The Dissertation Committee for Bin Zhang certifies that this is the approved version of the following dissertation:

IC Design for Reliability

Committee:

Michael Orshansky, Supervisor

Ari Arapostathis

Todd Arbogast

Jack Lee

David Pan

Nur Touba

IC Design for Reliability

by

Bin Zhang, B.S.; M.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May, 2009

Dedication

To my family

Acknowledgements

I am deeply grateful to my advisor, Prof. Michael Orshansky, for his guidance and support throughout these years. Prof. Orshansky is the one who introduced me into the world of CAD/VLSI. He has guided me, with great patience, through every step of conducting research, from formulating the problem to revising the research paper. This dissertation could not be in its current form without his assistance and encouragement. Thank you, Michael!

I would like to express my gratitude to the members of my dissertation committee, each of whom has given me guidance and important feedback. They are Prof. Ari Arapostathis, Prof. Todd Arbogast, Prof. Jack Lee, Prof. David Pan, Prof. Nur Touba. Their suggestions have made this dissertation better. Prof. Arapostathis has helped deriving the SRAM dynamic noise margin. Prof. Arbogast has spent a great amount of time with me on the differential equation solver for NBTI. Prof. Nur Touba has given his valuable comments on delay testing. Besides, I benefited from Prof. David Pan's optimization class and Prof. Nur Touba's dependable computing class. Prof. Jack Lee has made me aware of PBTI, an important failure mechanism. I am indebted to all of them.

There are several fellow students in the Robust IC Design Laboratory whom I would like to thank for their help. They are Wei-shen Wang, Murari Mani, Ashish Singh, Shayak Banerjee, and Ku He. I have benefited greatly from my discussions with them. I

would like to especially thank Wei-shen Wang for his great help with the many difficulties I have encountered in my research.

Finally, I am grateful to my family for their love and support. I am proud of all of them: my parents Dengyin Zhang and Youzhen Wen, sister Min Zhang, brother-in-law Yong Huang, and niece Yeshuang Huang.

IC Design for Reliability

Publication No._____

Bin Zhang, Ph.D.

The University of Texas at Austin, 2009

Supervisor: Michael Orshansky

As the feature size of integrated circuits goes down to the nanometer scale, transient and permanent reliability issues are becoming a significant concern for circuit designers. Traditionally, the reliability issues were mostly handled at the device level as a device engineering problem. However, the increasing severity of reliability challenges and higher error rates due to transient upsets favor higher-level design for reliability (DFR). In this work, we develop several methods for DFR at the circuit level.

A major source of transient errors is the single event upset (SEU). SEUs are caused by high-energy particles present in the cosmic rays or emitted by radioactive contaminants in the chip packaging materials. When these particles hit a N+/P+ depletion region of an MOS transistor, they may generate a temporary logic fault. Depending on where the MOS transistor is located and what state the circuit is at, an SEU may result in a circuit-level error. We analyze SEUs both in combinational logic and memories (SRAM). For combinational logic circuit, we propose FASER, a Fast Analysis tool of Soft ERror susceptibility for cell-based designs. The efficiency of FASER is achieved through its static and vector-less nature. In order to evaluate the impact of SEU on

SRAM, a theory for estimating dynamic noise margins is developed analytically. The results allow predicting the transient error susceptibility of an SRAM cell using a closed-form expression.

Among the many permanent failure mechanisms that include time-dependent oxide breakdown (TDDB), electro-migration (EM), hot carrier effect (HCE), and negative bias temperature instability (NBTI), NBTI has recently become important. Therefore, the main focus of our work is NBTI. NBTI occurs when the gate of PMOS is negatively biased. The voltage stress across the gate generates interface traps, which degrade the threshold voltage of PMOS. The degraded PMOS may eventually fail to meet timing requirement and cause functional errors. NBTI becomes severe at elevated temperatures. In this dissertation, we propose a NBTI degradation model that takes into account the temperature variation on the chip and gives the accurate estimation of the degraded threshold voltage.

