sub-threshold current is modeled as [15, 28]

$$i_{sub} = i_o \; exp(\frac{V_{GS} - V_t}{nV_T}) \; [1 - exp(\frac{-V_{DS}}{V_T})] \tag{3.30}$$

$$\text{where} \quad i_o = \mu_o \; C_{ox} \; (\frac{W}{L}) \; (V_T)^2 \; exp(1.8), \quad n = 1 + \frac{3T_{ox}}{W_{dm}}, \quad \text{and} \quad V_T = \frac{KT}{q} \tag{3.31}$$

where V_{GS} and V_{DS} are the transistor gate-to-source and drain-to-source voltages, respectively, V_t is the transistor threshold voltage, μ_o is the zero bias mobility, C_{ox} is the gate oxide capacitance, n is the sub-threshold swing coefficient, T_{ox} is the gate oxide thickness, W_{dm} is the maximum depletion layer width, V_T is the thermal voltage, K is the Boltzman constant, T is the temperature in ^{o}K, q is the electron charge, and W and L are the transistor channel width and length, respectively.

Therefore, the ratio between transistors M_{n2} and M_{p1} currents is given by

$$\frac{i_{n2_{sub}}}{i_{p1_{sub}}} = \frac{\mu_n \ (\frac{W}{L})_{n2}}{\mu_p \ (\frac{W}{L})_{p1}} \ \ exp(\frac{|V_{tp1}| - V_{tn2}}{n \ V_T}) \tag{3.32}$$

In typical CMOS technologies, the PMOS transistor threshold voltage, |V$_{tp}$|, and mobility, μ_p, are lower than V$_{tn}$ and μ_n , the NMOS transistor threshold voltage and mobility , respectively [22]. Thus, if the ratio, $\frac{\mu_n \ (\frac{W}{L})_{n2}}{\mu_p \ (\frac{W}{L})_{p1}}$, is much less than $exp(\frac{V_{tn2} - |V_{tp1}|}{n \ V_T})$, the NMOS transistor, M$_{n2}$, will be weaker than the PMOS transistor, M$_{p1}$, and the data node storing logic "0" (V$_B$ in Figure 3.19) is the most susceptible to particle strikes. Therefore, in sub-threshold SRAM critical charge modeling, the most susceptible node to particle strikes must be determined in advance. Sometimes, the critical charge values for the 1-to-0 flip and 0-to-1 flip cases are comparable, and both should be modeled and investigated.

1. **Critical Charge Model**

   In the following analysis, the first case when the ratio, $\frac{\mu_n \ (\frac{W}{L})_{n2}}{\mu_p \ (\frac{W}{L})_{p1}}$, is much larger than $exp(\frac{V_{tn2} - |V_{tp1}|}{n \ V_T})$ is investigated. Here, V$_A$ is more susceptible to particle strike than V$_B$. The other case, when V$_B$ is more susceptible to particle strike, is addressed later in this section. The particle strike is modeled by (3.3). The nodal current equation at node V$_A$ is written as

$$C_A \frac{dV_A}{dt} = [i_{p1_{sub}} - i_{n1_{sub}}] - i_{injected}(t) \tag{3.33}$$

   where C$_A$ is node V$_A$ capacitance; i$_{p1_{sub}}$ is the sub-threshold restoring current of the PMOS transistor, M$_{p1}$, which tries to pull-up V$_A$ to the supply voltage (V$_{DD}$); i$_{n1_{sub}}$ is the NMOS transistor, M$_{n1}$, sub-threshold current; and i$_{injected}$(t) is the injected current pulse given in (3.3).

   From (3.33), the values of Q and τ that equalize [i$_{p1_{sub}}$-i$_{n1_{sub}}$], and i$_{injected}$(t) currents are obtained. Hence, V$_A$ voltage attains a certain minimum value V$_{min}$. The time at which V$_{min}$ occurs is denoted by t$_{min}$ and given by

$$t_{min} = \tau \ln(\frac{Q}{\tau \ [i_{p1_{sub}} - i_{n1_{sub}}]}) \tag{3.34}$$

   By solving the differential equation in (3.33) and using (3.34), The value of V$_{min}$ is given by

$$V_{min} = V_{DD} - \ \frac{1}{C_A} \ (Q - [i_{p1_{sub}} - i_{n1_{sub}}] \ [t_{min} + \tau]) \tag{3.35}$$

The sub-threshold currents, $i_{p1_{sub}}$ and $i_{n1_{sub}}$, are given by (assuming $V_B \approx 0$):

$$\begin{aligned} i_{p1sub} &= i_{p1o}\ exp(\frac{V_{DD} - |V_{tp1}|}{nV_T})\ [1 - exp(\frac{-(V_{DD} - V_A)}{V_T})] \\ i_{n1sub} &= i_{n1o}\ exp(\frac{-V_{tn1}}{nV_T})\ [1 - exp(\frac{-V_A}{V_T})] \end{aligned} \quad (3.36)$$

$$\begin{aligned} \text{where}\quad i_{p1o} &= \mu_p\ C_{ox}\ (\frac{W}{L})_{p1}\ (V_T)^2\ exp(1.8), \\ i_{n1o} &= \mu_n\ C_{ox}\ (\frac{W}{L})_{n1}\ (V_T)^2\ exp(1.8) \end{aligned} \quad (3.37)$$

Since V_A changes from V_{DD} to V_{min} over the time interval [0, t_{min}], the currents $i_{p1_{sub}}$ and $i_{n1_{sub}}$ vary during the same time interval. In order to simplify the solution of the differential equation in (3.33), the currents, $i_{p1_{sub}}$ and $i_{n1_{sub}}$, are averaged over this time interval and are considered constants. Furthermore, V_{DD} and V_{min} are assumed to be greater than $3V_T$ ($\approx$ 75mV at room temperature). Therefore, the term [1-$exp(\frac{-V_A}{V_T})$] $\approx$ 1 and the term, [1-$exp(\frac{-(V_{DD}-V_A)}{V_T})$], is averaged over the time interval [0, t_{min}]. This average value is denoted by β_1 and given by

$$\beta_1 = \frac{1}{V_{DD} - V_{min}} \int_{V_{min}}^{V_{DD}} [1 - exp(\frac{-(V_{DD} - V_A)}{V_T})]dV_A \approx 1 - \frac{V_T}{V_{DD} - V_{min}} \quad (3.38)$$

Similar to Section 3.2, if V_{min} is slightly below the switching voltage of the second inverter, V_{M2}, V_B rises to logic "1", decreasing the restoring current, and resulting in a soft error. V_M is obtained by equating the inverter PMOS and NMOS sub-threshold currents, assuming that the input and output voltages equal V_M. Thus, V_M is given by (assuming V_M and $(V_{DD} - V_M) \geq 3V_T$)

$$V_M = \frac{1}{2}\ [V_{DD} - |V_{tp}| + V_{tn} + n\ V_T\ \ln(\frac{i_{po}}{i_{no}})] \quad (3.39)$$

Now, for the flipping case (i.e., $V_{min} < V_{M2}$), V_B voltage is assumed to stay around 0V for the time interval over which V_A is approaching V_{min} (i.e., t_{min}), and then starts to rise. Furthermore, V_A is assumed to remain constant at V_{min}, until V_B rises and exceeds the switching threshold of the first inverter, V_{M1}. These assumptions are summarized in the following equation [37, 67]

$$\begin{aligned}
&\text{for } 0 \le t \le t_{min} \begin{cases} V_A(t) : V_{DD} \rightarrow V_{min} \\ V_B(t) \approx 0V \end{cases} \\
&\qquad\qquad \text{and} \\
&\text{for } t_{min} \le t \le t_f \begin{cases} V_A(t) \approx V_{min} \\ V_B(t) : 0V \rightarrow V_{M1} \end{cases}
\end{aligned} \tag{3.40}$$

where t$_f$ is the flipping time at which V$_B$ hits V$_{M1}$. This assumption is validated by noticing that once V$_B$ hits V$_{M1}$, the positive feedback of the cell becomes strong enough to continue flipping the cell state. Equation (3.40) allows decoupling the cross-coupled inverters of the SRAM cell, as proposed in [37, 67].

From (3.35) and for a given τ, the value of Q that just cause V$_A$ to flip is obtained by equating V$_{min}$ with V$_{M2}$. As a result, Q is determined by

$$Q = C_A \ (V_{DD} - V_{M2}) + [i_{p1_{sub}} - i_{n1_{sub}}] \ [t_{min} + \tau] \tag{3.41}$$

By substituting (3.41) in (3.34), t$_{min}$ is calculated by solving the following equation:

$$t_{min} = \tau \ \ln(\gamma + t_{min}/\tau) \tag{3.42}$$

$$\text{where} \quad \gamma = 1 + \frac{C_A \ (V_{DD} - V_{M2})}{\tau \ [i_{p1_{sub}} - i_{n1_{sub}}]} \tag{3.43}$$

Equation (3.42) is a nonlinear equation that is solved numerically by using the Lambert W function (also called the Omega function), $\Omega(x)$ [132]. A more detailed definition of $\Omega(x)$ is given in Appendix A. t$_{min}$ is expressed as:

$$t_{min} = \tau \ [-\gamma - \Omega_{-1}(-\exp(-\gamma))] \tag{3.44}$$

Now, the objective is to find the flipping time t$_f$. t$_f$ is the sum of t$_{min}$, and the time delay that V$_B$ takes to rise from 0V to V$_{M1}$ (this time is denoted by t$_{up}$). This delay is driven by transistors M$_{p2}$ and M$_{n2}$, where their gate voltage V$_A$ is constant at V$_{M2}$ (Equation (3.40)). The nodal current equation at node V$_B$ is given by

$$C_B \frac{dV_B}{dt} = i_{p2_{sub}} - i_{n2_{sub}} \tag{3.45}$$

where C$_B$ is the capacitance of node V$_B$. The currents, i$_{p2_{sub}}$ and i$_{n2_{sub}}$, are given by

$$\begin{aligned} i_{p2sub} &= i_{p2o}\ exp(\frac{V_{DD} - V_{M2} - |V_{tp2}|}{nV_T})\ [1 - exp(\frac{-(V_{DD} - V_B)}{V_T})] \\ i_{n2sub} &= i_{n2o}\ exp(\frac{V_{M2} - V_{tn2}}{nV_T})\ [1 - exp(\frac{-V_B}{V_T})] \end{aligned} \tag{3.46}$$

$$\begin{aligned} \text{where} \quad i_{p2o} &= \mu_p\ C_{ox}\ (\frac{W}{L})_{p2}\ (V_T)^2\ exp(1.8), \\ i_{n2o} &= \mu_n\ C_{ox}\ (\frac{W}{L})_{n2}\ (V_T)^2\ exp(1.8) \end{aligned} \tag{3.47}$$

Similarly, since V$_B$ changes from 0V to V$_{M1}$ over the time interval [t$_{min}$, t$_f$], the currents, i$_{p2_{sub}}$ and i$_{n2_{sub}}$, vary during the same time interval. In order to simplify the solution of the differential equation in (3.45), the currents, i$_{p2_{sub}}$ and i$_{n2_{sub}}$, are averaged over this time interval and are considered constants. Furthermore, V$_{DD}$ and V$_M$ are assumed to be greater than 3V$_T$. Consequently, the term [1-$exp(\frac{V_{DD}-V_B}{V_T})$] $\approx$ 1 and the term, [1-$exp(\frac{-V_B}{V_T})$], is averaged over the time interval [t$_{min}$, t$_f$]. This average value is denoted by β_2 such that

$$\beta_2 = \frac{1}{V_{M1}} \int_0^{V_{M1}} [1 - exp(\frac{-V_B}{V_T})]dV_B \approx 1 - \frac{V_T}{V_{M1}} \tag{3.48}$$

By solving the differential equation in (3.45), the delay t$_{up}$ is expressed as

$$t_{up} = \frac{C_B\ V_{M1}}{[i_{p2_{sub}} - i_{n2_{sub}}]} \tag{3.49}$$

Thus, the critical charge, Q$_{critical}$, is obtained by using (3.13). The analytical formulas of the proposed critical charge model are summarized in Table 3.2.

2. **Statistical Critical Charge Variation Model**

From Table 3.2, it is evident that the critical charge, Q$_{critical}$, is dependent on the threshold voltages of transistors, M$_{p1}$, M$_{p2}$, M$_{n1}$, and M$_{n2}$, represented by V$_{tp1}$, V$_{tp2}$, V$_{tn1}$, and V$_{tn2}$, respectively. Similar to Section 3.2.2, the standard deviation of the critical charge variations is found as follows:

Table 3.2: Analytical formulas for the critical charge model

$Q_{critical} = Q\ (1\text{-exp}(\text{-}t_f\ /\ \tau))$
$Q = C_A\ (V_{DD} - V_{M2}) + [i_{p1_{sub}} - i_{n1_{sub}}]\ [t_{min} + \tau]$ $t_f = t_{min} + t_{up}$
$t_{min} = \tau\ [-\gamma - \Omega_{-1}(-\exp(-\gamma))], \quad \gamma = 1 + \frac{C_A\ (V_{DD}-V_{M2})}{\tau\ [i_{p1_{sub}}-i_{n1_{sub}}]}$ $t_{up} \quad = \frac{C_B\ V_{M1}}{[i_{p2_{sub}}-i_{n2_{sub}}]}$
$V_{M_{1,2}} = \frac{1}{2}\ [V_{DD} - \lvert V_{tp_{1,2}} \rvert + V_{tn_{1,2}} + n\ V_T\ \ln(\frac{i_{p_{1,2}o}}{i_{n_{1,2}o}})]$ $i_{p_{1,2}o} = \mu_p\ C_{ox}\ (\frac{W}{L})_{p_{1,2}}\ (V_T)^2\ exp(1.8),$ $i_{n_{1,2}o} = \mu_n\ C_{ox}\ (\frac{W}{L})_{n_{1,2}}\ (V_T)^2\ exp(1.8)$
$i_{p1sub} = \beta_1\ i_{p1o}\ exp(\frac{V_{DD}-\lvert V_{tp1} \rvert}{nV_T})$ $i_{n1sub} = i_{n1o}\ exp(\frac{-V_{tn1}}{nV_T})$ $i_{p2sub} = i_{p2o}\ exp(\frac{V_{DD}-V_{M2}-\lvert V_{tp2} \rvert}{nV_T})$ $i_{n2sub} = \beta_2\ i_{n2o}\ exp(\frac{V_{M2}-V_{tn2}}{nV_T})$ $\beta_1 = 1 - \frac{V_T}{V_{DD}-V_{M2}}, \qquad \beta_2 = 1 - \frac{V_T}{V_{M1}}$

$$\sigma_{Q_{critical}} = \{ (\frac{\partial Q_{critical}}{\partial V_{tp1}})^2 \sigma^2_{V_{tp1}} + (\frac{\partial Q_{critical}}{\partial V_{tp2}})^2 \sigma^2_{V_{tp2}} + (\frac{\partial Q_{critical}}{\partial V_{tn1}})^2 \sigma^2_{V_{tn1}} + (\frac{\partial Q_{critical}}{\partial V_{tn2}})^2 \sigma^2_{V_{tn2}} \}^{0.5} \quad (3.50)$$

where $\sigma_{Vtp1}, \sigma_{Vtp2}, \sigma_{Vtn1}$, and σ_{Vtn2} are the standard deviations of the threshold voltages, $V_{tp1}, V_{tp2}, V_{tn1}$, and V_{tn2}, respectively. This accurate model is valid for the same assumptions stated in Section 3.2.2.

It should be emphasized that the previous analysis is valid for the 1-to-0 flip, when V$_A$ is more susceptible to soft errors than V$_B$. This occurs when $\frac{\mu_n \ (\frac{W}{L})_{n2}}{\mu_p \ (\frac{W}{L})_{p1}}$ is much larger than $exp(\frac{V_{tn2}-|V_{tp1}|}{n \ V_T})$. However, when $\frac{\mu_n \ (\frac{W}{L})_{n2}}{\mu_p \ (\frac{W}{L})_{p1}}$ is much less than $exp(\frac{V_{tn2}-|V_{tp1}|}{n \ V_T})$, V$_B$ is more susceptible to soft errors than V$_A$. Therefore, the 0-to-1 flip case should be considered. Accordingly, the previous analysis can be repeated by replacing Equations (3.33) and (3.45) by the following differential equations at nodes A and B

$$C_B \frac{dV_B}{dt} = [i_{p2_{sub}} - i_{n2_{sub}}] + i_{injected}(t) \quad (3.51)$$

$$C_A \frac{dV_A}{dt} = [i_{p1_{sub}} - i_{n1_{sub}}] \quad (3.52)$$

Consequently, all the previously derived equations are used again for the 0-to-1 flip case by replacing C_A and C_B by C_B and C_A, respectively; the parameters of the transistors M_{p1}, M_{p2}, M_{n1}, and M_{n2} by the parameters of the transistors M_{n2}, M_{n1}, M_{p2}, and M_{p1}, respectively; and Q by -Q.

3.3.2 Approximate Model Assumptions and Derivations

In this section, this accurate model is approximated to account for the critical charge variations from a design perspective. The following assumptions are made to derive this approximate model.

-i- In sub-threshold SRAM cell, the flipping time, t$_f$, is larger than τ due to the lower supply voltages and smaller sub-threshold currents values (typically, $t_f/\tau \geq 3$). Therefore, the critical charge expression in (3.13) is approximated by Q$_{critical} \approx Q$ for the critical charge variability calculations.

-ii- The variation of the inverter threshold voltage V$_{M2}$, expressed in (3.39), which is dependent on V$_{tp2}$ and V$_{tn2}$ is calculated to be less than 2.3%, relative to its mean value. As a result, the variations in V$_{M2}$ are ignored, and V$_{M2}$ is assumed constant from the variations perspective. Therefore, $\sigma_{Q_{critical}}$ is dependent on only the variations in V$_{tp1}$ and V$_{tn1}$ through i$_{p1sub}$ and i$_{n1sub}$, respectively.

-iii- The current, i$_{n1sub}$, expressed in Table 3.2, is neglected with respect to i$_{p1sub}$, if V$_{tn1}$ $\geq 3nV_T$. This condition is always satisfied, since $3nV_T$ ≈125mV at room temperature for current CMOS technologies and the threshold voltages take on higher values than 125mV. Thus, the V$_{tp1}$ contribution to $\sigma_{Q_{critical}}$ dominates all other threshold voltages variations.

By adopting these assumptions, Q$_{critical}$ is approximated by the following equation:

$$Q_{critical} \approx \;\tau\; i_{p1_{sub}}\; |\Omega_{-1}(-\exp(-\gamma))| \tag{3.53}$$

Similarly, $\sigma_{Q_{critical}}$ is approximated by the following equation:

$$\begin{aligned}\sigma_{Q_{critical}} &= |\frac{\partial Q_{critical}}{\partial V_{tp1}}|\; \sigma_{V_{tp1}} \approx |\frac{\partial Q}{\partial V_{tp1}}|\; \sigma_{V_{tp1}} \\ &\approx \tau\; \frac{i_{p1_{sub}}}{n\; V_T}\; [1+(\gamma-1)\;(1+\theta)+\frac{t_{min}}{\tau}]\; \sigma_{V_{tp1}} \end{aligned} \tag{3.54}$$

$$\text{where} \quad \theta = \frac{\partial\; \Omega_{-1}(-\exp(-\gamma))}{\partial\gamma} = -\; \frac{\Omega_{-1}(-\exp(-\gamma))}{1+\Omega_{-1}(-\exp(-\gamma))} \tag{3.55}$$

Equation (3.54) is simplified further by using the formulas tabulated in Table 3.2, and $\sigma_{Q_{critical}}$ is approximated further by

$$\sigma_{Q_{critical}} \approx (\frac{\tau i_{p1_{sub}} \sigma_{V_{tp1}}}{nV_T})|\frac{\Omega_{-1}(-\exp(-\gamma))(\gamma+\Omega_{-1}(-\exp(-\gamma)))}{1+\Omega_{-1}(-\exp(-\gamma))}| \tag{3.56}$$

Thus, the relative critical charge variation, ($\sigma_{Q_{critical}}/\mu_{Q_{critical}}$), is given by the following equation:

$$\sigma_{Q_{critical}}/\mu_{Q_{critical}} \approx |(\frac{\sigma_{V_{tp1}}}{n\; V_T})\frac{\gamma+\Omega_{-1}(-\exp(-\gamma))}{1+\Omega_{-1}(-\exp(-\gamma))}| \tag{3.57}$$

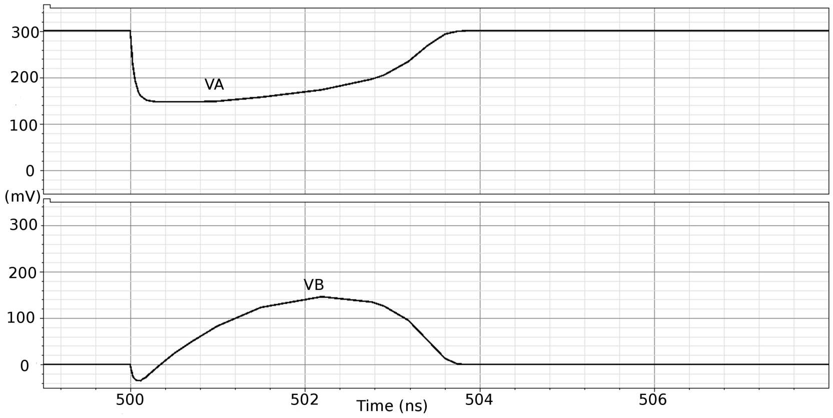


Figure 3.20: V_A and V_B voltages in the non-flipping case when V_A voltage falls down till it hits V_{min} then it recovers back to V_{DD}. In this case, V_B voltage does not hit V_{M1}, and therefore, V_A recovers. In this simulation, $V_{DD} = 0.3V$.

3.3.3 Results and Discussions

1. **Verification of the Models Assumptions**

   First, the assumptions in (3.40) are justified. Figure 3.20 shows V_A and V_B voltages in the non-flipping case. It is clear that since V_B voltage cannot hit V_{M1}, the SRAM cell recovers. However, in Figure 3.21, V_B hits V_{M1}, and hence, the SRAM cell exhibits a soft error. Moreover, Figure 3.21 depicts that the V_B voltage is approximately 0V, as long as the V_A voltage is falling. Once the V_A voltage reaches V_{min}, the V_A voltage stays constant at V_{min}, whereas the V_B voltage starts to rise to V_{M1}. These results ensure that the assumptions made in (3.40) are realistic. It should be mentioned that the minimum voltage, V_{min}, shown in Figure 3.21 at which V_A stays constant before flipping to 0V ($\approx$ 146.9mV), is slightly less than V_{min} shown in Figure 3.20 ($\approx$ 148.4mV) for the non-flipping case. This demonstrates that the flipping occurs, when V_{min} is less than V_{M2} ($V_{M2}\approx$ 147.8mV). In all the simulations, an industrial hardware-calibrated 65nm CMOS technology transistor model, whose technological parameters are listed in Table 3.1, is employed with the sub-threshold swing coefficient, $n = 1.6$.

2. **Verification of the Models Estimated Critical Charge**

   To verify the critical charge nominal value, and the critical charge variations models, the analytical models are compared to the simulation results from SPICE transient and Monte Carlo simulations. These simulations are performed to validate the nom-

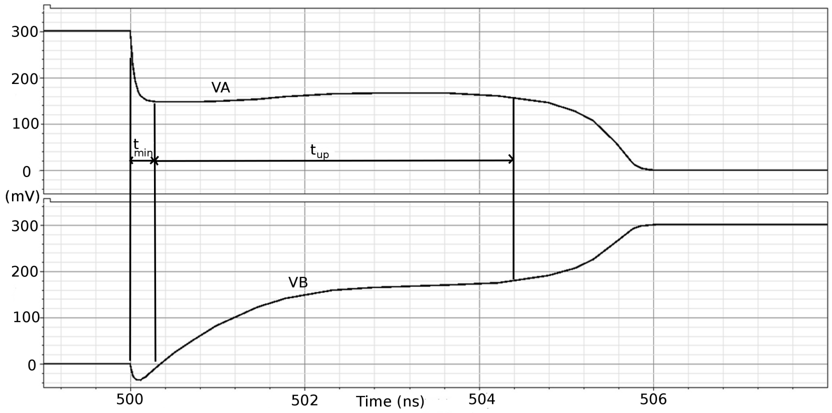


Figure 3.21: V_A and V_B voltages in the flipping case when V_B voltage hits V_{M1}, and hence, the SRAM cell exhibits a soft error. In this simulation, $V_{DD} = 0.3V$.

inal critical charge and the critical charge variability models for both the proposed accurate and approximate models.

In the following, the validation results for these models are presented. 5,000 Monte Carlo runs are used to provide a good accuracy in determining the critical charge mean and standard deviation. For each Monte Carlo run, the value of the current pulse charge, Q, which causes the cell to flip, is determined. Then, the simulations are repeated for different V_{DD} (from 0.15V to 0.3V) to find the effect of reducing V_{DD} on the critical charge mean and variations. The SRAM cell sizing shown in Table 3.1 is used in the simulation setups.

-i- **Nominal Critical Charge:**

Figure 3.22 displays the nominal critical charge which is obtained by using the transient simulations, $Q_{critical}$, and Monte Carlo simulations, $\mu_{Q_{critical}}$. Clear agreement between $Q_{critical}$ and $\mu_{Q_{critical}}$ justifies the linearity approximation assumption (i.e., process variations affect only on the critical charge variance (spread) and have no effect on its mean).

Figure 3.23 demonstrates the nominal critical charge value calculated from the proposed accurate and approximate model versions, and compared to the transient simulation results for different supply voltage values. It should be noted that the approximate model is proposed mainly for the WID variations estimation, although the model still shows an acceptable match for the nominal critical charge value. These results are obtained by using $\tau = 500$ps. This larger τ value, compared to the super-threshold case, is adopted in the sub-threshold SRAM design since the flipping time is larger than 4ns. It is evident from Figures 3.22

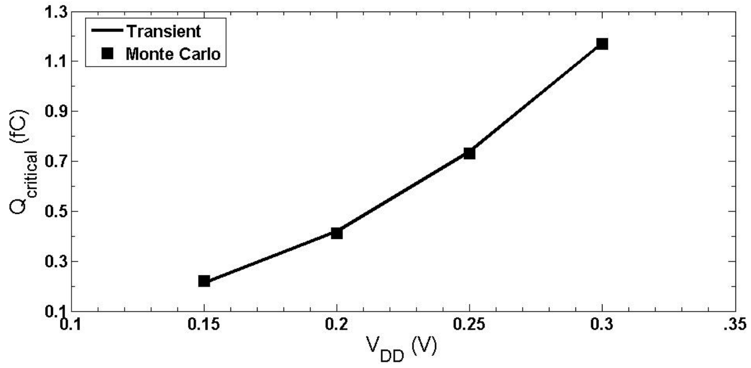


Figure 3.22: $Q_{critical}$ versus V$_{DD}$ from the transient simulations (when no variations are introduced) and from Monte Carlo simulations. Clear agreement between Q$_{critical}$ (obtained from transient simulations) and $\mu_{Q_{critical}}$ (obtained from Monte Carlo simulations) justifies the linearity approximation assumption.

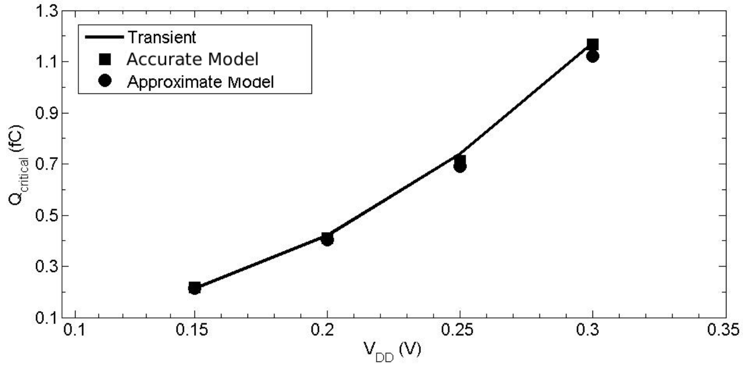


Figure 3.23: $Q_{critical}$ versus V$_{DD}$ from SPICE transient simulations. Also shown the results from the proposed accurate and approximate models.

and 3.23 that reducing the supply voltage decreases the critical charge, which is expected.

Figure 3.24 depicts the values of Q$_{critical}$, computed from the proposed models and from SPICE transient simulations for different current pulse widths (by varying τ from 250ps to 750ps). The approximate model results are not revealed in this figure since the approximate model is primarily introduced for the estimation of the WID variations. In Figure 3.24, It is observed that as the current pulse width increases (i.e., the diffusion current dominates), the critical charge increases.

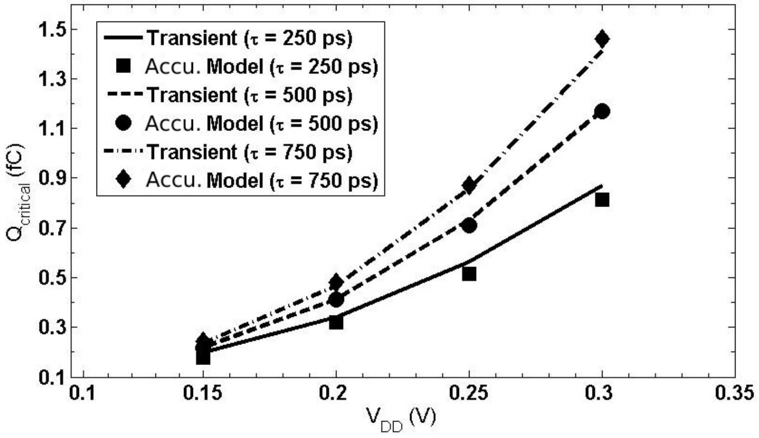


Figure 3.24: $Q_{critical}$ versus V$_{DD}$ for different values of τ (250ps, 500ps, and 750ps) from the transient simulations and from the proposed accurate model.

-ii- **Critical Charge Variations:**

Figure 3.25 shows the simulation results for $\sigma_{Q_{critical}}$ for different V$_{DD}$ values. Note that each data point represents $\sigma_{Q_{critical}}$ calculated from 5,000 Monte Carlo runs. Also, Figure 3.25 displays the results from the proposed models. The results of both models match those of the simulations. Figure 3.26 shows the critical charge standard deviation, $\sigma_{Q_{critical}}$, obtained from simulations and from the approximate model for different values of τ. It is demonstrated that as τ is reduced, the critical charge variations are reduced. Also, reducing the supply voltage, V$_{DD}$, reduces $\sigma_{Q_{critical}}$. It is important to mention that only the approximate model is used in all the following results and discussions.

3. **Critical Charge Variations Design Knobs**

From (3.53), Q$_{critical}$ is a function of τ, i$_{p1_{sub}}$, and γ. However, γ is also a function of τ, i$_{p1_{sub}}$, C$_A$, and V$_{DD}$. Therefore, to investigate the effect of these parameters on Q$_{critical}$, (3.53) is rewritten as follows by using the fact that $\gamma = 1 + \frac{C_A\ (V_{DD}-V_{M2})}{\tau\ i_{p1_{sub}}}$.

$$\begin{cases} \frac{Q_{critical}}{\tau\ i_{p1_{sub}}} \approx |\Omega_{-1}(-\exp(-\gamma))| \\ \qquad\qquad \text{for } (\tau\ i_{p1_{sub}}) = constant \\ \\ \frac{Q_{critical}}{C_A\ (V_{DD}-V_{M2})} \approx \frac{C_A\ (V_{DD}-V_{M2})}{\gamma-1} |\Omega_{-1}(-\exp(-\gamma))| \\ \qquad\qquad \text{for } (C_A\ (V_{DD}-V_{M2})) = constant \end{cases} \tag{3.58}$$

Figure 3.27.a plots (Q$_{critical}/(\tau\ i_{p1_{sub}})$) versus γ for a constant ($\tau\ i_{p1_{sub}}$), and illustrates

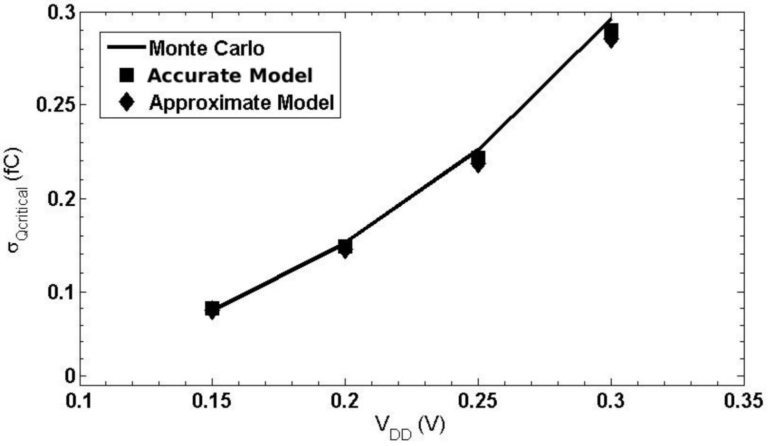


Figure 3.25: Critical charge variations, $\sigma_{Q_{critical}}$, versus V$_{DD}$ from Monte Carlo simulations and from the proposed accurate and approximate models.

that as γ increases, Q$_{critical}$ increases. Therefore, increasing C$_A$ and/or V$_{DD}$ increases γ, and accordingly, increases Q$_{critical}$. Figure 3.27.b plots (Q$_{critical}/(C_A\ (V_{DD}-V_{M2}))$) versus γ for a constant $(C_A\ (V_{DD}-V_{M2}))$ and shows that as γ increases, Q$_{critical}$ is reduced. Therefore, increasing τ and/or i$_{p1_{sub}}$ reduces γ, and accordingly, increases Q$_{critical}$. This result can be justified since increasing i$_{p1_{sub}}$, the transistor M$_{p1}$ restoring current, increases Q$_{critical}$.

Figure 3.24 portrays these results for V$_{DD}$ and τ, and demonstrates that increasing any of them increases Q$_{critical}$. Figure 3.28 illustrates the effect of C$_A$ and i$_{p1_{sub}}$ on Q$_{critical}$ and compares these results to SPICE transient simulation results. The subthreshold current, i$_{p1_{sub}}$, is varied by changing the width of transistor M$_{p1}$, W$_{p1}$. The capacitance, C$_A$, is varied by employing a variable coupling capacitor , C$_c$, between nodes V$_A$ and V$_B$. Then, The model capacitances, C$_A$ and C$_B$, are obtained by applying the Miller theorem as follows [37, 67].

$$C_A^{'} = C_A + 2C_c \quad \text{and} \quad C_B^{'} = C_B + 2C_c \tag{3.59}$$

Similarly, the same analysis, applied to Q$_{critical}$, is repeated for $\sigma_{Q_{critical}}$ and shown in Figures 3.29.a and 3.29.b. From these figures, $\sigma_{Q_{critical}}$ increases when any of the parameters τ, C$_A$, i$_{p1_{sub}}$, and V$_{DD}$ increases. From (3.56), $\sigma_{Q_{critical}}$ is proportional to $(\sigma_{V_{tp1}}\ i_{p1_{sub}})$, since $\sigma_{V_{tp1}}\ \ \alpha\ \ \frac{1}{\sqrt{W_{p1}}}$ and $i_{p1_{sub}}\ \ \alpha\ \ W_{p1}$, $\sigma_{Q_{critical}}$ is proportional to $\sqrt{W_{p1}}$. Therefore, increasing i$_{p1_{sub}}$ by increasing W_{p1}, reduces $\sigma_{V_{tp1}}$, but results in increasing $\sigma_{Q_{critical}}$.

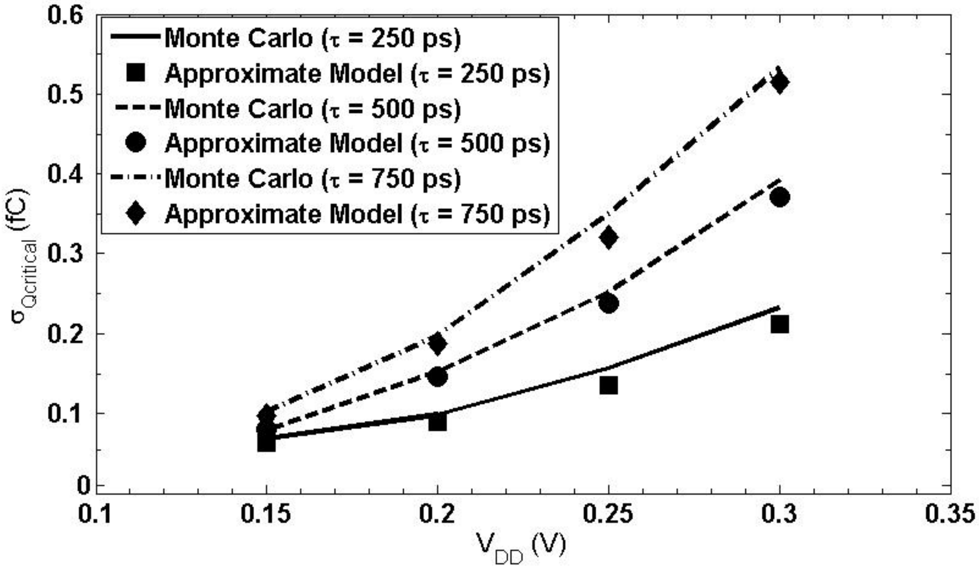


Figure 3.26: Critical charge variations $\sigma_{Q_{critical}}$ versus V$_{DD}$ for different values of τ (250ps, 500ps, and 750ps) from Monte Carlo simulation and from the proposed approximate model.

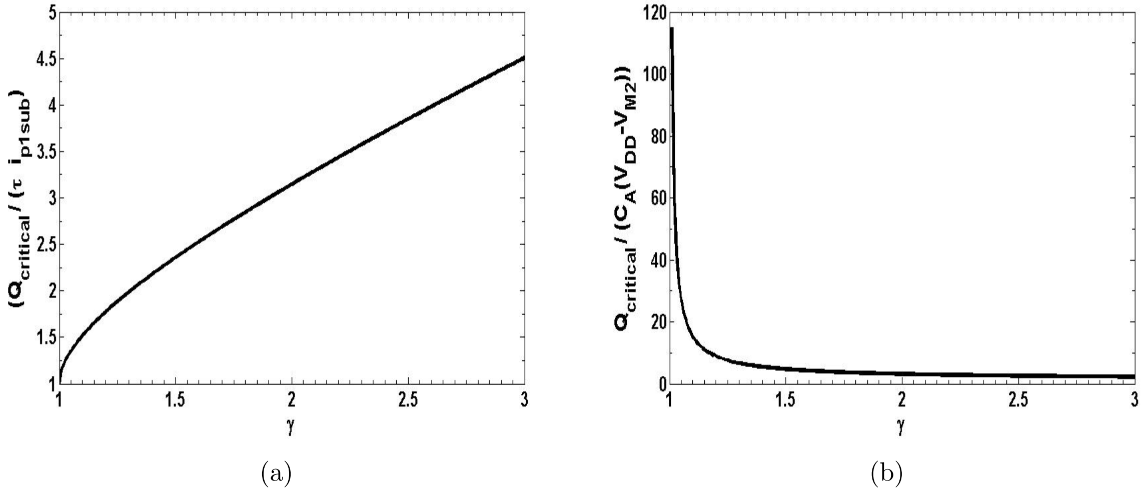


Figure 3.27: (a) (Q$_{critical}/(\tau\ i_{p1_{sub}})$) versus γ for a constant ($\tau\ i_{p1_{sub}}$) illustrating that as γ increases, Q$_{critical}$ increases and (b) (Q$_{critical}/(C_A\ (V_{DD} - V_{M2})$) versus γ for a constant ($C_A\ (V_{DD} - V_{M2})$) showing that as γ increases, Q$_{critical}$ is reduced in this case

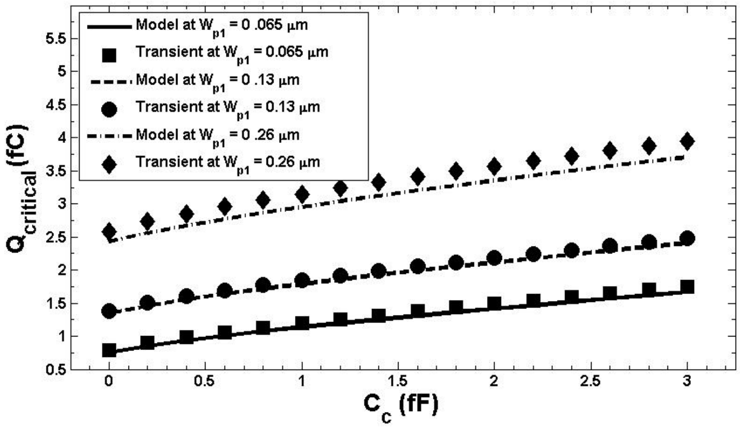


Figure 3.28: The effect of adding a coupling capacitor, C$_c$, for different values of W$_{p1}$ (0.065μ m, 0.13μ m ,and 0.26μ m) on Q$_{critical}$ from the proposed model and transient simulations. In this figure, V$_{DD}$ = 0.3V.

Figure 3.26 validates these results for V$_{DD}$ and τ, and shows that an increase in any of them, increases Q$_{critical}$. Figure 3.30 shows the effect of C$_A$ and W$_{p1}$ on $\sigma_{Q_{critical}}$, and compares these results to Monte Carlo simulation results.

By using (3.57), the relative critical charge variation ($\sigma_{Q_{critical}}/\mu_{Q_{critical}}$) is plotted versus γ for a constant ($\frac{\sigma_{V_{tp1}}}{n\ V_T}$), in Figure 3.31. According to this figure, increasing C$_A$ and/or V$_{DD}$ results in reducing ($\sigma_{Q_{critical}}/\mu_{Q_{critical}}$), whereas, ($\sigma_{Q_{critical}}/\mu_{Q_{critical}}$) is reduced by reducing τ. The effect of W$_{p1}$ on ($\sigma_{Q_{critical}}/\mu_{Q_{critical}}$) is different from its effect on either Q$_{critical}$ or $\sigma_{Q_{critical}}$. Although increasing W$_{p1}$ increases Q$_{critical}$ and $\sigma_{Q_{critical}}$, it results in reducing ($\sigma_{Q_{critical}}/\mu_{Q_{critical}}$) due to the dependence of $\sigma_{V_{tp1}}$ on W$_{p1}$ (Equation(3.57)). Figures 3.32 and 3.33 compare these results to Monte Carlo simulations.

4. **The Effect of the Temperature on the Critical Charge Relative Variations**

   Furthermore, the effect of the temperature, T, on ($\sigma_{Q_{critical}}/\mu_{Q_{critical}}$) is obtained by using (3.57), and is shown in Figures 3.34.a and 3.34.b for V$_{DD}$ equals 0.3V and 0.25V, respectively. These results are compared to Monte Carlo simulations. Figure 3.34.a shows that ($\sigma_{Q_{critical}}/\mu_{Q_{critical}}$) exhibits a minimum value at $T \approx 15\ ^oC$, when V$_{DD}$ = 0.3V. In addition, ($\sigma_{Q_{critical}}/\mu_{Q_{critical}}$) exhibits a minimum value at $T \approx 7\ ^oC$, when V$_{DD}$ = 0.25V. In the other cases, when V$_{DD}$ equals 0.2V and 0.15V, ($\sigma_{Q_{critical}}/\mu_{Q_{critical}}$) exhibits its minimum at $T < -30\ ^oC$ and are not shown in these figures. In general, as V$_{DD}$ is reduced, the temperature, at which ($\sigma_{Q_{critical}}/\mu_{Q_{critical}}$)

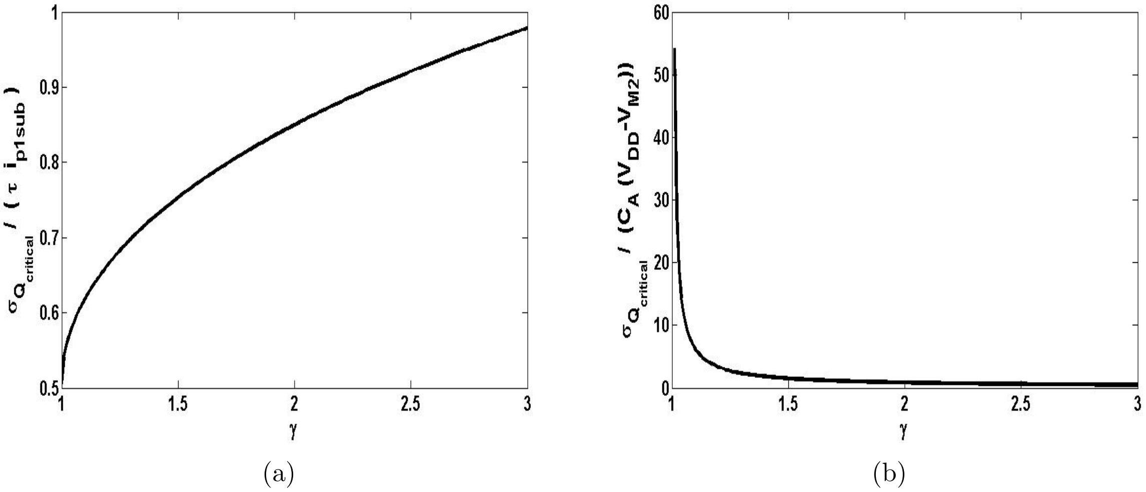


Figure 3.29: (a) $(\sigma_{Q_{critical}}/(\tau\ i_{p1_{sub}}))$ versus γ for a constant $(\tau\ i_{p1_{sub}})$ illustrating that as γ increases, $\sigma_{Q_{critical}}$ increases and (b) $(\sigma_{Q_{critical}}/(C_A\ (V_{DD} - V_{M2}))$ versus γ for a constant $(C_A\ (V_{DD} - V_{M2}))$ showing that as γ increases, $\sigma_{Q_{critical}}$ is reduced in this case.

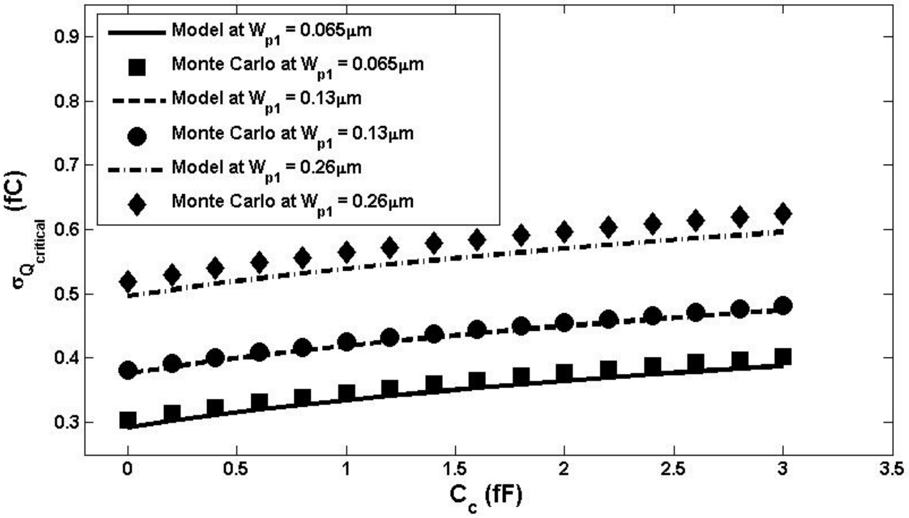


Figure 3.30: The effect of adding a coupling capacitor, C_c, for different values of W_{p1} (0.065μ m, 0.13μ m ,and 0.26μ m) on $\sigma_{Q_{critical}}$ from the proposed model and Monte Carlo simulations. In this figure, V_{DD} = 0.3V.

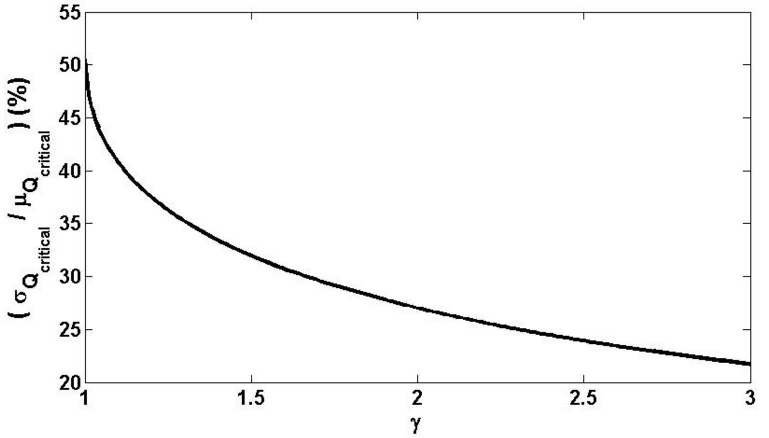


Figure 3.31: The relative critical charge variation, $(\sigma_{Q_{critical}}/\mu_{Q_{critical}})$, versus γ for a constant $(\frac{\sigma_{V_{tp1}}}{n\ V_T})$.

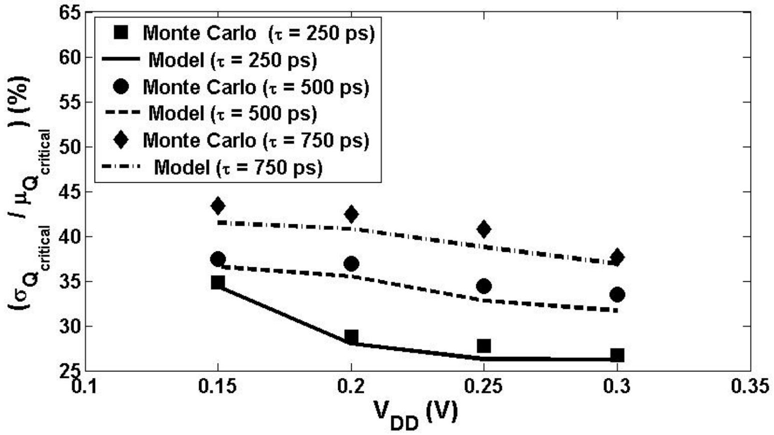


Figure 3.32: The relative critical charge variation, $(\sigma_{Q_{critical}}/\mu_{Q_{critical}})$, versus V$_{DD}$ for different values of τ (250ps, 500ps, and 750ps)from Monte Carlo simulations and from the proposed model.

is minimum, is reduced. This result is essential when the SRAM cells are used in applications with strict SER constraints such as space and satellite applications. Temperature control techniques can be employed to keep the temperature at the values that keep $(\sigma_{Q_{critical}}/\mu_{Q_{critical}})$ at its minimum value.