In order to account for the degradation of devices, traditional design methods add guard-bands to ensure that the circuit will function properly during its lifetime. However, the worst-case based guard-bands lead to significant penalty in performance. In this dissertation, we propose an effective macromodel-based reliability tracking and management framework, based on a hybrid network of on-chip sensors, consisting of temperature sensors and ring oscillators. The model is concerned specifically with NBTI-induced transistor aging. The key feature of our work, in contrast to the traditional tracking techniques that rely solely on direct measurement of the increase of threshold voltage or circuit delay, is an explicit macromodel which maps operating temperature to circuit degradation (the increase of circuit delay). The macromodel allows for cost-effective tracking of reliability using temperature sensors and is also essential for enabling the control loop of the reliability management system.

The developed methods improve the over-conservatism of the device-level, worst-case reliability estimation techniques. As the severity of reliability challenges continue to grow with technology scaling, it will become more important for circuit designers/CAD tools to be equipped with the developed methods.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The evolution of integrated circuits is driven by the trend of increasing operating frequency and greater functionality on a single chip. This is achieved through downscaling of the feature size of the devices on the chip. Downscaling increases process variation and leakage current, and makes the devices less reliable. In this dissertation, we are specifically interested in the device reliability issues. Downscaling threatens reliability through at least the following two ways. Firstly, smaller charge is stored on a scaled circuit node to represent a logic value which makes the circuit node more susceptible to noise [1]. Secondly, in practical scaling, the supply voltage does not scale as fast as the feature size because of the non-scaling of the subthreshold slope and the resulting leakage current [2], leading to higher device electrical fields which accelerate several failure mechanisms that we discuss below.

The impact of the above threats to reliability is the greater incidence of circuit faults. Based on whether a circuit can recover after the occurrence of a fault, we classify the faults as being transient or permanent. Faults may manifest themselves as functional errors of the circuits. Traditionally, the reliability issues were mostly addressed at the device level [3][4]. However, the increasing error rates due to downscaling favor higher-level deign for reliability (DFR). In this work, we focus on DFR at the circuit level. Specifically, we aim to develop analysis techniques that enable circuit designers to evaluate the error-susceptibility of a circuit and identify the problematic transistors, gates, or circuit blocks.

Error-susceptibility of a circuit is closely related to the physical nature of the faults. In the following, we give a brief introduction to the physics of the transient and permanent faults that we deal with in this work.

1.1 PHYSICAL MECHANISMS OF TRANSIENT AND PERMANENT FAULTS

1.1.1 Transient faults

There are two types of transient faults. The first is due to on-chip noise, and the second, referred to as single event upset (SEU), is due to extrinsic high-energy particles, such as neutrons and alpha particles, which are present in cosmic rays or emitted by radioactive contaminants in the chip packaging materials [5][6].

On-chip noise sources include power and ground network noise, substrate injection noise and capacitive coupling noise [7][8][9]. They cause the node voltages to deviate from their noise-free values. On-chip noise can be minimized by proper design, such as adding de-coupling capacitors to the power/ground networks, so that the noise level at any point on the chip is kept under control [10].

An SEU occurs when a high-energy particle hits a N+/P+ depletion region of an MOS transistor, temporarily forming a low-impedance path between the depletion region and the ground or power supply. This leads to a current pulse flowing through the low-impedance path, which disrupts charge stored on the node, leading to a fluctuation of the node voltage [5][6][11]. If the deviation of the node voltage exceeds the threshold value, a logic fault occurs and may propagate to the output. Due to the wide energy spectrum of the high-energy particles [1], it is impractical to design circuits that can prevent logic fault due to every SEU at every point on the chip [12]. As a result, it is inevitable that certain transient logic faults will be generated, propagated (in a combinational logic

circuit) and latched by register, causing soft errors. The IBM design goal for soft errors is 1 SDC (silent data corruption) per 1000 years in its Power4 system [13].

1.1.2 Permanent faults

Among important permanent faults are time-dependent oxide breakdown (TDDB), electro-migration (EM), hot carrier effect (HCE), negative bias temperature instability (NBTI). Their common feature is that permanent physical defects are gradually developed throughout the lifetime of the integrated circuit and that the emergence of these faults is accelerated by elevated temperature and voltage.