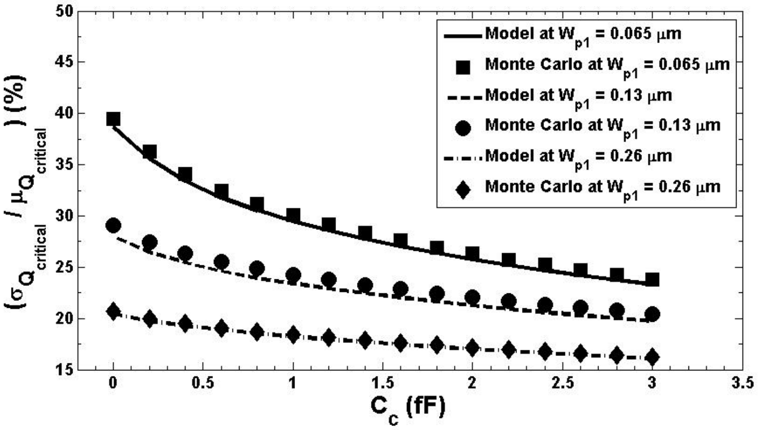


Figure 3.33: The effect of adding a coupling capacitor, C$_c$, for different values of W$_{p1}$ (0.065μ m, 0.13μ m ,and 0.26μ m) on ($\sigma_{Q_{critical}}/\mu_{Q_{critical}}$) from the proposed model and Monte Carlo simulations. In this figure, V$_{DD}$ = 0.3V.

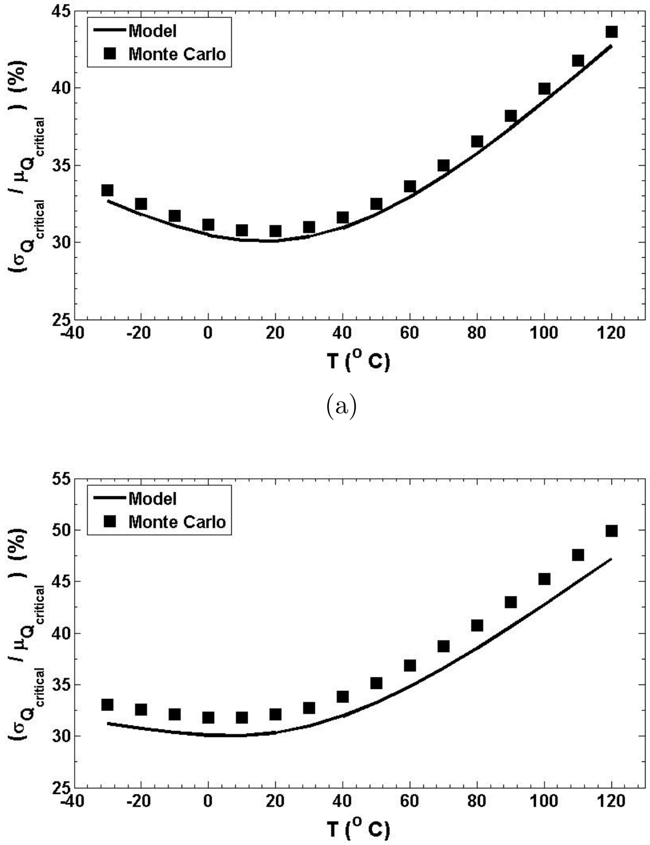
5. **The Effect of the Sub-Threshold Swing Coefficient on the Critical Charge Relative Variations**

   The effect of the sub-threshold swing coefficient, n, on ($\sigma_{Q_{critical}}/\mu_{Q_{critical}}$) is plotted in Figure 3.35 illustrating that increasing n results in reducing ($\sigma_{Q_{critical}}/\mu_{Q_{critical}}$). Therefore, increasing n can be used as a device optimization technique to mitigate the critical charge variability in sub-threshold SRAM cells. This sub-threshold device optimization is pivotal for applications with strict SER constraints.

6. **The Proposed Models Accuracy**

   In Figure 3.36, Q$_{critical}$, from the proposed accurate model, is plotted versus the transient simulation results for different values of τ, V$_{DD}$, W$_{p1}$, C$_A$, and T. The maximum error is 4.6%, and the average error is 2.1%. Figure 3.37 shows $\sigma_{Qcritical}$ from the approximate model plotted versus Monte Carlo simulation results for different values of τ, V$_{DD}$, W$_{p1}$, C$_A$, and T. The maximum error is 12.2%, and the average error is 5.4%. Good agreement between the proposed models and the simulation results justifies all the assumptions used to derive the models, as explained in Sections 3.3.1, and 3.3.2.

   As shown in the previous discussions, the proposed models are based on easily measurable parameters, which can be directly extracted from the measurements or technology information (i.e., C$_A$, C$_B$, σ_{V_t}, V$_t$, and n).



(a)

(b)

Figure 3.34: (a) $(\sigma_{Q_{critical}}/\mu_{Q_{critical}})$ versus T when V$_{DD}$ = 0.3V, showing that $(\sigma_{Q_{critical}}/\mu_{Q_{critical}})$ exhibits a minimum value at T = 15 ^{o}C and (b) $(\sigma_{Q_{critical}}/\mu_{Q_{critical}})$ versus T when V$_{DD}$ = 0.25V showing that $(\sigma_{Q_{critical}}/\mu_{Q_{critical}})$ exhibits a minimum value at T = 7 ^{o}C.

3.3.4 Design Insights

In this section, some design insights, extracted from the proposed models, are reported. Equations (3.53), (3.56), and (3.57) provide the following design insights:

-i- Increasing the supply voltage, V$_{DD}$, results in increasing both Q$_{critical}$ and $\sigma_{Q_{critical}}$. However, $\sigma_{Q_{critical}}/\mu_{Q_{critical}}$ is reduced by increasing V$_{DD}$. Therefore, the SER variations are readily minimized by the proper selection of the supply voltage, V$_{DD}$.

-ii- From the formulas in Table 3.2, the critical charge nominal value for the SRAM cell

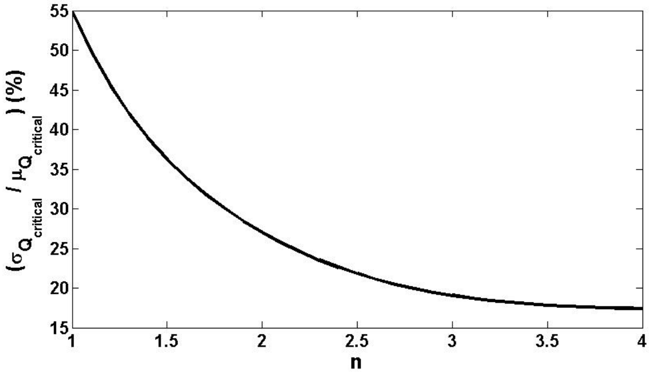


Figure 3.35: The effect of the sub-threshold swing coefficient, n, on $(\sigma_{Q_{critical}}/\mu_{Q_{critical}})$ for the case when V$_{DD}$ = 0.3V.

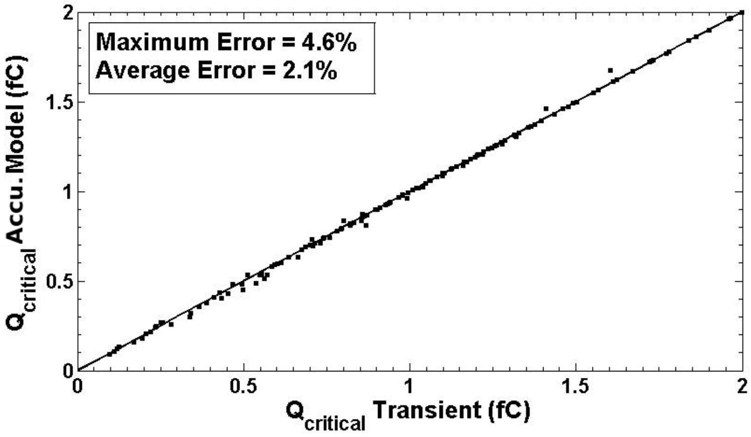


Figure 3.36: Q$_{critical}$ from the proposed accurate model is plotted versus the transient simulation results for different values of τ, V$_{DD}$, W$_{p1}$, C$_{A}$, and T.

is estimated, accurately, without time consuming transient simulations. Since Q$_{critical}$ exhibits an exponential relationship with the SER, Q$_{critical}$ should be designed high enough by proper circuit design to limit the SER.

-iii- Increasing the coupling capacitor, C$_{c}$, results in increasing Q$_{critical}$ and $\sigma_{Q_{critical}}$. However, $\sigma_{Q_{critical}}/\mu_{Q_{critical}}$ is reduced by increasing C$_{c}$. Since this coupling capacitor is one of the most common techniques to increase $Q_{critical}$ and mitigate soft errors, it should be designed carefully to achieve an acceptable $\sigma_{Q_{critical}}/\mu_{Q_{critical}}$ level.

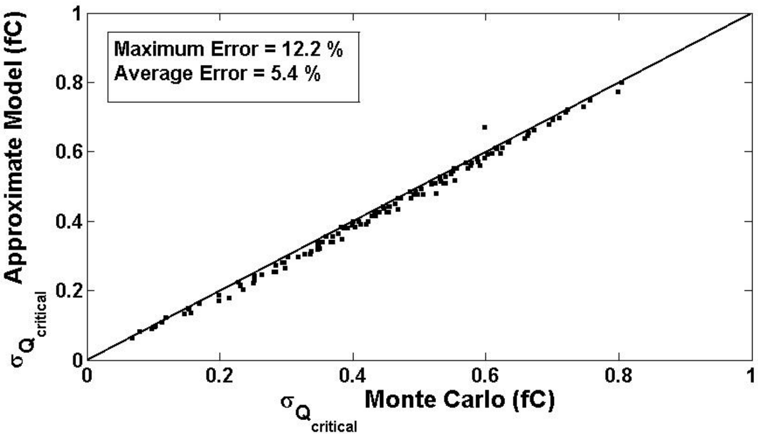


Figure 3.37: $\sigma_{Qcritical}$ from our approximate model plotted versus Monte Carlo simulation results for the same ranges of τ, V$_{DD}$, W$_{p1}$, C$_A$, and T.

-iv- The particle strike current pulse width , τ, affects the critical charge calculations. Wide current pulse models (large values of τ) result in a larger Q$_{critical}$, larger $\sigma_{Q_{critical}}$, and larger $\sigma_{Q_{critical}}/\mu_{Q_{critical}}$.

-v- In sub-threshold SRAM design, $\sigma_{Q_{critical}}$ is dominated by transistor M$_{p1}$ threshold voltage variations ($\sigma_{V_{tp1}}$). The contribution of ($\sigma_{V_{tp1}}$), calculated from the proposed accurate model, is more than 96% in all cases. Thus, the assumption used in deriving the approximate model is justified. This assumption is valid as long as t$_f$ is greater than 3τ, which is usually applicable in sub-threshold SRAM cells.

-vi- W$_{p1}$ is the only sizing parameter that affects the critical charge in this derivation case (for example, it is W$_{p2}$ if V$_B$ is at logic "1"). Increasing W$_{p1}$ results in increasing Q$_{critical}$ and $\sigma_{Q_{critical}}$ whereas increasing W$_{p1}$ results in reducing $\sigma_{Q_{critical}}/\mu_{Q_{critical}}$. Therefore, for a given Q$_{critical}$ variation constraint, the SRAM sizing can be designed to meet this constraint by using the proposed models at the design phase (before fabrication).

-vii- From (3.57), the relative critical charge variation, $\sigma_{Q_{critical}}/\mu_{Q_{critical}}$, exhibits a minimum value at a certain temperature, T. Temperature control techniques can be used to keep the temperature at the value that results in a minimum $\sigma_{Q_{critical}}/\mu_{Q_{critical}}$, and to limit the SER spread.

-viii- Increasing the sub-threshold swing coefficient, n, results in reducing $\sigma_{Q_{critical}}/\mu_{Q_{critical}}$. Consequently, the transistor can be optimized for sub-threshold operation to minimize $\sigma_{Q_{critical}}/\mu_{Q_{critical}}$.

3.3.5 Models Extension to Sub-Threshold Flip-Flops Circuits

Although this section has focused on the critical charge and its variability modeling for the sub-threshold SRAM cell, the proposed models can be extended to model them in sub-threshold flip-flops circuits. This is possible because all the flip-flops topologies consist of an embedded cross-coupled inverters as those in the SRAM cell. Also, the proposed models can be used for the asymmetric sub-threshold SRAM cells or flip-flops, in which $\mathrm{V}_{M1} \neq \mathrm{V}_{M2}$, as long as the models assumptions are satisfied (i.e., V_{M1}, V_{M2}, V_{DD}-V_{M1}, and V_{DD}-$\mathrm{V}_{M2} \geq 3\mathrm{V}_T$ for the accurate model, and $\mathrm{t}_f/\tau \geq 3$ and $\mathrm{V}_{tn1} \geq 3nV_T$ for the approximate model).

3.4 Summary

In this chapter, analytical critical charge variability models accounting for both D2D and WID variations, are proposed. The proposed models deal with the D2D variations, by using corner-based methods. Moreover, they deal with the WID variations, by using statistical techniques. The accuracy of the proposed models is validated by transient and Monte Carlo SPICE simulation results, for an industrial 65nm technology, over a wide range of supply voltages, particle strike induced current pulse widths, and coupling capacitors. The derived statistical models are scalable, bias dependent, and require only the knowledge of easily measurable parameters. Moreover, the models are very efficient, compared to Monte Carlo simulations. This makes them very useful in early design cycles, SRAM design optimization, and technology prediction. Finally, the proposed models can be extended for the flip-flops critical charge variability as well.

In the super-threshold SRAM models, it is shown that, the use of the coupling capacitor in the SRAM cell, as a soft error mitigation technique, is limited by the relative variations. The proposed models provide an analytical equation, to calculate the value of the coupling capacitor, that results in minimum relative variations. Finally, the proposed models show that, the PMOS transistors in the SRAM cell, are dominating the variations, and hence, the PMOS transistors must be designed, while taking the critical charge variations into account.

In the sub-threshold SRAM models, it is found that the relative critical charge variability exhibits a minimum at a certain temperature value. This result can be used by circuit designers to keep the temperature at this value, by using temperature control techniques, to minimize the relative critical charge variability. Moreover, the proposed models show that the transistor sub-threshold swing coefficient can be optimized to minimize the critical charge variability. These results are particulary relevant for applications with strict SER constraints.

Chapter 4

Comparative Analysis of Yield Improved Flip-Flops Using Statistical Gate Sizing

In this chapter, statistical gate sizing technique is used for improving the timing yield of the super-threshold flip-flops, used in high performance systems, and improving the power yield of the sub-threshold flip-flops, used in low power systems. Following that, a comparative analysis between different flip-flops topologies is introduced for both cases considering the yield improvement corresponding overheads. This comparative analysis will help flip-flops designers to select the best flip-flop topology that satisfies their design power, timing, and robustness requirements. In addition, the effect of the yield improvement on the soft errors immunity is discussed in this chapter.

This chapter is organized as follows. Comparative analysis of timing yield improved super-threshold flip-flops is displayed in Section 4.1. Section 4.2 presents a comparative analysis of power yield improved sub-threshold flip-flops. Finally, in Section 4.3, some conclusions are drawn.

4.1 Timing Yield Improvement of Super-Threshold Flip-Flops

As discussed in Chapter 2, the demand for higher performance has moved the clock frequencies up to multi-GHz in microprocessors and high performance VLSI applications. These increased clock frequencies lead to aggressive pipelining which means that hundreds of thousands of flip-flops are utilized to control the data flow under strict timing constraints.

A violation of the timing constraints at a flip-flop can result in latching incorrect data causing the overall system to malfunction [10]. In addition, deterministic gate sizing tools size the flip-flops circuits to optimize the Power-Delay-Product (PDP), as shown in Figure 2.28. However, due to process variations, a large number of circuits might not meet the target delay due to random process variations as shown in Figure 2.32. Therefore, the flip-flops have to be designed using statistical gate sizing tools to improve the timing yield [34, 35].

In this section, a comparative analysis of the process variations impact on flip-flops soft errors vulnerability for different flip-flops topologies is introduced as well. First, these flip-flops are sized using statistical gate sizing algorithm to achieve a timing yield of 99.87%. Following that, these flip-flops are fairly compared in terms of the required power and PDP overheads to achieve this timing yield improvement. Then, the impact of the process variations on the soft errors vulnerability of these flip-flops topologies is investigated.

4.1.1 Flip-Flops Selection

In this comparison, four different flip-flops have been selected representing different trade-off choices between performance and power dissipation. Figures 2.29 and 4.1 show the Transmission Gate Master-Slave Flip-Flop (TG-MSFF), which is used in IBM PowerPC 603 processor [133], and the Modified Clocked CMOS Master-Slave Flip-Flop (M-C^{2}MOS-MSFF), respectively. Both of them are implemented by cascading two complementary latches. This master-slave implementation results in robust flip-flop with good hold time behavior. Moreover, they are used in standard libraries [96] which makes it so important to include them in this comparison.

Figure 4.2 shows one of the fastest flip-flops which is called Semi-Dynamic Flip-Flop (SD-FF) [94]. This flip-flop can be considered as a pulsed latch since it samples the input data to the flip-flop output during a very short transparency period around the clock sampling edge. Accordingly, the input data may arrive after the clock edge which results in negative setup time. Therefore, this flip-flop is used in high performance VLSI applications due to its relatively short data-to-output delay (T$_{D-Q}$) at the expense of poor hold time behavior and excess power consumption. Figure 2.31 shows a Sense-Amplifier based Flip-Flop (SA-FF) with a NAND SR-latch [91], which is adopted in high performance WD21264 Alpha processors [134]. This flip-flop can be viewed as a compromise between the master-slave robustness and the pulsed latches high performance. These four selected flip-flops are selected as representatives of the flip-flops topologies. However, the timing yield improvement algorithm, presented in this section, can be adopted to any flip-flop circuit in the literature and is independent of the flip-flop circuit itself.

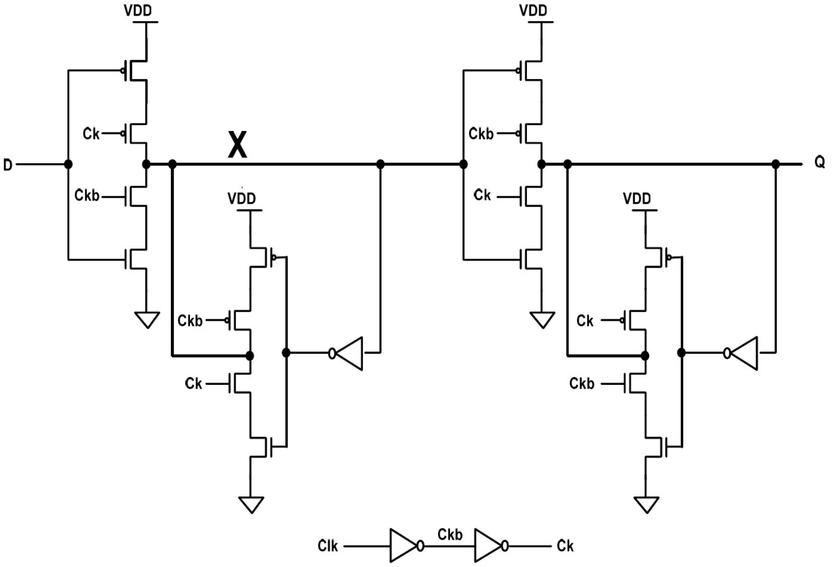


Figure 4.1: Modified Clocked CMOS Master-Slave Flip-Flop (M-C^{2}MOS-MSFF) [91]

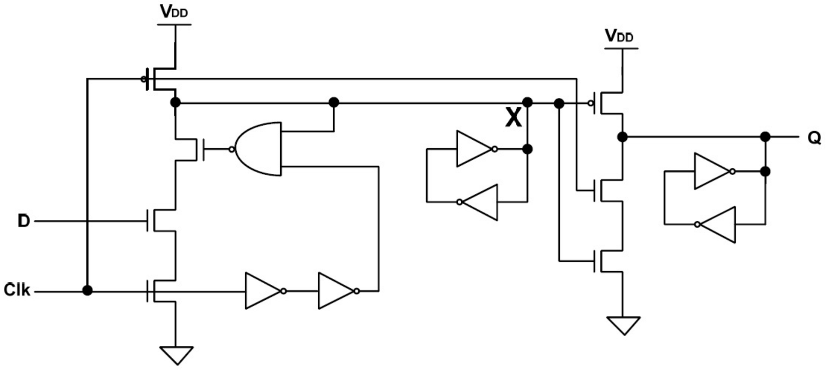


Figure 4.2: Pulsed Semi-Dynamic Flip-Flop (SD-FF) [94]

4.1.2 Simulation Setup

1. **Optimum PDP Sizing**

   All flip-flops are optimized for minimum PDP using an industrial 65nm CMOS technology transistor model, a 1V power supply, a typical process corner, a clock frequency of 1 GHz, and pseudorandom input data with 50% data activity. The measured PDP is obtained by multiplying the data-to-output delay (T_{D-Q}) and the total power which includes both internal power dissipation and local clock/data power dissipation [135]. The optimum setup time for each flip-flop is determined to achieve minimum PDP. The optimization process is conducted by using the CFSQP (C Version Feasible Sequential Quadratic Programming) optimization technique, implemented in Spectre-RF and explained in details in Appendix B. This algorithm is based on the Finite Difference Perturbation (FDP) method to determine how sensitive the PDP is to each device size, and hence, this algorithm is considered one of the sensitivity based gate sizing algorithms. Then, it provides the optimal sizing and setup time that achieve the minimum PDP.

2. **Impact of Process Variations on Flip-Flop Delay**

   Monte Carlo analysis, using the statistical transistor model reported in Appendix D, is conducted. The number of Monte Carlo analysis points used is 5,000 points to provide good accuracy. The delay, power, and PDP variability are then obtained.

3. **Functional Yield Improvement Using Setup Time Margin**

   The optimum setup time determined above for optimum PDP is obtained by using a typical process corner to minimize the PDP. This results in a poor functional yield, since the setup time constraint of some of the flip-flop simulated Monte Carlo points is violated. Typically, the functional yield of the flip-flops using this setup time ranges from 85% to 95%. A setup time margin is added to achieve a functional yield greater than 99.9% [96]. This setup time margin is determined by sweeping the setup time and calculating the functional yield and the mean delay. The setup time that achieves functional yield greater than 99.9% and minimum mean delay is selected.

4. **Timing Yield Improvement Using Gate Sizing**

   The delay variability is obtained from the Monte Carlo simulations by adopting the modified setup time. The timing yield of all the flip-flops at the target delay (assumed to be the optimal delay achieved at minimum PDP) is less than 50%. A simplified gate sizing algorithm is employed by using the CFSQP optimization technique, implemented in Spectre-RF and explained in details in Appendix B. Figure 4.3 represents the gate sizing algorithm flow diagram. It starts with a given delay constraint (A_o) and timing yield constraint (Y_o), where A_o is the optimal delay obtained at minimum PDP. Then, the gate sizing values obtained for the minimum PDP are used as an initial gate sizing values. Monte Carlo statistical analysis is then applied to obtain the delay variability. The standard deviation (σ) of the obtained delay distribution is calculated. Following that, the new delay constraint (A_o') is calculated by using the following equation:

$$A_o' = A_o - n * \sigma \tag{4.1}$$

where n is dependent on the target timing yield value (Y_o) and can be obtained from the normal distribution tables. For example, in this work, a timing yield of 99.87% (Y_o = 99.87%) is required which means that "n" must equal 3.0 from the normal distribution tables. Following the calculation of (A_o'), an optimization problem is solved by employing CFSQP to determine the new gate sizing that matches the delay (A_o') and minimizes the total power consumption such that the power overhead does not exceed a certain percentage of the original flip-flop power at optimal PDP (this percentage is chosen to be 60% in this work). These steps are repeated, until the timing yield and power overhead constraints are met.

It should be emphasized that the delay pdf changes after each iteration because the variations in the threshold voltage are a strong function of the transistor width [15, 34]. Figure 4.4 illustrates how this gate sizing algorithm improves the timing yield by moving the delay pdf to a shorter mean delay.

5. **Power and PDP Overheads**

   The objective of this step is to conduct Monte Carlo simulations again on the timing yield improved flip-flops to obtain the delay, power, and PDP variability as well as the power and PDP overheads required to achieve this timing yield improvement.

6. **Soft Error Modeling**

   The SER in circuits can be estimated by using (2.4). Therefore, $Q_{critical}$ can be used as a measure of the SER for different flip-flops topologies. In this work, the particle strike is modeled in Spectre-RF simulation program as a double exponential current source connected to the flip-flop circuit nodes, as expressed in (2.3). The parameters τ_r and τ_f equals 10ps and 200ps respectively. Q is varied iteratively to achieve the minimum amount of charge resulting in a bit flip at the output node. Thus, $Q_{critical}$ is calculated by:

$$Q_{critical} = [\int_0^{t_f} i_{injected}(t)dt]|_{minimum} \tag{4.2}$$

   where t_f refers to the flipping time of the output node and $i_{injected}$(t) is the current pulse model given in (2.3). The critical charge is calculated at all nodes of each flip-flop for the 1-to-0 flip and the 0-to-1 flip at the output node. Then, the node that has the smallest critical charge is chosen as the most susceptible node to soft errors. Following that, the same Monte Carlo setup is conducted, and the critical charge

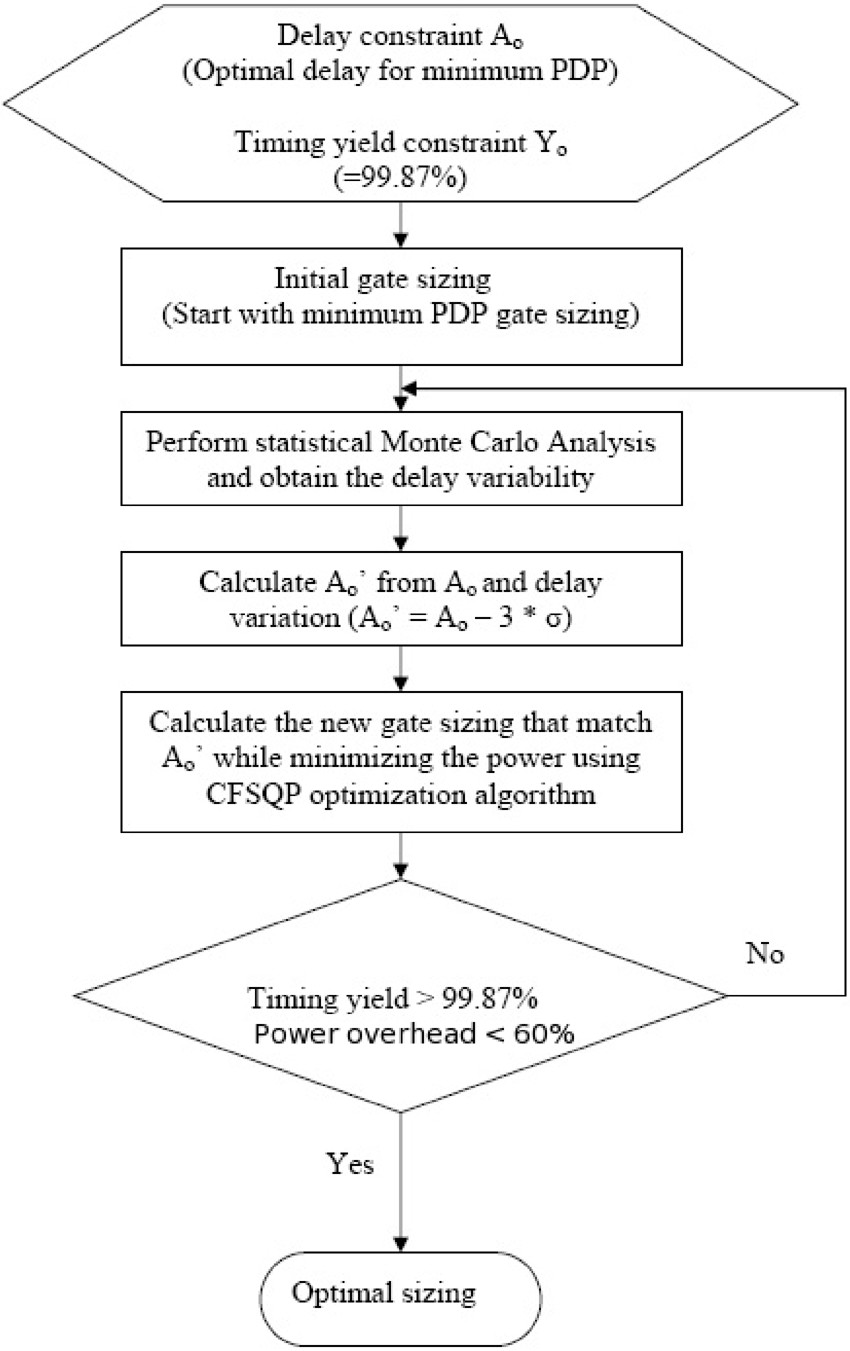

Delay constraint A$_o$
(Optimal delay for minimum PDP)
Timing yield constraint Y$_o$
(=99.87%)
Initial gate sizing
(Start with minimum PDP gate sizing)
Perform statistical Monte Carlo Analysis and obtain the delay variability
Calculate A$_o$' from A$_o$ and delay variation (A$_o$' = A$_o$ – 3 * σ)
Calculate the new gate sizing that match A$_o$' while minimizing the power using CFSQP optimization algorithm
Timing yield > 99.87%
Power overhead < 60%
No
Yes
Optimal sizing

Figure 4.3: Gate sizing algorithm flow diagram

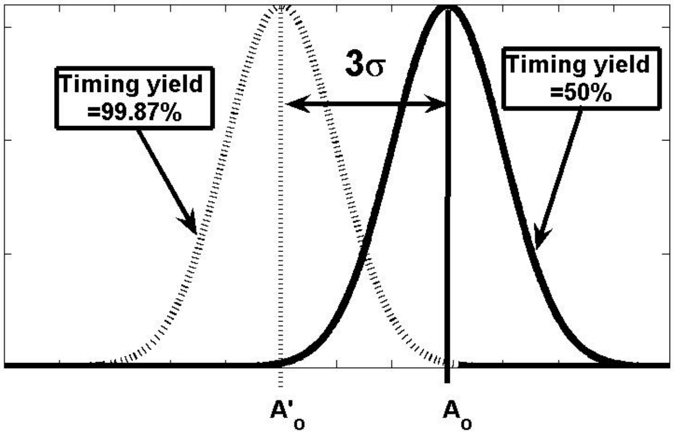


Figure 4.4: Timing yield improvement under process variations using statistical gate sizing

distribution is obtained. All these simulations are performed for the selected flip-flops when sized for minimum PDP as well as when sized for timing yield improvement.

4.1.3 Simulation Results and Discussions

1. **Timing yield Improvement and Required Power and PDP Overheads**

   Table 4.1 summarizes the simulation results for all the flip-flops. The comparison is performed for the improved timing yield flip-flops. The optimal T_{D-Q} delay is adopted as the target delay constraint for timing yield improvement for each flip-flop. According to Table 4.1, the SD-FF has 2.4X higher performance compared to the M-C^2MOS-MSFF at the expense of 1.4X higher power dissipation.

   Figure 4.5 shows the relative power and the relative PDP overheads of the improved timing yield flip-flops. According to the results in Figure 4.5, the SA-FF has a power overhead of 58.2% which is 1.7X higher than that of the TG-MSFF. Moreover, the SA-FF suffers 25.26% PDP overhead which is 2.8X higher than that of the TG-MSFF. The reason for this is that the SA-FF implementation utilizes a symmetric cross-coupled architecture which suffers from devices mismatch more than all other flip-flops.

   The M-C^2MOS-MSFF delay standard deviation increases from one iteration to the next. Consequently, this flip-flop requires several gate sizing algorithm iterations which increases the required power overhead to reduce the mean delay. The SD-FF has the same PDP overhead as the M-C^2MOS-MSFF while having 1.2X less power

Table 4.1: Simulation results for timing yield improved flip-flops topologies

		TG-MSFF	M-C^{2}MOS-MSFF	SD-FF	SA-FF
T_{D-Q} delay	Optimal (ps)	48.6	76.4	32.6	49.6
	Mean (ps)	40.7	57.7	26.1	39.4
	σ (%)	6.7	4.6	5.9	6.5
Power	Optimal (μW)	8.9	11.5	16.3	12.7
	Mean (μW)	11.7	17.7	23.7	20.2
	σ (%)	2.0	1.2	1.9	1.6
	Relative overhead (%)	30.9	53.9	44.8	58.2
PDP	Optimal (fJ)	0.43	0.88	0.53	0.63
	Mean (fJ)	0.47	1.02	0.62	0.80
	σ (%)	5.5	4.5	5.6	6.0
	Relative overhead (%)	9.0	15.8	15.7	25.3

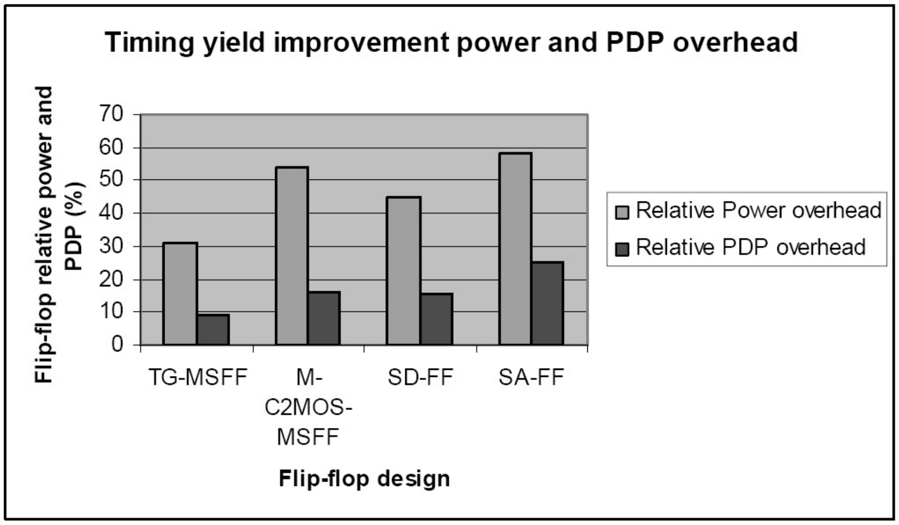


Figure 4.5: The relative power and PDP overheads due to timing yield improvement

overhead. The TG-MSFF exhibits the lowest power and PDP overheads of 30.87% and 9%. This advantage is due to the fact that its delay standard deviation decreased with iterations, and therefore, this flip-flop takes the lowest number of gate sizing algorithm iterations.

2. **Nominal Critical Charge**

Table 4.2 summarizes the nominal critical charge values and the corresponding nodes for the 1-to-0 and the 0-to-1 flips for the selected flip-flops. It should be noted that the SA-FF critical charge has to be determined for both nodes S and R because node R is more susceptible to soft errors in the case of a 1-to-0 flip while node S is more in the case of a 0-to-1 flip. The values of the nominal critical charge are obtained for two different flip-flops sizing scenarios. The first scenario is for the optimum PDP sized flip-flops whereas the second scenario is for the timing yield improved flip-flops.

According to the results in Table 4.2, the least vulnerable flip-flop to soft errors is SD-FF. It has the largest critical charge for both the 1-to-0 and the 0-to-1 flips in the two sizing scenarios. This advantage is due to its cross-coupled inverters connected at node X which fight to keep this node at its logic state. The SA-FF exhibits the smallest $Q_{critical}$ due to the SR latch since any error occurs at S or R results in flipping the output node immediately. Hence, this flip-flop has very small flipping time (t_f).

The master-slave flip-flops exhibit roughly the same critical charge nominal value which lies half-way between the SD-FF and the SA-FF critical charge values. The master-slave flip-flops exhibit a long flipping time, since, the error at node X in both master-slave flip-flops takes longer time to propagate to the output node. In addition, node X is in the hold mode, when it is connected to the back to back inverters, which reduces its susceptibility to soft errors. Figures 4.6 and 4.7 show how the timing yield improvement increases the soft errors immunity of all the flip-flops. This can be simply justified, since the timing yield improvement increases the aspect ratio of the devices, and hence, increases their nodal capacitances. Correspondingly, the critical charge value is increased.

For the 1-to-0 flip case, the critical charge increased due to timing yield improvement by a factor ranging from 1.4X to 5.8X. On the other hand, for the 0-to-1 flip case, this factor ranges from 1.4X to 1.6X. It should be noted that although, the largest power and PDP overheads to achieve this timing yield improvement occurs in the SA-FF, its critical charge increasing factor is still small.

3. **Critical Charge Distribution**

   The critical charge distributions for the selected flip-flops are tabulated in Table 4.3. It is shown that the TG-MSFF exhibits small critical charge variations for both the 1-to-0 and the 0-to-1 flips for both sizing scenarios. The reason for these small variations in TG-MSFF is that the soft error, occurs at node X, takes longer path to affect the output node (two inverters and a transmission gate). This long path exhibits averaging effect which results in random variations cancelation.

   The SA-FF suffers from higher critical charge variations. There are two main reasons for these higher variations. The first reason is due to the differential architecture used in the SA-FF which suffers from the transistor mismatch variations (WID variations).

Table 4.2: The nominal critical charge values for the selected flip-flops topologies

Flip-flop type		TG-MS-FF	M-C^2MOS-MSFF	SD-FF	SA-FF	
Most susceptible node		X	X	X	S	R
$Q_{critical}$ (fC) (1-to-0 flip)	Minimum PDP	2.91	3.51	6.70	0.72	0.22
	Improved timing yield	3.99	6.9	38.95	1.94	0.31
$Q_{critical}$ (fC) (0-to-1 flip)	Minimum PDP	3.94	3.15	4.35	0.71	1.85
	Improved timing yield	5.75	5.17	5.93	1.11	2.1

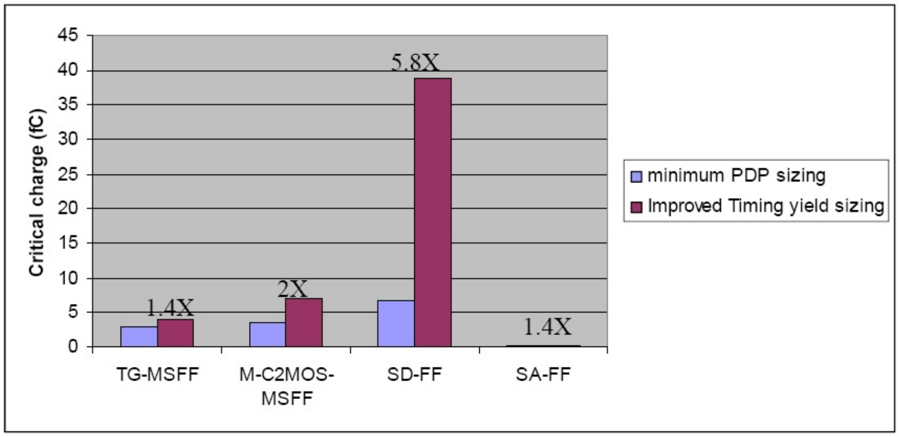


Figure 4.6: The critical charge increase due to timing yield improvement for the 1-to-0 flip.

The second reason is the short path from nodes S or R to the output node which exhibits small averaging.

It is obvious that the timing yield improvement increases the critical charge mean while reducing the critical charge variance. There are only two cases (shown in bold font in Table 4.3) in which the timing yield improvement is no longer capable of reducing the critical charge variations. These two cases related to the SD-FF which actually has the largest critical charge mean in the two sizing scenarios adopted.

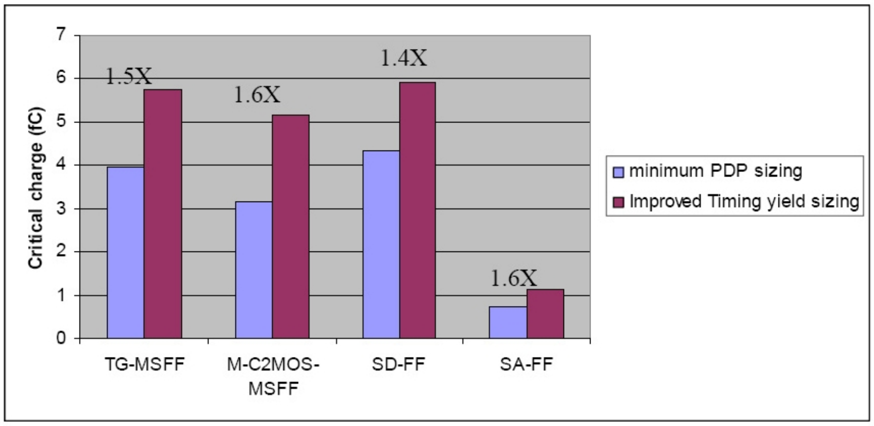


Figure 4.7: The critical charge increase due to timing yield improvement for the 0-to-1 flip.

Table 4.3: The critical charge mean and standard deviation for the two sizing cases

Flip-flop type			TG-MS-FF	M-C^2MOS-MSFF	SD-FF	SA-FF	
						S	R
$Q_{critical}$ (1-to-0 flip)	Minimum PDP	Mean (fC)	2.93	3.5	5.51	0.65	0.23
		σ(%)	14.6	38.4	**24.2**	36.9	24.8
	Improved Timing Yield	Mean (fC)	3.95	6.66	38.87	1.76	0.31
		σ(%)	10.9	18.6	**27.4**	25.3	13.9
$Q_{critical}$ (1-to-0 flip)	Minimum PDP	Mean (fC)	3.91	3.11	4.36	0.69	1.36
		σ(%)	6.6	16.9	**5.5**	9.6	24
	Improved Timing Yield	Mean (fC)	5.7	5.14	5.9	1.12	1.63
		σ(%)	5.4	8.7	**5.5**	7.4	15.2

Moreover, the critical charge variation increases in the first case by a factor of 1.1X while kept constant in the second case.

4.1.4 Design Insights

The discussion above shows that the timing yield improvement using gate sizing increases the soft errors immunity significantly at the expense of power overhead. The best flip-flop choice is the SD-FF for high performance applications at the expense of large power consumption. This flip-flop has the advantage of high soft errors immunity which is increased more with timing yield improvement sizing. The large power consumption of this flip-flop may exceed the allowed power budget. Therefore, the second candidate flip-flop is the M-C^2MOS-MSFF which yields a high soft errors immunity under timing yield improvement and lower power consumption as well at the expense of the increased delay.

If the M-C^2MOS-MSFF delay violates the timing yield constraints, the next recommended flip-flop is the TG-MSFF which has good power and performance metrics at the expense of less soft errors immunity although this soft errors immunity is improved significantly under timing yield improvement sizing. A soft error mitigation technique is necessary when selecting the SA-FF, such as those mentioned in Chapter 2, due to its poor soft errors immunity.

4.2 Power Yield Improvement of Sub-Threshold Flip-Flops

In modern digital synchronous systems, the flip-flops power dissipation represents a considerable fraction of the total power dissipation. Voltage supply scaling is one of the most promising power reduction techniques for flip-flops circuits [91, 136]. When the supply voltage, V_{DD}, is decreased below the transistor threshold voltage, V_t, the transistor is operating in the sub-threshold region [91]. Sub-threshold flip-flops are considered the most energy efficient solution for low power applications in which, performance is of secondary importance [136, 137]. Deterministic gate sizing tools size the sub-threshold flip-flops circuits to optimize the PDP. However, due to random process variations, a large number of flip-flops circuits might not meet the allowed power budget. Consider as an intuitive example, a flip-flop that is designed for optimum PDP, which exhibits a specific target power dissipation. Due to random process variations, the power dissipation, which is dominated by the sub-threshold leakage power that has an exponential relationship with V_t, is modeled by a log-normal distribution with the pdf shown in Figure 4.8. Here, 42% of the total number of flip-flops do not meet the desired target power constraint. Therefore, the flip-flops must be designed by using statistical sizing tools to improve the power yield.

Moreover, the utilization of statistical sizing tools for power yield improvement is more appropriate for an efficient and fair comparison of sub-threshold flip-flops, since the power yield is the main concern in low power applications (similar to the timing yield improvement in the high performance circuits in Section 4.1). This section provides a comparative analysis of power yield improvement under process variations of the same four flip-flops circuits utilized in Section 4.1, especially for the power variability and the required timing and PDP overheads for power yield improvement.

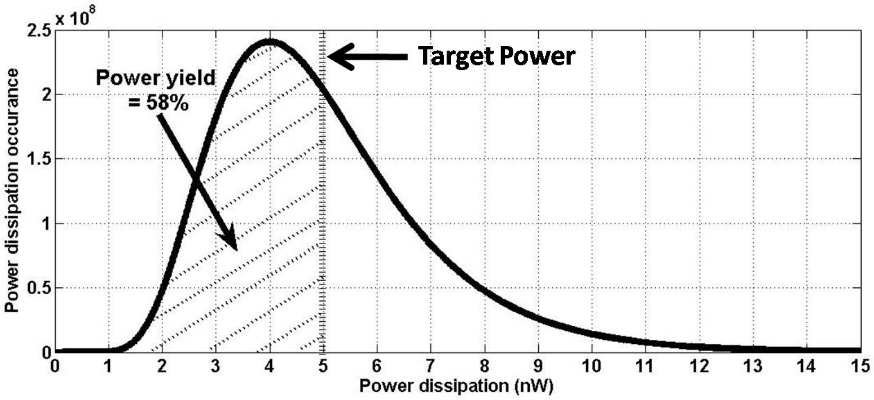


Figure 4.8: The power pdf due to process variations under deterministic gate sizing algorithms.

4.2.1 Flip-Flops Selection

The same four flip-flops, utilized in Section 4.1, are used to represent the various trade-off choices between performance and power dissipation. The SD-FF flip-flop circuit is modified from that in Figure 4.2 by adding two additional inverters into the delayed clock signal as shown in the dotted rectangle in Figure 4.9. This modification is to allow enough transparency period length for the sub-threshold flip-flop sampling.

4.2.2 Simulation Setup

The first three steps in the simulation setup for power yield improvement (i.e., optimal PDP sizing, impact of process variations on the flip-flop power, and functional yield improvement using setup time margin) are similar to those explained in the timing yield improvement in Section 4.1 except for the clock frequency which is reduced to 1MHz instead of 1GHz and the supply voltage, V_{DD}, is swept from 0.15V to 0.3V. Then, the following step is performed.

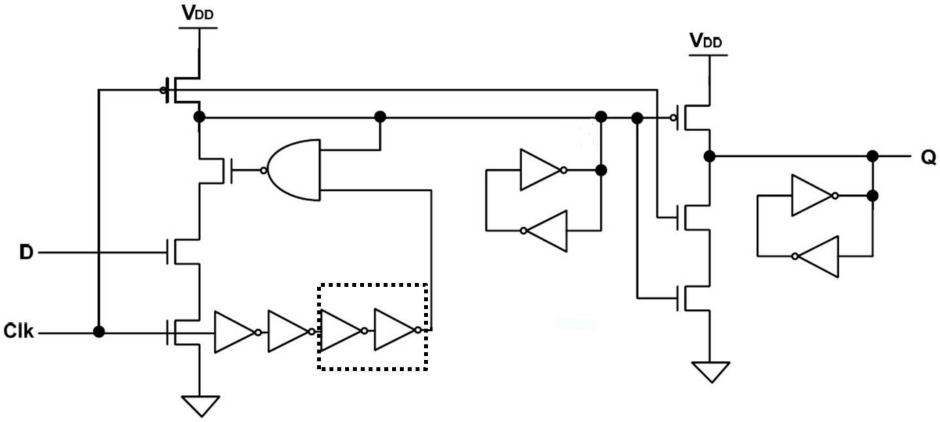


Figure 4.9: Pulsed Semi-Dynamic Flip-Flop (SD-FF) after the addition of two inverters into the delayed clock signal (dotted rectangle). This modification is to allow enough transparency period length for the sub-threshold SD-FF flip-flop sampling period.

Power Yield Improvement using Gate Sizing

The power variability is obtained from the Monte Carlo simulations by adopting the modified setup time. The same algorithm adopted in Section 4.1 is applied again starting with a given power constraint (P$_o$) and power yield constraint (Y$_o$), where (P$_o$) is the optimal power obtained at minimum PDP. Then, the gate sizing values obtained for the minimum PDP are used as an initial gate sizing values. Monte Carlo statistical analysis is then applied to obtain the power variability. The standard deviation (σ) of the obtained power distribution is calculated. Using the power log-normal distribution mean, P$_o$, and standard deviation, σ, the equivalent power's natural logarithm (lnP$_o$) normal distribution mean and variance, μ_{ln} and σ_{ln}, respectively, are given by [138]:

$$\mu_{ln} = \ln(\frac{P_o}{\sqrt{1+\frac{\sigma^2}{P_o^2}}}) \quad \text{and} \quad \sigma_{ln} = \sqrt{\ln(1+\frac{\sigma^2}{P_o^2})} \tag{4.3}$$

Following that, the geometric mean and standard deviation of the log-normal distribution, μ_g and σ_g are calculated as follows [138]:

$$\mu_g = \exp(\mu_{ln}) \quad \text{and} \quad \sigma_g = \exp(\sigma_{ln}) \tag{4.4}$$

In order to ensure that the power dissipation log-normal distribution integral from 0 to the desired power constraint P$_o$ equals the desired power yield Y$_o$, the power distribution pdf has to be shifted from P_o to $P_o{}'$ by using statistical gate sizing where $P_o{}'$ is given by [138]:

$$P_o^{'} = \frac{\mu_g}{(\sigma_g)^n} \tag{4.5}$$

where n is dependent on the target power yield value (Y_o) and can be obtained from the normal distribution tables (i.e., n = 3.0 for Y_o = 99.87%). Following the calculation of ($P_o^{'}$), an optimization problem is solved by employing CFSQP to determine the new gate sizing that matches the power ($P_o^{'}$) and minimizes the delay and PDP overheads. These steps are repeated, until the power yield constraint is achieved.