Time-dependent oxide breakdown (TDDB)

When the oxide layer of an MOS transistor is stressed by voltage across it, a high electrical field in the order of several MV/cm is created and can generate defects inside the oxide layer. When these defects form a low-impedance path across the oxide layer, an oxide breakdown is said to occur [14]. This breakdown is known as “hard breakdown”. However, in recent years, it was found that for ultra-thin oxides, less than 1 out of 1015 oxide breakdowns is “hard”, i.e., capable of destroying the transistor and even the entire chip, while the rest of the oxide breakdowns are “soft”. Soft breakdowns lead to the increase of the leakage current through the oxide layer [15]. If an integrated circuit is fault-tolerant to at least two soft breakdowns, the lifetime of the circuit can be about 1000 times that predicted by the traditional way [16]. Therefore, it is now suspected that TDDB is not a fundamental obstacle for continued feature size scaling [16].

Electro-migration (EM)

Electro-migration refers to the mass transport of metal atoms in the interconnects. Electrical current flowing through the interconnect involves the collective motion of electrons. The movement of the electrons exerts force on the conductor metal atoms,

causing them to be moved (depleted) from their original sites. As the sites of atom depletion pile up, the resistance of interconnect increases and may eventually result in an open circuit [17][18]. However, if the current can flow in both directions, the depleted sites can be partially re-filled with atoms when the current flows reversely [18].

Hot carrier effect (HCE)

When the drain-to-source voltage of an MOS transistor is biased in the saturation mode, the high electrical field near the drain accelerates the carriers (electrons or holes) in the channel, driving them to become energetic or "hot". The hot carrier effects can overcome the potential barrier at the Si-SiO_2 interface and penetrate into the oxide, generating localized traps near the drain, which degrade the current-voltage characteristics of the transistor. HCE only occurs when there is current flowing through the channel, which is only possible when the transistor is switching [19].

Negative bias temperature instability (NBTI)

NBTI occurs only in PMOS transistors. When the gate of a PMOS transistor is under negative bias, holes are accumulated in the channel, which breaks the Si-H bonds at the Si-SiO_2 interface and generate interface traps. The accumulation of interface traps leads to increased threshold voltage for the PMOS transistor [53]. Removal of the voltage stress can partially anneal the interface traps, but can never make the PMOS return to its "fresh" state [20].

Our research focuses on transient faults due to SEU and permanent faults due to NBTI. However, some techniques we have developed are also extendable to faults due to other failure mechanisms.

1.2 ANALYSIS OF SOFT ERROR SUSCEPTIBILITY FOR CELL-BASED DESIGNS

Reliability of commercial electronics with respect to the single event upsets (SEU) caused by extrinsic radiation is becoming a significant concern. Historically, the most significant impact of SEU was on memory units (latches, flip-flops, registers, and arrays). However, as the transistor feature size scales down, the error rate due to single event upsets in the combinational logic becomes substantial. It is predicted that by 2011, the soft error rate (SER) due to combinational logic may be comparable to that of the memory units [1]. Because of the increasing error rates in combinational circuitry, new tools and analysis methodologies are needed to ensure circuit reliability. System designers, micro-architects, and circuit designers need accurate prediction of error rates in the designed components. Having this capacity is a prerequisite for choosing the proper hardening strategy for the design.

A soft error may occur when a high-energy particle, typically, an alpha particle, or a neutron, hits the diffusion regions of an MOS transistor and produces charge that leads to a faulty transition. The pulse will cause an error only if it successfully propagates to the latching element and is latched at the clock arrival (sampling) time. There are several mechanisms that reduce the overall likelihood of the pulse producing an erroneous value at the memory units, making the actual SER substantially lower than the raw particle strike rate. In the literature [1][12][21][22][23][24], these mechanisms are referred to as electrical masking, logic masking, and latching-window masking.

In this work, we propose an efficient and accurate approach for SER analysis of cell-based designs. The efficiency is achieved by resorting to symbolic representation of the error pulses using binary decision diagrams (BDD). The accuracy is guaranteed by relying on the precise description of the non-linear gate transfer characteristics using the SPICE-based precharacterization of the cells in the library. In addition to the electrical properties of the cells, the logic structure of the circuit also has a significant impact on

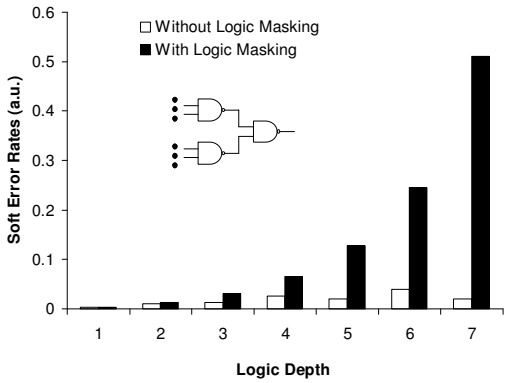


Figure 1.1: Soft error analysis ignoring logic masking can overestimate soft error rates by up to 25X.