Figure 4.10 illustrates how this gate sizing algorithm improves the power yield by shifting the power pdf to a shorter mean power. Finally, the associated delay and PDP overheads with the power yield improvement sizing scenario are calculated.

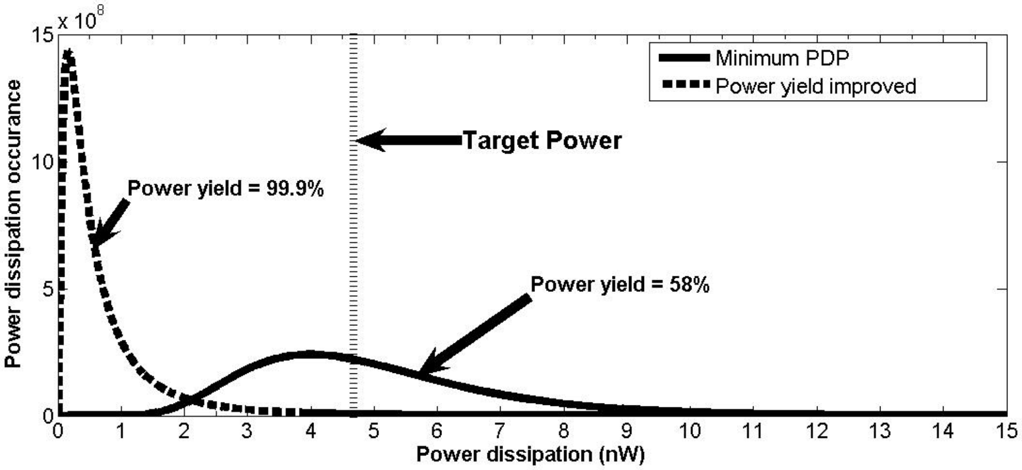


Figure 4.10: The power yield improvement under process variations employing gate sizing. The dotted pdf represents the power pdf of the power yield improved sub-threshold flip-flops (power yield = 99.87%) while the solid pdf represents the power pdf of the minimum PDP sub-threshold flip-flops (power yield = 58%)

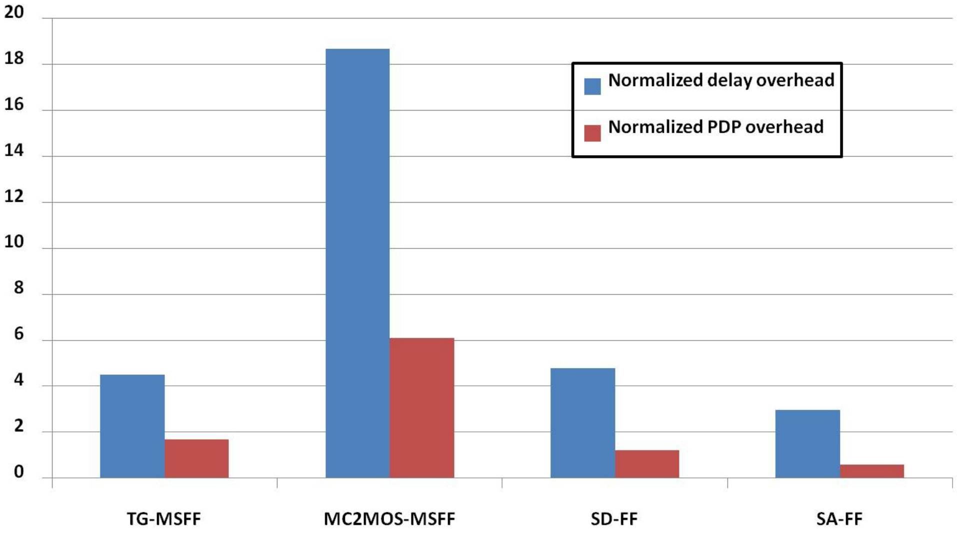
4.2.3 Simulation Results and Discussion

Figure 4.11 shows the delay and energy overheads for all flip-flops when V_{DD} = 0.15V. According to this figure, the power yield improved SA-FF exhibits the lowest delay and energy overheads among all other flip-flops and following it, comes the TG-MSFF. However, the absolute value of these overheads are lower in the TG-MSFF. For example, the delay overhead of the SA-FF (when V_{DD} = 0.15) is 3X while that of the TG-MSFF is 4.5X, however, the absolute delay of the SA-FF is 438 nsec while that of the TG-MSFF is 184 nsec. The M-C^{2}MOS-MSFF exhibits the largest overheads in all parameters. The delay and energy overheads of the M-C^{2}MOS-MSFF flip-flops are higher than that of the SA-FF by factors of 6.3X and 10.2X, respectively, when V$_{DD}$ = 0.15V.

Figure 4.12 portrays the optimal values for the power, delay, and PDP for each flip-flop at the minimum PDP point for different values of the supply voltage, V$_{DD}$. It is evident from Figure 4.12 that the TG-MSFF exhibits the lowest power, delay, and PDP among all other flip-flops. The SD-FF has the largest power and PDP. The M-C^{2}MOS-MSFF introduces the largest delay.

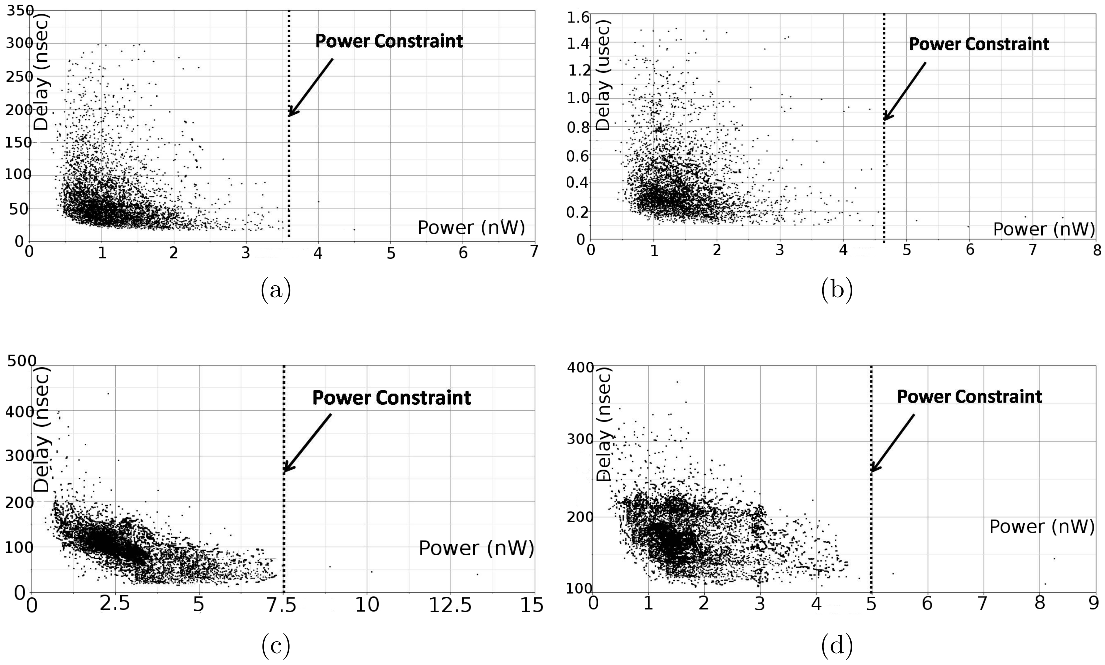
Figure 4.13 shows the simulation results after applying the power yield improvement technique for different values of the supply voltage, V$_{DD}$. Figure 4.13.a shows the new power dissipation mean, $P_o^{'}$ and Figures 4.13.b and 4.13.c show the delay and energy (PDP) values calculated after adopting the power yield improvement. It is clear from Figure 4.13 that the TG-MSFF is still showing the lowest values of power, delay, and PDP even after adopting the power yield improvement technique. The M-C^{2}MOS-MSFF exhibits the largest delay and PDP which means that this flip-flop requires large overheads to achieve the target power yield improvement. Therefore, the M-C^{2}MOS-MSFF is not recommended for sub-threshold operation as it requires large overheads to achieve the target power yield.

Figure 4.14 shows the delay versus the power space for the improved power yield flip-flops when V$_{DD}$ = 0.2V. It is evident that all flip-flops samples achieve a power yield larger than 99.9%. It should be highlighted that the tail of the log-normal distribution, utilized in the power yield improvement, is assumed to be within the 3σ design space. Practically, the tail behavior is very difficult to be estimated which makes the power yield improvement process more complicated. Also, the unexpected tail behavior might increase the design margin up to 5σ or larger to achieve safe and reliable flip-flops samples operation.



(a)

Figure 4.11: The power yield improvement associated normalized overheads for V$_{DD}$ = 0.15V. These overheads are normalized to their nominal values.



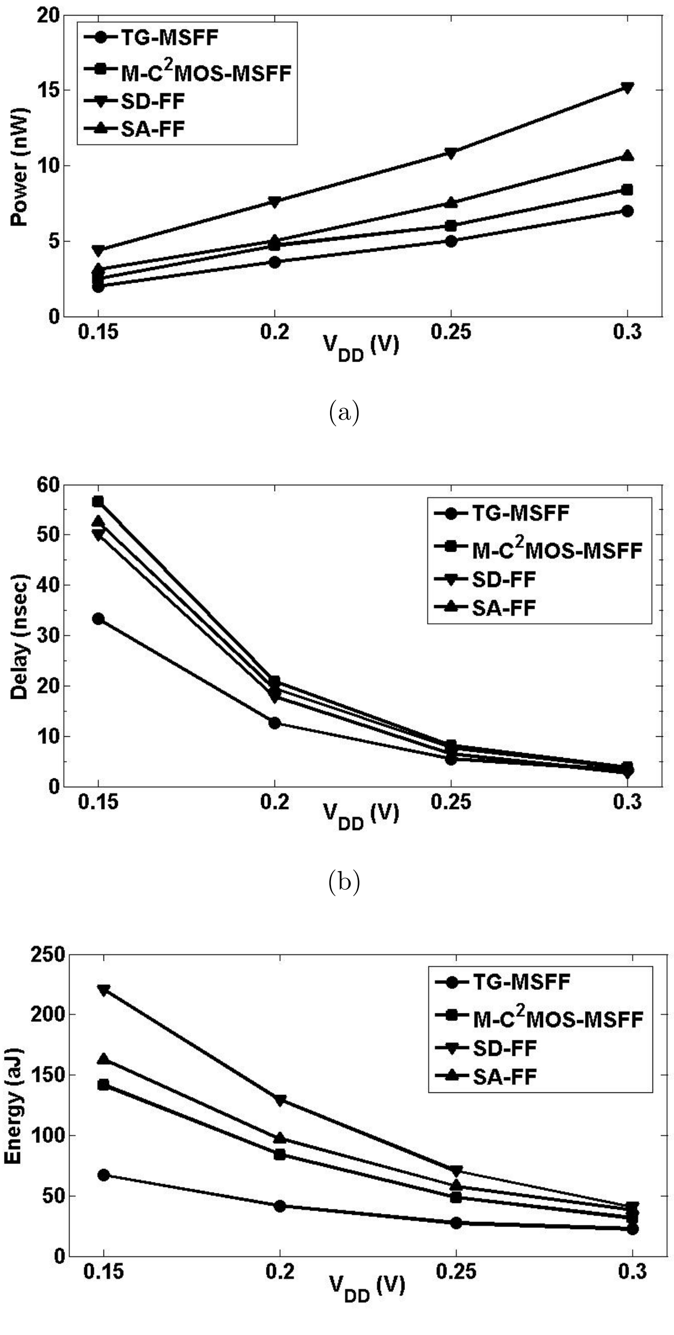
(a)

(b)

(c)

(d)

Figure 4.14: The delay-power scattered plot for (a) TG-MSFF, (b) M-C^{2}MOS-MSFF, (c) SD-FF, and (d) SA-FF

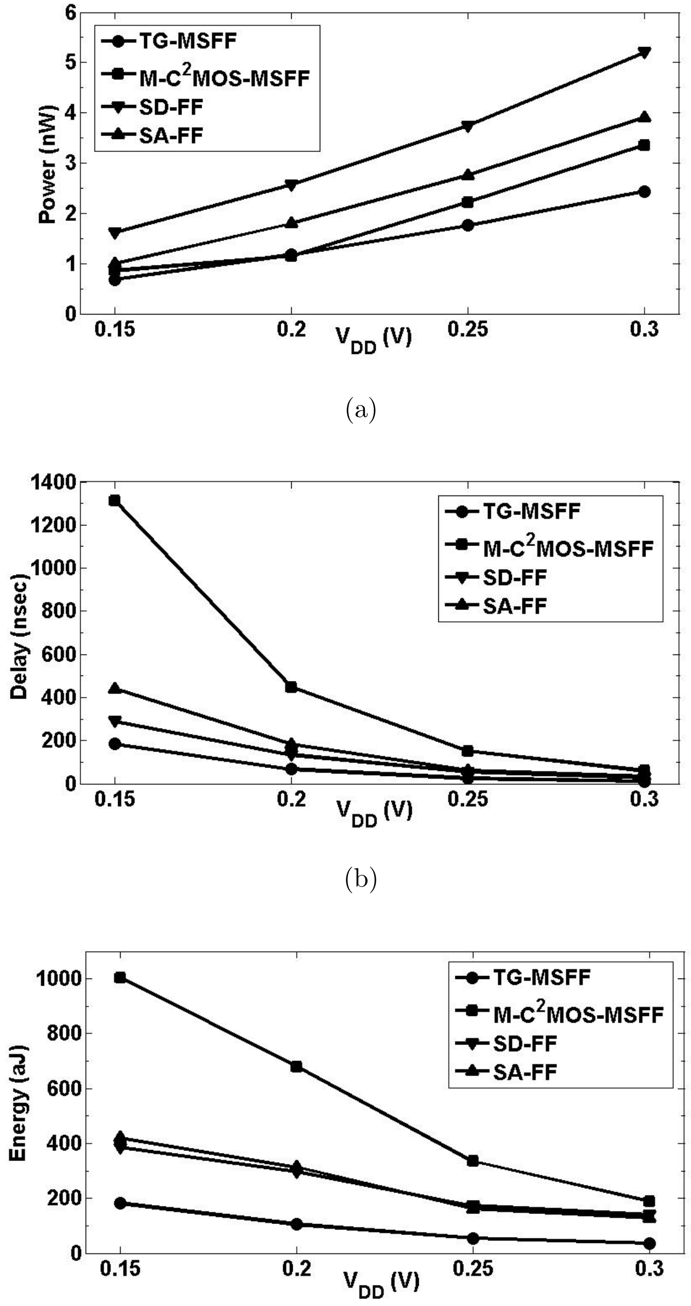


(a)

(b)

(c)

Figure 4.12: Minimum PDP simulation results for the four selected flip-flops (a) Optimal power (P_o), (b) Optimal delay, and (c) Optimal energy (PDP).



(a)

(b)

(c)

Figure 4.13: Power yield improved simulation results for the four selected flip-flops (a) Power mean (P_o'), (b) Delay mean, (c) Energy mean

4.3 Summary

In this chapter, a comparative analysis between commonly used flip-flops topologies is conducted, after performing yield improvement by using statistical gate sizing. First, the timing yield improved super-threshold flip-flops are compared for the required power and PDP overheads. Following that, the power yield improved sub-threshold flip-flops are compared for the required timing and PDP overheads.

For the super-threshold flip-flops, the super-threshold SA-FF suffers from device mismatch which results in power overhead of 1.7X higher than that of the super-threshold TG-MSFF, and PDP overhead of 2.8X higher than that of the super-threshold TG-MSFF. Moreover, the impact of this timing yield improvement on the soft errors immunity is investigated. Simulation results show that the timing yield improvement increases the soft errors immunity since, increasing the transistor sizing increases the nodal capacitances. However, The super-threshold SA-FF is the most vulnerable flip-flop to soft errors and its soft errors immunity is very poor even under timing yield improvement. The reason for that is due to its sensitivity to transistor mismatch (WID variations). The least vulnerable flip-flop to soft errors is super-threshold SD-FF with the highest soft errors immunity. This work recommends that the super-threshold SD-FF is the best choice for high soft errors immunity and high performance at the expense of large power. When the power budget is not met, super-threshold master-slave flip-flops are preferred. If the super-threshold SA-FF has to be used, soft error mitigation techniques are required for proper operation since the super-threshold SA-FF has a poor soft errors immunity.

For the sub-threshold flip-flops, the sub-threshold SA-FF exhibits the lowest overheads in delay and energy (PDP), however, the sub-threshold M-C^{2}MOS-MSFF flip-flop has the largest overheads. The sub-threshold M-C^{2}MOS-MSFF delay and energy overheads are higher than that of the sub-threshold SA-FF by factors of 6.3X and 10.2X, respectively, when V_{DD} = 0.15V. These results recommend the utilization of the sub-threshold SA-FF. In addition, the results show that the sub-threshold M-C^{2}MOS-MSFF flip-flop is not recommended to be used in the sub-threshold region for power yield improvement requirements.

In conclusion, this fair comparison between different flip-flops topologies show that the super-threshold SA-FF is not recommended due to its large power and energy overheads to achieve the timing yield improvement and the poor immunity to soft errors. However, the sub-threshold SA-FF is recommended due to its low timing and energy overheads to achieve the power yield improvement. In addition, the super-threshold SD-FF is highly recommended due to its low overheads for timing yield improvement whereas the sub-threshold M-C^{2}MOS-MSFF is not recommended to be used in the sub-threshold region for power yield improvement requirements. Finally, it has been shown that the timing yield improvement increases the soft errors immunity of the flip-flops circuits.

Chapter 5

ABB Circuits for Process Variations and NBTI Compensation

In this chapter, new ABB circuits are introduced for process variations and NBTI aging compensation to increase the circuits robustness and yield. These ABB circuits have the advantages of lower area overhead and resolution free operation, compared to previously published ABB circuits. A circuit block extracted from a real microprocessor critical path and an SRAM column are used as benchmarks for the proposed ABB circuits to prove their efficiency in compensating for process variations and NBTI impacts by using post layout simulation results and test chip measurements. This test chip is fabricated by using TSMC 65nm Triple-well CMOS technology process.

This chapter is organized as follows. In Section 5.1, an introduction about the ABB circuits is introduced. The proposed ABB circuits are presented in Section 5.2. Sections 5.3 and 5.4 prove the efficiency of the proposed ABB circuits, for compensating for process variations and NBTI aging impacts, by applying them to a high performance circuit block and an SRAM column, respectively. Finally, in Section 5.5, some conclusions are drawn.

5.1 Introduction

As discussed in Chapter 2, with continual CMOS technology scaling, power density has become a significant concern in microprocessor design due to the increasing chip density and clock frequencies [6, 46]. Power constraints of a microprocessor, which is dictated by the overall system thermal design, impact the system cost and the maximum operating frequency. Thus, the goal of a microprocessor designer is not only to achieve the maximum operating frequency, but also to satisfy the power constraints. Process variations result in a spread of the microprocessor operating frequencies and the associated leakage power, as portrayed in Figure 2.8. Therefore, some of the fabricated microprocessor chips are discarded because they are either too slow or highly leaky. There is a trade-off between the microprocessor speed and its leakage power consumption, which means that slow circuits are less leaky, and highly leaky circuits are fast. Accordingly, process variations result in a parametric yield loss.

Adaptive Body Bias (ABB) allows the tuning of the transistor V_t, by controlling the transistor body-to-source voltage, V_{BS}. A Forward Body Bias (FBB) reduces V_t, increasing the device speed at the expense of increased leakage power. Alternatively, a Reverse Body Bias (RBB) increases V_t, reducing the leakage power but slowing the device. Therefore, the impact of process variations is mitigated by using the ABB circuit [43, 139], as depicted in Figure 5.1. Practically, the implementation of the ABB is desirable to bias each device in a design independently, to mitigate D2D and WID variations. However, supplying so many separate voltages inside a die results in a large area overhead. On the other hand, using the same body bias for all devices on the same die limits their capability to compensate for WID variations. Thus, the granularity level of the ABB scheme is a trade-off between the target yield and the associated area overhead.

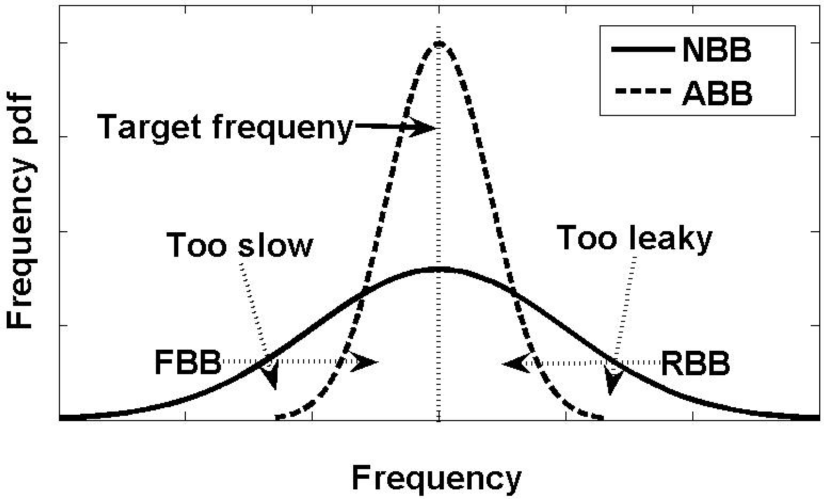


Figure 5.1: The effect of the ABB technique on the frequency distribution. NBB denotes the No Body Bias case

In the literature, researchers have attempted to use ABB to maximize the system clock frequency or minimize the leakage power. FBB is used in [31] for a 1GHz communication router in 150nm CMOS technology to maximize the clock frequency. The objective of the research work in [44, 140, 141] is to design a body bias generator circuit to compensate for process variations. Several optimization algorithms are presented in [45, 142] aiming at finding the optimal body bias voltages to minimize the leakage power. In [139], the optimal granularity level of the ABB scheme is discussed mathematically to achieve near-optimal performance and power characteristics. In [41], ABB is used by estimating the process parameters and using a digital controller to control the body bias. Finally, in [46], ABB is used to compensate for variations by maximizing the die frequency subject to a power constraint. In all the aforementioned research, the ABB circuit area overhead limits its capability to mitigate the WID variations by using the ABB circuit for each circuit block. For example, most of the previously published ABB circuits have a large area overhead because they consist of an Analog-to-Digital Converter (ADC) and/or a Digital-to-Analog Converter (DAC) in conjunction with a digital controller to achieve the required body bias control (i.e., these ABB circuits convert the estimated threshold voltages to digital by using the ADC. Then, the digital controller finds the optimal body bias voltages which are converted back to analog by using the DAC). Therefore, the ABB scheme area overhead should be reduced to allow more WID variations compensation, which is performed by the direct implementation of the body bias generation circuits in the proposed ABB circuits in this chapter (i.e., no ADC or DAC circuits are required). It should be noted that the ABB in [46], shown in Figure 2.9, is the only ABB circuit that has been fabricated and provides measurements results, displayed in Figure 2.10. This is the reason why the proposed ABB circuits, in this chapter, are compared to this ABB in [46].

In this chapter, a Direct ABB (D-ABB) circuit is proposed. It is based on V_t estimation circuits and direct adaptive control of the body bias, achieved by an on-chip direct controller circuit. This direct controller circuit generates the appropriate body bias voltage based on the V_t fluctuations by directly implementing the relationship between V_t and V_{BS}. The goal of the proposed D-ABB is to reduce the process variations impact by considering D2D and WID variations. This, in turn, improves the parametric yield for the clock frequency, dynamic power, and leakage power. This goal is achieved by using a direct controller circuit which exhibits lower area overhead compared to other ABB circuits [41, 46]. Following that, a Linear D-ABB (LD-ABB) circuit is introduced which has less layout area compared to the D-ABB at the expense of less process variations compensation. The D-ABB and LD-ABB circuits are applied to a high performance circuit block case study, extracted from a real microprocessor critical path to compensate for the process variations. Following that, the D-ABB circuit is applied to an SRAM column to reduce the NBTI and the process variations impact. The effectiveness of these ABB circuits is verified by using post layout simulation results and test chip measurements.

5.2 Proposed ABB Circuits

5.2.1 Proposed D-ABB Circuit

In the proposed D-ABB circuit, the effect of process variations on V_t is compensated by estimating the actual values of V_t, which are impacted by process variations, by using estimation circuits placed close to the critical path. Then, the direct controller generates the appropriate body bias voltage, V_{BS}, to mitigate the process variations impact. The direct controller is a direct implementation of the relationship between V_t and V_{BS}. In [22, 139], the relationship between V_t and V_{BS} for an NMOS transistor is given by:

$$V_t = V_{to} + \Delta V_t|_{BB} \qquad \text{and} \qquad \Delta V_t|_{BB} = \gamma(\sqrt{2\phi_F - V_{BS}} - \sqrt{2\phi_F}) \tag{5.1}$$

where V_{to} is the NMOS transistor threshold voltage at zero body bias (i.e., when $V_{BS} = 0$), $\Delta V_t|_{BB}$ is the body bias effect on V_t, γ is the body effect coefficient, and ϕ_F is is the difference between Fermi level and intrinsic level [22]. If V_{to} is increased due to process variations by $\Delta V_t|_{PV}$. Therefore, the body bias voltage, V_{BS}, compensates for this process variations by producing a threshold voltage change, $\Delta V_t|_{BB}$, that cancels out the process variations change, $\Delta V_t|_{PV}$ (i.e., $\Delta V_t|_{BB} =$ – $\Delta V_t|_{PV}$). The value of V_{BS} that compensates for the process variations change is given by:

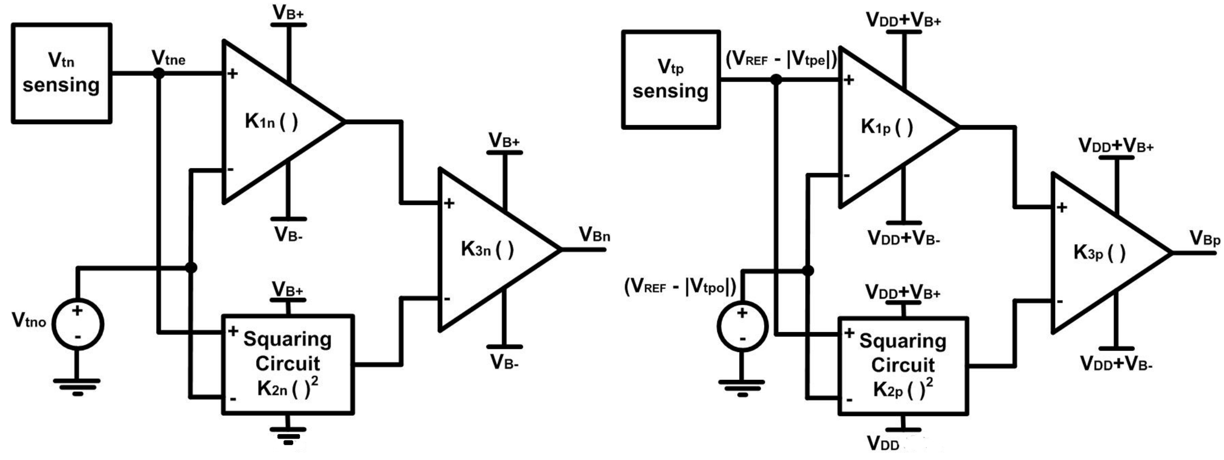
$$V_{BS} = \frac{2\sqrt{2\phi_F}}{\gamma} \times \Delta V_t|_{PV} - \frac{1}{\gamma^2}(\Delta V_t|_{PV})^2 \tag{5.2}$$

where $\Delta V_t|_{PV}$ is the difference between the estimated threshold voltage, V_{te}, which is impacted by the process variations, and the nominal threshold voltage, V_{to}. Similarly, for PMOS transistors, the same relationship in (5.2) is used by replacing V_{BS} by V_{SB}. Typically, the sources of the NMOS transistors are connected to the ground (zero voltage), and the sources of the PMOS transistors are connected to the supply voltage, V_{DD}. Therefore, the body bias voltages of the NMOS transistors, V_{Bn}, and the PMOS transistors, V_{Bp}, which result in process variations compensation, are given by:

$$V_{Bn} = \frac{2\sqrt{2\phi_{F_n}}}{\gamma_n}[V_{tne} - V_{tno}] - \frac{1}{\gamma_n^2}[V_{tne} - V_{tno}]^2 \tag{5.3}$$

$$V_{Bp} = V_{DD} - \frac{2\sqrt{2\phi_{F_p}}}{\gamma_p}[|V_{tpe}| - |V_{tpo}|] + \frac{1}{\gamma_p^2}[|V_{tpe}| - |V_{tpo}|]^2 \tag{5.4}$$

The proposed D-ABB circuit is depicted in Figures 5.2.a and 5.2.b for the bias voltages, V_{Bn} and V_{Bp}, respectively. A set of sensing circuits estimates the actual values of the threshold voltages, which are impacted by the process variations. The sensing circuit for the NMOS transistor, shown in Figure 5.2.a, outputs an estimate for the NMOS threshold voltage, denoted by V_{tne}. In the mean time, the sensing circuit for the PMOS transistor, shown in Figure 5.2.b, outputs an estimate for the PMOS threshold voltage, denoted by $V_{REF} - |V_{tpe}|$, where V_{REF} is a dc reference voltage. The estimated variables (i.e., V_{tne} and $V_{REF} - |V_{tpe}|$) are applied to a set of amplifiers and squaring circuits to produce the required bias voltages, which are capable of reducing the impact of the process variations.



(a) (b)

Figure 5.2: The D-ABB circuit for (a) NMOS transistors body bias control, V_{Bn}, and (b) PMOS transistors body bias control, V_{Bp}.

In Figure 5.2.a, the voltage source, V_{tno}, is a dc bias voltage representing the NMOS transistor nominal threshold voltage value at zero body bias. According to Figure 5.2.a and recalling (5.3), the gains K_{1n}, K_{2n}, and K_{3n} are given by:

$$K_{1n} \times K_{3n} = \frac{2\sqrt{2\phi_{F_n}}}{\gamma_n} \quad \text{and} \quad K_{2n} \times K_{3n} = \frac{1}{\gamma_n^2} \tag{5.5}$$

Accordingly, the amplifiers gains, K_{1n} and K_{3n}, and the squaring circuit gain, K_{2n}, are arbitrarily selected according to (5.5). Similarly, The voltage ($V_{REF} - |V_{tpo}|$), shown in Figure 5.2.b, is a dc bias voltage representing the difference between the reference voltage, V_{REF}, and the PMOS transistor nominal threshold voltage value. According to Figure 5.2.b and recalling (5.4), the gains K_{1p}, K_{2p}, and K_{3p} are given by:

$$K_{1p} \times K_{3p} = \frac{2\sqrt{2\phi_{F_p}}}{\gamma_p} \quad \text{and} \quad K_{2p} \times K_{3p} = -\frac{1}{\gamma_p^2} \tag{5.6}$$

The implementations of the sensing circuits, the amplifiers, and the squaring circuits are given in the following discussions.

1. **Sensing Circuits**

Sensing circuits are used to estimate the actual values of the threshold voltages of the NMOS and PMOS transistors, which are impacted by process variations. Figures 5.3 and 5.4 illustrate the sensing circuit implementations for the NMOS and PMOS transistors, respectively [41].

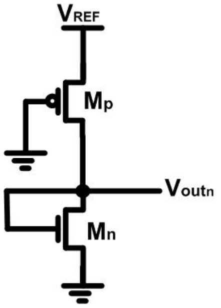


Figure 5.3: V_{tn} sensing circuit [41].

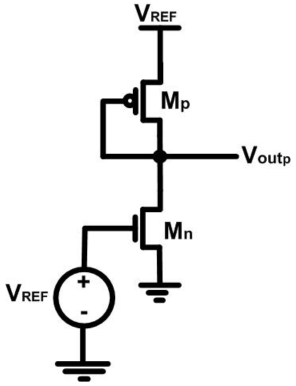


Figure 5.4: $|V_{tp}|$ sensing circuit [41].

In the NMOS threshold voltage sensing circuit, displayed in Figure 5.3, the PMOS transistor is sized with minimum area and acts as a current source. The NMOS transistor is a diode connected transistor, and V_{REF} is a reference voltage. By using the $\alpha-$power law model, introduced in [114], and equating the dc currents of the NMOS and PMOS transistors, the output voltage of this circuit, V_{outn}, is expressed as:

$$V_{outn} = V_{tn} + r_n \times [V_{REF} - |V_{tp}|] \text{ and } r_n = (\frac{k_{p'} \frac{W}{L}|_p}{k_{n'} \frac{W}{L}|_n})^{\frac{1}{\alpha}} \tag{5.7}$$

where V_{tn} and $|V_{tp}|$ are the threshold voltages, $k_{n'}$ and $k_{p'}$ are the technological parameters, and $\frac{W}{L}|_n$ and $\frac{W}{L}|_p$ are the sizes of the NMOS and the PMOS transistors, respectively. By sizing this circuit such that $\frac{W}{L}|_n >> \frac{W}{L}|_p$, (5.7) is rewritten as:

$$V_{outn} \approx V_{tn} \tag{5.8}$$

Therefore, the output voltage of the NMOS threshold voltage sensing circuit, shown in Figure 5.3, represents the actual NMOS transistor threshold voltage, which is impacted by process variations, and denoted by V_{tne}.

Alternatively, the PMOS threshold voltage sensing circuit is depicted in Figure 5.4. The NMOS transistor is sized with minimum area and acts as a current source, whereas the PMOS is a diode connected transistor. Similarly, by sizing this circuit such that $\frac{W}{L}|_p >> \frac{W}{L}|_n$, the output voltage of this circuit, V_{outp}, is given by:

$$V_{outp} \approx V_{REF} - |V_{tp}| \tag{5.9}$$

This output voltage is denoted by $V_{REF} - |V_{tpe}|$ and represents the actual PMOS transistor threshold voltage, which is impacted by process variations.

Figure 5.5 portrays V_{tne}, the output voltage of the NMOS sensing circuit, versus V_{tno} and Figure 5.6 displays the output voltage of the PMOS sensing circuit, $(V_{REF} - |V_{tpe}|)$, versus $(V_{REF} - |V_{tpo}|)$. These figures are obtained from SPICE simulations by sweeping the threshold voltage parameters of the industrial 65nm CMOS technology transistor model and using V_{REF} = 0.5V, r_n=r_p $\approx$ 0.075. Good agreements between the estimated threshold voltages values and their actual values, prove that the threshold voltage sensing circuits are effective, when used in nanometer technologies. The maximum error between the estimated threshold voltage values and their corresponding actual values is 4.5%, and the average error is 2.7%.

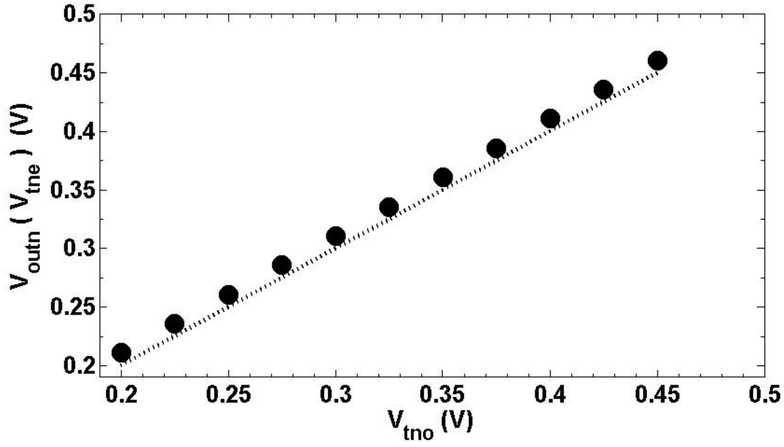


Figure 5.5: The output of the NMOS threshold voltage sensing circuit shown in Figure 5.3.

2. **Amplifier Circuit**

In the proposed D-ABB circuit in Figure 5.2, several amplifiers with various gains are required. Therefore, the differential amplifier shown in Figure 5.7 is utilized. The op-amp implemented in this amplifier is a two stage amplifier [18]. The first stage is configured in a differential pair topology to provide the high gain requirements. Typically, the second stage is configured as a common source stage to allow maximum output voltage swings [18]. The amplifier gain is given by R_F/R_I. According to

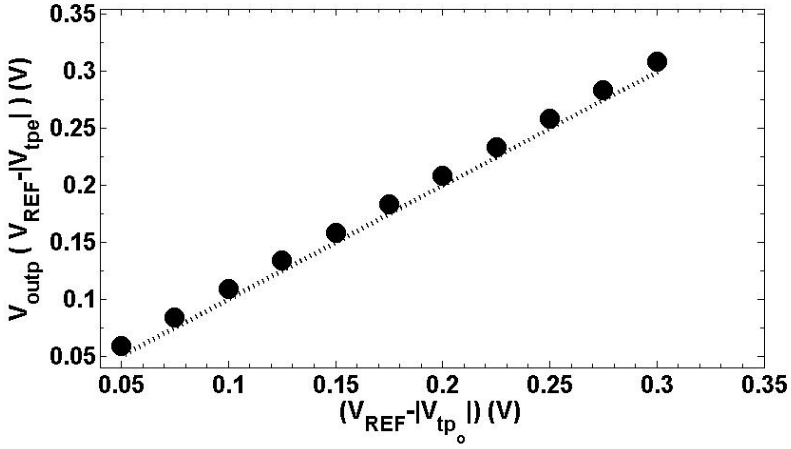


Figure 5.6: The output of the PMOS threshold voltage sensing circuit shown in Figure 5.4.

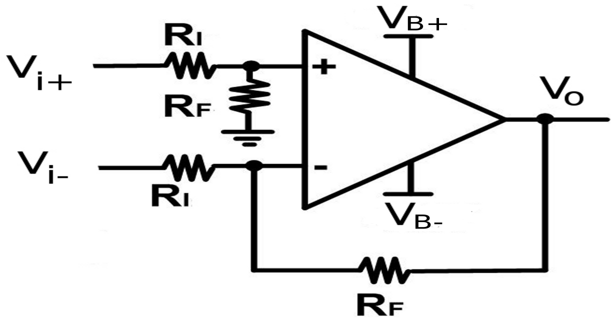


Figure 5.7: The amplifier circuit.

[17], the mismatch between the op-amp transistors threshold voltages is inversely proportional to the square root of the channel area (W*L). Thus, by designing all the amplifier and squaring circuit transistors widths larger than 195nm (the minimum width for 65nm transistor is 120nm) and lengths of 130nm (the minimum L for 65nm transistor is 60nm), this mismatch effect is reduced. In addition, the values of V_{tno} and $(V_{REF} - |V_{tpo}|)$ can be adjusted off-chip to compensate for this mismatch.

3. **Squaring Circuit**

   One of the essential building blocks in the D-ABB circuit, shown in Figure 5.2, is the squaring circuit. Several squaring circuits are reported in the literature [143–145].

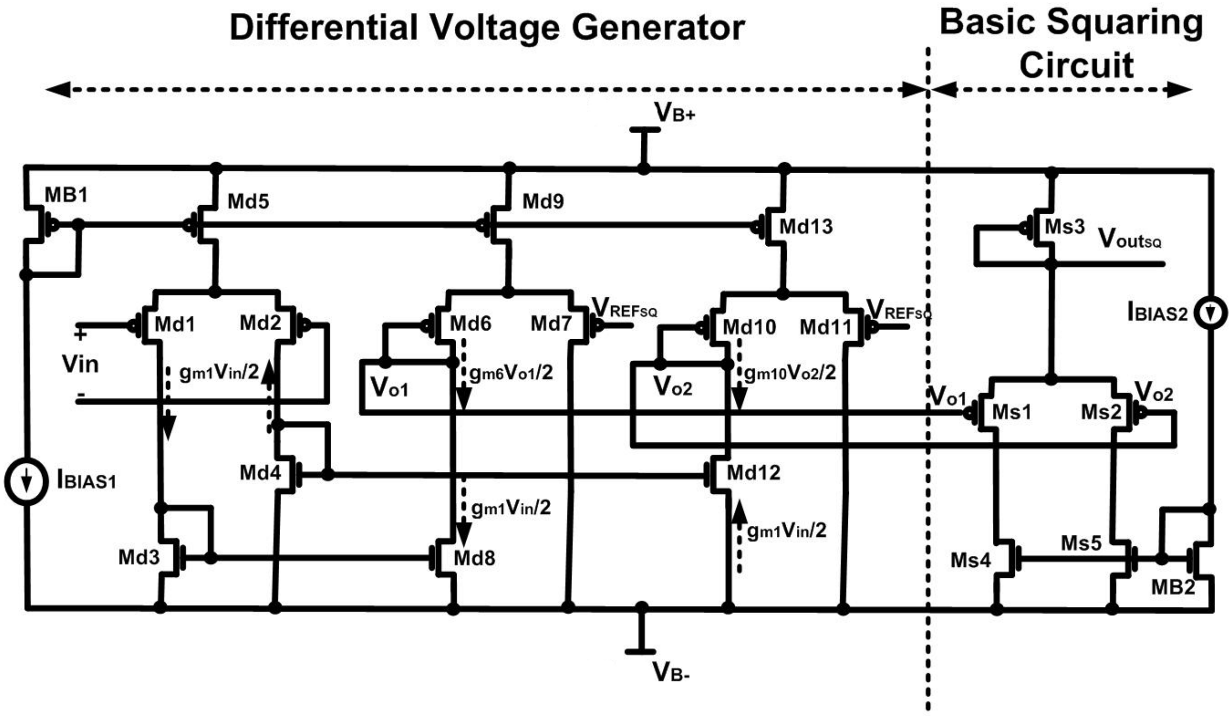


Figure 5.8: The squaring circuit.

Figure 5.8 depicts the squaring circuit used in the D-ABB circuit. The proposed squaring circuit consists of a differential voltage generator circuit and a basic common source differential pair squaring circuit. The differential voltage generator circuit is utilized to adjust the squaring circuit output voltage dc-offset and the squaring circuit gain. Assuming that the transistors pairs, (Md1 and Md2), (Md6, Md7, Md10, and Md11), (Md5, Md9, and Md13), (Md3 and Md8), and (Md4 and Md12), are matched. The small signal current flowing through Md1 is $g_{m1}V_{in}/2$ which is equal to the small signal current flowing through Md8 which is $g_{m6}V_{o1}/2$ due to the current mirror action between these transistors. Therefore, $V_{o1} = (g_{m1}/g_{m6})V_{in}$. Similarly, due to the current mirror action between transistors Md4 and Md12, the voltage V_{o2} is $-(g_{m1}/g_{m10})V_{in}$. Since transistors Md6, Md7, Md10, and Md11 are matched, the two output voltages, V_{o1} and V_{o2}, are given by:

$$V_{o1} = -V_{o2} = (g_{m1}/g_{m6})V_{in} \tag{5.10}$$

These two output voltages, V_{o1} and V_{o2}, have an equal common mode voltage, $V_{REF_{SQ}}$. When these two output voltages are applied to the basic squaring circuit, the resultant output voltage, $V_{out_{SQ}}$, is given by [144]:

$$V_{out_{SQ}} = \frac{(V_{B+} - |V_{tp}|)^2 - (V_{REF_{SQ}} + |V_{tp}|)^2}{2(V_{B+} - V_{REF_{SQ}} - 2|V_{tp}|)} + \frac{(g_{m1}/g_{m6})^2 \times V_{in}^2}{2(V_{B+} - V_{REF_{SQ}} - 2|V_{tp}|)} \quad (5.11)$$

where the transistors pairs, (Ms1 and Ms2) and (Ms4 and Ms5) are matched. It is evident that the squaring circuit output voltage dc-offset can be adjusted through $V_{REF_{SQ}}$, whereas the squaring circuit gain can be adjusted through the transconductance ratio, (g_{m1}/g_{m6}). Figure 5.9 displays the simulation results for the squaring circuit in Figure 5.8, where V_{in} is varied from -0.15V to 0.15V and the squaring circuit gain is 10.0.

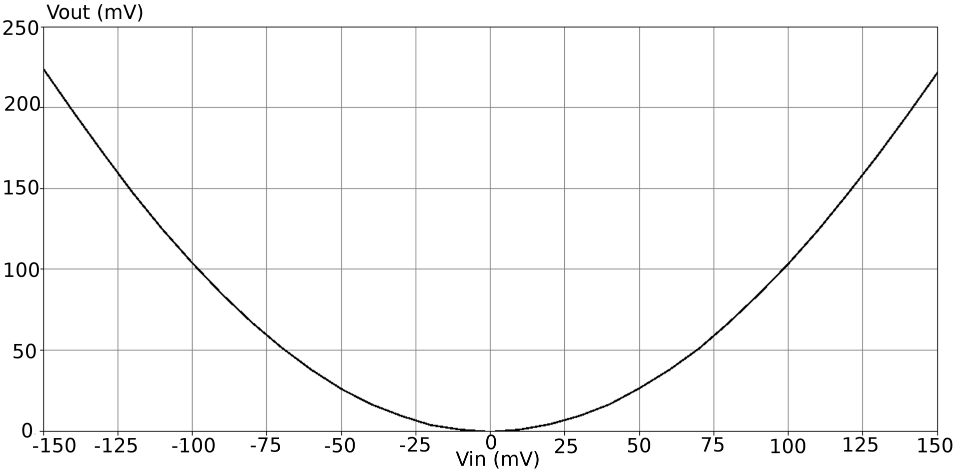


Figure 5.9: The simulated squaring circuit output with V_{in} is varied from -0.15V to 0.15V and the gain is 10.0.

4. **The D-ABB Circuit Design**

   The junction leakage current and the breakdown considerations determine the RBB voltage bound, while the FBB is limited by the sub-threshold leakage current and the forward biasing of the drain-bulk junction. According to [42, 146], the upper limit of the FBB voltage for latch-up free operation, in 65nm CMOS technology with V_{DD} ranges from 0.9V to 1.2V, is 0.6V. Also, SPICE simulations are conducted by sweeping the FBB voltage for NMOS and PMOS transistors. Simulation results show that the upper limits of the FBB voltage to prevent latch-up triggering for NMOS and PMOS transistors are 0.62V and 0.59V, respectively. Therefore, the maximum FBB voltage used in the D-ABB is set to 0.5V to ensure latch-up free operation in case of fluctuations of the FBB voltage around 0.5V. Accordingly, the FBB and the

RBB maximum voltages (i.e., V_{B+} and V_{B-}) are set to ±0.5V [46] (i.e., the body bias voltage changes around its normal value by ±0.5V).

Table 5.1: 65nm Technology information at $T = 120^oC$

	NMOS	PMOS
V_{to} (V)	0.342	-0.204
ϕ_F (V)	0.467	0.439
γ (dimensionless)	0.296	0.174

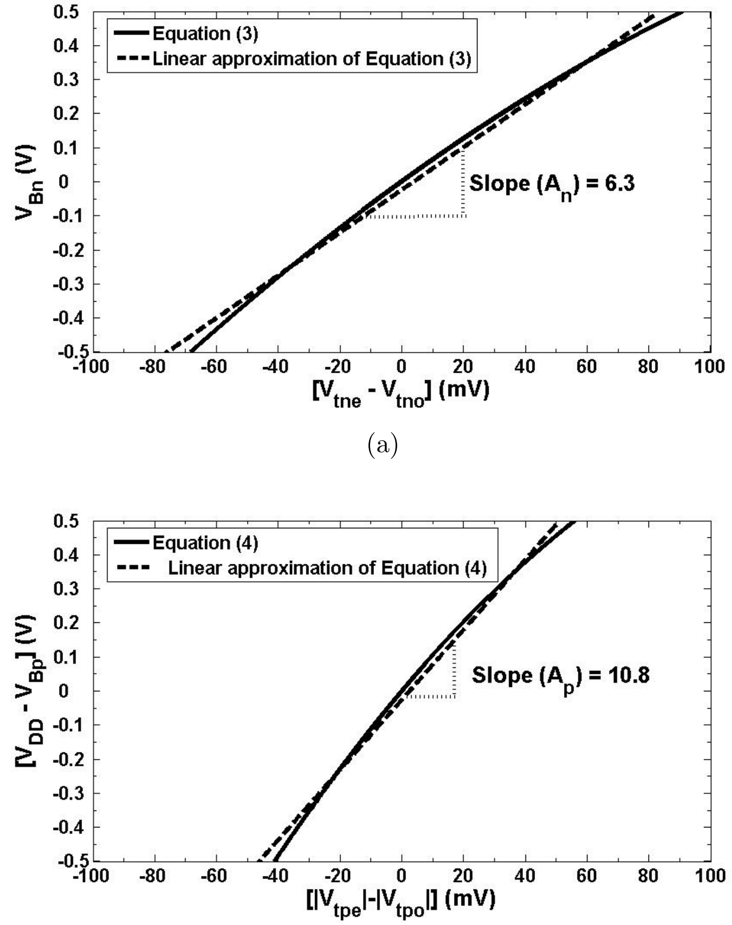
Using the technology information in Table 5.1, the D-ABB circuit is designed with V_{Bn} generation circuit parameters K_{1n}, K_{2n}, and K_{3n} equal 5.7, 10.0, and 1.14, respectively, and with V_{Bp} generation circuit parameters K_{1p}, K_{2p}, and K_{3p} equal -10.8, 33.0, and -1.0, respectively. All the above parameters are for T = 120^oC. It should be mentioned that the technology parameter ϕ_F is linearly proportional to the temperature T in ^oK, accordingly, the D-ABB design is performed at the worst case temperature T = 120^oC. The effect of the temperature on the D-ABB circuit performance will be discussed later in this chapter.

5.2.2 Proposed LD-ABB Circuit

Equations (5.3) and (5.4) are plotted in Figures 5.10.a and 5.10.b, for $-0.5V < V_{Bn} < 0.5V$ and $-0.5V < V_{DD} - V_{Bp} < 0.5V$, respectively. It is evident from Figures 5.10.a and 5.10.b that these equations can be linearized and approximated by:

$$\begin{aligned} V_{Bn} &= A_n \times [V_{tne} - V_{tno}] && (5.12) \\ V_{Bp} &= V_{DD} - A_p \times [V_{tpe} - V_{tpo}] && (5.13) \end{aligned}$$

where A_n and A_p are constant gains and equal 6.3 and 10.8, respectively, as obtained from Figures 5.10.a and 5.10.b. The same threshold voltage sensing circuits shown in Figures 5.3 and 5.4 are used in the LD-ABB circuit. The amplifier circuit, shown in Figure 5.11.a, is designed such that $R_{F_n}/R_{I_n} = 6.3$ whereas, the amplifier circuit, shown in Figure 5.11.b, is designed such that $R_{F_p}/R_{I_p} = 10.8$.



(a)

(b)

Figure 5.10: Linear approximation of (a) Equation (5.3) by (5.12) and (b) Equation (5.4) by (5.13).

5.2.3 The Effect of Process Variations and Temperature on the D-ABB and LD-ABB Circuits

A 5,000 point Monte Carlo analysis using the statistical transistor model reported in Appendix D is conducted. Simulation results reveal that the maximum ratio between the standard deviation of the sensing, amplifier, and squaring circuits parameters (i.e., gain, output voltage swing, and dc offset) to their mean values is 2.3%. In addition, the maximum change in the sensing, amplifier, and squaring circuits parameters, relative to their nominal values, is 1.2% over the temperature range -30^oC to 120^oC. Accordingly, the D-ABB and LD-ABB circuits are insensitive to process variations and temperature.

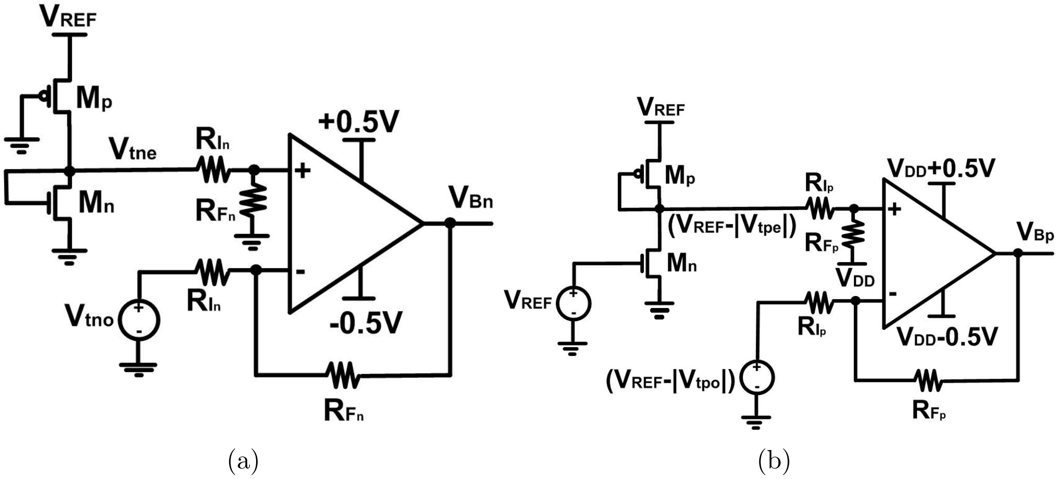


Figure 5.11: The LD-ABB for (a) NMOS body bias, V_{Bn} and (b) PMOS body bias, V_{Bp}.

5.3 Application of the ABB Circuits to High-Performance Circuits

In this section, the ABB circuits (i.e., the D-ABB and the LD-ABB) are adopted to high performance circuits (i.e., a circuit block, extracted from a real microprocessor critical path), to verify their effectiveness in compensating for process variations.

5.3.1 Test Circuit Description

The newly developed ABB circuits are applied to a circuit block, extracted from a real microprocessor critical path, to verify their effectiveness in process variations compensation. This circuit block consists of 15 CMOS gates including CMOS inverter gates, NAND gates, NOR gates, and Transmission gates, similar to the test circuits used in [41, 46]. Figure 5.12 portrays the test circuit, which consists of 30 critical paths, a global ABB circuit, and 30 local ABB circuits. The global ABB provides same bias voltages to all the die critical paths. Therefore, its effectiveness, in reducing WID variations, is limited. The distributed local ABB circuits supply different bias voltages to each critical path, achieving better results in reducing WID variations, at the expense of higher area overhead than that in the global ABB circuit.

This circuit block is selected to model the effect of the proposed ABB on the yield improvement of a real microprocessor design [41]. The figures of merit considered in this experiment are the clock frequency (F_{clk}), the dynamic power (P_{dyn}) of the circuit block when configured as a ring oscillator, and the leakage power (P_{leak}) of the circuit block

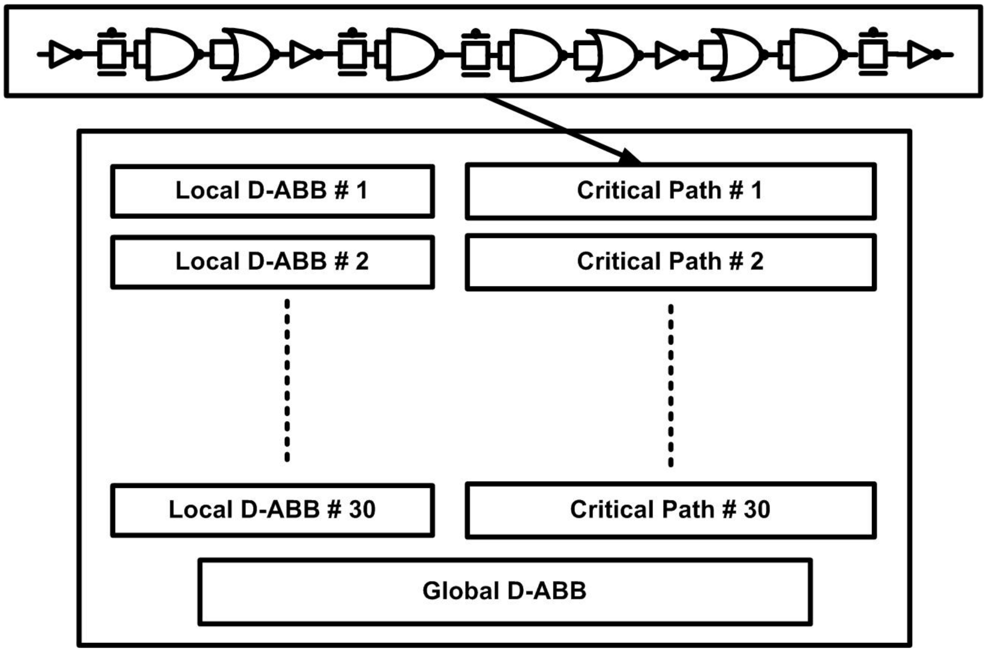


Figure 5.12: The test circuit used in the simulation setup

when operating in static conditions [41]. The circuit block and the ABB circuits are implemented by using the industrial hardware-calibrated 65nm CMOS technology reported in Appendix D. Table 5.1 shows the NMOS and PMOS transistor parameters, extracted from the transistor model. The supply voltage, V_{DD}, equals 1.0V and post layout simulations are conducted.

The effectiveness of the proposed ABB circuits is proved by showing their ability on reducing the D2D and WID variations. The impact of WID variations depends on the number of critical paths per die. In [46, 48], it is reported that when the number of critical paths per die exceeds 14, there is no significant change in the frequency distribution, as shown in Figure 2.11. Therefore, the test circuit used in this chapter has 30 critical paths per die which is sufficiently accurate for obtaining frequency distributions of real microprocessors which contain hundreds of critical paths.

5.3.2 Simulation Setup

First, the global ABB circuit is enabled and all the local ABB circuits are disabled. The global ABB sensing circuit is placed close to any critical path (critical path number 30 is selected in this test circuit). Based on the threshold voltage variations of this critical path, the global ABB provides the body bias voltages to all the die critical paths. Since the body bias voltages are determined based on the threshold voltage calculations of a single critical path, this global ABB circuit does not reduce the WID variations effectively.

Following that, the local ABB circuits are enabled and the global ABB is disabled. Each local ABB sensing circuit is placed close to its corresponding critical path, as shown in Figure 5.12, and supplies the appropriate body bias voltages to this critical path. Therefore, the use of the local ABB is very efficient in accounting for WID variations. The granularity level of the global ABB circuit is the whole die while the granularity level of the local ABB is the critical path. Therefore, the granularity level of the local ABB is smaller at the expense of more area overhead.

The Monte Carlo analysis generates 5,000 different dies. In each Monte Carlo statistical run (which is corresponding to a certain die), the die frequency is calculated as the minimum frequency of the die critical paths. Since the real microprocessor die contains hundreds of critical paths, the die power (i.e., the dynamic power and the leakage power) is calculated as the average power per critical path. This is performed by summing the critical paths powers and dividing by the number of critical paths per die.

5.3.3 The D-ABB Circuit

1. **Global D-ABB**

   In this case, the global D-ABB circuit is enabled and all the local D-ABB circuits are disabled. 5,000 point Monte Carlo analysis, with the same transistor statistical models reported in Appendix D, is conducted. Figure 5.13 depicts F_{clk}, P_{dyn}, and P_{leak} histograms for the No Body Bias (NBB) control case (Figures 5.13.a, 5.13.b, and 5.13.c, respectively) and for the global D-ABB control case (Figures 5.13.d, 5.13.e, and 5.13.f, respectively), when both D2D and WID variations are taken into account.

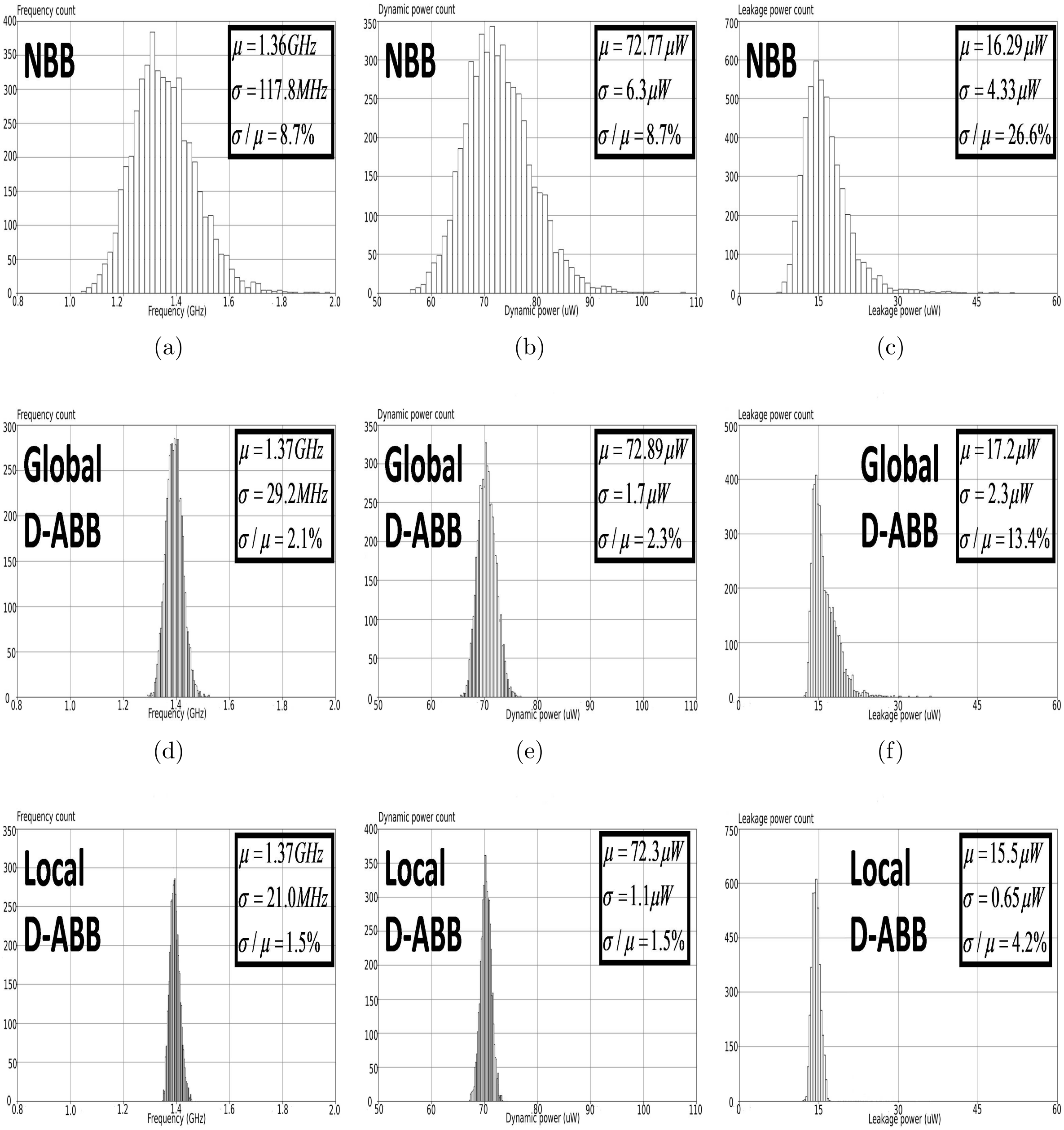
   The following observations are extracted for the global D-ABB control case:

   -i- The means of F_{clk}, P_{dyn}, and P_{leak} (i.e., $\mu_{F_{clk}}, \mu_{P_{dyn}}$, and $\mu_{P_{leak}}$), have a slight change between the NBB case and the global D-ABB case (i.e., the means are changed by a factor less than 1.06X for all design parameters). Therefore, the global D-ABB circuit does not affect the mean of the design parameters.

   -ii- The global D-ABB circuit reduces the standard deviations of F_{clk}, P_{dyn}, and P_{leak} (i.e., $\sigma_{F_{clk}}, \sigma_{P_{dyn}}$, and $\sigma_{P_{leak}}$), by factors of 4.0X, 3.7X, and 1.9X, respectively.

2. **Local D-ABB**

   In this case, the global D-ABB circuit is disabled and all the local D-ABB circuits are enabled. Figure 5.13 depicts F_{clk}, P_{dyn}, and P_{leak} histograms for the local D-ABB



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 5.13: Monte Carlo Histograms of F_{clk}, P_{dyn}, and P_{leak}, with No Body Bias (NBB) control (a,b,c), Global D-ABB control (d,e,f), and Local D-ABB control (g,h,i). Both D2D and WID variations are considered at a temperature $T = 120^oC$.

control case (Figures 5.13.g, 5.13.h, and 5.13.i, respectively), when both D2D and WID variations are taken into account.

The following observations are extracted for the local D-ABB control case:

-i- Similar to the global D-ABB, the local D-ABB circuits do not affect the means of F_{clk}, P_{dyn}, and P_{leak}.

-ii- The local D-ABB circuits achieve significantly more process variations reduction than that of the global D-ABB circuit. For example, F_{clk}, P_{dyn}, and P_{leak} standard deviations are reduced by applying the local D-ABB circuits by factors of 5.6X, 5.7X, and 6.7X, respectively.

5.3.4 The LD-ABB Circuit

1. **Global LD-ABB**

   Similarly, in this case, the global LD-ABB circuit is enabled and all the local LD-ABB circuits are disabled. The global LD-ABB circuit reduces the standard deviations of F_{clk}, P_{dyn}, and P_{leak} (i.e., $\sigma_{F_{clk}}$, $\sigma_{P_{dyn}}$, and $\sigma_{P_{leak}}$), by factors of 3.8X, 2.4X, and 1.5X, respectively. The LD-ABB simulation results histograms are not shown.

2. **Local LD-ABB**

   Also, in this case, the global LD-ABB circuit is disabled and all the local LD-ABB circuits are enabled. The local LD-ABB circuits achieve significantly more process variations reduction than that of the global LD-ABB circuit. For example, F_{clk}, P_{dyn}, and P_{leak} standard deviations are reduced by applying the local LD-ABB circuits by factors of 4.1X, 3.9X, and 2.3X. These results show that the performance of the LD-ABB is less than that of the D-ABB in compensating for process variations at the advantage of less area overhead.

5.3.5 The Effect of Temperature on the D-ABB and LD-ABB Circuits Performance

The proposed ABB circuits (i.e., D-ABB and LD-ABB) design is performed at a temperature $T = 120^{o}C$ which is the worst case condition for both the operating frequency and the leakage power. When the operating temperature decreases, V_t is increased [22], resulting in leakage power reduction. The reduction in the leakage power is large because the leakage power exhibits an exponential relationship with the temperature and V_t [22]. This V_t increase is sensed by the ABB circuits and the corresponding body bias voltages are generated.

Table 5.2: The mean and relative variations, σ/μ, values of F_{clk}, P_{dyn}, and P_{leak}, for $T = 0^oC$, $T = 60^oC$, and $T = 120^oC$, when D2D and WID variations are considered for the NBB, the local D-ABB and the local LD-ABB control scenarios

T =		120^oC	60^oC	0^oC
NBB	F_{clk} mean (GHz)	1.36	1.40	1.45
	F_{clk} σ/μ (%)	8.7	8.6	8.5
	P_{dyn} mean (μW)	72.8	74.8	77.6
	P_{dyn} σ/μ (%)	8.7	8.7	8.6
	P_{leak} mean (μW)	16.3	8.7	2.8
	P_{leak} σ/μ (%)	26.6	31.8	30.2
D-ABB	F_{clk} mean (GHz)	1.37	1.63	1.81
	F_{clk} σ/μ (%)	1.5	1.8	2.0
	P_{dyn} mean (μW)	72.3	72.9	75.2
	P_{dyn} σ/μ (%)	1.5	1.6	1.6
	P_{leak} mean (μW)	15.5	15.2	13.9
	P_{leak} σ/μ (%)	4.2	6.1	5.8
LD-ABB	F_{clk} mean (GHz)	1.37	1.61	1.77
	F_{clk} σ/μ (%)	2.2	2.5	2.9
	P_{dyn} mean (μW)	72.5	73.1	74.3
	P_{dyn} σ/μ (%)	2.3	2.4	2.6
	P_{leak} mean (μW)	15.9	15.6	14.7
	P_{leak} σ/μ (%)	11.6	13.8	9.4

The ABB circuits target is to compensate for the temperature variations effect by reducing V_t, and therefore, FBB is adopted to the die critical paths. Table 5.2 shows the mean values of F_{clk}, P_{dyn}, and P_{leak}, for $T = 0^oC$, $T = 60^oC$, and $T = 120^oC$, when both D2D and WID variations are considered for the NBB, the local D-ABB, and the local LD-ABB control scenarios. It is evident from Table 5.2 that the ABB circuits result in increasing F_{clk} and P_{leak}. The value of P_{leak} is kept less than its worst case value at $T = 120^oC$ whereas F_{clk} increases to larger values. Thus, at temperature values lower than $T = 120^oC$, the ABB circuits speed up the dies, keeping the leakage power less than its worst case value at $T = 120^oC$ (i.e., less than 16.3 μW).

5.3.6 ABB Circuits Design Considerations

Practically, there are several design considerations that should be addressed, when the proposed ABB circuits are to be fabricated. These design considerations are:

1. **Mixed Analog-Digital Design Considerations**

   Separate power supply and ground planes are routed for the analog components because analog components are very sensitive to disturbances in the supply voltage. Thus, a low noise analog power supply network is a stringent requirement for the proper operation of these analog components. Noise, due to variations in the power supply and ground (i.e., the noise resulting from the digital components switching), is coupled into the analog portion of the chip and is amplified along with the desired signal. This affects the functionality of the analog components [147]. Several techniques, to help prevent the digital switching noise from affecting the analog components, are discussed in [147]. This analog and digital supplies and grounds separation is also required for any ABB circuit such as those introduced in [41, 46].

2. **Area Overhead and Granularity Level Trade-off**

   The global D-ABB and global LD-ABB are very efficient for the D2D variations. However, for WID variations, there is a trade-off between the D-ABB and LD-ABB granularity level and the associated area overhead (i.e., the lower the granularity level is, the higher the associated area overhead). This trade-off exists in the proposed D-ABB and LD-ABB circuits and any ABB circuit such as those introduced in [41, 46]. However, the lower area overheads of the D-ABB and LD-ABB circuits allows lowering the granularity level while the total area overhead is similar to that in [41, 46].

   The ABB circuit is basically utilized for reducing the D2D and the systematic WID variations, such as channel length variations, that exhibit high spatial correlation (i.e., two devices separated by a close distance behave more similarly than two devices spaced farther apart). Accordingly, there is a trade-off between the ABB granularity level and the associated area overhead (i.e., the lower the granularity level is, the higher the associated area overhead and more systematic WID variations reduction). The ABB circuits are not efficient for random WID variations compensation because these random variations are spatially uncorrelated, as discussed in Chapter 2.

3. **Generation of the DC Supply Voltages for the D-ABB and LD-ABB**

   In the post layout simulations conducted in this section, the DC supply voltages, V_{B+}, V_{B-}, V_{DD}+V_{B+}, and V_{DD}+V_{B-}, are generated externally from an off-chip power supply. However, in real microprocessor design, these DC supply voltages are generated by using an on-chip DC-DC converter [148]. This DC-DC converter increases the required area overhead of the proposed D-ABB and LD-ABB circuits. However, the same area overhead is required for the ABB circuits in [41, 46] because a DC-DC converter is required in these circuits as well. Generally speaking, any ABB circuit must have analog driving bias amplifiers at its output to provide the body bias voltages. These analog driving amplifiers need DC supply voltages, which are generated by using the on-chip DC-DC converter.

5.3.7 Test Chip Results and Discussions

The test chip details are listed in Appendix C including the chip micrograph and the Printed Circuit Board (PCB) design. **Triple-well process** is utilized to allow the control of the NMOS transistors bias voltage. **35** test chips are packaged to account for the D2D variations. Each test chip contains two blocks dedicated for the D-ABB and the LD-ABB circuits, as shown in Figure 5.14. Each block of these two blocks represents the test circuit shown in Figure 5.12. Therefore, 32 critical paths are utilized representing the key circuit elements of a microprocessor critical path. The body bias terminals, V_{Bp} and V_{Bn}, of each critical path can be connected to either V_{DD} and Ground, V_{Bp} and V_{Bn} provided by the global ABB, V_{Bp} and V_{Bn} provided by the local ABB, and V_{Bp} and V_{Bn} generated externally (off chip). While testing the 35 chips, only 27 chips are working in the D-ABB block testing whereas no chip is working for the LD-ABB block testing. Accordingly, the test chip results shown in this section are only for the D-ABB circuit.

For each test chip, the 32 critical paths are configured in a ring oscillator structure for frequency measurements (which is measured by observing the output signal on the scope), or disabled for leakage power measurements (which is measured by reading the current supplied by the supply voltage, V_{DD}, and multiply this measured current by $V_{DD} = 0.9$V). It should be noted that the measured leakage power is the total power of the 32 critical paths. The chip frequency is the minimum of the 32 critical paths frequencies because these critical paths are meant to emulate a real microprocessor design.

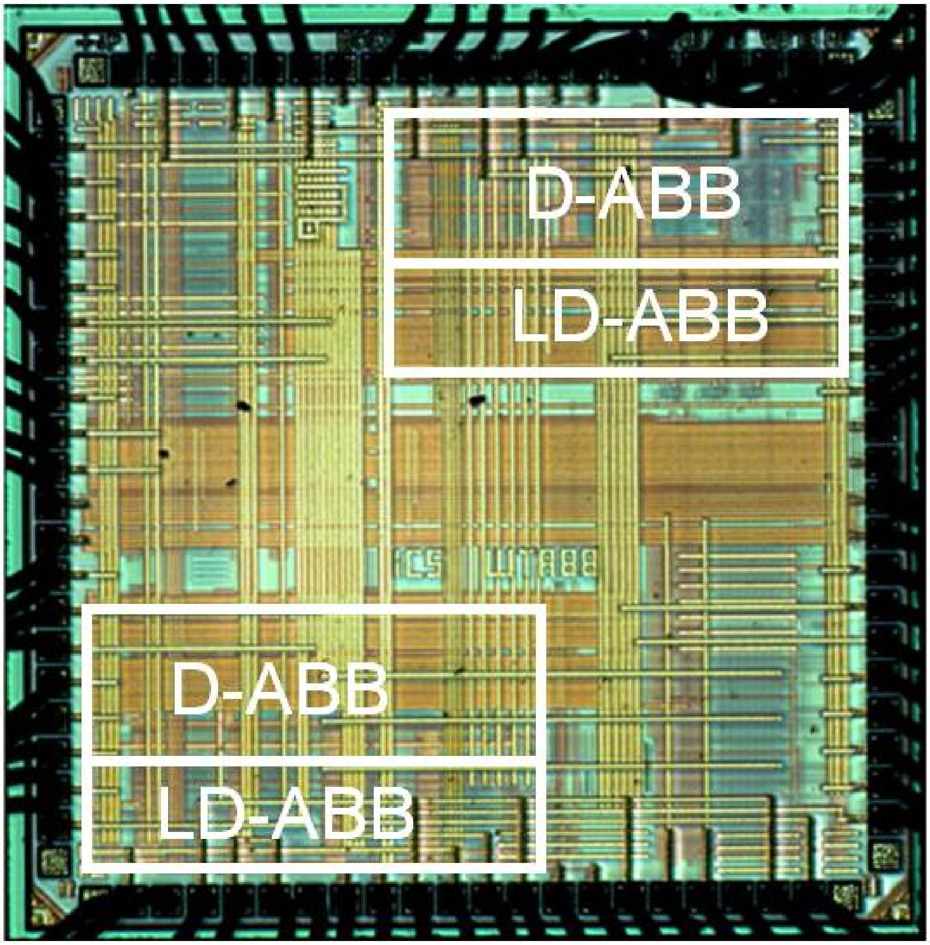


Figure 5.14: Test chip micrograph showing the D-ABB and the LD-ABB blocks

These results are obtained for the NBB (No Body Bias), global D-ABB, and local D-ABB. It should be emphasized that each ABB circuit output is connected to an analog buffer circuit (i.e., an operational amplifier configured as a unity-gain amplifier) to ensure low output impedance and to be able to drive the critical path body bias terminal. Figure 5.15 displays the measured frequency and leakage power for the 27 test chip dies. In this figure, there exist a significant variations in the frequency and leakage power due to D2D and WID variations. Each die is accepted if it satisfies the frequency and leakage power constraints. The frequency constraint is set to be 500MHz and represented by the vertical line shown in Figure 5.15, whereas the leakage power constraint is represented by the slanted leakage power line. The maximum allowed leakage power of each die can be calculated by using the following equation:

$$P_{leakage|_{max}} = P_{density|_{max}} \times Area|_{die} - \alpha|_{worst\ case} \times V_{DD}^2 \times C_{total} \times f \qquad (5.14)$$

where $P_{leakage|_{max}}$ is the maximum allowable leakage power, $P_{density|_{max}}$ is the maximum power density at a given temperature, $Area|_{die}$ is the die area containing the 32 critical paths excluding the ABB circuits, $\alpha|_{worstcase}$ is the worst case activity factor, C_{total} is the total switched capacitance, and f is the frequency. Therefore, the slanted leakage power line is representing (5.14). Assuming a worst case power density of 20 W/cm^2 [46] and using $Area|_{die} = 73 \times 124\ \mu m^2$, and $\alpha|_{worst\ case} \times V_{DD}^2 \times C_{total} = 1.496\ pF.V^2$. Equation (5.14) represents the slanted leakage power line. For any die to be accepted, it must have a frequency larger than 500MHz and a leakage power less than $P_{leakage_{max}}$ at its operating frequency. Therefore, in Figure 5.15, the number of dies that are accepted is 14 out of 27 dies (52% acceptance). The relative frequency variation, σ_f/μ_f, of this NBB case is 9.6%.

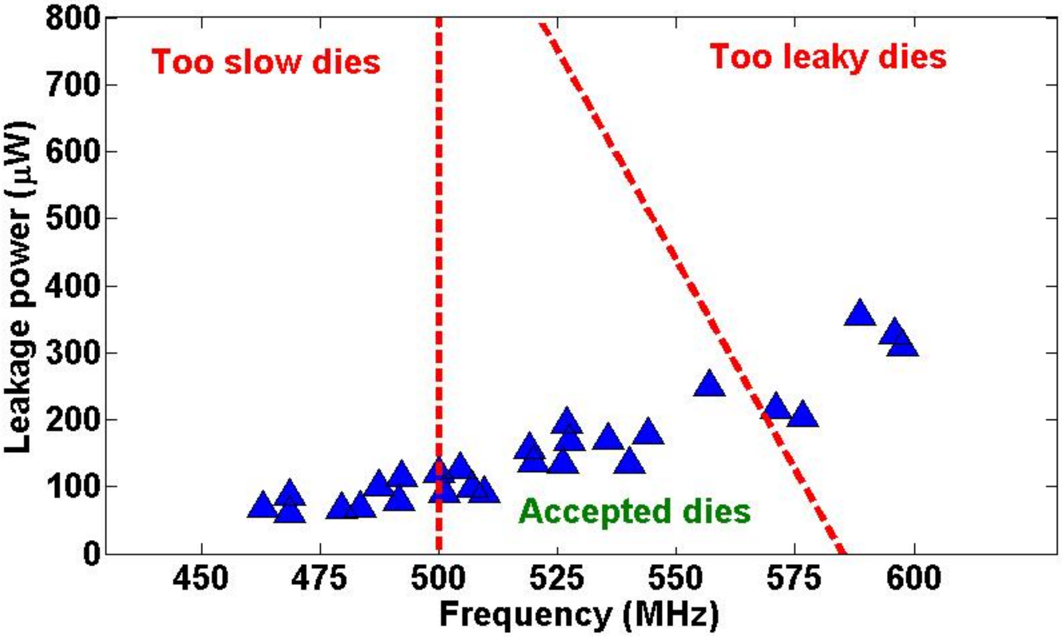


Figure 5.15: Measured frequency and leakage power for 27 dies for the NBB case

1. **Global D-ABB Testing Results**

   The objective of the D-ABB is to apply the optimum NMOS and PMOS transistors body bias voltages that maximize the die frequency while meeting the leakage power constraint. The DC voltages V_{tno} and $V_{REF} - |V_{tpo}|$ are swept for each die within ±30% of their nominal values and the corresponding die frequency and the leakage power are measured. The V_{tno} and $V_{REF} - |V_{tpo}|$ values that result in the highest die frequency with a leakage power satisfying the leakage power constraint is selected. There is only one die out of the 27 dies that achieve a leakage power higher than the maximum allowable leakage power for all the V_{tno} and $V_{REF} - |V_{tpo}|$ values as shown in Figure 5.16. It should be highlighted that the tweaking of V_{tno} and $V_{REF} - |V_{tpo}|$ results in finding the optimal body bias voltages for all the 32 critical paths. In other words, only these two DC sources are tweaked for the whole die critical paths. This V_{tno} and $V_{REF} - |V_{tpo}|$ tweaking is performed for both the global and the local D-ABB circuits.

The effect of the global D-ABB in compensating for process variations is portrayed in Figure 5.16. The number of dies that are accepted is 24 out of the 27 dies (89% acceptance). σ_f/μ_f of the global D-ABB equals 3.1%. Therefore, the global D-ABB adoption results in increasing the acceptance percentage from 52%, in the NBB case, to 89% and reducing σ_f/μ_f by a factor of 3.1X with respect to the NBB case.

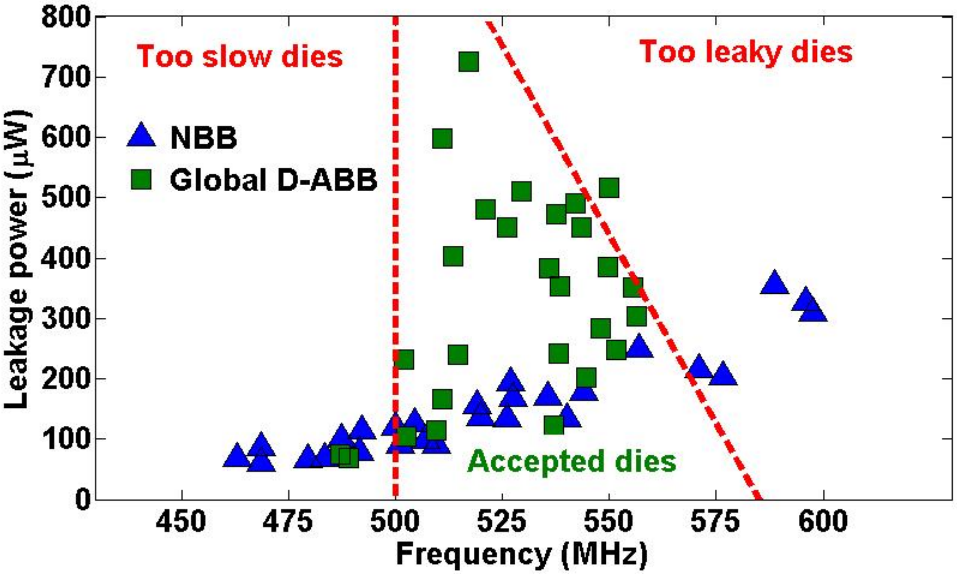


Figure 5.16: Measured frequency and leakage power for 27 dies for the NBB case and the global D-ABB case

2. **Local D-ABB Testing Results**

The local D-ABB is used to compensate for both D2D and systematic WID variations. It should be emphasized again that the ABB technique, in general, is not effective in reducing the random WID variations. The effect of the local D-ABB in compensating for process variations is portrayed in Figure 5.17. All the 27 dies are accepted (100% acceptance). σ_f/μ_f of the local D-ABB equals 0.79%. Therefore, the local D-ABB adoption results in increasing the acceptance percentage from 52%, in the NBB case, to 100% and reducing σ_f/μ_f by a factor of 12.2X with respect to the NBB case.

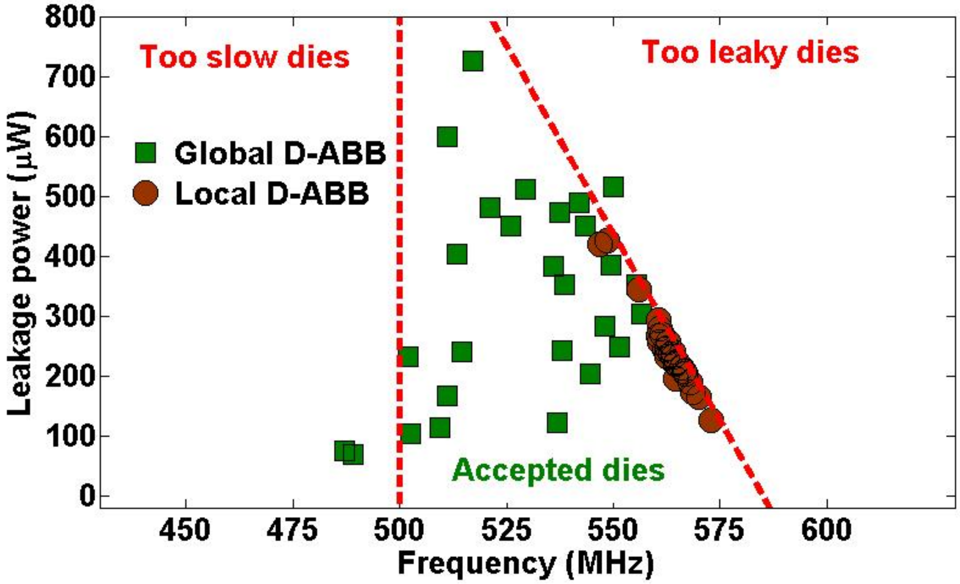


Figure 5.17: Measured frequency and leakage power for 27 dies for the global D-ABB case and the local D-ABB case

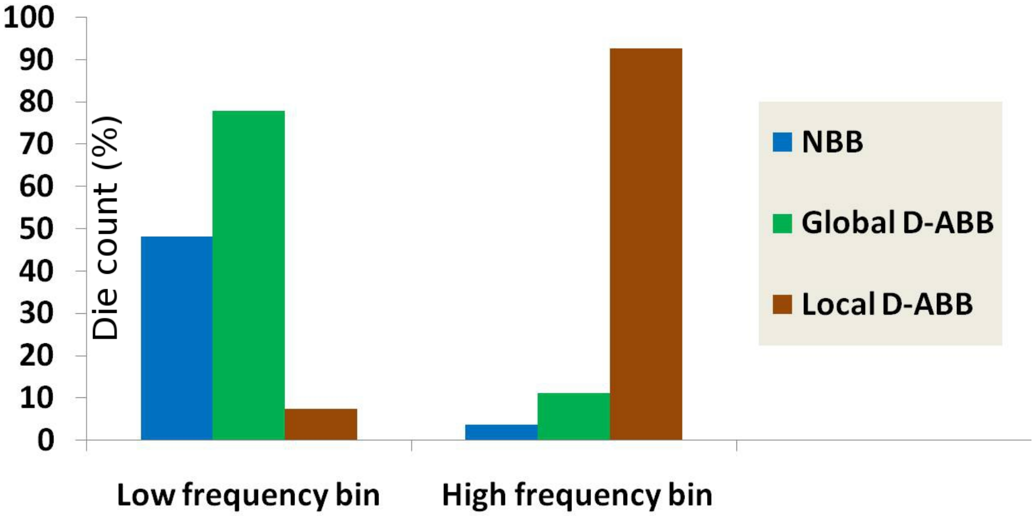


Figure 5.18: Histogram showing the number of accepted dies in each frequency bin for the NBB, global D-ABB, and local D-ABB cases. The low frequency bin contains accepted dies with frequencies less than 550MHz whereas high frequency bin contains accepted dies with frequencies higher than 550MHz.

Due to process variations, the frequency distribution is divided into several frequency bins. In this testing process, we divided the frequency distribution into two frequency bins (i.e., low frequency bin (500MHz $\leqslant$ frequency $<$ 550MHz) and high frequency bin (frequency $\geqslant$ 550MHz)). Only the dies that satisfy the frequency and leakage power constraints are placed in these frequency bins. Figure 5.18 displays a histogram showing the number of acceptable dies in each frequency bin for the NBB case and the global and local D-ABB cases. According to Figure 5.18, in the NBB case, only one die, out of the 14 accepted dies, is placed in the high frequency bin while the other 13 dies are placed in the low frequency bin (i.e., 4% of the 27 dies are placed in the high frequency bin). The global D-ABB increases the number of accepted dies to 24 dies. However, only 3 dies of these 24 accepted dies are placed in the high frequency bin (i.e., 11% of the 27 dies are placed in the high frequency bin). The number of accepted dies that are placed in the high frequency bin for the local D-ABB case is 25 dies (i.e., 93% of the 27 dies are placed in the high frequency bin).

Table 5.3 shows a comparison between the D-ABB and the ABB circuit presented in [46]. The ABB in [46] is selected because this is the only work in the literature that provides detailed ABB area overhead and test chip measurements. It is evident that the D-ABB exhibits less area overhead compared to the ABB in [46] by a factor of 143X given that the D-ABB is fabricated by using 65nm technology whereas the ABB in [46] is fabricated by using 150nm technology. The layout area of the die in [46] equals 4.5X5.3 mm^2 containing 21 subsites, Each subsite contains an ABB controller with a circuit block under test. Accordingly, the ABB area with the circuit block under test equals 4.5X5.3 mm^2 / 21 = 1.136 mm^2. In the D-ABB circuit, the total die area is 1.1X0.24 mm^2 which contains 32 D-ABB controllers with the circuit block under test. Thus, the ABB area with the circuit block under test equals 1.1X0.24 mm^2 /32 = 0.008 mm^2. These results are reported in Table 5.3. In [46], it is reported that the local ABB circuit exhibits an area overhead of 3%. This means that the proposed D-ABB circuit area overhead is 0.021% if used at the same granularity level as the ABB introduced in [46]. However, the main achievement of the proposed D-ABB circuit is that it can be used at a finer granularity level.

The ABB in [46] results in more reduction of the process variations by factors of 1.3X and 1.6X for the global ABB and local ABB circuits, respectively, when using 5-bit DAC resolution. However, the D-ABB is superior to the ABB in [46] in process variations reduction by factors of 1.1X and 1.5X for the global ABB and local ABB circuits, respectively, when 3-bit DAC resolution is utilized. Accordingly, one of the main advantages of the D-ABB, in addition to the low area overhead, is that it is a resolution free ABB circuit since no ADC or DAC circuits are utilized in its implementation. However, the ABB in [46] complexity and overhead increase with the required resolution in the body bias voltage. Finally, it should be noted that the V_t variations in 65nm technology is larger than that in the 150nm technology. It is stated in ITRS [6] that $\sigma_{V_t}/\mu_{V_t} = 7\%$, for 150nm technology, and $= 12\%$, for 65nm technology. The average power consumption of the proposed D-ABB circuit is measured to be 132μW.

Table 5.3: Chip measurement results and Performance Comparison with the ABB in [46]

	D-ABB	**ABB in [46]**	
		3-bit	5-bit
Technology	65nm	150nm	
σ_{V_t}/μ_{V_t}	12%	7%	
Number of test chip dies	27	62	
Number of critical paths/die	32	21	
Die area (mm^2)	1.1X0.24	4.5X5.3	
ABB area (mm^2)	0.008	1.14	
High frequency bin accepted dies (local ABB)	93%	66%	99%
σ_f/μ_f **reduction factor (global ABB)**	3.1X	2.8X	4.1X
σ_f/μ_f **reduction factor (local ABB)**	12.2X	8.2X	19.5X
Resolution limitations	Unlimited	Limited	

5.4 Application of the ABB Circuits to SRAM Cells

The objective of this section is to apply the proposed D-ABB circuits to the SRAM array to compensate for both NBTI aging and process variations (D2D and systematic WID variations). Only the D-ABB simulation results are displayed in this section. Thus, ABB in this section refers to the D-ABB circuit. The LD-ABB circuit exhibits less NBTI and process variations reduction than that of the D-ABB, as discussed in Section 5.3.

5.4.1 Proposed ABB Circuit

In the proposed ABB circuit, the effect of NBTI on $|V_{tp}|$ is compensated by estimating the actual value of $|V_{tp}|$, which is impacted by NBTI, by using an estimation circuit. Then, the analog controller generates the appropriate body bias voltage, V_{Bp}, to mitigate the NBTI impact. The analog controller is a direct implementation of the relationship between $|V_{tp}|$ and V_{Bp}. The body bias voltages of the PMOS transistor, V_{Bp}, which result in NBTI compensation, is given by (5.4):

$$V_{Bp} = V_{DD} - \frac{2\sqrt{2\phi_F}}{\gamma}[|V_{tp_{stressed}}| - |V_{tpo}|] + \frac{1}{\gamma^2}[|V_{tp_{stressed}}| - |V_{tpo}|]^2 \tag{5.15}$$

where $|V_{tp_{stressed}}|$ is the PMOS transistor threshold voltage which is impacted by NBTI aging and process variations. The proposed ABB circuit is depicted in Figure 5.19 for the bias voltage V_{Bp}. The sensing circuit, shown in Figure 5.19, is used to estimate the actual value of $|V_{tp}|$, which is impacted by the NBTI under full stress (the worst case NBTI effect). This sensing circuit outputs an estimate for the PMOS threshold voltage, denoted by V_{out} $= r$ $(V_{DD} - |V_{tp_{stressed}}|)$ which is applied to an amplifier circuit and a squaring circuit to produce the required bias voltage, which is capable of reducing the impact of NBTI.

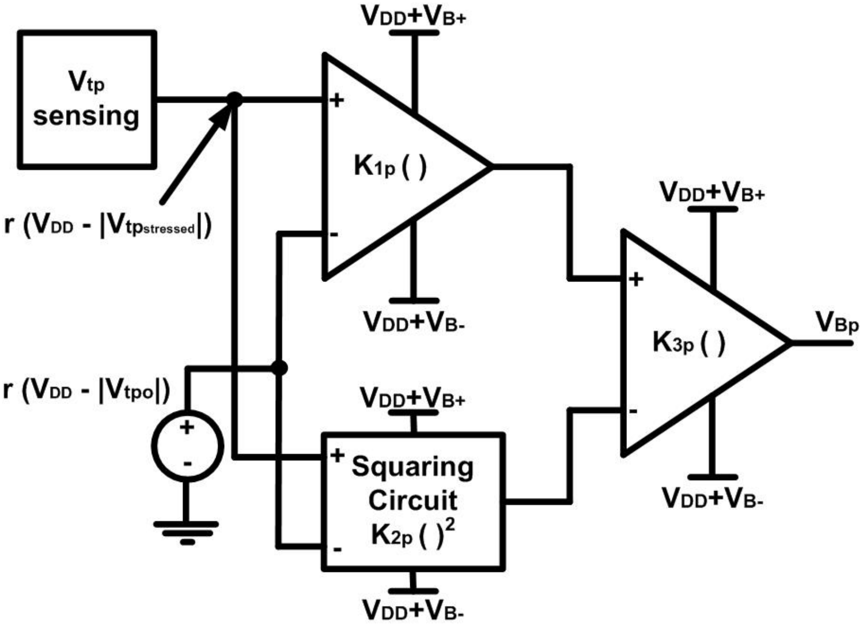


Figure 5.19: The proposed ABB circuit for NBTI compensation

In Figure 5.19, the voltage source of the value r $(V_{DD} - |V_{tpo}|)$ is a DC bias voltage representing the ratio r multiplied by the difference between the supply voltage, V_{DD}, and the PMOS transistor nominal threshold voltage value at zero body bias. According to Figure 5.19 and recalling (5.15), the gains K_{1p}, K_{2p}, and K_{3p} are given by:

$$K_{1p} \times K_{3p} = \frac{2\sqrt{2\phi_F}}{\gamma \times r} \quad \text{and} \quad K_{2p} \times K_{3p} = -\frac{1}{\gamma^2 \times r^2} \tag{5.16}$$

It should be mentioned that the proposed ABB is capable of compensating for both the NBTI and process variations (D2D and systematic WID variations) impacts.

NBTI results on increasing $|V_{tp}|$ with aging time, whereas systematic process variations result in increasing or decreasing $|V_{tp}|$ by a certain amount. Accordingly, if the resultant $|V_{tp}|$ due to NBTI and variations is increased, FBB is supplied by the ABB circuit. On the other hand, if the resultant $|V_{tp}|$ due to NBTI and variations is decreased, RBB is supplied by the ABB circuit.

The sensing circuit, displayed in Figure 5.20, is used to estimate the actual value of the threshold voltage of the PMOS transistor, which is impacted by NBTI under static DC stress. In this circuit, the PMOS transistor is sized with the same sizing as the SRAM PMOS transistor and the NMOS transistor is a native transistor. Native transistors are manufactured without additional threshold voltage implantation in the channel area and thus exhibit a natural threshold voltage in the manufacturing process. This natural threshold voltage is typically around 0V [149]. The minimum size of the native transistor as introduced by the 65nm CMOS technology is 500nm/300nm which is adopted in this sensing circuit.

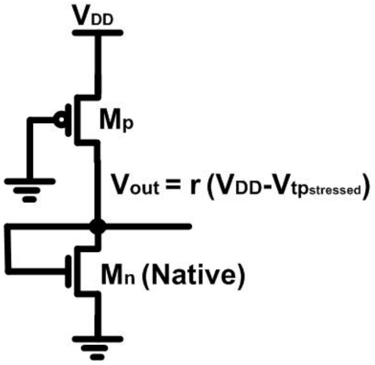


Figure 5.20: The PMOS transistor $|V_{tp}|$ sensing circuit.

By using the $\alpha-$power law model, introduced in [114], and equating the DC currents of the NMOS and PMOS transistors, the output voltage of this circuit, V_{out}, is expressed as:

$$\begin{aligned} V_{out} &= V_{tn} + r \times [V_{DD} - |V_{tp_{stressed}}|] \\ &\approx r \times [V_{DD} - |V_{tp_{stressed}}|] \text{ and } r = \left(\frac{k_{p'}\frac{W}{L}|_p}{k_{n'}\frac{W}{L}|_n}\right)^{\frac{1}{\alpha}} \end{aligned} \tag{5.17}$$

where $k_{n'}$ and $k_{p'}$ are the technological parameters, and $\frac{W}{L}|_n$ and $\frac{W}{L}|_p$ are the sizes of the NMOS and the PMOS transistors, respectively. It should be noted that the native NMOS transistor threshold voltage, V_{tn}, is assumed to be 0V in (5.17) [149]. Figure 5.21 displays the output voltage of the sensing circuit, V_{out}, versus $(V_{DD} - |V_{tpo}|)$. This figure is obtained from SPICE simulations by sweeping the threshold voltage of the transistor model and using V_{DD} = 1.0V and r = 0.54. Good agreements between the estimated threshold voltage values and their actual values, prove that the threshold voltage sensing circuit is effective. The maximum error between the estimated threshold voltage values and their corresponding actual values is 5.4%, and the average error is 3.2%. The amplifier and the squaring circuits designs are the same as that presented in Section 5.2.

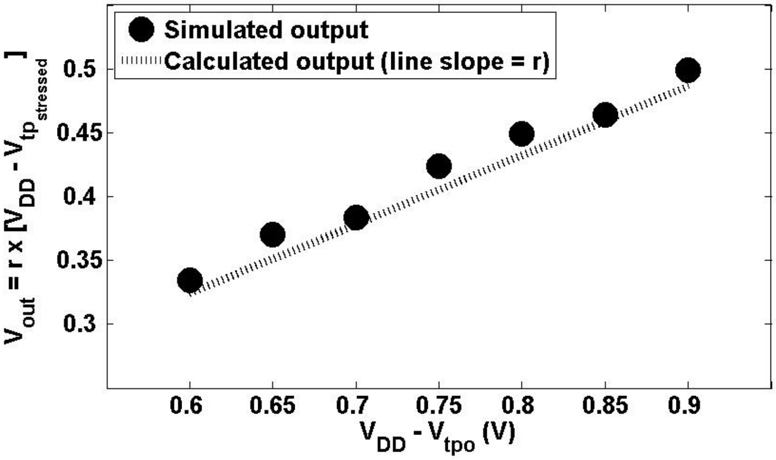


Figure 5.21: The output of the PMOS threshold voltage sensing circuit shown in Figure 5.20.

5.4.2 Simulation Results and Discussions

In the following simulation results, the layout of a 512 6T SRAM cells column is utilized with V_{DD} = 1V, referring to an industrial hardware-calibrated 65nm CMOS transistor model. This model card includes the process variations and the NBTI stress effects which are declared by the manufacturer to be Silicon verified. The reliability analysis is performed by using Cadence RelXpert, Virtuoso Spectre, and Virtuoso UltraSim tools. The ABB circuit is designed with K_{1p}, K_{2p}, and K_{3p} equal -1.8, 10.0, and -10.6, respectively. All the above parameters are calculated at T = 120^oC and r = 0.54. The FBB and the RBB maximum voltages (i.e., $V_{DD} + V_{B+}$ and $V_{DD} + V_{B-}$) are set to 1V±0.5V, as explained in Section 5.2.

The effectiveness of the proposed ABB circuit in mitigating the NBTI stress impact and the process variations is examined by performing post layout simulations for the SRAM column without the ABB circuit (NBB case) and with the ABB circuit (ABB case). This effectiveness is measured by examining the SRAM parameters such as SNM, read failure probability, WM, write failure probability, sub-threshold leakage, and $Q_{critical}$, as explained in Chapter 2. In these simulations, the temperature used is T = 120^oC with input signal probability $p_L = p_R = 0.5$, with the aging time changes from 0 to 10 years. p_L and p_R denote the probability that transistors M_{pL} and M_{pR}, shown in Figure 2.18, are ON, respectively.

1. **SNM**

   Figure 5.22.a shows the HOLD SNM degradation percentage versus aging time for the No Body Bias (NBB) case and the ABB case. It is evident that the ABB circuit not only keep the HOLD SNM constant but also improves it with aging up to 5 years aging time. This is because the ABB sensing circuit is represented by a DC stressed PMOS transistor whereas the SRAM PMOS transistors exhibit 50% stress probability because $p_L = p_R = 0.5$. Accordingly, the ABB circuit provides more FBB than required which improves the HOLD SNM. After 5 years aging, the ABB case exhibits some HOLD SNM degradation because the ABB is limited to a body bias voltage of 0.5V. Accordingly, the NBTI $|V_{tp}|$ increase is larger than the $|V_{tp}|$ reduction amount supplied by the ABB when the body bias voltage becomes 0.5V. However, this HOLD SNM degradation percentage is 1% at 10 years aging compared to 4.3% for the NBB case.

   Similarly, the READ SNM degradation percentage, displayed in Figure 5.22.b, exhibits improvement for the ABB case up to 5 years aging time. At 10 years aging time, the ABB case READ SNM degradation percentage is 4.3X less compared to that of the NBB case. Also, it should be noted that the READ SNM is more sensitive to NBTI than the HOLD SNM. For example, the READ SNM degradation percentage at 10 years aging time is 10.9% whereas the HOLD SNM degradation percentage is 4.3% at the same aging time.

2. **Read Failure Probability**

   The ABB circuit adoption helps in mitigating both the NBTI and the process variations impact on the PMOS transistors. Accordingly, at zero aging time, the ABB adoption reduces the read failure probability from 0.07% to 0.03% (i.e., the number of SRAM cells that fail in the read operation is reduced from 734 cells to 356 cells in a 1Mb SRAM block) as portrayed in Figure 5.22.c. The ABB circuit improves the read failure probability up to 5 years aging time. At 10 years aging time, the ABB case shows a reduction in the read failure probability by a factor of 6.4X compared to that of the NBB case. The number of Monte Carlo points in these simulations is 10,000.

3. **WM and Write Failure Probability**

   The WM and the write failure probability are improved with NBTI which is shown in Figures 5.22.d and 5.22.e. The WM is increased at an aging time of 10 years by 5.7% and the write failure probability is reduced from 0.4% at zero aging time to 0.1% at 10 years aging. Unfortunately, the ABB circuit adoption results in WM degradation (due to the $|V_{tp}|$ compensation) in the first 5 years. From 5 years to 10 years aging time, the ABB circuit allows some increase of the WM but still less than that of the NBB case. For example, at 10 years aging time, the WM is increased by 1.6% for the ABB case whereas it is increased by 5.7% for the NBB case.

   In the mean time, the write failure probability is reduced by the ABB circuit as well due to the process variations compensation effect. The write failure probability is reduced at zero aging time due to the ABB adoption by a factor of 3.9X and at 10 years aging time by a factor of 1.6X as shown in Figure 5.22.e. In addition, the ABB adoption results in lower write failure probability over all the aging time period.

4. **Sub-Threshold Leakage**

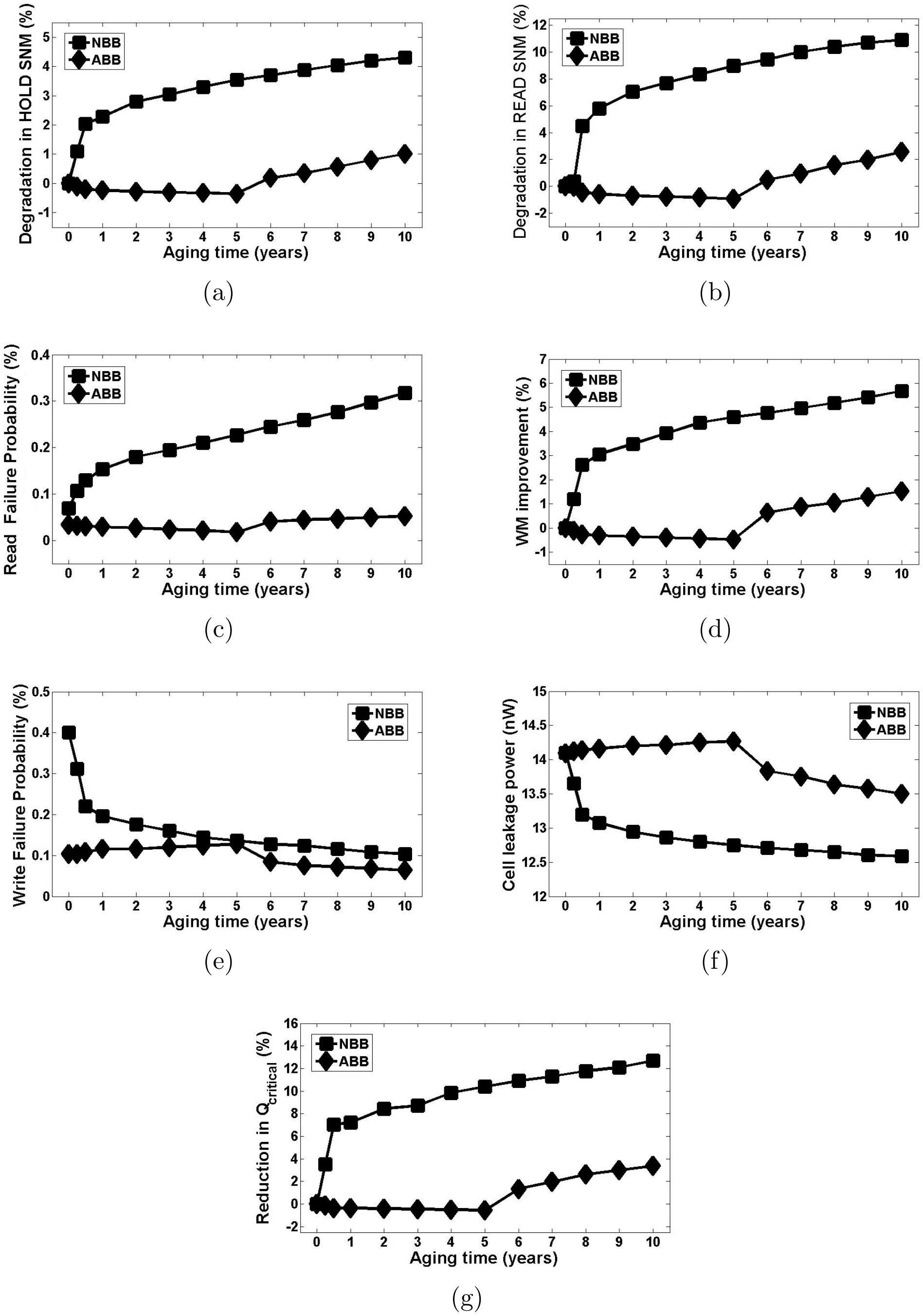
   Figure 5.22.f displays the SRAM cell leakage power for both the NBB and the ABB cases. As reported in [77] and explained in Chapter 2, the leakage power decreases with NBTI aging and this reduction is maximized when $p_L = p_R = 0.5$. Correspondingly, the NBTI effect results in reducing the leakage current by 10.8% over 10 years aging time. The ABB increases the leakage power in the first 5 years. Following that, the ABB reduces the leakage power for aging time larger than 5 years. However, this leakage power reduction is still less than that in the NBB case.

5. $Q_{critical}$

   The critical charge, $Q_{critical}$, is calculated only for the 1-to-0 flip which is affected by the NBTI and is much smaller than the 0-to-1 flip as mentioned in Section 3.2.2. $Q_{critical}$ is calculated by applying an exponential current pulse at node V_L given by (2.3) and $Q_{critical}$ is calculated by using (4.2). Figure 5.22.g portrays that $Q_{critical}$ decreases with NBTI aging time and reaches up to 12.7% reduction at an aging time of 10 years. The ABB adoption increases $Q_{critical}$ in the first 5 years and then

$Q_{critical}$ decreases. At an aging time of 10 years, the ABB reduces $Q_{critical}$ degradation percentage by a factor of 3.8X as displayed in Figure 5.22.g.

Moreover, it should be noted that the NBTI impact grows faster for the first year aging time. It has been shown by simulations that the $|V_{tp}|$ increase due to NBTI is 31.5mV at 1 year aging time whereas it becomes 37.4mV at 2 years aging time. At 5 years aging time, the $|V_{tp}|$ increase becomes 47mV which is close to the maximum ABB compensation. At 10 years aging time, the $|V_{tp}|$ increase becomes 56mV.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

Figure 5.22: Post layout simulation results for the No Body Bias (NBB) case and the ABB case versus aging time at T = 120^oC and $p_L = p_R = 0.5$ considering (a) HOLD SNM degradation percentage, (b) READ SNM degradation percentage, (c) Read failure probability, (d) WM improvement, (e) Write failure probability, (f) Leakage power, and (g) The 1-to-0 flip critical charge reduction

5.4.3 Factors Affecting the Proposed ABB Performance

1. **The Effect of Temperature on the ABB Performance**

   The ABB design is performed at a temperature $T = 120^oC$ which is the worst case operating condition for the SRAM cells. When the operating temperature decreases, $|V_{tp}|$ increases by $(\Delta|V_{tp}|_T + \Delta|V_{tp}|_{NBTI})$, where $\Delta|V_{tp}|_T$ is the $|V_{tp}|$ increase due to temperature decrease and $\Delta|V_{tp}|_{NBTI}$ is the $|V_{tp}|$ increase due to NBTI. Decreasing the operating temperature results in increasing $\Delta|V_{tp}|_T$ [22] and decreasing $\Delta|V_{tp}|_{NBTI}$ (because K_{DC} is a function of temperature) [88]. This $|V_{tp}|$ change is sensed by the ABB sensing circuit and the corresponding body bias voltage is generated. Therefore, the ABB circuit compensates also for temperature variations. Table 5.4 shows the effect of the temperature on the ABB performance in mitigating NBTI and process variations impacts. It is evident from this table that the ABB circuit is working effectively as the temperature varies.

2. **The Effect of Unequal Gate Input Probabilities on the ABB Performance**

   All the above simulation results are performed by using equal gate input probabilities (i.e., $p_L = p_R = 0.5$). The effect of unequal gate input probabilities on the proposed ABB performance is tabulated in Table 5.5.

   The gate input probabilities $p_L = 0$ and $p_R = 1$ means that V_L = '0' and V_R = '1' over the 10 years aging time. This results in maximum NBTI degradation in the right PMOS transistor, M_{pR}, and no degradation in M_{pL}. Accordingly, the highest SNM degradation, the highest read failure probability, and highest $Q_{critical}$ reduction occur in this situation for the NBB and the ABB cases. On the other side, the highest WM improvement and the lowest write failure probability occur in this case. Since the leakage current is measured through the OFF transistor (i.e., M_{pL} in this case) which is not impacted by NBTI, the leakage power equals the same value at zero aging time (i.e., leakage power = 14.1 nW) with no leakage reduction. The leakage current increases by 30% for the ABB case because the ABB circuit applies the maximum body bias voltage for both PMOS transistors although the transistor M_{pL}, through which the leakage current is calculated in this case, is not impacted by NBTI. This forward body bias adoption for M_{pL} results in increasing the leakage current of the cell.

   The gate input probabilities $p_L = 0.25$ and $p_R = 0.75$ results in unequal degradation in the two PMOS transistors (i.e., M_{pL} is less degraded than M_{pR}). Accordingly, the SRAM parameters are dependent on the SRAM cell data status at 10 years aging time. Therefore, the SRAM parameters are calculated for the SRAM status when V_L = '0' and V_R = '1' and also when V_L = '1' and V_R = '0'. Following that, the SNM, WM, and $Q_{critical}$ is calculated as the minimum of these two statuses whereas

Table 5.4: Post layout simulation results for $T = 0^oC$, $T = 60^oC$, and $T = 120^oC$ considering the NBB and the ABB cases at an aging time of 10 years with $p_L = p_R = 0.5$

T =		120^oC	60^oC	0^oC
NBB	HOLD SNM degradation (%)	4.3	3.9	3.8
	READ SNM degradation (%)	10.9	8.7	6.7
	READ failure probability (%)	0.32	0.02	0
	WM improvement (%)	5.7	5.3	5.0
	Write failure probability (%)	0.1	0.08	0.02
	Leakage power (nW)	12.59	1.85	0.11
	$Q_{critical}$ reduction (%)	12.7	8.1	7.6
ABB	HOLD SNM degradation (%)	1.0	0.96	0.94
	READ SNM degradation (%)	2.6	2.2	1.4
	READ failure probability (%)	0.05	0	0
	WM improvement (%)	1.5	1.4	1.4
	Write failure probability (%)	0.06	0.05	0
	Leakage power (nW)	13.5	1.95	0.15
	$Q_{critical}$ reduction (%)	3.4	1.4	1.3

the leakage power is calculated as the maximum of these two statuses. The failure probabilities are calculated by using the following equation [77] where the failure probability is denoted by FP:

$$FP = FP|_{V_L='0' \text{ and } V_R='1'} \times p_R + FP|_{V_L='1' \text{ and } V_R='0'} \times p_L \tag{5.18}$$

It is evident from Table 5.5 that the ABB circuit reduces the NBTI and process variations impacts effectively for the unequal input gate probabilities. In addition,

Table 5.5: Post layout simulation results for different gate input probabilities considering the NBB and the ABB cases at an aging time of 10 years with $T = 120^oC$

p_L/p_R		**0.5/0.5**	**0/1 or 1/0**	**0.25/0.75 or 0.75/0.25**
NBB	HOLD SNM degradation (%)	4.3	7.3	5.3
	READ SNM degradation (%)	10.9	14.1	12.2
	READ failure probability (%)	0.32	0.4	0.33
	WM improvement (%)	5.7	7.6	4.6
	Write failure probability (%)	0.1	0.06	0.1
	Leakage power (nW)	12.59	14.1	12.8
	$Q_{critical}$ reduction (%)	12.7	15.5	14
ABB	HOLD SNM degradation (%)	1.0	3.5	2.3
	READ SNM degradation (%)	2.6	5.8	4.5
	READ failure probability (%)	0.05	0.09	0.07
	WM improvement (%)	1.5	3.1	0.22
	Write failure probability (%)	0.06	0.04	0.08
	Leakage power (nW)	13.5	18.3	14.1
	$Q_{critical}$ reduction (%)	3.4	9.9	5.6

the cases ($p_L = 1$ and $p_R = 0$) and ($p_L = 0.75$ and $p_R = 0.25$) provide similar results to the cases ($p_L = 0$ and $p_R = 1$) and ($p_L = 0.25$ and $p_R = 0.75$), respectively, due to the SRAM symmetry.

3. **The Adoption of the Proposed ABB to Larger SRAM Arrays**

In all the above simulation results, the proposed ABB circuit is adopted to a 512 SRAM cells column. The same simulation results are obtained when the proposed ABB circuit is adopted to a 1024 SRAM cells array (i.e., 2 columns, each column consists of 512 SRAM cells). However, the proposed ABB circuit fails in providing the correct body bias voltage when adopted to 3 SRAM columns array (i.e., 1536 SRAM cells).

Accordingly, the ABB circuit output must be buffered to ensure low output impedance and to be able to drive a larger number of SRAM cells. The voltage buffer is implemented by an operational amplifier as a unity-gain amplifier. The buffered ABB output is able to drive up to 11 SRAM columns, each column consists of 512 SRAM cells.

Correspondingly, several buffers are utilized to supply the output of the proposed ABB circuit to cover the whole SRAM array. For example, each buffer output is applied to a 4K SRAM array (i.e., 8 SRAM columns, each column consists of 512 SRAM cells). Adopting only one ABB circuit with multiple buffers (Global ABB circuit) is unable to compensate for the systematic WID variations. In order to compensate for these WID variations, one buffered ABB circuit should be adopted to each 4K SRAM array (Local ABB circuits) as shown in Figure 5.23.

Post layout simulations are conducted again for the 4K SRAM array with a buffered ABB circuit and the results are approximately the same as the results obtained when the unbuffered ABB circuit is adopted to a 512 SRAM cells column (the difference between these simulation results is less than 0.5%).

Figure 5.23 displays the layout of a 4K SRAM array with the buffered ABB circuit. The layout area of the 4K SRAM is 6065 μm^2 whereas the buffered ABB layout area is 359 μm^2. Thus, the SRAM area is increased by 5.9% with the adoption of local ABB circuits with a granularity level of 4K SRAM array. Increasing the granularity level (i.e., 16K SRAM array) by using one ABB circuit with 4 voltage buffers, results in reducing the area overhead.

Therefore, there is a trade-off between the granularity level used and the associated area overhead. Also, the ability of the ABB circuit in compensating for WID systematic variations is reduced by increasing the granularity level.

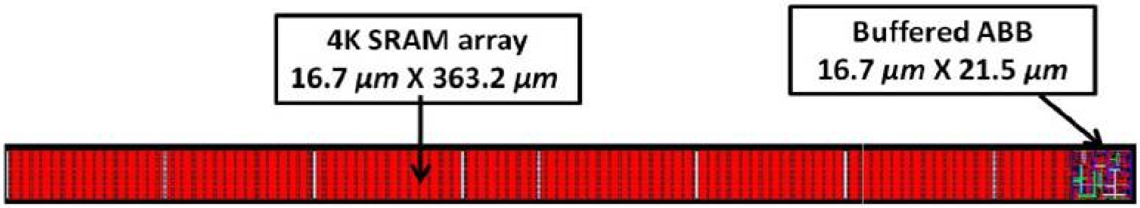


Figure 5.23: Layout of the proposed buffered ABB adopted to a 4K SRAM array (8 columns, each column consists of 512 SRAM cells)

5.5 Summary

The proposed D-ABB and LD-ABB circuits consist of threshold voltage sensing circuits and a direct controller that generates the required body bias voltages to compensate for process variations. The main advantage of the proposed D-ABB and LD-ABB circuits is the low area overhead compared to the previous state-of-art ABB techniques [41, 46]. Therefore, they can be used at a smaller granularity level. In addition, no ADC or DAC is required in the proposed D-ABB and LD-ABB circuits implementations. Accordingly, the proposed ABB circuits are resolution free compared to the previous state-of-art ABB techniques [41, 46]. The effectiveness of the proposed D-ABB in process variations compensation, when used globally and locally, is proved by using post layout simulation results and test chip measurements. The lower area overhead of the proposed D-ABB and LD-ABB circuits helps in utilizing them at smaller granularity levels to increase the circuit robustness and yield.

The proposed D-ABB circuit is also adopted to reduce the impacts of the NBTI aging and process variations on the SRAM cells. Post layout simulation results, referring to an industrial hardware-calibrated 65nm CMOS technology transistor model, show that the proposed ABB compensates effectively for NBTI and process variations. For example, the proposed ABB reduces the read failure probability from 0.32% to 0.05% and the SNM degradation from 10.9% to 2.6% at 10 years aging time. In addition, the proposed ABB enhances the soft errors immunity of the SRAM cell by reducing the critical charge degradation from 12.7% to 3.4% at 10 years aging time. Accordingly, the adoption of the D-ABB to the SRAM array improves the SRAM robustness and yield.

Chapter 6

Negative Capacitance Circuits for Statistical Yield Improvement

In this chapter, new negative capacitance circuits are developed, for the first time, to statistically improve the timing yield of high performance circuits and the read access yield of an SRAM column. The highly capacitive nodes of the wide fan-in high performance dynamic circuits and SRAM bitlines limit the performance of these circuits. In addition, process variations mitigation by statistical gate sizing increases this capacitance further and fails in achieving the target yield improvement. The proposed negative capacitance circuits reduce the overall parasitic capacitance of these highly capacitive nodes. These negative capacitance circuits are adopted to wide fan-in dynamic circuits for timing yield improvement up to 99.87% and to SRAM arrays for read access yield improvement up to 100%. The area and power overheads of these new negative capacitance circuits are amortized over the large die area of the microprocessor and the SRAM array. The effectiveness of the new negative capacitance circuits is verified by post layout simulation results and test chip measurements using 65nm CMOS technology.

This chapter is organized as follows. In Section 6.1, an introduction about the statistical yield improvement techniques is introduced. The proposed negative capacitance circuits are presented in Section 6.2. Sections 6.3 and 6.4 prove the efficiency of the proposed negative capacitance circuits, for yield improvement, by applying them to a high performance circuit block and an SRAM column, respectively. Finally, in Section 6.5, some conclusions are drawn.

6.1 Introduction

As discussed in Chapter 2, statistical design aims at changing the circuit parameters at the design phase statistically to reduce the impact of process variations and increase the circuit robustness and yield. The statistical parametric yield improvement begins with sizing the circuit for a target parameter, Φ_o, which may represent the timing delay or the power consumption. Following that, a Monte Carlo statistical analysis is conducted to calculate the standard deviation of the Φ variability, σ. The main idea of the parametric yield improvement is to shift the Φ pdf, centered around its mean Φ_o, to a new Φ pdf with a mean $\Phi_o^{'}$. The relationship between $\Phi_o^{'}$ and Φ_o is given by:

$$\Phi_o^{'} = \Phi_o \pm n \; \sigma \tag{6.1}$$

where σ is the standard deviation of the Φ variability around Φ_o, and "n" depends on the desired parametric yield, Y_o, and is obtained from the normal distribution tables. As shown in Figure 6.1, if the parametric yield, Y_o, represents the percentage of samples that have $\Phi \quad \leq \quad \Phi_o$ (i.e., when Φ represents the timing delay), the '$-$' sign is used in (6.1), whereas if the parameter yield, Y_o, represents the percentage of samples that have $\Phi \quad \geq \quad \Phi_o$ (i.e., when Φ represents the frequency), the '+' sign is used in (6.1).

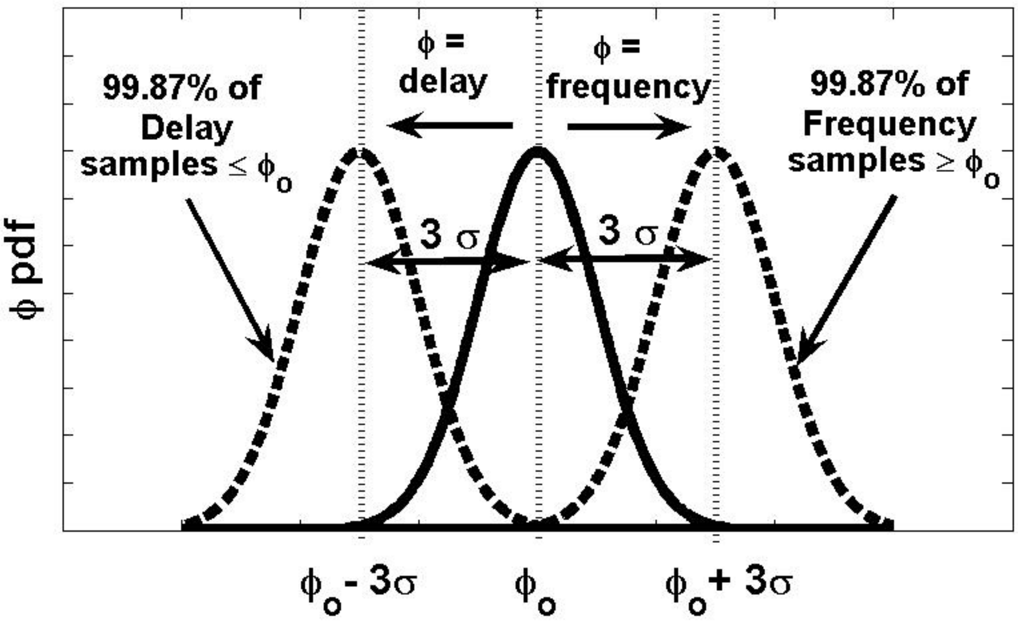


Figure 6.1: Statistical yield improvement for different circuit parameters.

In chapter 2, The parameter Φ is the delay and its reduction from Φ_o to $\Phi_o^{'}$ is accomplished by iterative statistical gate sizing. In this chapter, the reduction (or increase) of the parameter Φ to Φ_o is accomplished by adding a negative capacitance circuit at highly capacitive nodes to reduce the total parasitic capacitance. Two different benchmark circuits are selected in this chapter: (1) the highly capacitive output node of a wide fan-in dynamic OR gate and (2) the highly capacitive bitline of an SRAM column. In the first case, Φ is the delay whereas in the second case, Φ is the differential voltage generated between the SRAM column bitlines. The utilization of statistical gate sizing with these highly capacitive output nodes circuits increases this capacitance further and fails to achieve the target yield.

6.2 Proposed Negative Capacitance Circuits

In this section, two negative capacitance circuit implementation techniques are explained. The first technique is based on the Miller equivalent of a non-inverting amplifier with a feedback capacitance. The second technique is based on the Negative Impedance Converter (NIC) loaded with a capacitance.

6.2.1 Miller Effect Based Negative Capacitance Circuit

The negative capacitance circuit is designed by using a capacitance, C_F, connected between the input and output terminals of a non-inverting amplifier with gain A as displayed in Figure 6.2.a. Applying the Miller effect on this circuit results in the equivalent circuit in Figure 6.2.b. The input equivalent capacitance of this circuit, C_{NEG}, is given by:

$$C_{NEG} = C_F\ (1 - A) \tag{6.2}$$

Therefore, when the amplifier gain, A, is larger than unity, C_{NEG} takes on negative values and a negative capacitance circuit is developed.

In (6.2), the negative capacitance C_{NEG} is achieved only and only if the amplifier gain, A, is constant and independent of frequency (i.e., the amplifier bandwidth is sufficiently larger than the benchmark circuit maximum operating frequency). This means that the closed loop response of the amplifier in the time domain is faster than that of the benchmark circuit. Unfortunately, the requirement to increase the speed of the amplifier results in increasing the power overhead of the negative capacitance circuit. The non-inverting amplifier is designed by using (1) a differential-pair amplifier and (2) a two-inverters buffer amplifier as follows:

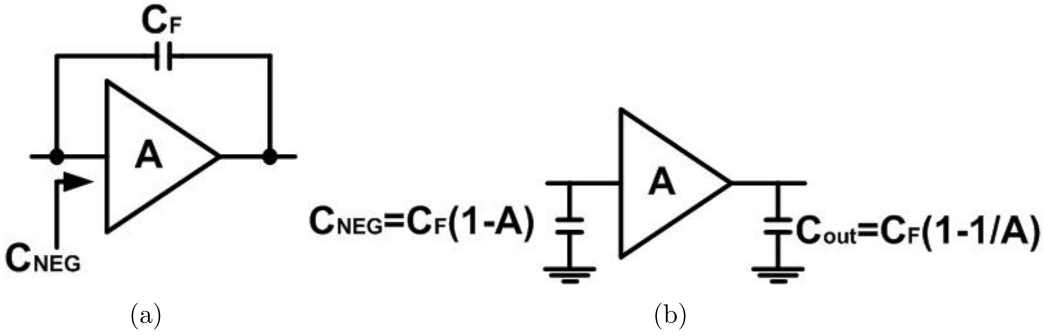


Figure 6.2: (a) The negative capacitance implementation using a non-inverting amplifier with a feedback capacitance and (b) The Miller equivalent circuit of (a) [150].

1. **Differential-Pair Amplifier Based Negative Capacitance (DA-NC) Circuit**

   The differential amplifier, used in Chapter 5 in the implementation of the D-ABB circuit and shown in Figure 5.7, is employed with a feedback capacitance, C_F, to implement the Miller effect based negative capacitance circuit. The gain, A, is controlled by the resistors ratio R_F/R_I. Considering the effect of the input capacitance of the DA-NC circuit, C_{M1}, (6.2) is rewritten as $C_{NEG} - C_{M1} = C_F\ (1 - A)$. For example, to achieve a negative capacitance, $C_{NEG} = -\ 2$ fF with an input capacitance $C_{M1} = 1$ fF, if the feedback capacitance, $C_F = 1$ fF, the required differential-pair amplifier gain, A is 4.0. In addition, the amplifier bandwidth must be sufficiently larger than the benchmark circuit operating frequency to achieve a negative capacitance according to (6.2) which results in increasing the negative capacitance power overhead and reducing the amplifier gain due to the constant gain-bandwidth product of the amplifier.

2. **Buffer Based Negative Capacitance (B-NC) Circuit**

   Figure 6.3 displays the implementation of a digital two-inverters buffer based negative capacitance (B-NC) circuit. The gain of this two-inverters buffer, A_b, is illustrated in Figure 6.4.a versus the buffer input voltage (V_{in}), where V_{DD_H} is the buffer supply voltage, V_M is the buffer threshold voltage (i.e., the voltage value that results if the buffer input and output terminals are connected together), V_{IL} is the maximum buffer input voltage that results in zero output voltage, and V_{IH} is the minimum buffer input voltage that results in an output voltage equals to the supply voltage, V_{DD_H}. The maximum gain, $A_{b_{max}}$, occurs at the buffer threshold voltage, V_M. Figure 6.4.b displays the corresponding total capacitance (including the negative capacitance), C'_{out}, as a function of V_{in}, where C_{inv} is the input capacitance of the buffer circuit, and C_{out} is the total capacitance without the negative capacitance adoption.

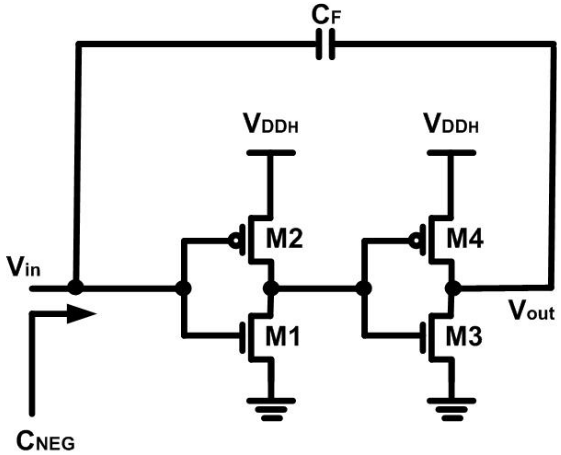


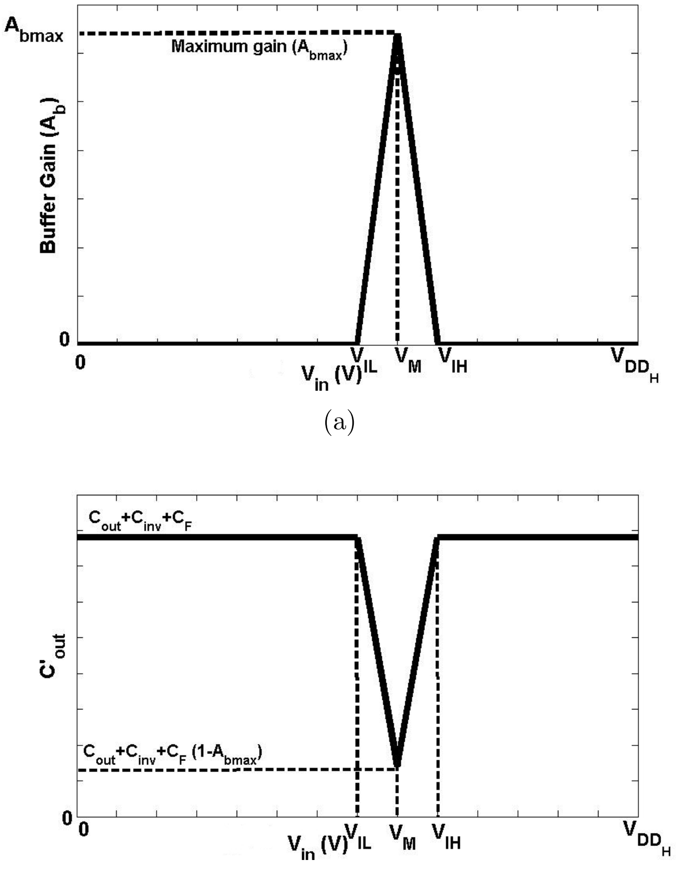
Figure 6.3: The proposed two-inverters buffer based negative capacitance (B-NC) circuit implementation.

To calculate the relationship between V_{DD_H}, the buffer supply voltage, and V_{DD_L}, the benchmark circuit supply voltage, the well-known alpha-power law model for the transistors current [114] is adopted. According to this model, the threshold voltage, V_M, is given by (3.15)[114]. Thus, by equating V_M to V_{DD_L} (to have the maximum gain at V_{DD_L} to allow the highly capacitive node to see the maximum negative capacitance, when it starts to discharge from V_{DD_L}), the buffer supply voltage, V_{DD_H}, is given by:

$$V_{DD_H} = (1 + \frac{1}{r}) \times V_{DD_L} + |V_{tp}| - \frac{V_{tn}}{r} \tag{6.3}$$

Unfortunately, the B-NC circuit exhibits a large power due to the static power which occurs because the maximum input voltage to the inverter equals V_{DD_L} is unable to turn the inverter PMOS transistor completely OFF as the inverter supply voltage is V_{DD_H} [91]. In order to reduce this static power, the difference between V_{DD_H} and V_{DD_L} should be designed slightly larger than $\mid V_{tp} \mid$.

From (6.3), the difference V_{DD_H}-V_{DD_L} $= (V_{DD_L} - V_{tn})/r + |V_{tp}|$ which implies that by increasing r and using high-V_t buffer transistors, this difference is designed close to $\mid V_{tp} \mid$. For example, if standard-V_t transistors are adopted with V_{tn}=0.342V and $\mid V_{tp} \mid$=0.204V as provided by the 65nm CMOS technology model and assuming r = 2, the difference V_{DD_H}-V_{DD_L} = 0.43V which is higher than $\mid V_{tp} \mid$ by a factor of 2.1X. However, the difference V_{DD_H}-V_{DD_L} = 0.56V (close to $\mid V_{tp} \mid$ value of 0.54V), when high-V_t transistors are utilized with V_{tn}=0.59V, $\mid V_{tp} \mid$=0.54V and r = 10.



(a)

(b)

Figure 6.4: (a)The two-inverters buffer gain, A_b, versus the input voltage, V_{in} and (b) The total capacitance, C'_{out}, when the B-NC is adopted.

Moreover, the B-NC circuit provides a higher gain and correspondingly lower feedback capacitance requirement than that of the DA-NC circuit. However, the B-NC circuit is only suitable when dual supply voltage (dual-V_{DD}) and dual threshold voltage (dual-V_t), are available. It should be noted that the speed of the buffer amplifier must be faster than that of the benchmark circuit to achieve a negative capacitance based on (6.2). Since the utilization of high-V_t transistors (to reduce the power consumption) results in reducing the speed of the buffer amplifier, thus, the buffer transistor sizes should be increased to compensate for this speed reduction. This, in turn, results in increasing the negative capacitance power overhead and more capacitive loading (i.e., larger buffer input capacitance, C_{inv}). Therefore, there is a trade-off between the buffer speed requirement and the associated power overhead.

6.2.2 Negative Impedance Converter (NIC) Based Negative Capacitance Circuit

The Negative Impedance Converter (NIC) is a two-port circuit whose input impedance is the negative of the load impedance at its output port, as shown in Figure 6.5 [150, 151]. When the NIC circuit is loaded with a capacitance, C$_L$, at its output node, an equivalent negative capacitance is seen at the input node. Therefore, a negative capacitance, C$_{NEG}$, circuit is implemented. The value of this C$_{NEG}$ is given by:

$$C_{NEG} = -\beta \; C_L \tag{6.4}$$

where β is dependent on the NIC circuit implementation.

Similarly, in (6.4), the negative capacitance C$_{NEG}$ is achieved only and only if the factor β is constant with frequency (i.e., the NIC bandwidth is sufficiently larger than the benchmark circuit maximum operating frequency). This means that the closed loop response of the NIC in the time domain is faster than that of the benchmark circuit. Unfortunately, the requirement to increase the speed of the NIC results in increasing the power overhead of the negative capacitance circuit.

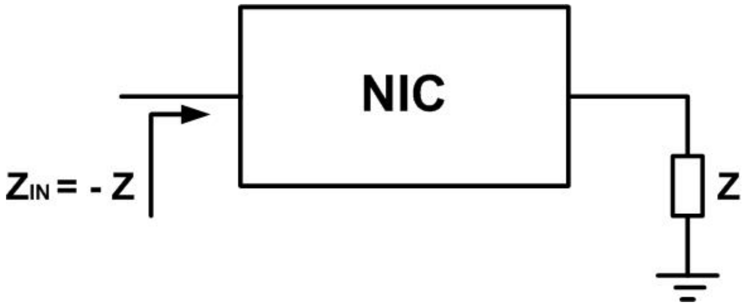


Figure 6.5: The NIC based negative capacitance implementation [151].

The NIC based negative capacitance circuit is implemented by using the Positive Second Generation Current Conveyor (CCII+) [151, 152]. This Current Conveyor based Negative Capacitance (CC-NC) circuit block is shown in Figure 6.6.a, and its implementation is illustrated in Figure 6.6.b.

The CCII+ is a three terminal analog circuit with terminals X, Y, and Z. The function of the CCII+ is defined as [152]:

$$\begin{bmatrix} v_x \\ i_y \\ i_z \end{bmatrix} = \begin{bmatrix} 0 & +1 & 0 \\ 0 & 0 & 0 \\ +1 & 0 & 0 \end{bmatrix} \begin{bmatrix} i_x \\ v_y \\ v_z \end{bmatrix} \tag{6.5}$$

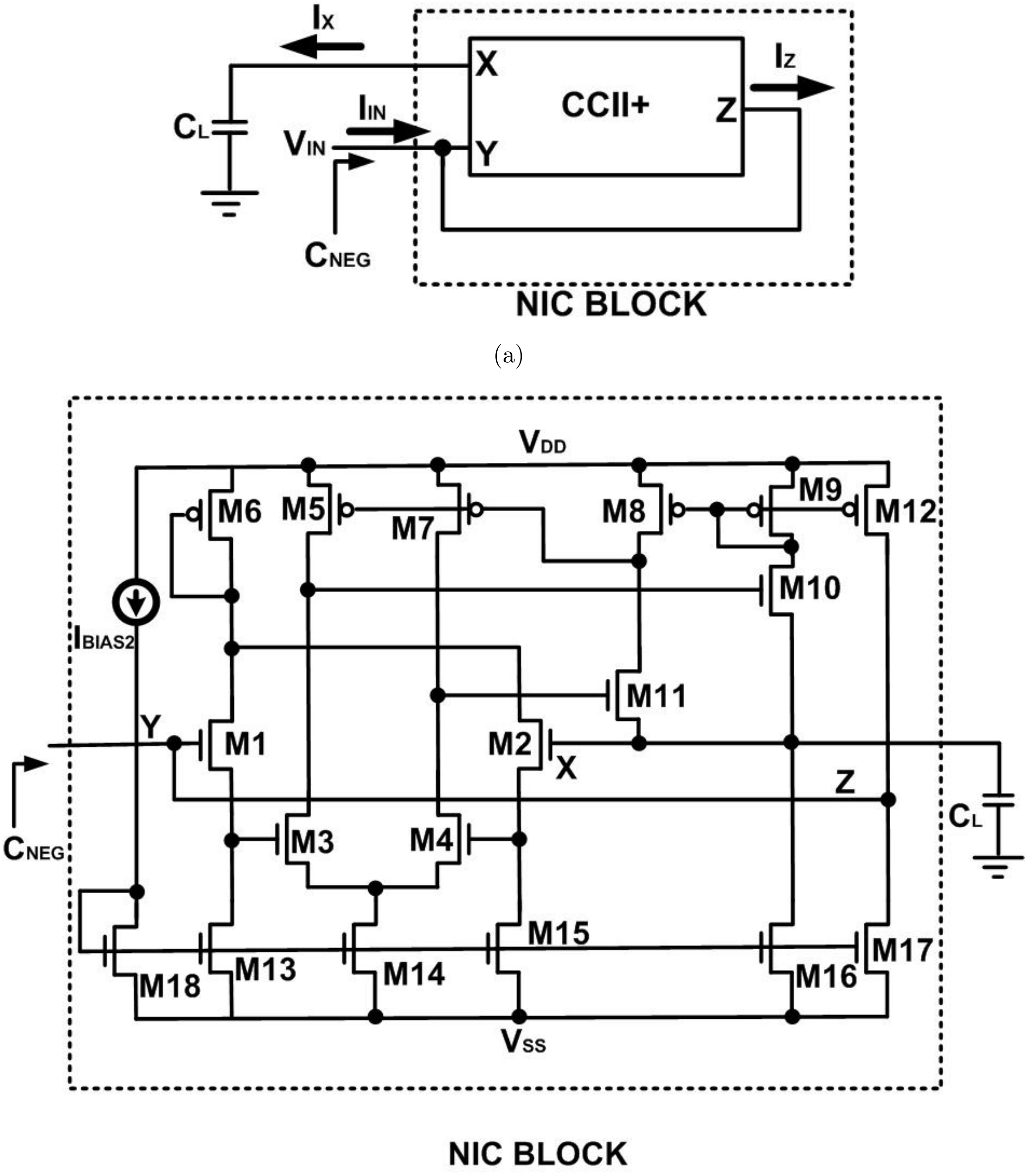
Therefore, the CCII+ performs a voltage conveying action from terminal Y to terminal X (i.e., $V_X = V_Y$), and a current conveying action from terminal X to terminal Z (i.e., $I_Z = I_X$). In addition, no current is flowing into terminal Y. If ε_v, the error in conveying the voltage at node Y to node X, and ε_i, the error in conveying the input current at node X to node Z, are considered, the input voltage, in Figure 6.6.a, is given by: $V_{IN} = V_Y = V_X/(1-\varepsilon_v) = I_X/[sC_L(1-\varepsilon_v)]$. The input current is given by: $I_{IN} = -I_Z = -I_X(1-\varepsilon_i)$, since the current I_Y=0. Correspondingly, the input impedance at terminal Y is given by: $Z_{IN} = V_{IN}/I_{IN} = -1/[sC_L(1-\varepsilon_v)(1-\varepsilon_i)] = -1/[sC_{NEG}]$. Thus, the value of C_{NEG} is given by: $C_{NEG} = -(1-\varepsilon_v)(1-\varepsilon_i)C_L$. Hence, the constant, β, is given by:

$$\beta = (1-\varepsilon_v)\ (1-\varepsilon_i) \tag{6.6}$$

where ε_v is the error in conveying the voltage at node Y to node X, and ε_i is the error in conveying the input current at node X to node Z. Considering the effect of the input capacitance of the CC-NC circuit, C_Y, (6.4) is rewritten as $C_{NEG} - C_Y = -\beta\ C_L$.

The main advantage of the CCII+ over the differential-pair amplifier is that the CCII+ does not exhibit a constant gain-bandwidth product [153], whereas the differential-pair amplifier does. For example, if the differential-pair amplifier gain-bandwidth product is 5GHz, and the input signal frequency is 5GHz, the maximum gain, achieved by this amplifier, is limited to unity and the DA-NC fails to implement a negative capacitance circuit. Accordingly, the CC-NC is the best alternative for high frequency input signals, because the CCII+ is not prone to the constant gain-bandwidth product limitation. However, the CC-NC power dissipation and area are larger than those in the DA-NC. In addition, from (6.2) and (6.4), the capacitance, C_L, used in the CC-NC, is larger than the capacitance, C_F, used in the DA-NC to implement the same negative capacitance value.

Moreover, the CCII+ speed must be faster than the benchmark circuit speed to achieve a negative capacitance according to (6.4) which results in increasing the negative capacitance power overhead and making the factor β less than unity. Accordingly, a larger C_L value is required to compensate the reduction of the factor β. The advantage of the CC-NC over the DA-NC is that the reduction of the factor β of the CCII+ (due to the large bandwidth requirements) is much less than the reduction of the gain A of the amplifier because the CCII+ is not prone to the constant gain-bandwidth product limitation while the differential amplifier is.



(a)

(b)

Figure 6.6: (a) The NIC based negative capacitance circuit using CCII+ and (b) The circuit implementation of this negative capacitance [152].

6.3 Application of the Negative Capacitance Circuits to High-Performance Dynamic Circuits

Dynamic gates are preferred in the design of high-performance modules in modern microprocessors due to the relatively high speed of dynamic gates compared with that of standard CMOS gates. One of the important applications of high performance dynamic circuits is the register files. These register files are one of the most essential modules in the critical path of modern microprocessors [154, 155]. The basic operation of a register file is to store temporary and intermediate variables that are being used in the execution of a sequence of instructions. Figure 6.7 depicts the block diagram of the Intel Pentium 4 processor architecture [154, 156]. In this processor, two register files are employed in the data path, which are marked with a dotted box in Figure 6.7. These register files are the integer register file, denoted by Integer RF, and the floating point register file, denoted by FP RF. Data are read from or written to these register files with each instruction execution. Therefore, fast register file architectures are crucial in achieving a high-performance operation in microprocessors [154].

Figure 6.8.a demonstrates the block diagram of a simplified register file, which is composed of an array of registers, a read port, and a write port. Typically, register files have multiple read and write ports and also have many more registers. Read and write ports are generally implemented by using multiplexers (MUXs) and de-multiplexers (DE-MUXs) circuits, respectively. Typically, these MUXs and DE-MUXs circuits are realized by utilizing OR and inverter gates, as shown in Figure 6.8.b. This figure illustrates a simple 4X1 MUX with 4-input lines (D_0,D_1,D_2, and D_3), 2-bit address lines (S_0 and S_1), and one output (Out). Therefore, a register file with 2^n registers requires n-bit address lines, and hence, $(n + 1)$-input intermediate OR gates and a 2^n-input output OR gate. As a result, for large register files, wide fan-in dynamic OR gates are required for address coding/decoding. Also, the propagation delay of the wide fan-in dynamic OR gate increases linearly with fan-in [157]. This makes the wide dynamic OR gate an excellent choice for the implementation of high-performance modules.

As discussed in Chapter 2, the wide fan-in dynamic OR gate suffers from noise sensitivity due to the sub-threshold leakage current flowing through the evaluation pull down network [40], which increases with technology scaling. Accordingly, noise immunity has become a great concern, especially, in the design of wide fan-in gates. According to [40, 158, 159], the noise immunity is quantified by using the Unity Noise Gain (UNG) metric. The UNG is defined as the amplitude of input noise V_{noise} that causes an equal amplitude noise pulse at the OR gate output node F (i.e., UNG = V_{noise} such that V_{noise}=V_F), as displayed in Figure 6.9. UNG captures the critical input noise strength, as any noise pulse larger than UNG is amplified due to the nonlinear behavior of the transistor [159]. Thus, all the inputs X1-X64 are driven by noise pulses with the same duration of 100 psec

[158] and varying amplitude. The pulse amplitude is swept till a glitch, with the same amplitude of the inputs, occurs at the output node F.

The increased sub-threshold leakage current flowing through the evaluation pull down network forces the circuit designers to upsize the keeper transistor to improve the circuit robustness and noise immunity, at the expense of larger delay and power dissipation, when the output node is discharging to ground. In addition, due to the increased process variations in scaled technologies, the dynamic circuit delay exhibits a substantial variability around its nominal value. This delay variability results in violating the timing constraints, and correspondingly, causes a timing yield loss. In this section, the proposed negative capacitance circuits are adopted to statistically improve the timing yield under process variations.

In this section, a negative capacitance circuit is connected to the highly capacitive output node, G, of a wide fan-in 64-input dynamic OR gate, as shown in Figure 6.9. This negative capacitance connection reduces node G parasitic capacitance, and correspondingly, improves the timing yield without changing the gate sizing. The timing yield improvement is achieved by calculating the amount of the negative capacitance that should be added to the parasitic capacitance, at node G to shift the dynamic OR gate delay pdf center from A$_o$ to $A_o^{'}$ without affecting the gate sizing. In addition, the effect of the proposed negative capacitance circuits on the circuit robustness and noise immunity is discussed.

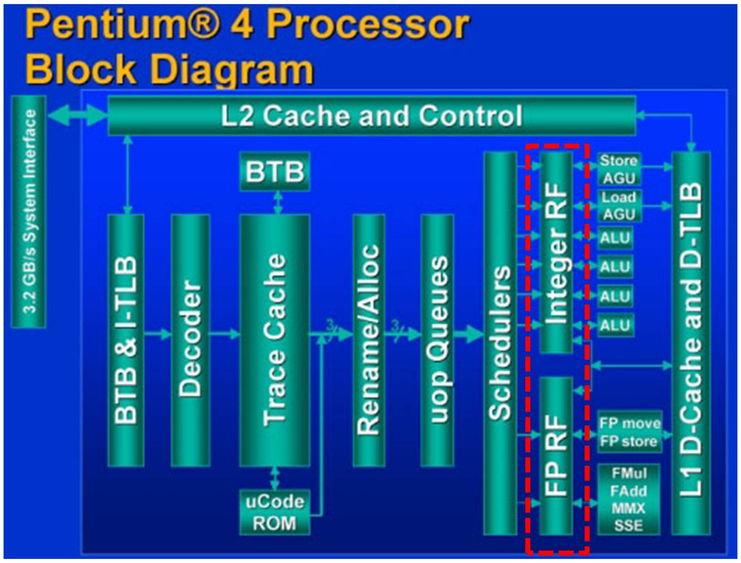


Figure 6.7: The Intel Pentium 4 processor block diagram. [156].

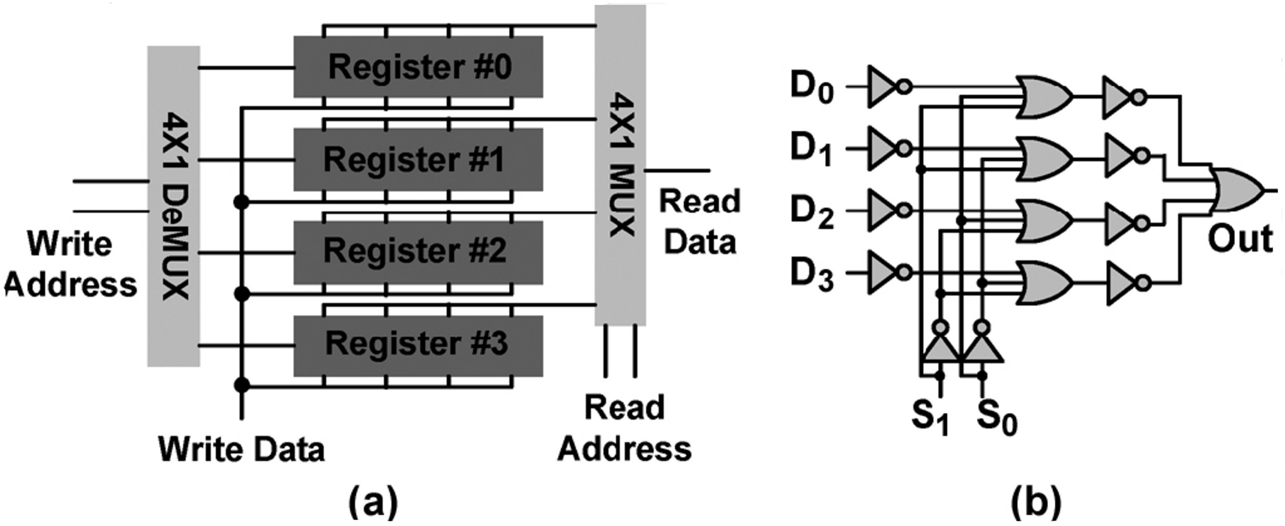


Figure 6.8: (a) Block diagram of a simplified register file and (b) Read port implemented using 4 X 1 MUX [154].

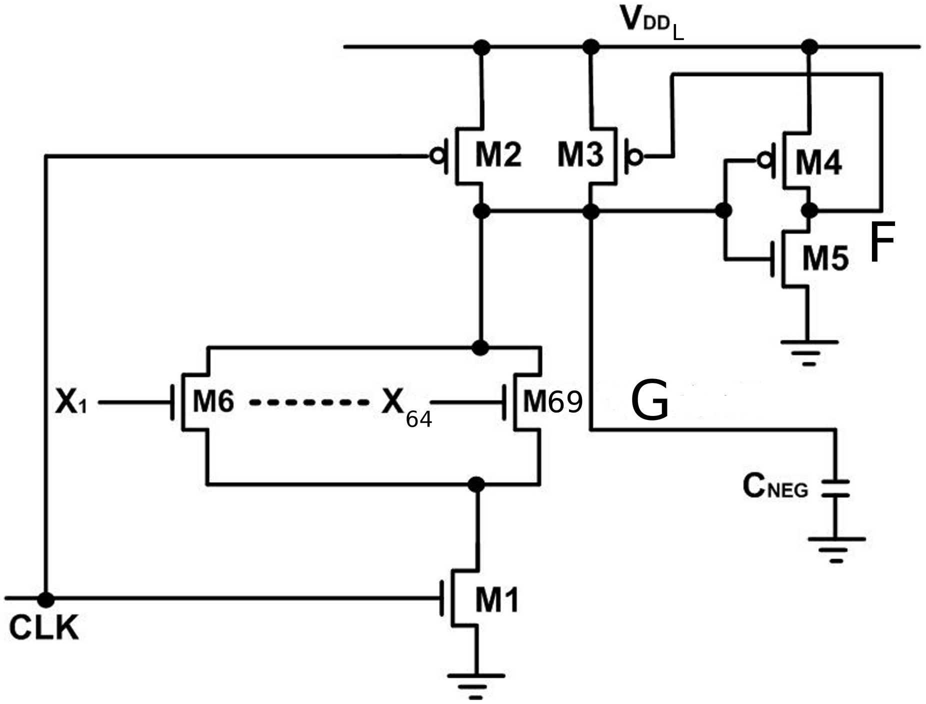


Figure 6.9: The 64-input dynamic OR gate circuit with a negative capacitance employed at the highly capacitive output node, G. Transistor M3 is the keeper transistor.

6.3.1 Statistical Timing Yield Improvement using Negative Capacitance

Assume that the dynamic OR gate circuit is designed such that the nominal delay is the target delay, A_o (This can be performed at any design corner), the circuit delay, A_o, is given by [91]:

$$A_o = \zeta \; C_{out} \tag{6.7}$$

where ζ is a proportionality constant, dependent on the output resistance of the dynamic OR circuit, and C_{out} is the circuit parasitic output capacitance.

Due to the process variations, this dynamic OR gate circuit delay is normally distributed around this nominal value. Therefore, the resulting timing yield is close to ∼50%, as shown in Figure 6.10. It should be noted that the impact of the process variations on C_{out} is neglected with respect to the impact on the output resistance [89]. This is due to the fact that the output resistance depends on the transistors threshold voltage, the main source of the variability [27, 33, 89]. To improve the timing yield, the delay variability, which is centered around A_o, is shifted to a new center $A_o^{'}$ given by:

$$A_o^{'} = A_o - n \; \sigma^{'} \tag{6.8}$$

where $\sigma^{'}$ is the standard deviation of the delay variability around $A_o^{'}$, and "n" is dependent on the desired timing yield, Y_o. It should be mentioned that (6.8) can be used only when $\sigma^{'}$ is known in advance, which is satisfied when the negative capacitance circuit is adopted. In statistical gate sizing [34, 35, 145], (6.8) should not be used because $\sigma^{'}$ is not known in advance.

Figure 6.10 illustrates how the timing yield is improved by shifting the delay pdf to a shorter mean delay, $A_o^{'}$. The delay, A_o, is reduced by adding a negative capacitance at the output node of the dynamic OR gate. The addition of the negative capacitance, C_{NEG}, at the circuit output node results in a modified output capacitance, $C_{out}^{'}$, which is given by:

$$C_{out}^{'} = C_{out} + C_{NEG} \tag{6.9}$$

and accordingly, the modified circuit delay, $A_o^{'}$, is expressed as:

$$A_o^{'} = \zeta \; C_{out}^{'} \tag{6.10}$$

By using (6.7-6.10), the negative capacitance, C_{NEG}, which achieves the desired timing yield improvement, is expressed as:

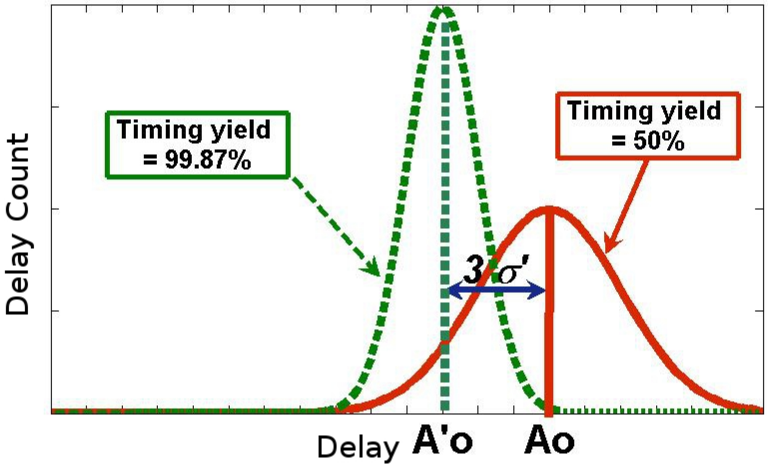


Figure 6.10: The timing yield improvement is obtained by shifting the delay pdf center from A$_o$ to A$_o^{'}$. In this case, $A_o^{'} = A_o - 3\ \sigma^{'}$ for a timing yield of 99.87%.

$$C_{NEG} = \frac{-\ n\ \sigma^{'}}{\zeta} \tag{6.11}$$

From (6.7), the delay A$_o$ variability, ΔA$_o$ = $\Delta\zeta$ C$_{out}$, and from (6.10), the delay A$^{'}{}_o$ variability, ΔA$^{'}{}_o$ = $\Delta\zeta$ C$^{'}{}_{out}$, assuming that the capacitances, C$_{out}$ and C$^{'}{}_{out}$, are constants from the variability perspective. From (6.9), C$^{'}{}_{out}$ < C$_{out}$ because C$_{NEG}$ has a negative value. Therefore, ΔA$^{'}{}_o$ < ΔA$_o$ which explains why the adoption of the negative capacitance reduces the delay variability, if the negative capacitance circuit variations are neglected. The ratio between $\sigma^{'}$ and σ is obtained by computing:

$$\frac{\sigma^{'}}{\sigma} = \frac{\Delta A_o^{'}}{\Delta A_o} = \frac{C_{out}^{'}}{C_{out}} = 1 + \frac{C_{NEG}}{C_{out}} \tag{6.12}$$

From (6.11) and (6.12), C_{NEG} is given by:

$$C_{NEG} = \frac{-\frac{n\ \sigma}{\zeta}}{1 + \frac{n\ \sigma}{A_o}} \tag{6.13}$$

It should be noted that the negative capacitance, C$_{NEG}$, exhibits some variations due to its circuit implementation. However, these variations contribution to $\sigma^{'}$ is ignored in (6.12) and (6.13) to have an initial guess for the required value of C_{NEG}. This contribution should be calculated because the negative capacitance circuit has different implementations such as DA-NC, CC-NC, and B-NC circuits. If the negative capacitance circuit variability is taken into account, the delay variability, is expressed as follows:

$$\frac{\Delta A_o^{'}}{A_o^{'}} = \frac{\Delta\zeta}{\zeta} + \frac{\Delta C_{NEG}}{C_{out} + C_{NEG}} \tag{6.14}$$

Since the factor $\frac{\Delta\zeta}{\zeta} = \frac{\Delta A_o}{A_o}$, the standard deviation $\sigma^{'}$ is given by:

$$(\frac{\sigma^{'}}{A_o^{'}})^2 = (\frac{\sigma}{A_o})^2 + \underbrace{\frac{(\sigma_{C_{NEG}}/C_{NEG})^2}{((C_{out} + C_{NEG})/C_{NEG})^2}}_{C_{NEG} \quad variability} \tag{6.15}$$

The negative capacitance circuit should be designed such that its variability contribution to $\sigma^{'}$ is small (i.e., the contribution of the factor ζ is dominant). Thus, (6.13) is valid and the negative capacitance contribution to $\sigma^{'}$ is neglected. However, if this contribution is not neglected and taken into account, (6.13) is used as an initial guess and then the following algorithm is adopted:

1) Calculate the initial value of C_{NEG} by using (6.13).

2) Conduct Monte Carlo simulations while the negative capacitance circuit is adopted.

3) Determine the values of A$^{'}{}_o$ and $\sigma^{'}$.

4) Calculate the value of the timing yield. If the timing yield is greater than or equal to 99.87%, the target timing yield improvement is achieved and the algorithm stops. If the timing yield is less than 99.87%, calculate the new C_{NEG} by using the new $\sigma^{'}$ value by using (6.11).

5) Repeat steps 2-4 above

This negative capacitance, C$_{NEG}$, is designed by using the DA-NC, the CC-NC, and the B-NC circuits. In the following section, the three negative capacitance circuits are adopted for the 64-input dynamic OR gate, and the simulation results are discussed. In the design of the three negative capacitance circuits, our objective is to reduce the area/power overheads of these circuits and also their variations contribution, to avoid the use of the iterative solution.

6.3.2 Simulation Results and Discussions

The 64-input dynamic OR gate in Figure 6.9 is utilized as a benchmark case study to verify the proposed timing yield improvement technique. The parasitic capacitances at the intermediate node G are large due to the 64 NMOS transistors diffusion capacitances. Therefore, the negative capacitance, C_{NEG}, is connected to this node, as shown in Figure 6.9. The dynamic OR gate is designed with a nominal high-to-low delay at node G, A$_o$, of 433 psec, UNG of 466mV, and an associated total average power dissipation of 62.8 μW at temperature T = 120^oC with a layout area of 135.8 μm^2. The supply voltage used in the dynamic OR gate design, V_{DD_L}, is 0.8V [160]. The OR gate design is performed by using post layout simulations referring to an industrial hardware-calculated 65nm CMOS technology. Low-V_t transistors are utilized for the OR gate transistors to achieve high performance. The total capacitance at the node G of the OR gate, C$_{out}$, is calculated from the circuit layout, by using the CALIBRE tool provided by Mentor Graphics, and equals 15.1 fF (this layout capacitance includes all the parasitic capacitances coupled to node G). Therefore, the constant ζ, defined in (6.7), equals 28.68 psec/fF.

A 5,000 point Monte Carlo analysis, using the statistical transistor model reported in Appendix D, is conducted. A typical histogram of the OR gate high-to-low delay at node G is shown in Figure 6.11.a. The standard deviation of this delay, σ, is 58.5 psec. Accordingly, the required negative capacitance value, C$_{NEG}$, for a timing yield Y$_o$ = 99.87% (i.e., "n" = 3.0) is obtained from (6.13) and equals - 4.4 fF. Then, the values of A$_o^{'}$ and $\sigma^{'}$ are calculated as 308.2 psec and 41.6 psec, respectively. This negative capacitance is implemented by using the three proposed negative capacitance circuits (i.e., DA-NC, B-NC, and CC-NC). In the following, post layout SPICE transient and Monte Carlo simulations of the OR gate alone and the OR gate with the negative capacitance circuits (DA-NC, CC-NC, and B-NC) connected to node G, are conducted to calculate the delay, the delay variability, and the power overhead.

The adoption of the negative capacitance circuits to the dynamic wide fan-in OR gate is utilized for timing yield improvement of the high-to-low delay at node G. However, the precharge delay (low-to-high delay at node G) is also affected by the negative capacitance adoption. Therefore, the DA-NC and the B-NC circuits are disabled during the precharge phase by adding a tail NMOS transistor driven by the clock signal (This is because the Miller based negative capacitance circuit realizes a positive capacitance when the voltage at node G is rising from 0 to $V_{DD_L}/2$). This results in a slight increase in the precharge delay by 4% and 6% (i.e., the precharge delay increases from 93ps to 97ps and 99ps) due to the DA-NC and the B-NC circuits, respectively. The power and area overheads of these circuits are calculated including this tail transistor. The CC-NC circuit does not suffer from this problem and reduces the precharge delay by 15% (i.e., the precharge delay decreases from 93ps to 81ps). This is because the CC-NC realizes a negative capacitance when the voltage at node G is falling from V_{DD_L} to 0 or rising from 0 to V_{DD_L}.

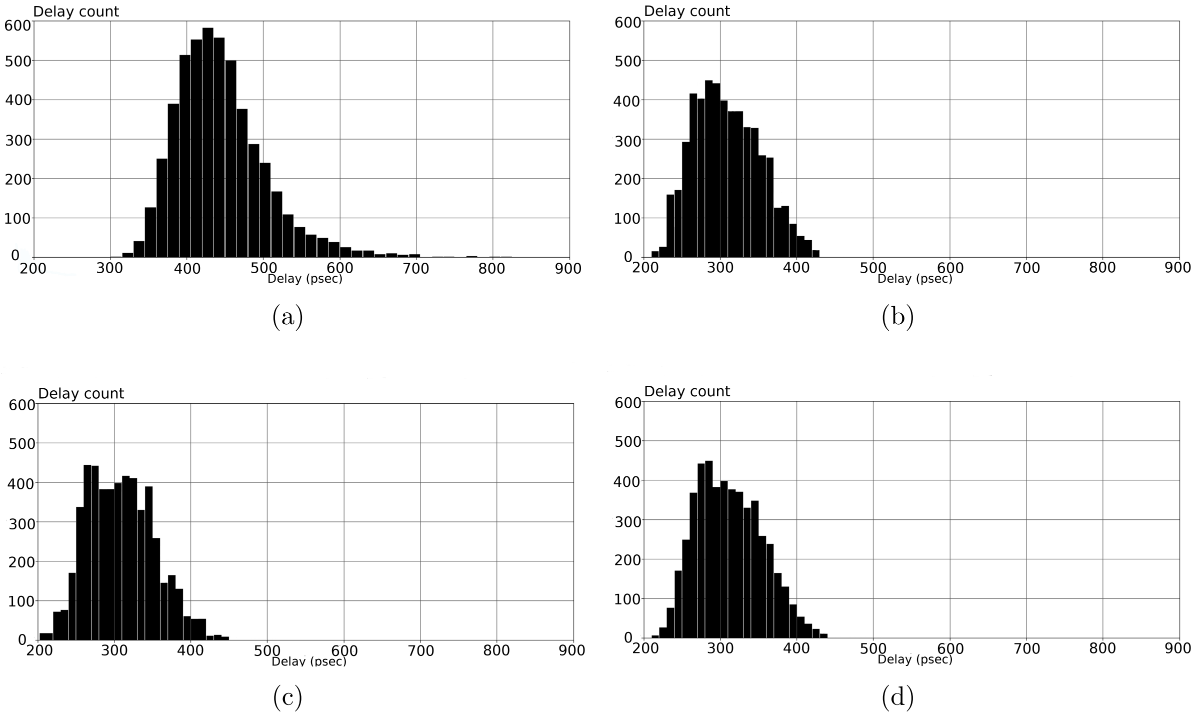


Figure 6.11: Delay histograms for the 64-input dynamic OR gate (a) Before employing the negative capacitance (μ= 433 psec and σ= 58.5 psec), After employing (b) The DA-NC circuit (μ= 315.7 psec and σ= 43.6 psec), (c) The CC-NC circuit (μ= 310 psec and σ= 44.9 psec), and (d) The B-NC circuit (μ= 312.5 psec and σ= 43.9 psec).

1. **DA-NC Circuit**

   In Figure 6.9, the 64-input dynamic OR gate highly capacitive output node, G, is initially pre-charged to the OR gate supply voltage, V_{DD_L}. Then, depending on the inputs, the node G is either maintained at logic '1' or pulled down to ground (logic '0'). The idea of the DA-NC circuit is to allow node G to see a reduced capacitance (Due to the negative capacitance circuit adoption), when discharging from V_{DD_L} to $V_{DD_L}/2$. Therefore, the fall time is reduced due to this negative capacitance adoption. When the output node G reaches $V_{DD_L}/2$, the positive feedback of the keeper transistor (i.e., transistors M3-M5 in Figure 6.9) circuit becomes strong enough to continue discharging node G to ground. Accordingly, the input dynamic range (i.e., the range of the input voltage over which the amplifier exhibits a linear gain) of the differential-pair amplifier, should include the range from V_{DD_L} to $V_{DD_L}/2$.

The DA-NC circuit is adopted to implement the required negative capacitance of - 4.4 fF. The input capacitance of the DA-NC circuit, C_{M1}, is calculated from the layout by using the CALIBRE tool provided by Mentor Graphics and equals 0.5 fF (this means that the actual negative capacitance to be implemented is -4.9 fF to compensate for C_{M1}). The feedback capacitance, C$_F$, is implemented by using a Metal-Insulator-Metal Capacitor (MIM-CAP). The value of C$_F$ is chosen to be 4.9 fF, with an area of $1.96\mu m^2$. Accordingly, the required amplifier gain, A, is 2.0. The amplifier bandwidth, over which the gain, A, is independent of frequency, is calculated to be 3.4 GHz which is sufficiently larger than the NOR gate maximum operating frequency of 2.5 GHz (this maximum operating frequency is calculated based on a high-to-low delay of 308.2psec and a precharge delay of 93psec). It should be noted that if the amplifier gain A = 3.0, for instance, the amplifier bandwidth becomes less than 2.5 GHz and the NOR gate delay is not reduced as expected by adopting the negative capacitance circuit.

From (6.2), the negative capacitance variability is given by:

$$\frac{\Delta C_{NEG}}{C_{NEG}} = \frac{\Delta C_F}{C_F} + \frac{\Delta A}{A-1} \tag{6.16}$$

Thus, the standard deviation of C_{NEG} is given by [89]:

$$(\frac{\sigma_{C_{NEG}}}{C_{NEG}})^2 = (\frac{\sigma_{C_F}}{C_F})^2 + \frac{(\sigma_A/A)^2}{((A-1)/A)^2} \tag{6.17}$$

Monte Carlo simulation results reveal that the ratio of $\sigma^{'}/A^{'}_o$ is calculated from (6.15), by using (6.17), and equals 14.3% whereas if the negative capacitance variability contribution (the second term in (6.15)) is ignored, the ratio of $\sigma^{'}/A^{'}_o$ equals 13.5% (this value is obtained by calculating $\sigma/A_o = 58.5$ psec/ 433 psec). Accordingly, the negative capacitance variation is ignored in this negative capacitance implementation and (6.13) is applied because the error in $\sigma^{'}/A^{'}_o$ when the negative capacitance variability contribution is ignored equals 5.6% (i.e., (14.3-13.5)/14.3) and also because the resulting timing yield is larger than the target yield of 99.87%.

A typical histogram of the OR gate high-to-low delay at node G is shown in Figure 6.11.b. It is evident from Figure 6.11.a and Figure 6.11.b that the DA-NC circuit shifts the delay pdf as required. Figure 6.11.b indicates that 100% of the dynamic OR gate samples have delays less than the target delay of 433 psec. In addition, the adoption of the negative capacitance reduces the delay standard deviation from 58.5 psec to 43.6 psec (25.5% variability reduction). The DA-NC circuit exhibits a total average power consumption of 58.3 μW, including the biasing circuit power consumption. The total power dissipation of the OR gate with the negative capacitance circuit is 118.7 μW and the total layout area equals 239.9 μm^2.

2. **CC-NC Circuit**

   The CCII+ based negative capacitance (CC-NC) circuit in Figure 6.6 is utilized to implement the required negative capacitance of - 4.4 fF. The input capacitance of the CC-NC circuit, C_Y, is calculated from the layout by using the CALIBRE tool provided by Mentor Graphics and equals 1.2 fF (this means that the actual negative capacitance to be implemented is -5.6 fF to compensate for C_Y). In order to design the CCII+ bandwidth sufficiently larger than the OR gate maximum operating frequency of 2.5GHz, the factor $\beta \approx 0.86$. Accordingly, the load capacitance C$_L$ = 6.5 fF implemented by using a MIM-CAP, with an area of 2.7μm^2. Following that, the CCII+ circuit is designed by using V$_{DD}$=0.8V,V$_{SS}$=0V, and I$_{BIAS2}$=20μA.

   From (6.4), the negative capacitance variability is given by:

   $$\frac{\Delta C_{NEG}}{C_{NEG}} = -\frac{\Delta C_L}{C_L} - \frac{\Delta\beta}{\beta} \tag{6.18}$$

   Thus, the standard deviation of C_{NEG} is given by [89]:

   $$(\frac{\sigma_{C_{NEG}}}{C_{NEG}})^2 = (\frac{\sigma_{C_L}}{C_L})^2 + (\frac{\sigma_\beta}{\beta})^2 \tag{6.19}$$

   Monte Carlo simulation results reveal that the ratio of $\sigma^{'}/A^{'}_o$ is calculated from (6.15), by using (6.19), and equals 14.6% whereas if the negative capacitance variability contribution is ignored, the ratio of $\sigma^{'}/A^{'}_o$ equals 13.5%. Accordingly, the negative capacitance variation is ignored in this negative capacitance implementation and (6.13) is applied because the error in $\sigma^{'}/A^{'}_o$ when the negative capacitance variability contribution is ignored equals 7.5% (i.e., (14.6-13.5)/14.6) and also because the resulting timing yield is larger than the target yield of 99.87%.

   A typical histogram of the OR gate high-to-low delay at node G is shown in Figure 6.11.c. It is evident from Figure 6.11.c and Figure 6.11.a that the CC-NC circuit shifts the delay pdf as required. Figure 6.11.c shows that 99.88% of the dynamic OR gate samples have delays less than the target delay of 433 psec. In addition, the delay standard deviation is reduced from 58.5 psec to 45 psec (23% variability reduction). The CC-NC circuit exhibits a total average power consumption of 118.6 μW. This power consumption is larger than that of the DA-NC circuit. The total power dissipation of the OR gate with the negative capacitance circuit is 176.2 μW and the total layout area equals 299.2 μm^2.

3. **B-NC Circuit**

   Similar to the DA-NC circuit, the B-NC circuit should allow node G to see a reduced capacitance (Due to the negative capacitance circuit adoption), when discharging from V_{DD_L} to $V_{DD_L}/2$. Therefore, the fall time is reduced due to this negative capacitance. To achieve this fall time reduction, the buffer maximum gain, $A_{b_{max}}$, is designed to occur at an input voltage of V_{DD_L}. Thus, the buffer threshold voltage, V_M, should be designed equal to V_{DD_L}. According to Figure 6.4.b, the output capacitance, C'_{out}, is averaged over the input voltage from $V_{DD_L}/2$ to V_{DD_L} and given by:

$$C'_{out} = C_{out} + \underbrace{C_{inv} + C_F(1 + A_{b_{max}}(\frac{V_{IL}}{V_{DD_L}} - 1))}_{C_{NEG}} \tag{6.20}$$

   The values of V_{IL}, and $A_{b_{max}}$ are calculated by conducting SPICE DC analysis. The values of C_{out}, and C_{inv} are obtained from the layout parasitic extractions by using the CALIBRE tool provided by Mentor Graphics.

   The two-inverters buffer based negative capacitance (B-NC) circuit shown in Figure 6.3 is employed to implement the required negative capacitance of - 4.4 fF. The B-NC is designed with high-V_t transistors (V_{tn} = 0.59V and $\mid V_{tp} \mid$ = 0.54V as provided by the 65nm CMOS technology transistor model files), the ratio r = 10, and accordingly, V_{DD_H} = 1.36V by using (6.3). The value of α is calculated by fitting the Log (I$_D$)-Log (V$_{GS}$) characteristics to the alpha-power model and equals 1.25. The post layout SPICE DC simulations reveal that V_{IL} = 0.71V and $A_{b_{max}}$ = 32. The value of the buffer input capacitance, C_{inv}, is calculated by using the layout parasitic extraction CALIBRE tool and equals 0.96 fF. Thus, in order to achieve a negative capacitance C_{NEG} = - 4.4fF, the value of C_F = 2.1 fF by recalling (6.20). This feedback capacitance is implemented by using a MIM-CAP, with an area of $0.9\mu m^2$.

   The transient response of the buffer circuit is measured by using post layout transient simulations and it is found that the decaying slope of the buffer circuit output waveform around V$_{DD_H}$/2 equals -3.7 V/nsec whereas the decaying slope of the NOR gate output waveform around V$_{DD_L}$/2 equals -1.2 V/nsec. Accordingly, the speed of the buffer circuit is faster than that of the NOR gate to ensure that (6.2) can be applied.

   From (6.20), the B-NC negative capacitance is expressed as:

$$
\begin{aligned}
C_{NEG} &= C_{inv} + C_F \;\; \theta \\
\text{where } \; \theta &= 1 + A_{b_{max}} (\frac{V_{IL}}{V_{DD_L}} - 1)
\end{aligned}
\tag{6.21}
$$

Thus, from (6.21), the negative capacitance variability is given by:

$$
\frac{\Delta C_{NEG}}{C_{NEG}} = \frac{C_{NEG} - C_{inv}}{C_{NEG}} \;\; (\frac{\Delta C_F}{C_F} + \frac{\Delta \theta}{\theta})
\tag{6.22}
$$

Thus, the standard deviation of C_{NEG} is given by [89]:

$$
(\frac{\sigma_{C_{NEG}}}{C_{NEG}})^2 = (\frac{C_{NEG} - C_{inv}}{C_{NEG}})^2 \;\; [(\frac{\sigma_{C_F}}{C_F})^2 + (\frac{\sigma_\theta}{\theta})^2]
\tag{6.23}
$$

Monte Carlo simulation results reveal that the ratio of $\sigma^{'}/A^{'}_o$ is calculated from (6.15), by using (6.23), and equals 14.8% whereas if the negative capacitance variability contribution is ignored, the ratio of $\sigma^{'}/A^{'}_o$ equals 13.5%. Accordingly, the negative capacitance variation is ignored in this negative capacitance implementation and (6.13) is applied because the error in $\sigma^{'}/A^{'}_o$ when the negative capacitance variability contribution is ignored equals 8.8% (i.e., (14.8-13.5)/14.8) and also because the resulting timing yield is larger than the target yield of 99.87%.

A typical histogram of the OR gate high-to-low delay at node G is shown in Figure 6.11.d. It is evident from Figure 6.11.d and Figure 6.11.a that the B-NC circuit shifts the delay pdf as required. Figure 6.11.d signifies that 99.94% of the dynamic OR gate samples have delays less than the target delay of 433 psec. In addition, the delay standard deviation is reduced from 58.5 psec to 43.9 psec (25% variability reduction). The B-NC circuit exhibits a total average power consumption of 16.9 μW. This power consumption is smaller than that of the DA-NC and the CC-NC circuits by factors of 3.5X and 7X, respectively. The total power dissipation of the OR gate with the negative capacitance circuit is 75.9 μW and the total layout area equals 169.1 μm^2.

4. **Noise Immunity**

   Post layout simulation results show that the UNG of the OR gate without the negative capacitance adoption equals 466mV whereas the adoption of the DA-NC, CC-NC, and B-NC reduces the UNG value to 444mV, 435mV, and 449mV, respectively. Accordingly, the adoption of the negative capacitance circuit results in slightly reducing the OR gate noise immunity.

The simulation results for the three negative capacitance circuits are tabulated in Table 6.1. The following observations are extracted from Table 6.1:

-i- All the three negative capacitance circuits achieve higher yield values than the target yield of 99.87% for both the delay at node G and node F constraints. The delay histograms at node F are not shown, since they exhibit similar results to the ones in Figure 6.11. In addition, the adoption of the DA-NC, CC-NC, and B-NC circuits reduces the OR gate delay variability at node G by 25.5%, 23%, and 25%, respectively. Also, at node F, the adoption of the DA-NC, CC-NC, and B-NC circuits reduces the OR gate delay variability by 23%, 20%, and 26%, respectively. The delay standard deviations at node F are less than those at node G due to the averaging effect. This averaging effect results in averaging and reducing the random variations through the output OR gate inverter.

-ii- The B-NC circuit power dissipation is less than that of the DA-NC and the CC-NC by factors of 3.5X and 7X, respectively. The total power of the OR gate with the DA-NC, CC-NC, and B-NC circuits adopted equals 118.7μW, 176.2μW, and 75.9μW, respectively. Also, the associated power overheads of the DA-NC, CC-NC, and B-NC circuits are 55.9μW, 113.4μW, and 13.1μW, respectively (by subtracting the power dissipation of the OR gate without the negative capacitance circuit adoption from the total power dissipation of the OR gate with the negative capacitance circuit adopted).

-iii- The best negative capacitance circuit is the B-NC circuit, as long as the technology supports the dual supply voltage (dual-V_{DD}) and the dual threshold voltage (dual-V_t) requirements. This circuit dissipates the lowest power consumption and utilizes the smallest added capacitance (2.1 fF).

-iv- If the dual-V_t and the dual-V_{DD} are not supported by the CMOS technology, the second candidate is the DA-NC circuit, because it exhibits less power dissipation and added capacitance than that of the CC-NC. If the OR gate frequency is high such that the B-NC and DA-NC circuits speeds are not faster than the OR gate speed, the CC-NC is the best choice, because it is not prone to the gain-bandwidth product limitation at the expense of higher power dissipation and added capacitance value than those of the DA-NC and the B-NC circuits.

-v- The proposed negative capacitance circuits slightly reduce the noise immunity of the dynamic OR gate. It is shown that the adoption of the DA-NC, CC-NC, and B-NC circuits reduces the UNG by 4.7%, 6.7%, and 3.6%, respectively.

Table 6.1: Post layout simulation results for the dynamic 64-input OR gate without C_{NEG} and with DA-NC, CC-NC, and B-NC circuits adopted. Simulations are performed at a temperature T = 120^oC (worst case delay).

	Without C_{NEG}	**With C_{NEG}**		
		DA-NC	**CC-NC**	**B-NC**
Node G delay				
μ **(psec)**	433	315.7	310	312.5
σ **(psec)**	58.5	43.6	44.9	43.9
σ/μ **(%)**	13.5	13.8	14.5	14
Node F delay				
μ **(psec)**	442.2	326.9	324.3	320
σ **(psec)**	53.6	41.2	42.7	39.6
σ/μ **(%)**	12.1	12.6	13.2	12.4
OR gate power				
μ **(μW)**	62.8	60.4	57.6	59
σ **(μW))**	3.9	4.4	3.9	4
σ/μ **(%)**	6.2	7.0	6.8	6.8
C_{NEG} **power**				
μ **(μW)**		58.3	118.6	16.9
σ **(μW))**		2.8	5.7	1.5
σ/μ **(%)**		5	4.8	8.9
Total Power (μW)	62.8	118.7	176.2	75.9
Total Area (μm^2)	135.8	239.9	299.2	169.1
UNG (mV)	466	444	435	449
Capacitance added (fF)		4.9	6.5	2.1
Technology Considerations				Dual-V_{DD} Dual-V_t

6.3.3 Test Chip Results and Discussions

The test chip details are listed in Appendix C including the chip micrograph and the PCB design. **35** test chips are packaged to account for the D2D variations. Each test chip contains two blocks dedicated for the negative capacitance circuits. Each block of these two blocks contains 32 critical paths, each critical path consists of a 64-input wide fan-in OR gate, CMOS inverters, transmission gates, and several pass transistors (representing a 32X1 MUX to select one output at a time out of the 32 paths). In the first block, the DA-NC negative capacitance circuit is adopted whereas the B-NC negative capacitance circuit is adopted in the second block, as shown in Figure 6.12.

While testing the 35 chips, only 23 chips are working for both the DA-NC and B-NC blocks testing. For each test chip, the 32 critical paths delays are measured by applying an external clock signal of frequency 200 MHz with $V_{DDL} = 0.8$V. Accordingly, the following results represent the delay and total power of the 23 chips where the delay is the maximum delay of the 32 critical paths within the test chip die and the total power is the sum of the 32 critical paths powers. These results are presented for both the DA-NC circuit and the B-NC circuit. The CC-NC negative capacitance circuit is not implemented in this test chip due to chip area and time limitations in submitting the test chip layout. Therefore, only post layout simulation results are available for the CC-NC case.

Figure 6.13 portrays the 23 test chip dies delay and power before the adoption of the negative capacitance circuits (NO-NC case). The measured delay in this figure is the average delay of the precharge delay and the high-to-low delay. The measured power is the sum of the dynamic power (when the clock signal is applied) and the leakage power (when the clock signal is '0'). It should be noted that the target delay is chosen to be 1.85 nsec (the mean of the NO-NC case) which includes the delay of the peripheral circuits delay (i.e., MUXs and buffers delays). The relative variation of the delay, $\sigma/\mu|_{delay}$, is 6.5% and the number of accepted dies is 10 out of 23 dies (43.5% acceptance rate).

1. **The B-NC Circuit Test Results**

   The B-NC circuit is utilized to improve the dynamic OR gate timing yield. For each die, the B-NC circuit supply voltage, V_{DDH}, is swept from 1.0V to 1.4V in 0.1V steps. The V_{DDH} value that results in the lowest die delay and power is selected. Figure 6.13 shows that all the 23 dies are accepted when the B-NC circuit is adopted (100% acceptance rate) and $\sigma/\mu|_{delay}$ is 5.7% which means a 1.1X reduction compared to the NO-NC case. Moreover, the mean power of the 23 dies increases by 28% with the adoption of the B-NC circuit (i.e., the mean power increases from 1.11 mW to 1.42 mW). The die layout area is increased by 14% with the adoption of the B-NC circuit.

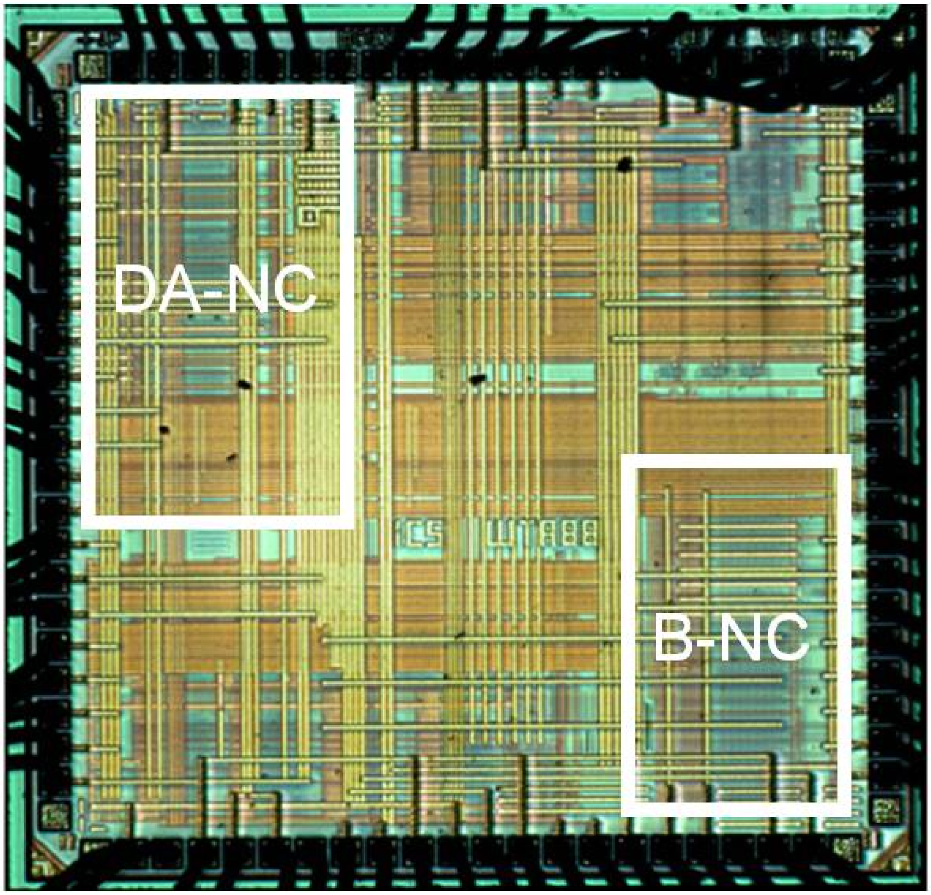


Figure 6.12: Test chip micrograph showing the DA-NC and the B-NC blocks

2. **The DA-NC Circuit Test Results**

   The DA-NC circuit is enabled to improve the timing yield. For each die, the DA-NC op-amp bias current is swept within 30% of its design value. The bias current level that results in the lowest die delay and power is selected. Figure 6.13 shows that all the 23 dies are accepted when the DA-NC circuit is adopted (100% acceptance rate) and $\sigma/\mu|_{delay}$ is 5.8% which means a 1.1X reduction compared to the NO-NC case and approximately same reduction as that occurs in the B-NC case. Moreover, the mean power of the 23 dies increases by 55% with the adoption of the DA-NC circuit (i.e., the mean power increases from 1.11 mW to 1.72 mW). The die layout area is increased by 31% with the adoption of the DA-NC circuit.

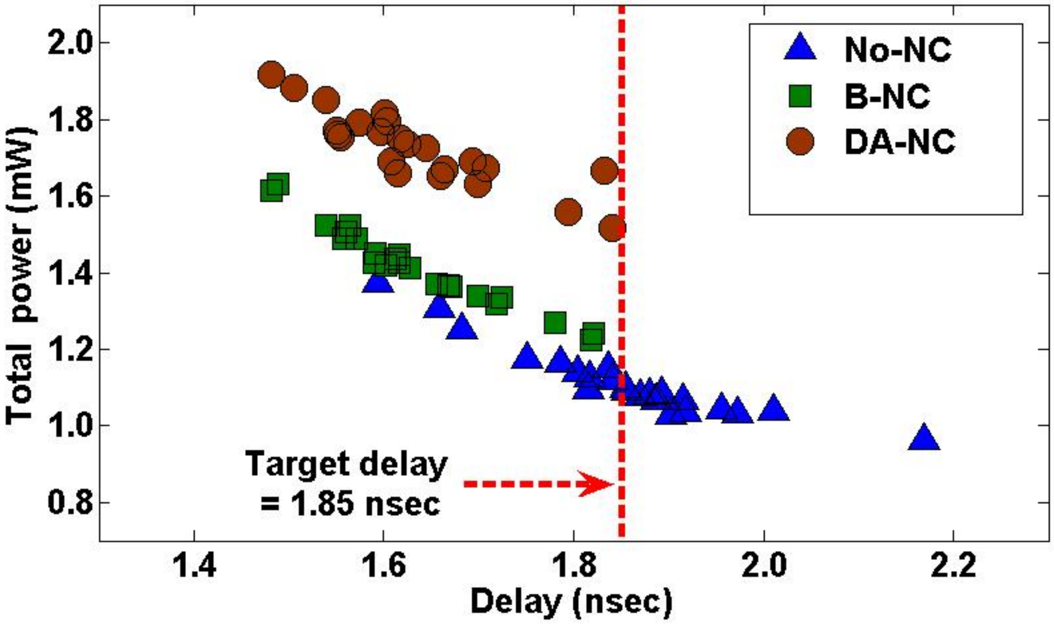


Figure 6.13: Measured delay and power for the 23 test chip dies for the NO Negative Capacitance (NO-NC) case, the B-NC case, and the DA-NC case.

6.4 Application of the Negative Capacitance Circuits to SRAM Cells

SRAM cells have the smallest device sizes on the chip. Thus, SRAM cells show the largest sensitivity to different sources of WID random variations such as RDF and LER [24, 82]. Process variations in logic circuits cause delay spread [14, 116] which reduces the parametric yield, whereas, for SRAM cells, process variations cause the memory to functionally fail, which reduces the functional yield. With lower supply voltages and higher variations, statistical design methodologies are important to improve the SRAM yield with minimum overhead. There are several SRAM cell failure types such as read access failure (i.e., incorrect read operation), read stability failure (i.e., cell flips when accessed), and write stability failure (i.e., cell is not written within the write window) [8, 81, 82], as discussed in details in Chapter 2.

In the SRAM read operation, the column bitlines (BL and BLB) are precharged to V_{DD} and the row wordline (WL) is enabled (WL ='1'), connecting the internal nodes of the cell to the bitlines. In this read operation, one of the bitlines discharges via the node storing '0' whereas the other bitline remains at V_{DD}. Accordingly, a differential voltage

(V_{Δ}) is developed between the two bitlines (i.e., one bitline voltage remains at V_{DD} and the other discharges to $(V_{DD} - V_{\Delta})$). This bitlines differential voltage, V_{Δ}, is a function of the read current (I_{read}), the bitline capacitance (C_{BL}), and the time during which the WL is enabled. To ensure correct read operation, the Sense Amplifier (SA) is enabled using a control signal (SAE) after a sufficient differential signal V_{Δ} is developed (i.e., $V_{\Delta} \geq$ SA input offset voltage ($V_{SAoffset}$)), which is amplified by the SA to a digital output level as shown in Figure 6.14 [23, 161].

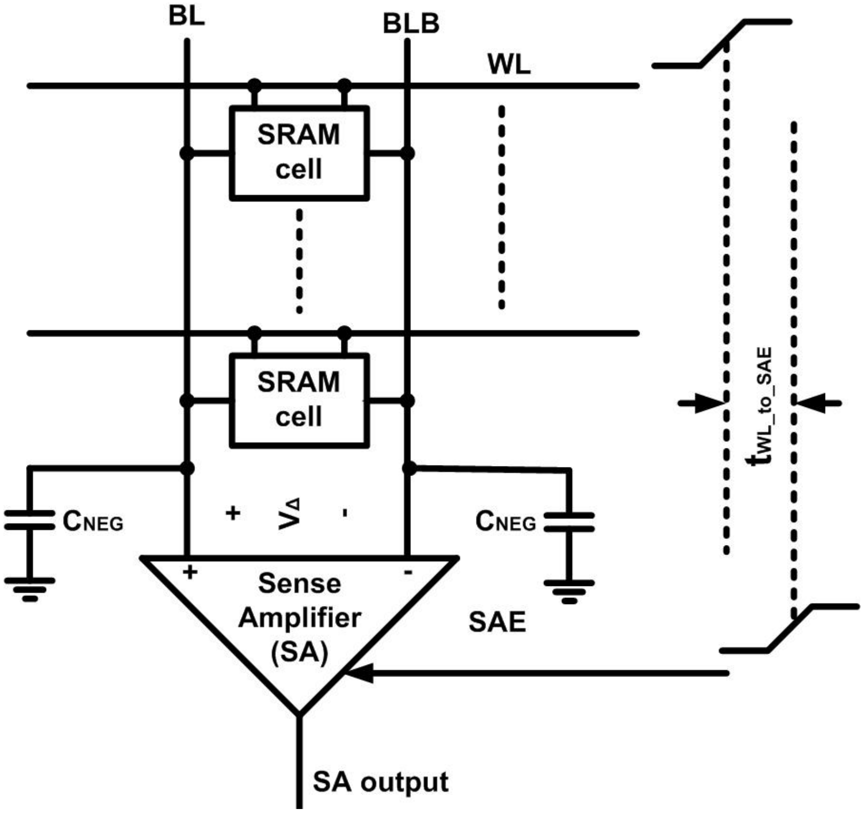


Figure 6.14: Simplified SRAM cell read path.

The delay difference between the WL rising and the SAE rising is called SA read sensing window ($t_{WL-to-SAE}$). $t_{WL-to-SAE}$ has a direct impact on the memory performance as it contributes a large percentage of the memory access time. As $t_{WL-to-SAE}$ increases, V_{Δ} increases, which reduces the probability of read failure due to SA input offset voltage, $V_{SAoffset}$. Hence, it is desirable to reduce $t_{WL-to-SAE}$ as long as correct read operation is ensured (i.e., $V_{\Delta} \geq V_{SAoffset}$). Therefore, there is a strong tradeoff between yield and performance for SRAM, which is one of the most important design decisions for memory designers [23, 161, 162].

Due to the small size of the SRAM bitcell and the inverse proportionality between the WID random variations standard deviation and the square root of the device area [15, 17], the SRAM cell read current, I_{read}, shows large WID random variations [24, 81], and follows a normal distribution [163]. From the SRAM design perspective, I_{read} determines the time required to develop sufficient V_{Δ} before enabling the SA. I_{read} variation is considered one of the largest sources of yield loss in SRAM cells [23, 81, 161]. According to [161], the voltage V_{Δ} is given by:

$$\begin{aligned} V_{\Delta} &= (I_{read} - N * I_{leakage})\frac{t_{WL-to-SAE}}{C_{BL}} \\ &\geq V_{SAoffset} \quad \text{for proper read operation} \end{aligned} \tag{6.24}$$

where $N * I_{leakage}$ are the leakage currents from all the other N cells connected between the same bitlines and exhibit large WID random variations as well. The variations of I_{read} and $N * I_{leakage}$ result in V_{Δ} spread which makes some of the SRAM cells do not develop a sufficient V_{Δ} and accordingly, the read access yield is reduced. The leakage current, $I_{leakage}$, exhibits a log-normal distribution (not a normal distribution) with Vt variations. However, the usage of the central limit theorem [164] helps to model the sum of the leakage of a sufficiently large number of SRAM cells as a normal distribution [9]. In [163], 16 SRAM cells are a sufficient number to validate these results. Therefore, I_{read} and $N * I_{leakage}$ have normal distributions which makes V_{Δ} follows a normal distribution as well referring to (6.24).

The statistical gate sizing has been performed for the SRAM in [163] to maximize the total yield considering the stability, leakage power, read current, and area constraints. For some specific constraints, the maximum achievable yield is 99.8%, as reported in [163], which means that 2097 SRAM cells are going to fail, in a 1 Mb SRAM array, when the resulting optimal SRAM sizes are applied. Moreover, adding another constraint for the voltage V_{Δ} makes it more difficult to find the optimal sizes that maximize the yield. Therefore, the adoption of the negative capacitance should be performed in conjunction with the statistical gate sizing to relax the yield maximization constraints.

In this section, a negative capacitance circuit is connected to each bitline of a 512 SRAM cells column, as shown in Figure 6.14. This negative capacitance connection reduces the bitlines parasitic capacitance, and correspondingly, improves the read access yield without changing the SRAM gate sizing. This technique can be combined with the statistical gate sizing for higher SRAM yield values.

6.4.1 Statistical Read Access Yield Improvement using Negative Capacitance

Assume that the SRAM cells are designed such that the V_{Δ} mean (i.e., A$_o$) equals $V_{SAoffset}$ (This can be performed at any design corner). (6.24) is rewritten as:

$$A_o = \zeta \ / C_{BL} \tag{6.25}$$

where ζ is a proportionality constant, dependent on I_{read}, $N * I_{leakage}$, $t_{WL-to-SAE}$, and C$_{BL}$ is the bitline parasitic capacitance. Due to the process variations, V_{Δ} is normally distributed around A$_o$. It should be noted that the impact of the process variations on C$_{BL}$ is neglected with respect to the impact on ζ [89]. This is due to the fact that ζ depends on the transistors threshold voltage, the main source of the variability [27, 33, 89]. To improve the read access yield, the V_{Δ} variability, which is centered around A$_o$, is shifted to a new center A$_o^{'}$ given by:

$$A_o^{'} = A_o + n \ \sigma^{'} \tag{6.26}$$

where $\sigma^{'}$ is the standard deviation of the delay variability around A$_o^{'}$, and "n" is dependent on the desired read access yield, Y$_o$. In logic circuits timing yield improvement, "n" equals 3.0 to have a yield of 99.87%. However, in SRAM design, a yield of 99.87% in a 1.0 Mb SRAM block means that 1363 SRAM cells are expected to fail. Therefore, "n" is selected to be 5.0 for SRAM design achieving a yield of 99.99997% (i.e., 0.3 cells are expected to fail in a 1.0 Mb SRAM block).

A$_o$ is increased to A$_o^{'}$ by connecting a negative capacitance to the SRAM bitline. The addition of the negative capacitance, C$_{NEG}$, to the bitline parasitic capacitance results in a modified bitline capacitance, C$_{BL}^{'}$, which is given by:

$$C_{BL}^{'} = C_{BL} + C_{NEG} \tag{6.27}$$

and accordingly, the modified V_{Δ} pdf center, A$_o^{'}$, is expressed as:

$$A_o^{'} = \zeta \ / C_{BL}^{'} \tag{6.28}$$

From (6.25), the delay A$_o$ variability, ΔA$_o$ = $\Delta\zeta$ / C$_{BL}$, and from (6.28), the delay A$'_o$ variability, ΔA$'_o$ = $\Delta\zeta$ / C$'_{BL}$, assuming that the capacitances, C$_{BL}$ and C$'_{BL}$, are constants from the variability perspective. From (6.27), C$'_{BL}$ < C$_{BL}$ because C$_{NEG}$ has a negative value. Therefore, ΔA$'_o$ > ΔA$_o$. The ratio between $\sigma^{'}$ and σ is obtained by computing:

$$\frac{\sigma}{\sigma'} = \frac{\Delta A_o}{\Delta A_o'} = \frac{C_{BL}'}{C_{BL}} = 1 + \frac{C_{NEG}}{C_{BL}} \tag{6.29}$$

From (6.25), (6.28), and (6.29), it is evident that the adoption of the negative capacitance circuit results in increasing A_o' and σ' by the same factor compared to A_o and σ, respectively, even when the variation of the negative capacitance circuit is negligible.

By using (6.25-6.29), the negative capacitance, C$_{NEG}$, which achieves the desired read access yield improvement, is expressed as (when C$_{NEG}$ circuit variability is considered):

$$C_{NEG} = \frac{-\,n\ \sigma'}{A_o'} C_{BL} \tag{6.30}$$

and when C$_{NEG}$ variability is neglected (i.e., $\frac{\sigma'}{A_o'} = \frac{\sigma}{A_o}$), C$_{NEG}$ is given by:

$$C_{NEG} = \frac{-\,n\ \sigma}{A_o} C_{BL} \tag{6.31}$$

It should be noted that C$_{NEG}$ variations contribution to σ' is ignored in (6.31) to have an initial guess for the required value of C_{NEG}. This contribution should be calculated because the negative capacitance circuit has different implementations such as DA-NC, and B-NC circuits. If the negative capacitance circuit variability is taken into account, V_Δ variability, is expressed as follows:

$$(\frac{\sigma'}{A_o'})^2 = (\frac{\sigma}{A_o})^2 + \underbrace{\frac{(\sigma_{C_{NEG}}/C_{NEG})^2}{((C_{BL}+C_{NEG})/C_{NEG})^2}}_{C_{NEG}\quad variability} \tag{6.32}$$

Once again, the negative capacitance circuit should be designed such that its variability contribution to σ' is small. Thus, (6.31) is valid and the negative capacitance contribution to σ' is neglected. However, if this contribution is not neglected and taken into account, (6.31) is used as an initial guess and then the following algorithm is adopted:

1) Calculate the initial value of C_{NEG} by using (6.31).

2) Conduct Monte Carlo simulations while the negative capacitance circuit is adopted.

3) Determine the values of A$'_o$ and σ'.

4) Calculate the value of the read access yield. If the read access yield is greater than or equal to 99.99997%, the target read access yield improvement is achieved and the algorithm stops. If the read access yield is less than 99.99997%, calculate the new C_{NEG} by using the new $\sigma^{'}$ value by using (6.30).

5) Repeat steps 2-4 above

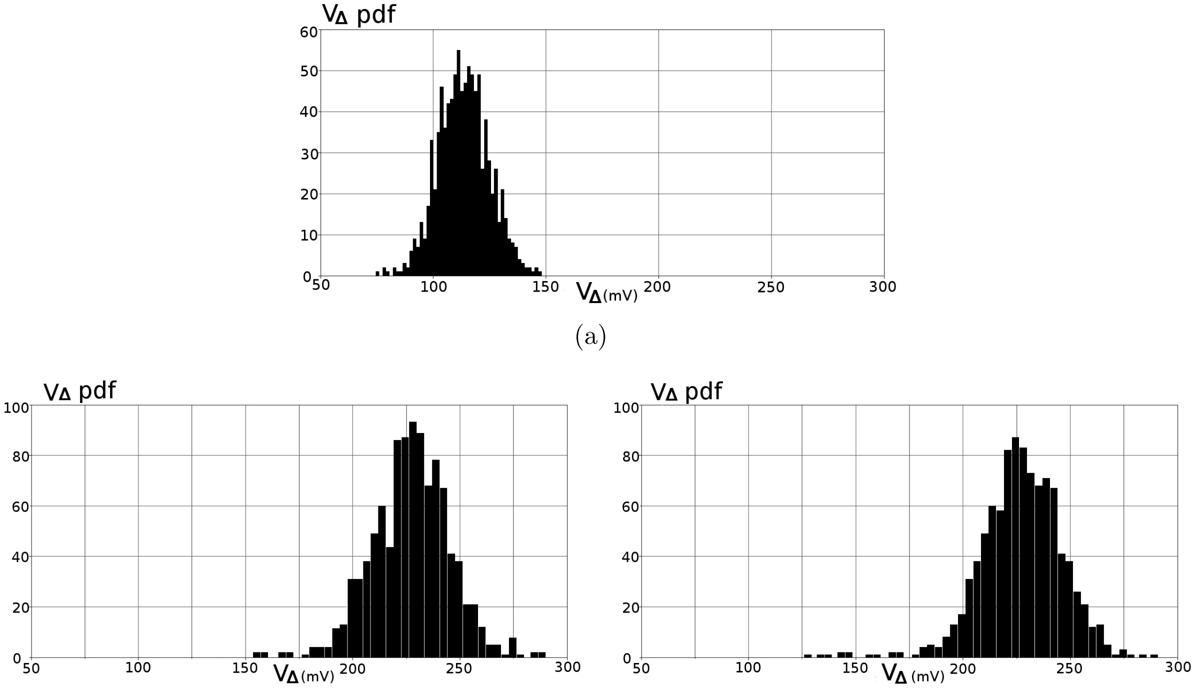
This negative capacitance, C_{NEG}, is designed by using the DA-NC, and the B-NC circuits. The CC-NC circuit is not adopted here due to its large layout area overhead which makes it difficult to be included in the SRAM array. Moreover, the SRAM speed is normally less than the high performance circuits speed which makes the DA-NC and B-NC are more appropriate for the SRAM read access yield improvement.

6.4.2 Simulation Results and Discussions

The 512 SRAM cells column shown in Figure 6.14 is utilized as a benchmark circuit to verify the proposed read access yield improvement technique. The parasitic bitlines capacitances are large due to the 512 cells connected to the bitlines. Therefore, the negative capacitance, C_{NEG}, is connected to each bitline, as shown in Figure 6.14. The SRAM cell is designed to achieve a nominal V_{Δ} = 110mV. The SA input offset voltage is assumed to be $V_{SAoffset}$ = 110mV which is considered to be the target A_o (i.e., A_o = 110mV). The supply voltage used in the SRAM design, V_{DD_L}, is 0.9V. The SRAM design is performed by using post layout simulations referring to an industrial hardware-calculated 65nm CMOS technology. The bitline capacitance, C_{BL}, is calculated from the SRAM layout, by using the CALIBRE tool provided by Mentor Graphics, and equals 46.1 fF (this layout capacitance includes all the parasitic capacitances coupled to the bitlines). Therefore, the constant ζ, defined in (6.25), equals 5.1 V.fF.

Monte Carlo analysis using the statistical transistor model reported in Appendix D is conducted. A typical histogram of V_{Δ} is shown in Figure 6.15.a. The standard deviation of V_{Δ} around A_o, σ, is 11.2 mV. Accordingly, the required negative capacitance value, C_{NEG}, for a read access yield Y_o = 99.99997% (i.e., "n" = 5.0) is obtained from (6.13) and equals - 23.5 fF. Then, the values of $A_o^{'}$ and $\sigma^{'}$ are calculated as 224.4mV and 22.9mV, respectively. This negative capacitance is implemented by using the DA-NC and B-NC circuits.

The adoption of the negative capacitance circuits to the SRAM bitlines is utilized for read access yield improvement of the bitlines discharge during the read operation. However, the precharge delay is also affected by the negative capacitance circuit adoption. Therefore, the DA-NC and the B-NC circuits are disabled during the precharge phase by adding a tail NMOS transistor driven by the precharge signal (This is because the Miller based



(a)

(b)

(c)

Figure 6.15: V_{Δ} histograms for the 512 SRAM column using Monte Carlo analysis (a) Before employing the negative capacitance (μ= 110.2mV and σ= 11.2mV), (b) After employing the DA-NC circuit (μ= 228mV and σ= 23.9mV), and (c) After employing the B-NC circuit (μ= 220mV and σ= 25.5mV).

negative capacitance circuit realizes a positive capacitance when the voltage at the bitlines is charging. This results in a slight increase of the precharge delay by 4.1% and 5.3% due to the DA-NC and the B-NC circuits, respectively.

1. **DA-NC Circuit**

   In Figure 6.14, the bitlines are initially pre-charged to the SRAM supply voltage, V_{DD_L}. Then, during the read operation, one of the bitlines is discharged. The idea of the DA-NC circuit is to allow the discharging bitline to see a reduced capacitance (due to the negative capacitance circuit adoption) when discharging from V_{DD_L} to $(V_{DD_L} - V_{\Delta})$. This reduced capacitance increases the sensed bitlines differential voltage, V_{Δ}, according to (6.24). Accordingly, the input dynamic range (i.e., the range of the input voltage over which the amplifier exhibits a linear gain) of the differential-pair amplifier, should include the range from V_{DD_L} to $(V_{DD_L} - V_{\Delta})$.

In addition, the amplifier speed must be sufficiently faster than the SRAM bitline discharging speed to achieve a negative capacitance, which results in increasing the negative capacitance power overhead and reducing the amplifier gain due to the constant gain-bandwidth product of the amplifier.

The DA-NC circuit is adopted to implement the required negative capacitance of - 23.5 fF. The input capacitance of the DA-NC circuit, C_{M1}, is calculated from the layout by using the CALIBRE tool provided by Mentor Graphics and equals 0.5 fF (this means that the actual negative capacitance to be implemented is -24 fF to compensate for C_{M1}). The feedback capacitance, C$_F$, is implemented by using a MIM-CAP. The value of C$_F$ is chosen to be 9.6 fF, with an area of $3.84\mu m^2$. Accordingly, the required amplifier gain, A, is 3.5 which is designed by using $R_F = 1.75K\Omega$ and $R_I = 0.5K\Omega$.

The negative capacitance variability is given by:

$$\frac{\Delta C_{NEG}}{C_{NEG}} = \frac{\Delta C_F}{C_F} + \frac{\Delta A}{A-1} \tag{6.33}$$

Thus, the standard deviation of C_{NEG} is given by:

$$(\frac{\sigma_{C_{NEG}}}{C_{NEG}})^2 = (\frac{\sigma_{C_F}}{C_F})^2 + \frac{(\sigma_A/A)^2}{((A-1)/A)^2} \tag{6.34}$$

Monte Carlo simulation results reveal that the ratio of $\sigma^{'}/A_o^{'}$ is calculated from (6.32), by using (6.34), and equals 10.9% whereas if the negative capacitance variability contribution (the second term in (6.32)) is ignored, the ratio of $\sigma^{'}/A_o^{'}$ equals 10.2% (this value is obtained by calculating σ/A_o = 11.2mV/ 110mV). Accordingly, the negative capacitance variation is ignored in this negative capacitance implementation and (6.13) is applied because the error in $\sigma^{'}/A_o^{'}$ when the negative capacitance variability contribution is ignored equals 6.4% (i.e., (10.9-10.2)/10.9) and also because the resulting read access yield is larger than the target yield of 99.99997%.

A typical histogram of V_Δ is shown in Figure 6.15.b. It is evident from Figure 6.15.a and Figure 6.15.b that the DA-NC circuit shifts V_Δ pdf as required. Figure 6.15.b indicates that 100% of the SRAM samples have V_Δ larger than $V_{SAoffset}$. Moreover, the DA-NC circuit exhibits a total average power consumption of 41.1 μW, including the biasing circuit power consumption with layout area of 46.5 μm^2. The amplifier bandwidth, over which the gain, A, is independent of frequency, is calculated to be 1.8 GHz which is sufficiently larger than the SRAM bitline discharging frequency.

2. **B-NC Circuit**

Figure 6.16 displays the total capacitance at the bitline (including the negative capacitance), C'_{BL}, as a function of V_{in}, where C_{inv} is the input capacitance of the buffer circuit, and C_{BL} is the bitline capacitance without the negative capacitance adoption.

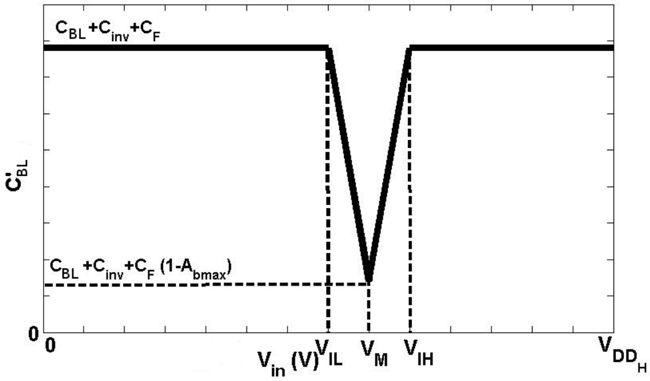


Figure 6.16: The total bitline capacitance, C'_{BL}, when the B-NC is adopted.

Similar to the DA-NC circuit, the B-NC circuit should allow the discharging bitline to see a reduced capacitance (Due to the negative capacitance circuit adoption), when discharging from V_{DD_L} to (V_{DD_L}-V_Δ). Therefore, V_Δ increases due to this reduced bitlines capacitance. To achieve this V_Δ increase, the buffer maximum gain, $A_{b_{max}}$, is designed to occur at an input voltage of V_{DD_L}. Thus, the buffer threshold voltage, V_M, should be designed equal to V_{DD_L}. According to Figure 6.16, the bitline capacitance, C'_{BL}, is averaged over the input voltage from (V_{DD_L}-V_Δ) to V_{DD_L} and given by:

$$C'_{BL} = C_{BL} + \underbrace{C_{inv} + C_F(1 - \beta A_{b_{max}})}_{C_{NEG}} \tag{6.35}$$

where

$$\beta = \begin{cases} \frac{V_{DDL} - V_{IL}}{2V_\Delta} & V_{IL} \geq (V_{DDL} - V_\Delta) \\ 1 - \frac{V_\Delta}{2(V_{DDL} - V_{IL})} & V_{IL} \leq (V_{DDL} - V_\Delta) \end{cases}$$

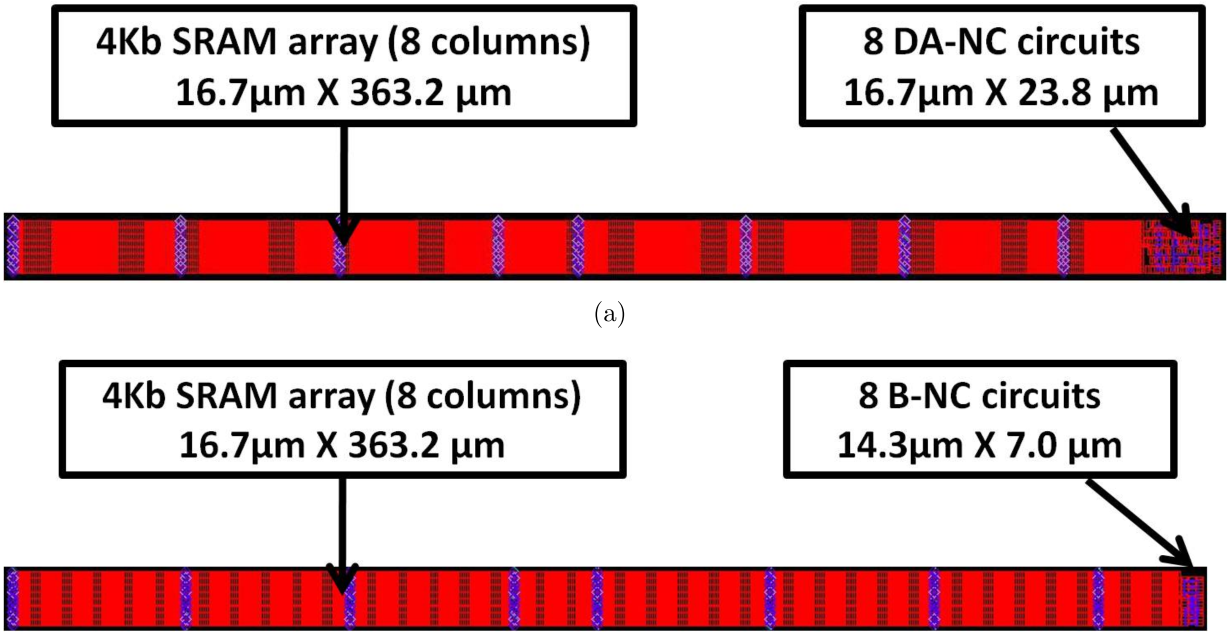
The value of V_{IL}, and $A_{b_{max}}$ are calculated by conducting SPICE DC analysis. The values of C_{BL}, and C_{inv} are obtained from the layout parasitic extractions by using the CALIBRE tool provided by Mentor Graphics. The relationship between V_{DD_H}, the buffer supply voltage, and V_{DD_L}, the SRAM supply voltage, is the same as that obtained in (6.3).

The two-inverters buffer based negative capacitance (B-NC) circuit is employed to implement the required negative capacitance of - 23.5 fF. The B-NC is designed with high-V_t transistors (V_{tn} = 0.59V and $\mid V_{tp} \mid$ = 0.54V as provided by the 65nm CMOS technology transistor model files), the ratio $r = 10$, and accordingly, V_{DD_H} = 1.47V by using (6.3). The value of α is calculated by fitting the Log (I_D)-Log (V_{GS}) characteristics to the alpha-power model and equals 1.25. The post layout SPICE DC simulations reveal that V_{IL} = 0.79V and $A_{b_{max}} = 30$. The value of the buffer input capacitance, C_{inv}, is calculated by using the layout parasitic extraction CALIBRE tool and equals 0.96 fF. Thus, in order to achieve a negative capacitance C_{NEG} = - 23.5fF, the value of C_F = 2.8 fF by recalling (6.35). This feedback capacitance is implemented by using a MIM-CAP with an area of $1.2\mu m^2$.

From (6.35), the B-NC negative capacitance, C_{NEG}, standard deviation is given by (6.23) where $\theta = 1 - A_{b_{max}}(\frac{V_{DDL}-V_{IL}}{2V_{\Delta}})$ since $V_{IL} > V_{DDL} - V_{\Delta}$. Monte Carlo simulation results reveal that the ratio of $\sigma^{'}/A^{'}_{o}$ is calculated from (6.32) and equals 11.1% whereas if the negative capacitance variability contribution is ignored, the ratio of $\sigma^{'}/A^{'}_{o}$ equals 10.2%. Accordingly, the negative capacitance variation is ignored in this negative capacitance implementation and (6.31) is applied because the error in $\sigma^{'}/A^{'}_{o}$ when the negative capacitance variability contribution is ignored equals 8.1% (i.e., (11.1-10.2)/11.1) and also because the resulting read access yield is larger than the target yield of 99.99997%.

A typical histogram of V_{Δ} is shown in Figure 6.15.c. It is evident from Figure 6.15.a and Figure 6.15.c that the B-NC circuit shifts V_{Δ} pdf as required. Figure 6.15.c indicates that 100% of the SRAM samples have V_{Δ} larger than $V_{SAoffset}$. Moreover, the B-NC circuit exhibits a total average power consumption of 10.4 μW with layout area of 13.3 μm^2. The transient response of the buffer circuit is measured by using post layout transient simulations and it is found that the decaying slope of the buffer circuit output waveform equals -3.2 V/nsec whereas the decaying slope of the SRAM discharging bitline waveform equals -0.85 V/nsec. Accordingly, the speed of the buffer circuit is faster than the bitline discharge speed.

Figure 6.17 shows the layout of a 4Kb SRAM array layout with the adoption of the negative capacitance circuits (DA-NC in Figure 6.17.a and B-NC in Figure 6.17.b). The area overheads of the DA-NC and the B-NC circuits are 13% and 3.4%, respectively. It should be noted that the 4Kb SRAM array consists of 8 columns which means it required the use of 16 negative capacitance circuits. Accordingly, the area overhead calculated here is for the use of the 16 negative capacitance circuits (only 8 of them are shown in Figure 6.17.



(a)

(b)

Figure 6.17: 4Kb SRAM array with the adoption of (a) The DA-NC circuit and (b) The B-NC circuit

The following observations and design guidelines are extracted from the simulation results:

-i- The adoption of the two proposed negative capacitance circuits results in improving the SRAM read access yield from 61.9% to 100%.

-ii- The power consumption and the layout area overhead of the DA-NC is larger than that of the B-NC by factors of 4X and 3.5X, respectively.

-iii- The recommended negative capacitance circuit is the B-NC circuit, as long as the technology supports the dual supply voltage (dual-V_{DD}) and the dual threshold voltage (dual-V_t) requirements. This circuit dissipates lower power consumption and exhibits lower area overhead than that of the DA-NC. If the dual-V_t and the dual-V_{DD} are not supported by the CMOS technology, the DA-NC circuit can be used.

-iv- The adoption of the negative capacitance circuits for read access yield improvement increases the area overhead of the SRAM column. Nevertheless, this area overhead can still be small since the additional area is amortized over the large size of the memory macro size.

The effect of the negative capacitance circuit on the SRAM parameters such as Static Noise Margin (SNM), Write Margin (WM), read access time, write access time, read failure probability, and write failure probability, has been investigated as well. It is found that the adoption of the negative capacitance circuit does not affect the WM, write failure probability, and write access time, because all these parameters are affected mainly by the internal SRAM nodes capacitances. The read access time and the read failure probability are approximately not affected, as well, since in this SRAM architecture, a fixed $t_{WL-to-SAE}$ is assumed. Also, the SNM is not affected by the adoption of the negative capacitance circuit. In addition, the effect of the mismatch between the bitlines capacitances of the BL and the BLB should be taken into account. This mismatch can be mitigated by the proper design of the negative capacitance circuit, especially, if the negative capacitance circuit variations contribution to the total differential voltage variations is ignored.

6.5 Summary

In this chapter, new negative capacitance circuits are developed to reduce the effects of process variations on wide fan-in dynamic circuits. Post layout simulation results and test chip measurements, using the 65nm CMOS technology, show that the adoption of these negative capacitance circuits results in improving the timing yield to values larger than 99.87% by reducing the delay mean and variability. In addition, these negative capacitance circuits are used to reduce the effects of process variations on SRAM arrays. Post layout simulation results show that the adoption of these negative capacitance circuits results in improving the SRAM read access yield from 61.9% to 100% by increasing the bitlines differential voltage V_{Δ}.

The recommended negative capacitance circuit is the B-NC circuit when the technology supports the dual supply voltage (dual-V_{DD}) and the dual threshold voltage (dual-V_t) requirements. This circuit dissipates the lowest power and area overheads and utilizes the smallest added capacitance. If the dual-V_t and the dual-V_{DD} are not supported by the CMOS technology, the second candidate is the DA-NC circuit, because it exhibits less power dissipation and added capacitance than that of the CC-NC. If the benchmark circuit frequency is high such that the B-NC and DA-NC circuits speeds are not faster than the benchmark circuit speed, the CC-NC is the best choice, because it is not prone to the gain-bandwidth product limitation at the expense of higher power dissipation, area overhead, and added capacitance value than those of the DA-NC and the B-NC circuits.

Chapter 7

Conclusion

In this chapter, we summarize our research contributions in Section 7.1 and discuss future research directions in Section 7.2.

With technology scaling, the expected higher sensitivities to variations, radiation induced soft errors, aging degradations, and noise, make the design of a robust circuit is extremely challenging for future CMOS technologies. In this thesis, we studied the challenges of robust design in variation-sensitive digital circuits including SRAM, flip-flops, and high performance circuits. This research work has contributed to new techniques to address process variability, soft errors, and aging degradations on the nanometer circuits.

7.1 Summary of Contributions

1. **Critical Charge Variability Modeling**

   We have proposed analytical critical charge variability models, accounting for both D2D and WID variations, of super-threshold SRAM cells and sub-threshold SRAM cells. The proposed models deal with the D2D variations, by using corner-based methods and deal with the WID variations, by using statistical techniques. The accuracy of the proposed models is validated by using transient and Monte Carlo SPICE simulation results, for an industrial hardware-calibrated 65nm CMOS technology. The derived statistical models are scalable, bias dependent, and require only the knowledge of easily measurable parameters. Moreover, the models are very efficient, compared to Monte Carlo simulations. This makes them very useful in early design cycles, SRAM design optimization, and technology prediction. Also, the proposed models can be extended for the flip-flops critical charge variability as well.

In the super-threshold SRAM models, it is shown that, the use of the coupling capacitor in the SRAM cell, as a soft error mitigation technique, is limited by the relative variations. The proposed models provide an analytical equation, to calculate the value of the coupling capacitor, that results in minimum and maximum relative variations. Finally, the proposed models show that, the PMOS transistors in the SRAM cell, are dominating the variations, and hence, the PMOS transistors must be designed, while taking the critical charge variations into account.

In the sub-threshold SRAM models, it is found that the relative critical charge variability exhibits a minimum at a certain temperature value. This result can be used by circuit designers to keep the temperature at this value, by using temperature control techniques, to minimize the relative critical charge variability. Moreover, the proposed models show that the transistor sub-threshold swing coefficient can be optimized to minimize the critical charge variability. These results are particulary relevant for applications with strict SER constraints.

Therefore, the proposed models address the impact of the process variations on the soft error mitigation techniques decisions and also provide useful design insights that can be adopted to reduce the critical charge variability, especially due to random WID variations.

2. **Comparative Analysis of Yield Improved Flip-Flops**

   We have conducted a comparative analysis between commonly used flip-flops topologies, after performing yield improvement by using statistical gate sizing. First, the timing yield improved super-threshold flip-flops are compared for the required power and PDP overheads. Following that, the power yield improved sub-threshold flip-flops are compared for the required timing and PDP overheads.

   For the super-threshold flip-flops, this comparative analysis recommends that the super-threshold SD-FF is the best choice for high soft errors immunity and high performance at the expense of large power. When the power budget is not met, super-threshold master-slave flip-flops are preferred. If the super-threshold SA-FF has to be used, soft error mitigation techniques are required for proper operation since the super-threshold SA-FF has a poor soft errors immunity. In addition, it has been shown that the timing yield improvement increases the soft errors immunity since, increasing the transistor sizing increases the nodal capacitances.

   For the sub-threshold flip-flops, this comparative analysis results recommend the utilization of the sub-threshold SA-FF. In addition, the results show that the sub-threshold M-C^{2}MOS-MSFF flip-flop is not recommended to be used in the sub-threshold region for power yield improvement requirements.

   From a design decision perspective, this fair comparison between different flip-flops topologies show that the super-threshold SA-FF is not recommended due to its large power and energy overheads to achieve the timing yield improvement and the poor immunity to soft errors. However, the sub-threshold SA-FF is recommended due to its low timing and energy overheads to achieve the power yield improvement. In addition, the super-threshold SD-FF is highly recommended due to its low overheads for timing yield improvement whereas the sub-threshold M-C^{2}MOS-MSFF is not recommended to be used in the sub-threshold region for power yield improvement requirements. Finally, it has been shown that the timing yield improvement increases the soft errors immunity of the flip-flops circuits.

3. **New Adaptive Body Bias (ABB) Circuits**

   We have proposed new ABB circuits (namely, D-ABB and LD-ABB) that consist of threshold voltage sensing circuits and a direct controller that generates the required body bias voltages to compensate for process variations. The proposed D-ABB and LD-ABB circuits are attractive mainly in two ways. First, the proposed ABB circuits exhibit low area overhead that facilitates the adoption of them at smaller granularity levels to increase their capability in reducing the process variations. Second, no ADC or DAC is required in the proposed D-ABB and LD-ABB circuits implementations. Accordingly, the proposed ABB circuits are resolution free compared to the previous state-of-art ABB techniques. The effectiveness of the proposed D-ABB and LD-ABB in process variations compensation, when used globally and locally, is proved by using post layout simulation results and test chip measurements, using TSMC 65nm Triple-well CMOS technology.

   The proposed D-ABB circuit is also adopted to reduce the impacts of the NBTI aging and process variations on the SRAM cells. Post layout simulation results, referring to an industrial hardware-calibrated 65nm CMOS technology transistor model, show that the proposed ABB compensates effectively for NBTI and process variations. In addition, the proposed ABB enhances the soft errors immunity of the SRAM cell by reducing the critical charge degradation with aging. Accordingly, the adoption of the D-ABB to the SRAM array improves the SRAM robustness and yield.

4. **New Negative Capacitance Circuits**

   We have developed new three negative capacitance circuits (namely, DA-NC, CC-NC, and B-NC) that can be connected to the highly capacitive output nodes of variation-sensitive circuits such as the output node of the wide fan-in dynamic OR gate and the SRAM bitlines. Post layout simulation results and test chip measurements, using TSMC 65nm CMOS technology, show that the adoption of these negative capacitance circuits results in improving the timing yield of dynamic wide fan-in OR gate by reducing the delay mean and variability. In addition, these negative capacitance circuits are used to reduce the effects of process variations on SRAM arrays. Post layout simulation results show that the adoption of these negative capacitance circuits results in improving the SRAM read access yield by increasing the bitlines differential voltage V_{Δ}.

Among the three developed negative capacitance circuits, the recommended one is the B-NC circuit when the technology supports the dual supply voltage (dual-V_{DD}) and the dual threshold voltage (dual-V_t) requirements. This circuit dissipates the lowest power and area overheads and utilizes the smallest added capacitance. If the dual-V_t and the dual-V_{DD} are not supported by the CMOS technology, the second candidate is the DA-NC circuit, because it exhibits less power dissipation and added capacitance than that of the CC-NC. If the benchmark circuit frequency is high such that the B-NC and DA-NC circuits speeds are not faster than the benchmark circuit speed, the CC-NC is the best choice, because it is not prone to the gain-bandwidth product limitation at the expense of higher power dissipation, area overhead, and added capacitance value than those of the DA-NC and the B-NC circuits.

7.2 Future Research Directions

The current technology trends show that process variations, soft errors, and aging degradation mechanisms will increase further with technology scaling and more research is required in the area of robust circuit design. More emphasis on statistical design techniques is required to enable the design of robust and yield improved circuits. In the following, we outline some future research directions along these lines based on the work presented in this thesis.

The proposed critical charge variability models can be used to develop an automated soft error rate variability prediction tool. A computer program in C programming language could be used in this purpose. The tool would enable estimating the change in the soft error rate performance and variability when the SRAM cell is designed by varying different transistor parameters. The design insights gained while developing the critical charge variability models can be used to develop the critical charge expressions for other types of SRAM cells (i.e., 4T SRAM cell). This would enable comparisons of the soft errors robustness of those cells with that of the 6T cell under process variations, different operating environments (voltage, temperature, etc.), and power budgets. In addition, the proposed analytical models can be combined with the other SRAM cell parameters models such as SNM, to statistically size the SRAM cell to achieve the highest yield and robustness.

In the area of flip-flops topologies comparison, it is very crucial to develop a statistical framework for flip-flops design taking all the flip-flops metrics in considerations such as setup time, latency delay, hold time, leakage power, total power consumption, and layout area. Then, solving an optimization problem that takes in consideration the impact of process variations, soft errors, and aging mechanisms on the flip-flops yield and robustness. The work provided in this thesis represents an initial step for this statistical framework, where only the total power consumption and the latency delay are considered and the impacts of process variations and soft errors are taken into account. Such a statistical framework, when developed, will be very beneficial to flip-flops designers in selecting the higher yield and more robust flip-flops topology that satisfies their designs constraints.

The proposed ABB circuits should be adopted to more benchmark circuits targeting the impacts of process variations and NBTI aging. In addition, the other aging degradation mechanisms such as HCI and PBTI should be addressed. We foresee the adoption of these ABB circuits to the sub-threshold benchmark circuits is highly desirable to mitigate the increased process variations in the sub-threshold region. In addition, the design of the ABB circuits needs more research to come out with less area overhead amplifiers and squaring circuits that mainly help in using these ABB circuits at lower granularity levels. Moreover, investigating the effect of the ABB circuits in compensating for random variations such as RDF and LER, is very important, especially with the large increase of these random variations with technology scaling. This may be performed by using very small granularity levels which can be done by using the proposed low area ABB circuits.

In the area of negative capacitance circuits, more research is required to develop lower area and power overheads negative capacitance circuits, especially, the NIC based negative capacitance circuits. Similar to the new ABB circuits, we foresee the adoption of these negative capacitance circuits to the sub-threshold benchmark circuits is highly desirable to mitigate the increased process variations in the sub-threshold region. In addition, the negative capacitance technique can be combined with other yield improvement techniques such as statistical gate sizing and ABB, to further improve the yield and increase the circuit robustness. Finally, the adoption of the ABB circuits is recommended in applications where the spatial correlated variations are dominating such as high performance applications with high V_{DD} in which the systematic channel length variations (systematic variations) are dominating. Whereas, the negative capacitance circuits are recommended for applications with highly capacitive nodes such as the SRAM bitlines, wide fan-in OR gates, and long interconnect lines.

APPENDICES

Appendix A

The Lambert W Function

In mathematics, the Lambert W function, named after Johann Heinrich Lambert, also called the Omega function $\Omega(x)$, is the inverse function of $f(x) = x\ exp(x)$ and x is any complex number. If x is real and $\{\exp(-1) \le x < 0\}$, two possible real values of $\Omega(x)$ exist: The branch, satisfying $\{-1 \le \Omega(x)\}$, is denoted by $\Omega_0(x)$ and is called the principal branch of $\Omega(x)$, and the other branch, satisfying $\{\Omega(x) \le -1\}$, is denoted by $\Omega_{-1}(x)$. If x is real and $\{x \ge 0\}$, there is a single real value for $\Omega(x)$ which also belongs to the principal branch, $\Omega_0(x)$. Both real branches $\Omega_0(x)$ and $\Omega_{-1}(x)$, for real x, are plotted in Figure A.1 [132]. The real branch, $\Omega_{-1}(x)$, is used in the proposed models.

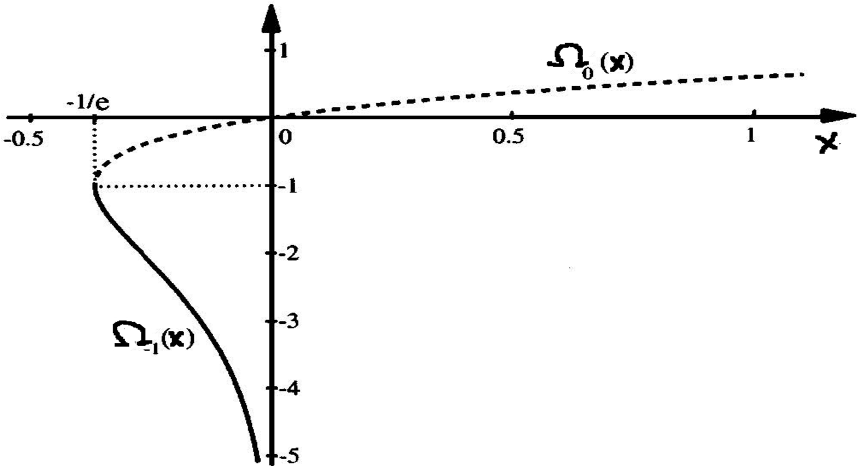


Figure A.1: Two real branches of the Omega function. Solid line: $\Omega_{-1}(x)$ defined for $\{\exp(-1) \le x < 0\}$. Dashed line:$\Omega_0(x)$ defined for $\{\exp(-1) \le x < \infty\}$. The two branches meet at point (exp(−1), -1) [132].

Appendix B

The CFSQP Optimization Package

CFSQP is a set of C functions for the minimization of the maximum of a set of smooth objective functions subject to general smooth constraints. If the initial guess provided by the user is infeasible for some inequality constraint or some linear equality constraint, CFSQP first generates a feasible point for these constraints; subsequently the successive iterates generated by CFSQP all satisfy these constraints. Nonlinear equality constraints are turned into inequality constraints (to be satisfied by all iterates) and the maximum of the objective functions is replaced by an exact penalty function which penalizes nonlinear equality constraint violations only. When solving problems with many sequentially related constraints (or objectives), such as discretized Semi-Infinite Programming (SIP) problems, CFSQP gives the user the option to use an algorithm that efficiently solves these problems, greatly reducing computational effort. The user has the option of either requiring that the objective function decrease at each iteration after feasibility for nonlinear inequality and linear constraints has been reached (monotone line search), or requiring a decrease within at most four iterations (non-monotone line search). The user must provide functions that define the objective functions and constraint functions and may either provide functions to compute the respective gradients or require that CFSQP estimate them by forward finite differences [165].

CFSQP is an implementation of two algorithms based on Sequential Quadratic Programming (SQP), modified so as to generate feasible iterates. In the first one (monotone line search), a certain Armijo type arc search is used with the property that the step of one is eventually accepted, a requirement for super-linear convergence. In the second one the same effect is achieved by means of a ”non-monotone” search along a straight line. The merit function used in both searches is the maximum of the objective functions if there is no nonlinear equality constraints, or an exact penalty function if nonlinear equality constraints are present. More details about the mathematical representation of the optimization problem can be found at [165].

Appendix C

Test Chip Details

A 1.7mm X 1.7mm test chip (CMC Run Code ICSWTABB) has been taped out through CMC (Canadian Microelectronics Corporation) Microsystems. The test chip micrograph, layout, and bonding diagram are shown in Figures C.1, C.2, and C.3, respectively. The test chip has been fabricated by using triple-well TSMC 65nm CMOS technology. Triple-well technology is required to allow the body biasing of the NMOS transistors of each critical path isolated from the substrate. This test chip consists of two main parts: (1) Proposed ABB circuits (i.e., the D-ABB and the LD-ABB circuits) applied to the high performance test circuit displayed in Figure 5.12 in Chapter 5, and (2) Proposed Negative capacitance circuits (i.e., the DA-NC, and the B-NC circuits) adopted to the wide fan-in dynamic OR gate, discussed in Chapter 6.

In order to model the D2D variations, large number of packaged chips should be tested (i.e., 62 chips are used in [46]). However, since these test chips are fabricated in the university environment, only a few packaged chips are available. In this testing process, **35** packaged chips are available to account for the D2D variations.

The test chip is packaged in a 80-pin Ceramic Flat Package (CFP80) available through CMC. The DC voltages for the test chip are generated using potentiometers implemented on an external breadboard. Large coupling capacitors are added with these potentiometers to minimize the supply noise. The PCB Test Fixture (CFP80TF), provided by the CMC Microsystems for testing integrated circuits operating at frequencies up to 10.0GHz that have been packaged in the CFP80 package supplied by CMC, is used in this testing process. This CFP80TF, shown in Figure C.4, allows clamping the test chip to the test fixture without soldering. This clamping reduces the difficulty of soldering/resoldering 35 packaged chips during the testing process at the expense of lower operating frequency of 3.0GHz. This 3.0GHz operating frequency is satisfactory since the maximum frequency expected from the tested circuits is 1.0GHz. All the high frequency signals are provided through SMA connectors.

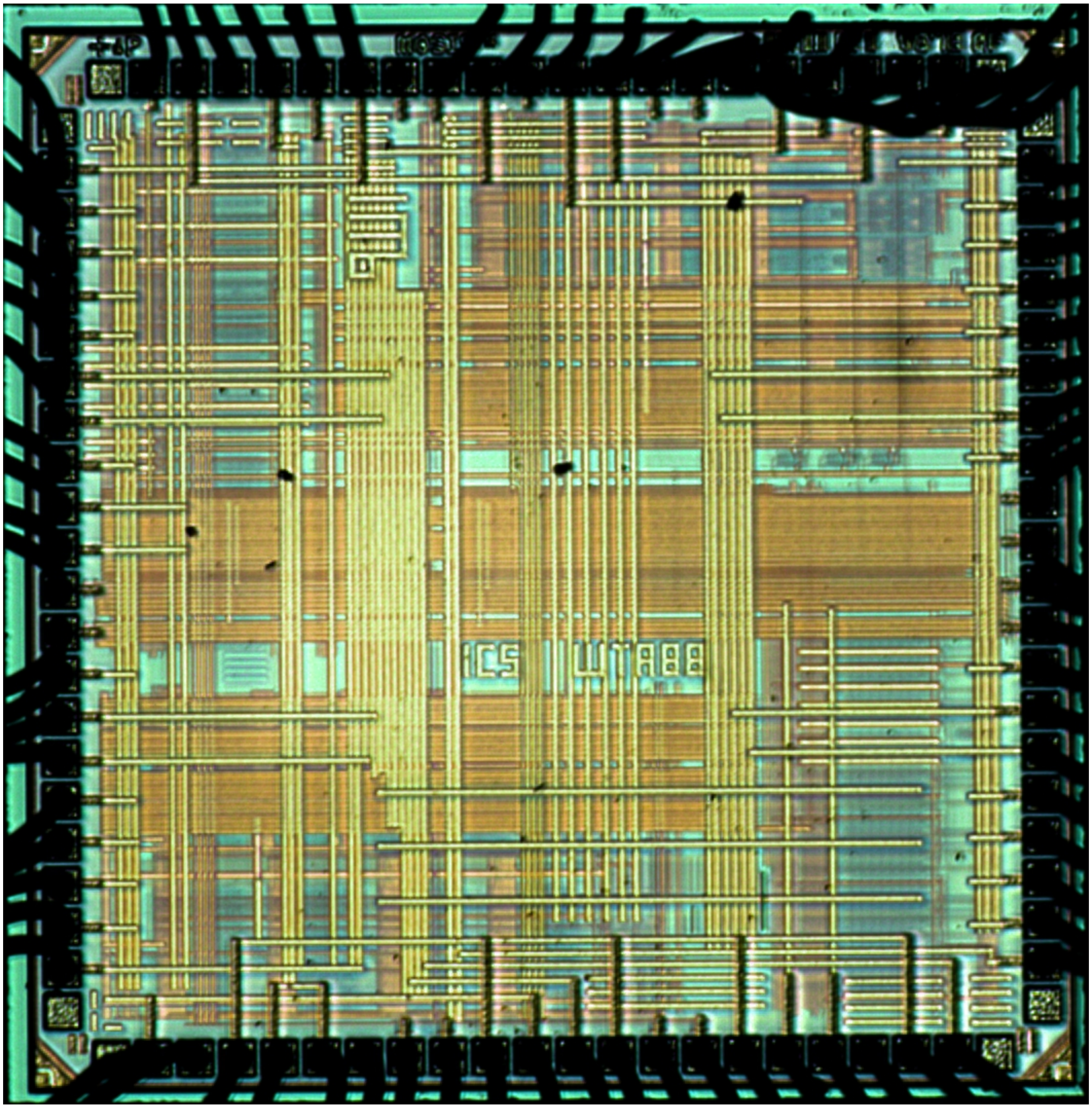


Figure C.1: Test chip micrograph.

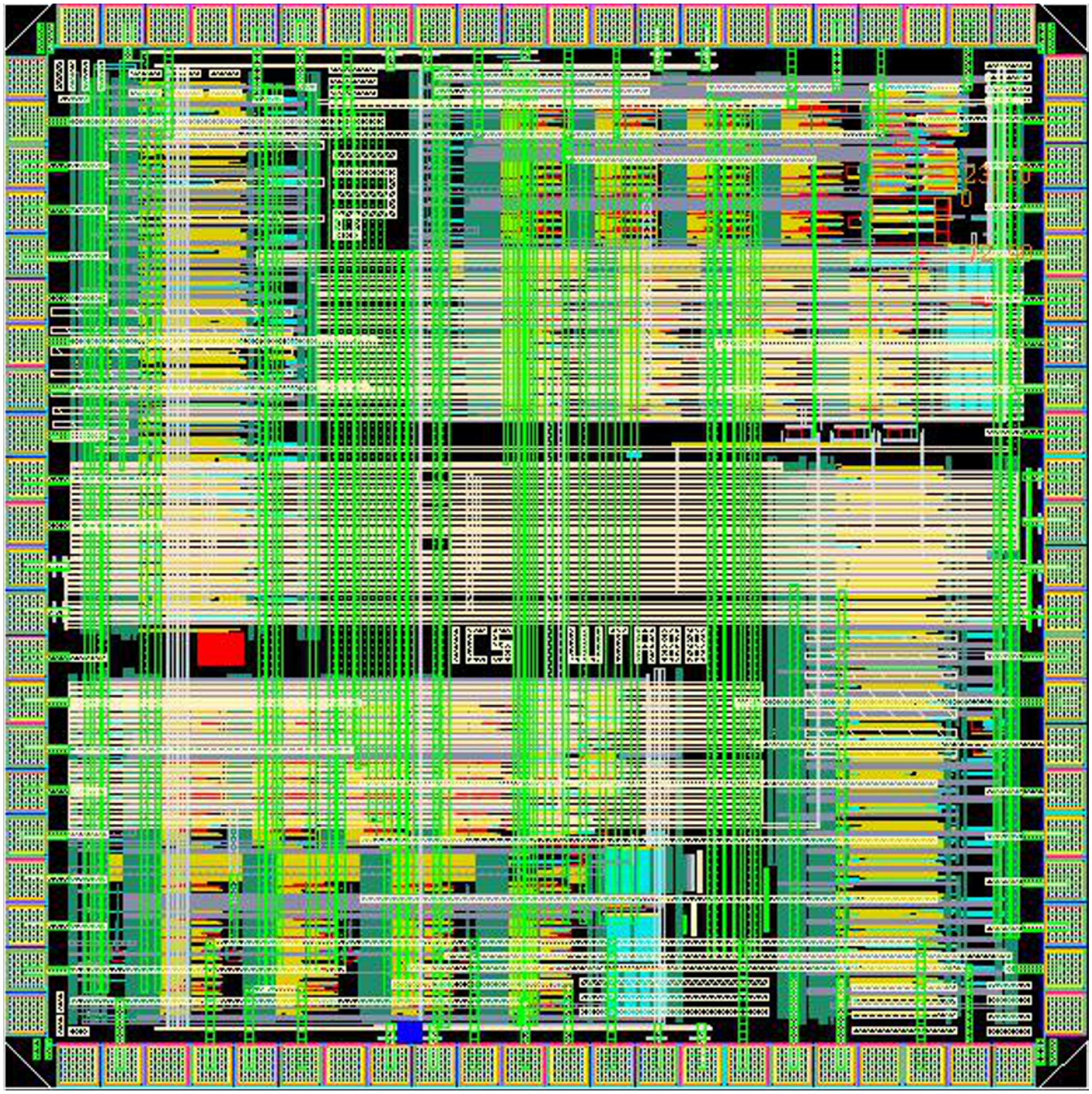


Figure C.2: Test chip layout.

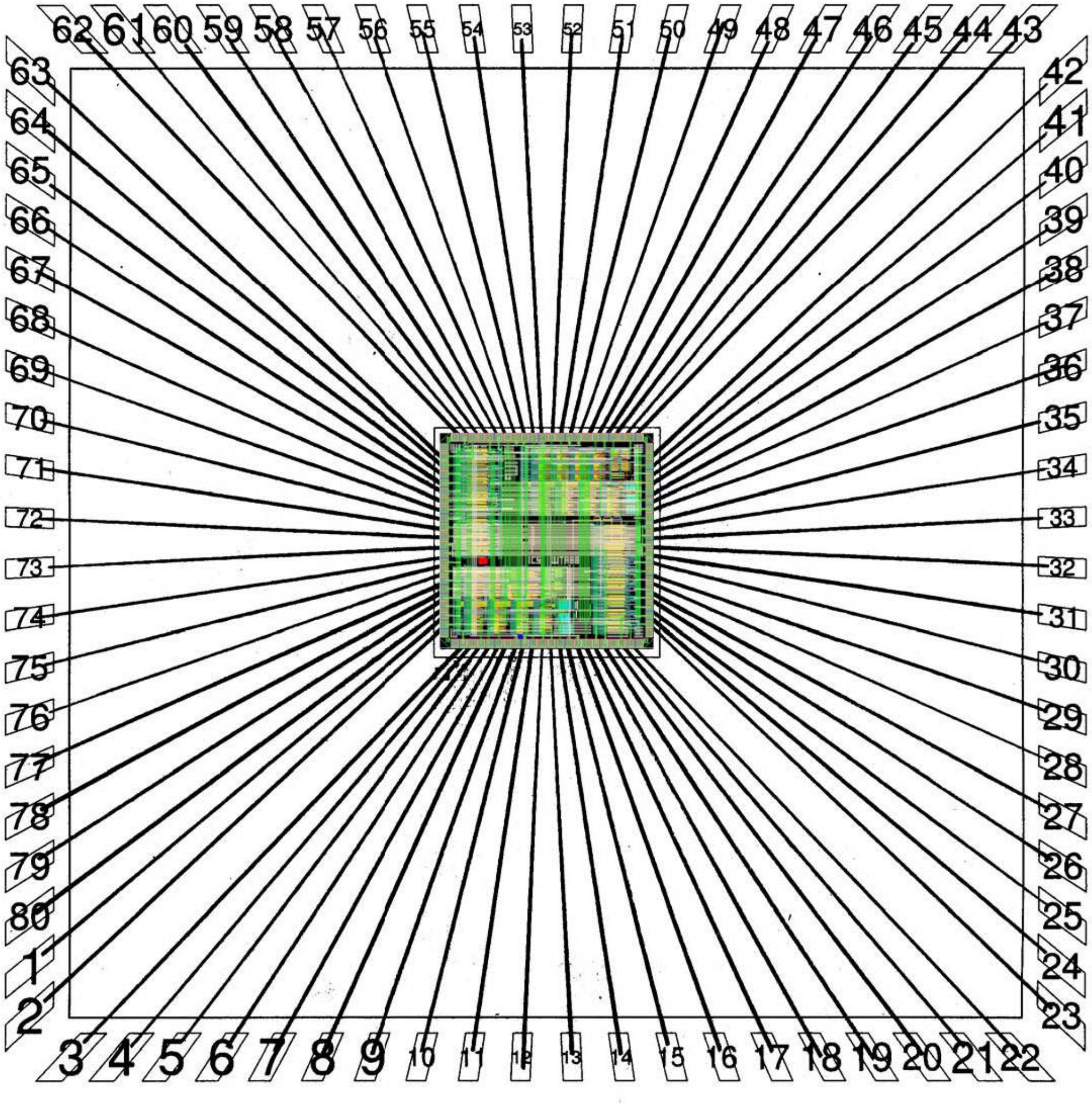


Figure C.3: Bonding diagram of the test chip. Package type: CFP80.

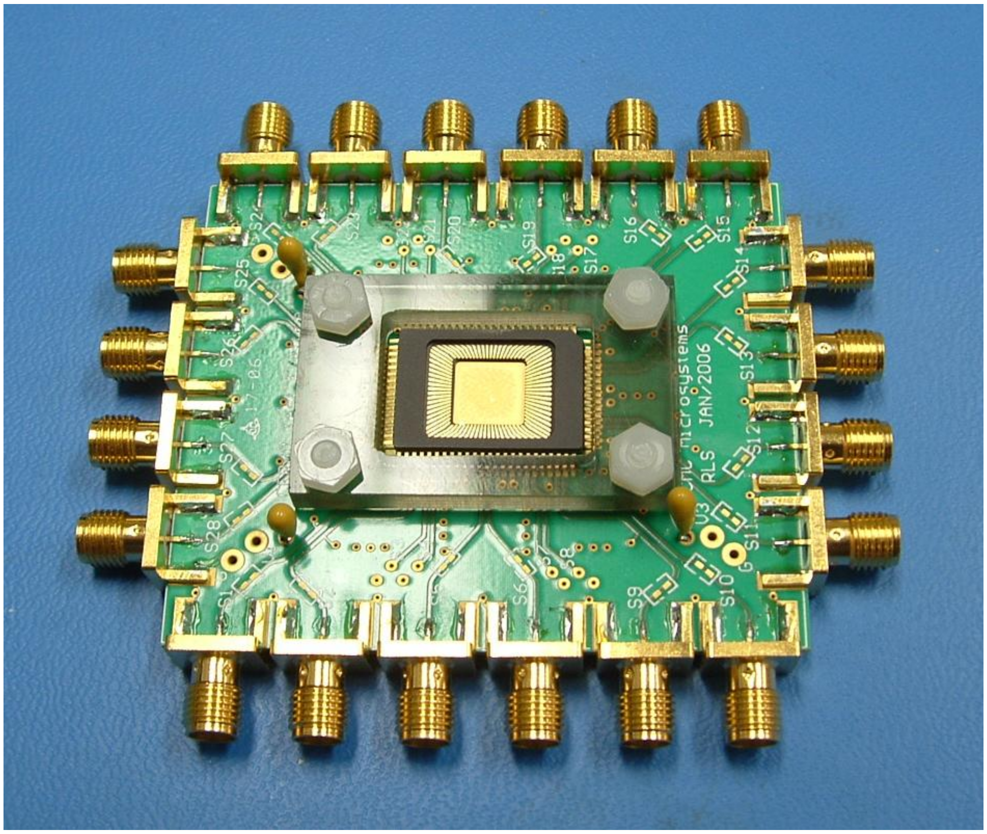


Figure C.4: PCB Test Fixture (CFP80TF) for test chip measurements.

Appendix D

Industrial Hardware-Calibrated Statistical Transistor Model

The industrial hardware-calibrated 65nm CMOS technology transistor statistical models are used to investigate the process variations impact. In [166, 167], it has been demonstrated that the utilization of statistical transistor models is capable of accounting for both D2D and WID variations. A very good fitting with the measured data is reported in [166, 167], not only for the mean and standard deviation values, but also for the correlation between NMOS and PMOS transistors data. These statistical models are available in the design kits provided by the manufacturer (i.e., STMicroelectronics and TSMC). The process variations (D2D and WID variations) are included in the transistor design kit and declared by the manufacturer to be Silicon verified. In this design kit, several process parameters are treated as variants such as the threshold voltage, mobility, drain-to-source resistance, Drain-Induced-Barrier-Lowering (DIBL) coefficient, all junction capacitances, and doping concentration. For example, the threshold voltage, V_t, is varied within the $\pm 3\sigma$ design space with standard deviation to mean ratio,$(\sigma/\mu)_{Vt} \approx 12\%$. Also, in this design kit, the WID variations (mismatch effect) are modeled as inversely proportional to the transistor area (W*L) [17]. These statistical models are used in all Monte Carlo simulations conducted in this thesis.

Bibliography

[1] S. Bhunia, and S. Mukhopadhyay, "Low-Power Variation-Tolerant Design in Nanometer Silicon," *Springer*, 2011. xiii, xiv, 1, 4, 9, 27, 29, 46, 47, 48

[2] M. Stanisavljevic, A. Schmid, and Y. Leblebici, "Reliability of Nanoscale Circuits and Systems: Methodologies and Circuit Architectures," *Springer*, 2011. 1

[3] J. Tschanz, K. Bowman, and V. De, "Variation-Tolerant Circuits: Circuit Solutions and Techniques," *Proceedings of the IEEE Design Automation Conference (DAC'05)*, pp. 762–763, 2005. xii, 2, 12, 13, 15

[4] H. Masuda, S. Ohkawa, A. Kurokawa, and M. Aoki, "Challenge: Variability Characterization and Modeling for 65-nm to 90-nm Processes," *Proceedings of the IEEE Custom Integrated Circuits Conference (CICC'05)*, pp. 593–599, 2005. 52

[5] B. Wong, A. Mittal, Y. Cao, and G. W. Starr, "Nano-CMOS Circuit and Physical Design," *Wiley-Interscience*, 2004. 8, 9, 14

[6] ITRS Web-site, "The International Technology Roadmap for Semiconductors." `http://public.itrs.net`, 2010. xiii, 2, 3, 30, 31, 32, 121, 146

[7] S. Sapatnekar, "Timing," *New York: Springer-Verlag*, 2004. 2, 8, 11, 13, 16

[8] K. Agarwal and S. Nassif, "Statistical Analysis of SRAM Cell Stability," *Proceedings of the Design Automation Conference (DAC'06)*, pp. 57–62, 2006. 2, 30, 32, 33, 184

[9] S. Mukhopadhyay, H. Mahmoodi, and K. Roy, "Modeling of Failure Probability and Statistical Design of SRAM Array for Yield Enhancement in Nanoscaled CMOS," *IEEE Transactions on Computer Aided Design (TCAD) of Integrated Circuits and Systems*, vol. 24, pp. 1859–1879, December 2005. 2, 59, 186

[10] P.R. Gada, W.R. Roberts, and D. Velenis, "Effects of Parameter Variations on Timing Characteristics of Clocked Registers," *Proceedings of the International Conference on Electro Information Technology*, pp. 1–4, 2005. 3, 43, 102

[11] R.C. Baumann, "Soft Errors in Commercial Semiconductor Technology: Overview and Scaling Trends," *IEEE Reliability Physics Tutorial Notes, Reliability Fundamentals*, vol. 121. 3

[12] A. Cataldo, EE Times, "SRAM Soft Errors Cause Hard Network Problems." 3

[13] A. Chandrakasan, W. J. Bowhill, and F. Fox, "Design of High Performance Microprocessor Circuits," *IEEE Press*, 2001. 7, 8, 9

[14] A. Srivastava, D. Sylvester, and D. Blaauw, "Statistical Analysis and Optimization for VLSI: Timing and Power," *Series on Integrated Circuits and Systems, Springer*, 2005. 8, 11, 13, 16, 184

[15] Y. Taur and T. H. Ning, "Fundamentals of Modern VLSI Devices," *New York, NY, USA: Cambridge University Press*, 1998. xii, 9, 11, 12, 13, 79, 105, 186

[16] T. Mizuno, J. Okumtura, and A. Toriumi, "Experimental Study of Threshold Voltage Fluctuation due to Statistical Variation of Channel Dopant Number in MOSFET's," *IEEE Transactions on Electron Devices*, vol. 41, pp. 2216–2221, November 1994. 9

[17] M. Pelgrom, A. Duinmaijer, and A. Welbers, "Matching Properties of MOS Transistors," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 24, pp. 1433–1439, October 1989. 9, 128, 186, 209

[18] B. Razavi, "Design of Analog CMOS Integrated Circuits," *McGraw-Hill*, 2000. 9, 127

[19] T.-C. Chen, "Where is CMOS Going: Trendy Hype Versus Real Technology," *Digest of Technical Papers of the International Solid-State Circuits Conference (ISSCC'06)*, pp. 22–28, 2006. xii, 10, 11

[20] D. Frank, R. Dennard, E. Nowak, P. Solomon, Y. Taur, and H. S. Wong, "Device Scaling Limits of Si MOSFETs and Their Application Dependencies," *Proceedings of IEEE*, vol. 89, pp. 259–288, March 2001. xii, 10, 14

[21] J. Luo, S. Sinha, Q. Su, J. Kawa, and C. Chiang, "An IC Manufacturing Yield Model Considering Intra-Die Variations," *Proceedings of the IEEE Design Automation Conference (DAC'06)*, pp. 749–754, 2006. 11

[22] W. Liu, "MOSFET Models for SPICE Simulation Including BSIM3v3 and BSIM4," *John Wiley & Sons, Inc*, 2001. 11, 27, 28, 80, 124, 137, 154

[23] M. H. Abu-Rahma, and M. Anis, “A Statistical Design-Oriented Delay Variation Model Accounting for Within-Die Variations,” *IEEE Transactions on Computer Aided Design (TCAD) of Integrated Circuits and Systems*, vol. 27, pp. 1983–1995, November 2008. 11, 16, 52, 59, 185, 186

[24] A. Asenov, A.R. Brown, J.H. Davies, S. Kaya, and G. Slavcheva, “Simulation of Intrinsic Parameter Fluctuations in Decananometer and Nanometer-Scale MOSFETs,” *IEEE Transactions on Electron Devices*, vol. 50, pp. 1837–1852, September 2003. 11, 184, 186

[25] J. A. Croon, W. Sansen, and H. E. Maes, “Matching Properties of Deep Sub-Micron MOS Transistors,” *Springer*, 2005. xii, 11, 13

[26] M. Popovich, A. V. Mezhiba, and E. G. Friedman, “Power Distribution Networks with On-Chip Decoupling Capacitors,” *Springer*, 2008. 13

[27] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De, “Parameter Variations and Impact on Circuits and Micro-Architecture,” *Proceedings of the IEEE Design Automation Conference (DAC’03)*, pp. 338–342, 2003. xii, 14, 15, 171, 187

[28] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, “Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits,” *Proceedings of IEEE*, vol. 91, pp. 305–327, Feburary 2003. xii, 14, 15, 79

[29] Z. Chen, M. Johnson, L. Wei, and K. Roy, “Estimation of Standby Leakage Power in CMOS Circuits Considering Accurate Modeling of Transistor Stacks,” *Proceedings of the IEEE International Symposium on Low Power Electronics and Design (ISLPSD’98)*, pp. 239–244, 1998. 14

[30] S. Narendra, V. De, S. Borkar, D. Antoniadis, and A.Chandrakasan, “Full-Chip Sub-Threshold Leakage Power Prediction and Reduction Techniques for Sub-0.18μm CMOS,” *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 39, pp. 501–510, Feburary 2004. 14

[31] S. Narendra, V. De, S. Borkar, D. Antoniadis, and A. Chandrakasan, “Full-Chip Sub-Threshold Leakage Power Prediction Model for Sub-0.18μm CMOS,” *Proceedings of the International Symposium on Low Power Electronics and Design (ISLPED’02)*, pp. 19–23, 2002. 14, 122

[32] S. Nassif, “Waiting for the Post-CMOS Godot,” *Keynote Speaker Slides in the Great Lakes Symposium on VLSI (GLSVLSI’11)*, pp. 1–40, 2011. xii, 15, 16

[33] S. Borkar, T. Karnik, and V. De, "Design and Reliability Challenges in Nanometer Technologies," *Proceedings of the IEEE Design Automation Conference (DAC'04)*, pp. 75–75, 2004. 15, 171, 187

[34] S. H. Choi, B. C. Paul, and K. Roy, "Novel Sizing Algorithm for Yield Improvement under Process Variation in Nanometer Technology," *Proceedings of the IEEE Annual Design Automation Conference (DAC'04)*, pp. 454–459, 2004. 16, 17, 44, 102, 105, 171

[35] A. Agarwal, K. Chopra, and D. Blaauw, "Statistical Timing Based Optimization Using Gate Sizing," *Proceedings of the IEEE Conference on Design, Automation, and Test in Europe (DATE'05)*, pp. 400–405, 2005. 16, 17, 44, 102, 171

[36] H. Fukui, M. Hamaguchi, H. Yoshimura, H. Oyamatsu, F. Matsuoka, T. Noguchi, T. Hirao, H. Abe, S. Onoda, T. Yamakawa, T. Wakasa, and T. Kamiya, "Comprehensive Study on Layout Dependence of Soft Errors in CMOS Latch Circuits and its Scaling Trend for 65nm Technology Node and Beyond," *Digest of Technical Papers in VLSI Circuits Symposium*, pp. 222–223, 2005. 17, 45

[37] S. M. Jahinuzzaman, M. Sharifkhani, and M. Sachdev, "Investigation of Process Impact on Soft Error Susceptibility of Nanometric SRAMs Using a Compact Critical Charge Model," *Proceedings of the IEEE International Symposiums on Quality Electronic Design (ISQED'08)*, pp. 207–212, 2008. 26, 36, 52, 53, 54, 55, 56, 57, 62, 78, 79, 81, 82, 91

[38] T. Chen and S. Naffziger, "Comparison of Adaptive Body Bias (ABB) and Adaptive Supply Voltage (ASV) for Improving Delay and Leakage under the Presence of Process Variation," *IEEE Transactions on Very Large Scale Integration (TVLSI) Systems*, vol. 11, pp. 888–899, October 2003. 17

[39] B. H. Calhoun and A. P. Chandrakasan, "Static Noise Margin Variation for Sub-Threshold SRAM in 65-nm CMOS," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 41, pp. 1673–1679, July 2006. 78

[40] F. Frustaci, P. Corsonello, S. Perri, and G. Cocorullo, "High-Performance Noise-Tolerant Circuit Techniques for CMOS Dynamic Logic," *IET Journal of Circuits, Devices, and Systems*, vol. 2, pp. 537–548, December 2008. 17, 168

[41] M. Olivieri, G. Scotti, and A. Trifiletti, "A Novel Yield Optimization Technique for Digital CMOS Circuits Design by Means of Process Parameters Run-Time Estimation and Body Bias Active Control," *IEEE Transactions on Very Large Scale Integration (TVLSI) Systems*, vol. 13, pp. 630–638, May 2005. xviii, 17, 123, 126, 133, 134, 139, 140, 158

[42] A. Hokazono, S. Balasubramanian, K. Ishimaru, H. Ishiuchi, T-J K. Liu, and C. Hu, "MOSFET Design for Forward Body Biasing Scheme," *IEEE Transactions on Electron Device Letters*, vol. 27, pp. 387–389, May 2006. 130

[43] J. Gregg, and T. W. Chen, "Post Silicon Power/Performance Optimization in the Presence of Process Variations Using Individual Well-Adaptive Body Biasing," *IEEE Transactions on Very Large Scale Integration (TVLSI) Systems*, vol. 15, pp. 366–376, March 2007. 122

[44] X. He, S. Al-Kadry, and A. Abdollahi, "Adaptive Leakage Control on Body Biasing for Reducing Power Consumption in CMOS VLSI Circuits," *Proceedings of the International Symposium on Quality Electronic Design (ISQED'09)*, pp. 465–470, 2009. 122

[45] S. V. Kumar, C. H. Kim, and S. S. Sapatnekar, "Body Bias Voltage Computation for Process and Temperature Compensation," *IEEE Transactions on Very Large Scale Integration (TVLSI) Systems*, vol. 16, pp. 249–262, March 2008. 17, 123

[46] J. Tschanz, J. Kao, S. Narendra, R. Nair, D. Antoniadis, A. Chandrakasan, and V. De, "Adaptive Body Bias for Reducing Impacts of Die-To-Die and Within-Die Parameter Variations on Microprocessor Frequency and Leakage," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 37, pp. 1396–1402, November 2002. xi, xii, 17, 18, 19, 121, 123, 131, 133, 134, 139, 140, 142, 145, 146, 158, 204

[47] J. Tschanz, S. Narendra, R. Nair, and V. De, "Effectiveness of Adaptive Supply Voltage and Body Bias for Reducing Impact of Parameter Variations in Low Power and High Performance Microprocessors," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 38, pp. 826–829, May 2003. 18

[48] K. Bowman, S. Duvall, and J. 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 (JSSC)*, vol. 37, pp. 183–190, February 2002. 19, 20, 134

[49] K. Bowman, S. Duvall, and J. Meindl, "Impact of Die-To-Die and Within-Die Parameter Fluctuations on the Maximum Clock Frequency Distribution," *Proceedings of the IEEE International Solid-State Circuits Conference (ISSCC'01)*, pp. 278–279, 2001. xii, 20

[50] K. Bowman and J. Meindl, "Impact of Within-Die Parameter Fluctuations on Future Maximum Clock Frequency Distributions," *Proceedings of the IEEE Conference on Custom Integrated Circuits (CICC'01)*, pp. 229–232, 2001. 19

[51] P. Shivakumar, S. W. Keckler, D. Burger, M. Kistler, and L. Alvisi, “Modeling the Effect of Technology Trends on the Soft Error Rate of Combinational Logic,” *Proceedings of the International Conference on Dependable Systems and Networks*, pp. 389–398, 2002. 21, 22, 51

[52] R. Phelan, Product manager at ARM, “Solutions for Soft Errors in System on Chip Designs.” 21

[53] R. Baumann, “Soft Errors in Advanced Computer Systems,” *IEEE Design and Test of Computers*, vol. 22, pp. 258–266, June 2005. 21, 22

[54] A. Pavlov, and M. Sachdev, “CMOS SRAM Circuit Design and Parametric Test in Nano-Scaled Technologies,” *Springer*, 2008. xiii, 21, 22, 24, 25, 31, 35, 36

[55] V. Degalahal, N. Vijaykrishnan, and M.J. Irwin, “Analyzing Soft Errors in Leakage Optimized SRAM Design,” *Proceedings of the IEEE International Conference on VLSI Design*, pp. 227–233, 2003. 21, 22, 38, 54, 79

[56] R. C. Baumanm, “Soft Errors in Advanced Semi-conductor Devices-Part I: the Three Radiation Sources,” *IEEE Transactions on Device and Materials Reliability*, vol. 1, pp. 17–22, March 2001. 21, 22, 45, 51

[57] R. Baumann, “Radiation-Induced Soft Errors in Advanced Semiconductor Technologies,” *IEEE Transactions on Device and Materials Reliability*, vol. 5, pp. 305–316, September 2005. xiii, 23, 24, 26, 36

[58] P. Hazucha, and C. Svensson, “Impact of CMOS Technology Scaling on the Atmospheric Neutron Soft Error Rate,” *IEEE Transactions on Nuclear Science*, vol. 47, pp. 2586–2594, December 2000. 22, 24, 51

[59] J. M. Palau, G. Hubert, K. Coulie, B. Sagnes, M.-C. Calvet, and S. Fourtine, “Device Simulation Study of the SEU Sensitivity of SRAMs to Internal Ion Tracks Generated by Nuclear Reactions,” *IEEE Transactions on Nuclear Science*, vol. 48, pp. 225–231, April 2001. 22, 23, 52, 53, 57

[60] T. Heijmen, D. Giot, and P. Roche, “Factors that Impact the Critical Charge of Memory Elements,” *Proceedings of the IEEE International On-Line Testing Symposium (IOLTS’06)*, pp. 57–62, 2006. 23, 51, 54, 68

[61] T. Karnik, B. Boechel, K. Soumyanath, V. De, and S. Borkar, “Scaling Trends of Cosmic Rays Induced Soft Errors in Static Latches Beyond 0.18μ,” *Digest of Technical Papers in VLSI Circuits Symposium*, pp. 61–62, 2001. 24

[62] T. Heijmen, "Analytical Semi-Empirical Model for SER Sensitivity Estimation of Deep-Submicron CMOS Circuits," *Proceedings of the IEEE International On-Line Testing Symposium (IOLTS'05)*, pp. 3–8, 2005. 24

[63] S. Mukherjee, "Architecture Design for Soft Errors," *Elsevier Inc.*, 2008. 25

[64] C. Chen and A. Somani, "Fault-Containment in Cache Memories for TMR Redundant Processor Systems," *IEEE Transactions on Computers*, vol. 48, pp. 386–397, April 1999. 26

[65] Q. Ding, R. Luo, and Y. Xie, "Impact of Process Variation on Soft Error Vulnerability for Nanometer VLSI Circuits," *Proceedings of the International Application Specific Integrated Circuits Conference (ASICON'05)*, pp. 1023–1026, 2005. 26, 52

[66] Q. Ding, R. Luo, H. Wang, H. Yang, and Y. Xie, "Modeling The Impact of Process Variation on Critical Charge Distribution," *Proceedings of the IEEE International System on Chip (SOC06)*, pp. 243–246, 2006. 26, 36, 52

[67] S. M. Jahinuzzaman, M. Sharifkhani, and M. Sachdev, "An Analytical Model for Soft Error Critical Charge of Nanometric SRAMs," *IEEE Transactions on Very Large Scale Integration (TVLSI) Systems*, vol. 17, pp. 1187–1195, September 2009. 26, 36, 52, 53, 54, 55, 56, 57, 62, 78, 79, 81, 82, 91

[68] Z. Qi, J. Wang, A. Cabe, S. Wooters, T. Blalock, B. Calhoun, and M. Stan, "SRAM-Based NBTI/PBTI Sensor System Design," *Proceedings of the IEEE Design Automation Conference (DAC'10)*, pp. 849–852, 2010. 27

[69] D. K. Schroder and J. F. Babcock, "Negative Bias Temperature Instability: Road to Cross in Deep Sub-Micron Silicon Semiconductor Manufacturing," *Journal of Applied Physics*, vol. 94, pp. 1–18, July 2003. 27

[70] D. Bhaduri, S. K. Shukla, P. S. Graham, and M. B. Gokhale, "Reliability Analysis of Large Circuits Using Scalable Techniques and Tools," *IEEE Transactions on Circuits and Systems I (TCAS-I)*, vol. 54, pp. 2447–2460, November 2007.

[71] S. Mahapatra, P. B. Kumar, and M. A. Alam, "Investigation and Modeling of Interface and Bulk Trap Generation During Negative Bias Temperature Instability of P-MOSFETs," *IEEE Transactions on Electronic Devices*, vol. 51, pp. 1371–1379, September 2004.

[72] A. Vassighi and M. Sachdev, "Thermal and Power Management of Integrated Circuits," *Springer*, 2006.

[73] A. S. Goda and G. Kapila, "Design for Degradation: CAD Tools for Managing Transistor Degradation Mechanisms," *Proceedings of the IEEE International Symposiums on Quality Electronic Design (ISQED'05)*, pp. 416–420, 2005.

[74] B. C. Paul, K. Kang, H. Kufluoglu, M. A. Alam, and K. Roy, "Impact of NBTI on the Temporal Performance Degradation of Digital Circuits," *IEEE Transactions on Electronic Devices*, vol. 26, pp. 560–562, August 2005. 27

[75] S. V. Kumar, C. H. Kim, and S. S. Sapatnekar, "Impact of NBTI on SRAM Read Stability Design for Reliability," *Proceedings of the IEEE International Symposiums on Quality Electronic Design (ISQED'06)*, pp. 213–218, 2006. 27

[76] K. K. Kim, W. Wang, and K. Choi, "On-Chip Aging Sensor Circuits for Reliable Nanometer MOSFET Digital Circuits," *IEEE Transactions on Circuits and Systems II (TCAS-II)*, vol. 57, pp. 798–802, October 2010. 27, 28

[77] K. Kang, H. Kufluoglu, K. Roy, and M. A. Alam, "Impact of Negative Bias Temperature Instability in Nanoscale SRAM Array: Modeling and Analysis," *IEEE Transactions on Computer Aided Design (TCAD) of Integrated Circuits and Systems*, vol. 26, pp. 1770–1781, October 2007. 27, 28, 33, 36, 37, 38, 151, 155

[78] T-H. Kim, R. Persaud, and C. H. Kim, "Silicon Odometer: An On-Chip Reliability Moniltor for Measuring Frequency Degradation of Digiltal Circuits," *Digest of Technical Papers in VLSI Circuits Symposium*, pp. 122–123, 2007. 28

[79] J. Keane, T-H Kim, and C. H. Kim, "An On-Chip NBTI Sensor for measuring pMOS Threshold Voltage Degradation," *IEEE Transactions on Very Large Scale Integration (TVLSI) Systems*, vol. 18, pp. 947–956, June 2010. 28

[80] T. P. Haraszti, "CMOS Memory Circuits," *Kluwer Academic Publishers*, 2000. 30

[81] S. Mukhopadhyay, H. Mahmoodi, and K. Roy, "Statistical Design and Optimization of SRAM Cell for Yield Enhancement," *Proceedings of the International Conference on Computer Aided Design (ICCAD'04)*, pp. 10–13, 2004. 30, 32, 33, 184, 186

[82] R. Heald and P. Wang, "Variability in Sub-100nm SRAM Designs," *Proceedings of the International Conference on Computer Aided Design (ICCAD'04)*, pp. 347–352, 2004. 30, 32, 33, 35, 184

[83] M. Yamaoka, and T. Kawahara, "Operating-Margin-Improved SRAM With Column-at-a-Time Body-Bias Control Technique," *Proceedings of the European Solid State Circuits Conference (ESSCIRC'07)*, pp. 396–399, 2007. 32, 33

[84] E. Seevinck, F. J. List, and J. Lohstroh, “Static Noise Margin Analysis of MOS SRAM Cells,” *IEEE Journal of Solid State Circuits (JSSC)*, vol. 22, pp. 748–754, October 1987. 33

[85] A. Bhavnagarwala, S. Kosonocky, C. Radens, K. Stawiasz, R. Mann, Q. Ye, and K. Chin, “Fluctuation Limits and Scaling Opportunities for CMOS SRAM Cells,” *Proceedings of the International Electron Devices Meeting (IEDM’05)*, pp. 659–662, 2005. xiii, 33, 34

[86] K. Kim, J-J Kim, and C-T Chuang, “Asymmetrical SRAM Cells with Enhanced Read and Write Margins,” *Proceedings of the International Symposium on VLSI Technology, Systems and Applications*, pp. 1–2, 2007. 35, 78

[87] D. F. Heidel, P. W. Marshall, J. A. Pellish, K. P. Rodbell, K. A. LaBel, J. R. Schwank, S. E. Rauch, M. C. Hakey, M. D. Berg, C. M. Castaneda, P. E. Dodd, M. R. Friendlich, A. D. Phan, C. M. Seidleck, M. R. Shaneyfelt, and M. A. Xapsos, “Single-Event Upsets and Multiple-Bit Upsets on a 45nm SOI SRAM,” *IEEE Transactions on Nuclear Science,*, vol. 56, pp. 3499–3504, December 2009. 35

[88] H. Singh and H. Mahmoodi, “Analysis of SRAM Reliability Under Combined Effect of NBTI, Process and Temperature Variations in Nano-Scale CMOS,” *Proceedings of the International Conference on Future Information Technology*, pp. 1–4, 2010. 37, 154

[89] H. Mostafa, M. Anis, and M. Elmasry, “A Design-Oriented Soft Error Rate Variation Model Accounting for Both Die-to-Die and Within-Die Variations in Sub-Micron CMOS SRAM Cells,” *IEEE Transactions on Circuits and Systems I (TCAS-I)*, vol. 57, pp. 1298–1311, June 2010. 38, 171, 176, 177, 179, 187

[90] B. Voss and M. Glesner, “A Low Power Sinusoidal Clock,” *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS’01)*, vol. 4, pp. 108–111, 2001. xiii, 38, 39, 40, 41, 42, 43, 44

[91] J. M. Rabaey, A. Chandrakasan, and B. Nikolic, “Digital Integrated Circuits, Design Prespective,” *Prentice Hall*, 2002. xiii, xvii, 39, 40, 42, 102, 103, 112, 163, 171

[92] M. Alioto, E. Consoli, and G. Palumbo, “Analysis and Comparison in the Energy-Delay-Area Domain of Nanometer CMOS Flip-Flops: Part I - Methodology and Design Strategies,” *IEEE Transactions on Very Large Scale Integration (TVLSI) Systems*, vol. 19, pp. 725–736, May 2011. 41, 43

[93] M. Alioto, E. Consoli, and G. Palumbo, “Analysis and Comparison in the Energy-Delay-Area Domain of Nanometer CMOS Flip-Flops: Part II Results and Figures

of Merit," *IEEE Transactions on Very Large Scale Integration (TVLSI) Systems*, vol. 19, pp. 737–750, May 2011. xiii, 41, 42, 43, 44

[94] F. Klass, C. Amir, A. Das, K. Aingaran, C.Truong, R. Wang, A. Mehta, R. Heald, and G. Yee, "A New Family of Semidynamic and Dynamic Flip-Flops with Embedded Logic for High-Performance Processors," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 34, pp. 712–716, May 1999. xvii, 43, 102, 103

[95] F. Ishihara, and B. Nikolic, "Level-Conversion for Dual-Supply System," *IEEE Transactions on Very Large Scale Integration (TVLSI) Systems*, vol. 12, pp. 185–195, Feburary 2004. 43

[96] M. Hansson, and A. Alvandpour, "Comparative Analysis of Process Variation Impact on Flip-Flop Power-Performance," *Proceedings of the IEEE International Symposiums on Circuits and Systems (ISCAS'07)*, pp. 3744–3747, 2007. 44, 102, 104

[97] T. Heijmen, P. Roche, G. Gasiot, K. R. Forbes, and D. Giot, "A Comprehensive Study on the Soft-Error Rate of Flip-Flops From 90-nm Production Libraries," *IEEE Transactions on Device and Materials Reliability*, vol. 7, pp. 84–96, March 2007. 45

[98] H. K. Alidash, and V. G. Oklobdzija, "Low-Power Soft Error Hardened Latch," *Journal of Low Power Electronics*, vol. 6, pp. 1–9, January 2010. 45

[99] R. Ramanarayanan, V. Degalahal, N. Vijaykrishnan, M. J. Irwin, and D. Duarte, "Analysis of Soft Error Rate in Flip-Flops and Scannable Latches," *Proceedings of the Application Specific Integrated Circuits (ASIC'03) Conference*, pp. 231–234, 2003. 45, 51

[100] T. Heijmen, "Soft Error Vulnerability of sub-100-nm Flip-Flops," *Proceedings of the IEEE International On-Line Testing Symposium (IOLTS'08)*, pp. 247–252, 2008. 45, 51

[101] S. Mitra, N. Seifert, M. Zhang, Q. Shi, and K. S. Kim, "Robust System Design with Built-In Soft-Error Resilience," *IEEE Journal for Computers*, vol. 38, pp. 43–52, March 2005. xiv, 46

[102] H. Dadgour, and K. Banerjee, "Aging-Resilient Design of Pipelined Architectures using Novel Detection and Correction Circuits," *Proceedings of the IEEE Conference on Design, Automation, and Test in Europe (DATE'10)*, pp. 244–249, 2010. 46

[103] K. Ramakrishnan, X. Wu, N. Vijaykrishnan, and Y. Xie, "Comparative Analysis of NBTI Effects on Low Power and High Performance Flip-Flops," *Proceedings of the IEEE International Conference on Computer Design (ICCD'08)*, pp. 200–205, 2008.

[104] H. Abrishami, S. Hatami, B. Amelifard, and M. Pedram, "NBTI-Aware Flip-Flop Characterization and Design," *Proceedings of the Great Lakes Symposium on VLSI (GLSVLSI'08)*, pp. 29–34, 2008. 46

[105] M. Alioto, G. Palumbo, and M. Pennisi, "Understanding the Effect of Process Variations on the Delay of Static and Domino Logic," *IEEE Transactions on Very Large Scale Integration (TVLSI) Systems*, vol. 18, pp. 697–710, May 2010. xiv, 46, 47

[106] F. Wang, Y. Xie, R. Rajaraman, and B. Vaidyanathan, "Soft Error Rate Analysis for Combinational Logic Using An Accurate Electrical Masking Model," *Proceedings of the IEEE International Conference on VLSI Design*, pp. 165–170, 2007. 48

[107] W. Wang, S. Yang, S. Bhardwaj, S. Vrudhula, F. Liu, and Y. Cao, "The Impact of NBTI Effect on Combinational Circuit: Modeling, Simulation, and Analysis," *IEEE Transactions on Very Large Scale Integration (TVLSI) Systems*, vol. 18, pp. 173–183, February 2010. 48

[108] E. H. Cannon, A. J. K. Osowski, R. Kanj, D. D. Reinhardt, and R. V. Joshi, "The Impact of Aging Effects and Manufacturing Variation on SRAM Soft-Error Rate," *IEEE Transactions on Device and Materials Reliability*, vol. 8, pp. 145–152, March 2008. 52

[109] T. Heijmen , and B. Kruseman, "Alpha-Particle-Induced SER of Embedded SRAMs Affected by Variations in Process Parameters and by The Use of Process Options," *Solid-State Electronics*, vol. 49, pp. 1783–1790, November 2005. 52

[110] Y.Z. Xu, H. Puchner, A. Chatila, O. Pohland, B. Bruggeman, B. Jin, D. Radaelli and S. Daniel, "Process Impact on SRAM Alpha-Particle SEU Performance," *Proceedings of the IEEE International Reliability Physics Symposiums*, pp. 294–299, 2004. 52, 53, 57

[111] B. Zhang, A. Arapostathis, S. Nassif, and M. Orshansky, "Analytical Modeling of SRAM Dynamic Stability," *Proceedings of the IEEE International Conference on Computer Aided Design (ICCAD'06)*, pp. 315–322, 2006. 52, 53, 57

[112] G.R. Srinivasan, P.C. Murley, and H. K. Tang, "Accurate, Predictive Modeling of Soft Error Rate due to Cosmic Rays and Chip Alpha Radiation," *Proceedings of the IEEE International Reliability Physics Symposiums*, pp. 12–16, 1994. 54

[113] R. C. Jaeger, R. M. Fox, and S. E. Diehl, "Analytic Expressions for the Critical Charge in CMOS Static RAM Cells," *IEEE Transactions on Nuclear Science*, vol. 30, pp. 4616–4619, December 1983. 57

[114] T. Sakurai, and A. Newton, “Alpha-Power Law MOSFET Model and its Applications to CMOS Inverter Delay and Other Formulas,” *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 25, pp. 584–594, April 1990. 57, 126, 148, 163

[115] H. Masuda, S. Okawa, and M. Aoki, “Approach for Physical Design in Sub-100nm Era,” *Proceedings of the IEEE International Symposiums on Circuits and Systems (ISCAS’05)*, pp. 5934–5937, 2005. 59

[116] M. Eisele, J. Berthold, D. Schmitt-Landsiedel, and R. Mahnkopf, “The Impact of Intra-Die Device Parameter Variations on Path Delays and on the Design for Yield of Low Voltage Digital Circuits,” *IEEE Transactions on Very Large Scale Integration (TVLSI) Systems*, vol. 5, pp. 360–368, December 1997. 184

[117] Y. Cao, and L. T. Clark, “Mapping Statistical Process Variations Towards Circuit Performance Variability: an Analytical Modeling Approach,” *Proceedings of the IEEE Design Automation Conference (DAC’05)*, pp. 658–663, 2005.

[118] M. R. de Alba-Rosano, and A. D. Garcia-Garcia, “Measuring Leakage Power in Nanometer CMOS 6T SRAM Cells,” *Proceedings of the IEEE International Conference on Reconfigurable Computing and FPGAs*, pp. 1–7, 2006.

[119] S. V. Walstra, and C. Dai, “Circuit-Level Modeling of Soft Errors in Integrated Circuits,” *IEEE Transactions on Device and Materials Reliability*, vol. 5, pp. 358–364, September 2005. 59

[120] J. T. Horstmann, U. Hilleringmann, and K. Goser, “Correlation Analysis of the Statistical Electrical Parameter Fluctuations in 50nm MOS Transistors,” *roceedings of the European Solid-State Devices Conference*, pp. 512–515, 1998. 59

[121] G. Hubert, N. Buard, C. Weulersse, T. Carriere, M.-C. Palau, J.-M. Palau, D. Lambert, J. Baggio, F. Wrobel, F. Saigne, R. Gaillard, “A Review of DASIE Code Family Contribution to SEU/MBU Understanding,” *Proceedings of the International Online Test Symposiums (IOLTS’05)*, pp. 87–94, 2005. 68

[122] A. Wang, B. H. Calhoun, and A. P. Chandrakasan, “Sub-Threshold Design for Ultra Low-Power Systems,” *Springer*, 2006. 78

[123] N. Verma and A. P. Chandrakasan, “A 256kb 65nm 8T Subthreshold SRAM Employing Sense-Amplifier Redundancy,” *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 43, pp. 141–149, January 2008. 78

[124] B. H. Calhoun and A. P. Chandrakasan, “A 256-kb 65-nm Sub-Threshold SRAM Design for Ultra-Low-Voltage Operation,” *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 42, pp. 680–688, March 2007. xv, 78, 79

[125] A. Wang and A. P. Chandrakasan, "A 180mV FFT Processor Using Sub-Threshold Circuit Techniques," *Proceedings of the IEEE International Solid-State Circuits Conference (ISSCC'04)*, pp. 292–293, 2004.

[126] H. Li, J. Mundy, W. Patterson, D. Kazazis, A. Zaslavsky, and R. I. Bahar, "A Model for Soft Errors in the Sub-Threshold CMOS Inverter," *Proceedings of the Workshop on System Effects of Logic Soft Errors (SELSE2)*, 2006.

[127] N. Verma, J. Kwong, and A. P. Chandrakasan, "Nanometer MOSFET Variation in Minimum Energy Sub-Threshold Circuits," *IEEE Transactions on Electron Devices*, vol. 55, pp. 163–174, January 2008. 78

[128] N. Kim, K. Flautner, D. Blaauw, and T. Mudge, "Circuit and Micro-Architectural Techniques for Reducing Cache Leakage Power," *IEEE Transactions on Very Large Scale Integration (TVLSI) Systems*, vol. 12, pp. 167–184, Feburary 2004. 78

[129] K. Osada, J. L. Shin, M. Khan, Y. Liou, K.Wang, K. Shoji, K. Kuroda, S. Ikeda, and K. Ishibashi, "Universal-Vdd 0.65V–2.0V 32-kB Cache Using A Voltage-Adapted Timing-Generation Scheme and a Lithographically Symmetrical Cell," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 36, pp. 1738–1744, November 2001. 78

[130] F. Moradi, D. T. Wisland, S. Aunet, H. Mahmoodi, and T. V. Cao, "65-nm Sub-Threshold 11T SRAM for Ultra Low Voltage Applications," *Proceedings of the IEEE International Systems-On-Chip (SOC'08) Conference*, pp. 113–118, 2008. 78

[131] S. Lin, Y. Kim, and F. Lombardi, "A 32-nm SRAM Design for Low Power and High Stability," *Proceedings of the Midwest Symposium on Circuits and Systems (MWSCAS'08)*, pp. 422–425, 2008. 78

[132] F. Chapeau-Blondeau and A. Monir, "Numerical Evaluation of the Lambert W Function and Application to Generation of Generalized Gaussian Noise with Exponent 1/2," *IEEE Transactions on Signal Processing*, vol. 50, pp. 2160–2165, September 2002. xx, 82, 202

[133] G. Gerosa, S. Gary, C. Dietz, D. Pham, K. Hoover, J. Alvarez, H. Sanchez, P. Ippolito, T. Ngo, S. Litch, J. Eno, J. Golab, N. Vanderschaaf, and J. Kahle, "A 2.2 W, 80 MHz Superscalar RISC Microprocessor," *IEEE Journal of Solid State Circuits (JSSC)*, vol. 29, pp. 1440–1454, December 1994. 102

[134] U. Ko, A. M. Hill, and P. T. Balsara, "Design Techniques for High Performance, Energy Efficient Control Logic," *Proceedings of the International Symposium on Low Power Electronics and Design (ISLPED'96)*, pp. 97–100, 1996. 102

[135] V. Stojanovic, and V. G. Oklobdzija, "Comparative Analysis of Master-Slave Latches and Flip-Flops for High Performance and Low Power Systems," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 34, pp. 536–548, April 1999. 104

[136] H. P. Alstad, S. Aunet, "Three Subthreshold Flip-Flop Cells Characterized in 90nm and 65nm CMOS Technology," *Proceedings of the IEEE Workshop on Design and Diagnostics of Electronic Circuits and Systems (DDECS'08)*, pp. 1–4, 2008. 112

[137] A. Wang, and A. Chandrakasan, "A 180 mV Sub-threshold FFT Processor Circuits," *IEEE Journal for Solid State Circuits (JSSC)*, vol. 40, pp. 310–319, January 2005. 112

[138] E. L. Crow, and K. Shimizu, "Lognormal Distribution: Theory and Applications," *Statistics, Textbooks, and Monographs*, 1988. 114, 115

[139] S. H. Kulkarni, D. M. Sylvester, and D. Blaauw, "Design-Time Optimization of Post-Silicon Tuned Circuits Using Adaptive Body Bias," *IEEE Transactions on Computer-Aided Design (TCAD) of Integrated Circuits and Systems*, vol. 27, pp. 481–494, March 2008. 122, 123, 124

[140] K. Kang, S. P. Park, K. Kim, and K. Roy, "On-Chip Variability Sensor Using Phase-Locked Loop for Detecting and Correcting Parametric Timing Failures," *IEEE Transactions on Very Large Scale Integration (TVLSI) Systems*, vol. 18, pp. 270–280, Feburary 2010. 122

[141] B. Choi and Y. Shin, "Lookup Table-Based Adaptive Body Biasing of Multiple Macros," *Proceedings of the International Symposium on Quality Electronic Design (ISQED'07)*, pp. 533–538, 2007. 122

[142] M. Mani, A. K. Singh, and M. Orshansky, "Joint Design-Time and Post-Silicon Minimization of Parametric Yield Loss using Adjustable Robust Optimization," *Proceedings of the International Conference on Computer Aided Design (ICCAD'06)*, pp. 19–26, 2006. 123

[143] R. Hidayat, K. Dejhan, P. Moungnoul, and Y. Miyanaga, "OTA-Based High Frequency CMOS Multiplier and Squaring Circuit," *Proceedings of the International Symposium on Intelligent Signal Processing and Communication Systems*, pp. 1–4, 2008. 128

[144] B. Boonchu and W. Surakampontorn, "A New nMOS Four-Quadrant Analog Multiplier," *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS'05)*, pp. 1004–1007, 2005. 129

[145] H. Mostafa, M. Anis, and M. Elmasry, "Comparative Analysis of Timing Yield Improvement under Process Variations of Flip-Flops Circuits," *Proceedings of IEEE International Symposiums on Very Large Scale Integration (ISVLSI '09)*, pp. 133–138, 2009. 128, 171

[146] S. Lakshminarayanan, J. Joung, G. Narasimhan, R. Kapre, M. Slanina, J. Tung, M. Whately, C-L. Hou, W-J. Liao, S-C. Lin, P-G. Ma, C-W. Fan, M-C. Hsieh, F-C. Liu, K-L.Yeh, W-C. Tseng, and S.W. Lu, "Standby Power Reduction and SRAM Cell Optimization for 65nm Technology," *Proceedings of the IEEE International Symposium on Quality Electronic Design (ISQED'09)*, pp. 471–475, 2009. 130

[147] Actel Application Note AC204, "Designing Clean Analog PLL Power Supply in a Mixed-Signal Environment." 139

[148] J-W Yang, P-T Huang, and W. Hwang, "On-Chip DC-DC Converter with Frequency Detector for Dynamic Voltage Scaling Technology," *Proceedings of the IEEE Asia Pacific Conference on Circuits and Systems (APCCAS'06)*, pp. 667–670, 2006. 140

[149] S-L Chen and M-D Ker, "A New Schmitt Trigger in a 0.13-μm 1/2.5-V CMOS Process to Receive 3.3-V Input Signals," *IEEE Transactions on Circuits and Systems II (TCAS-II)*, vol. 52, pp. 361–365, July 2005. 148, 149

[150] D. J. Comer, D. T. Comer, J. B. Perkins, K. D. Clark, and A. P. C. Genz, "Bandwidth Extension of High-Gain CMOS Stages Using Active Negative Capacitance," *Proceedings of IEEE International Conference on Electronics, Circuits, and Systems (ICECS'06)*, pp. 628–631, 2006. xix, 162, 165

[151] E. Yuce, "Negative Impedance Converter With Reduced Nonideal Gain and Parasitic Impedance Effects," *IEEE Transactions on Circuits and Systems I (TCAS-I)*, vol. 55, pp. 276–283, Feburary 2008. xix, 165

[152] H. Mostafa and A. M. Soliman, "Novel Accurate Wideband CMOS Current Conveyor," *German Frequenz Journal of Engineering and Telecommunications*, vol. 60, pp. 11–12, July 2006. xix, 165, 167

[153] S. Minaei, O. K. Sayin, and H. Kuntman, "A New CMOS Electronically Tunable Current Conveyor and Its Application to Current-Mode Filters," *IEEE Transactions on Circuits and Systems I (TCAS-I)*, vol. 53, pp. 1448–1457, July 2006. 166

[154] H. F. Dadgour, and K. Banerjee, "A Novel Variation-Tolerant Keeper Architecture for High-Performance Low-Power Wide Fan-In Dynamic OR Gates," *IEEE Transactions on Very Large Scale Integration (TVLSI) Systems*, vol. 18, pp. 1567–1577, November 2010. xix, 168, 170

[155] W. Hwang, R. V. Joshi, and W. H. Henkels, “A 500-MHz, 32-Word x 64-Bit, Eight-Port Self-Resetting CMOS Register File,” *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 34, pp. 56–67, January 1999. 168

[156] Online article, “Intel Pentium IV 1.4 GHz Review, Part 1: Processor Architecture and Platform Overview.” xix, 168, 169

[157] G. Yee and C. Sechen, “Clock-Delayed Domino for Adder and Combinational Logic Design,” *Proceedings of the IEEE International Conference on Computer Design (ICCD’96)*, pp. 332–337, 1996. 168

[158] C. J. Akl and M. A. Bayoumi, “Single-Phase SP-Domino: A Limited-Switching Dynamic Circuit Technique for Low-Power Wide Fan-in Logic Gates,” *IEEE Transactions on Circuits and Systems II (TCAS-II)*, vol. 55, pp. 141–145, Feburary 2008. 168, 169

[159] L. Wang, R. K. Krishwamurthy, K. Soumyanath, and N. R. Shanbhag, “An Energy-Efficient Leakage-Tolerant Dynamic Circuit Technique,” *Proceedings of the IEEE International Conference on Application Specific Integrated Circuits and System on Chip (ASIC/SOC’00)*, pp. 221–225, 2000. 168

[160] F. Moradi, and A. Peiravi, “An Improved Noise-Tolerant Domino Logic Circuit for High Fan-in Gates,” *Proceedings of the International Conference on Microelectronics (ICM’05)*, pp. 116–121, 2005. 174

[161] T. Shakir, D. Rennie, and M. Sachdev, “Integrated Read AssistSense Amplifier Scheme for High Performance Embedded SRAMs,” *Proceedings of the International Midwest Symposium on Circuits and Systems (MWSCAS’10)*, pp. 137–140, 2010. 185, 186

[162] M. H. Abu-Rahma, M. Anis, and S. S. Yoon, “Reducing SRAM Power Using Fine-Grained Wordline Pulsewidth Control,” *IEEE Transactions on Very Large Scale Integration (TVLSI) Systems*, vol. 18, pp. 356–364, March 2010. 185

[163] V. Gupta and M. Anis, “Statistical Design of the 6T SRAM Bit Cell,” *IEEE Transactions on Circuits and Systems I (TCAS-I)*, vol. 57, pp. 93–104, January 2010. 186

[164] A. Papoulis and S. U. Pillai, “Probability, Random Variables and Stochastic Process,” *New York: McGraw-Hill*, 2001. 186

[165] C. Lawrence, J. L. Zhou, and A. L. Tits, “User’s Guide for CFSQP Version 2.5: A C Code for Solving (Large Scale) Constrained Nonlinear (Minimax) Optimization Problems, Generating Iterates Satisfying All Inequality Constraints.” 203

[166] Q. Zhang, J. J. Liou, J. McMacken, K. Stiles, J. Thomson, and P. Layman, “An Efficient and Practical MOS Statistical Model for Digital Applications,” *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS’00)*, pp. 433–436, 2000. 209

[167] T. S. Gotarredona and B. L. Barranco, “A New 5-Parameter MOS Transitors Mismatch Model,” *Proceedings of IEEE International Conference of Electronics, Circuits, and Systems (ICECS’99)*, pp. 315–318, 1999. 209