the SER. Failing to account for logic masking may over-estimate SER by order of magnitude (25X for a tree-structured circuit with logic depth 7), as illustrated in Figure 1.1. It is evident that as the logic depth increases, logic masking plays a more important role. Accurately yet efficiently accounting for the reduction of error rate likelihood due to these masking mechanisms is the focus of this work.

Prior work in this area has concentrated on modeling and describing the particle interactions at the very low nuclear level [26], performing device-level simulations to predict the electrical response of individual transistors to a particle strike [27], and performing circuit-simulation of a small set of gates to model the propagation of pulses [1]. Several authors have addressed the problem of SER analysis for general combinational logic [12][21][22]. Accurately estimating the SER due to particle strikes on combinational logic gates represents a significant computational challenge. The primary reason is that SER de-ratings due to electrical, logic, and latching-window masking are all input vector dependent. Existing techniques approach this problem by explicitly enumerating all input vectors, or a set of randomly picked input vectors

[12][21][22]. However, the size of the input vector space is exponential in the number of primary inputs, and for circuits with a large number of primary inputs these techniques usually take hours, or even days, to achieve reasonable accuracy [22].

1.2 ANALYTICAL MODELING OF SRAM DYNAMIC STABILITY

Verifying stability of SRAM-based memory arrays is an essential design task. For memory design in nanometer scale technologies, the traditionally-used static stability analysis may no longer be sufficient. The overriding reason is that the power-minimization driven supply voltage scaling dangerously reduces design options. This is especially so when static noise margins (SNM) [38] are used. Checking stability using dynamic noise margins requires complex time-dependent stability analysis but permits a more aggressive flexible design without jeopardizing reliability. Dynamic noise margins (DNM) take into account spectral and time-dependent properties of the specific noise patterns. These include the on-chip noise sources such as power and ground network noise, substrate injection noise, capacitive coupling, as well as the extrinsic transient noise such as those due to single event upsets (SEU) [11][7][8][9]. Single event upsets are becoming especially troublesome for memory arrays (much more than for combinational logic with its natural masking mechanisms) and dynamic stability analysis is essential for predicting bit error rates accurately. An additional factor for the growing importance of DNM analysis is the adoption of SOI devices for SRAM cells because the floating-body effect leads to a dynamic variation of transistor strengths during read and write operations [40].

Stability analysis is concerned with identifying the maximum possible unintended violations that will not cause the stored state to be lost. Traditional static stability analysis finds the noise margins by identifying the maximum amplitude of a voltage deviation on an input node that can be tolerated. Static analysis ignores the fact that not all transient

noise affecting a sensitive node of an SRAM cell will cause the state to flip. There is a range of noise patterns that will cause only a temporary disturbance of the internal node voltage. Whether a transient noise signal is benign (will not cause a flip) or malignant (will cause a flip) depends not only on the amplitude of the signal but also on its duration (or pulse width). In addition, the outcome of the temporary disturbance is also determined by the electrical and geometric parameters of the cell. This work demonstrates how nonlinear study can be used to develop an analytical theory of the bi-stable cross-coupled inverter system affected by transient noise.

A well-developed and diverse theory exists for analysis of SRAM cell stability based on the notion of SNM [38]. SNM analysis ignores the transient aspect of noise and posits a constant noise signal is present at a node. Despite the wealth of methods, it has been shown the various criteria are equivalent [39], including coincidence of equilibrium states, the small signal closed-loop gain of the SRAM being unity, the setting of the Jacobian of the Kirchhoff equations of the SRAM to zero. The equivalency extends to the most "famous" criterion that is familiar to many SRAM circuit designers and is based on the graphical method of inscribing the maximum square between normal and mirrored voltage transfer characteristics [39].

While there exists rich theory for static stability analysis of SRAM cells, extending it to dynamic analysis has proven difficult. The importance of transient noise susceptibility of SRAM cells has been of course realized [11]. However, all of the prior works rely on simulation techniques at the device level [41], transistor level [27], or in a mixed mode of device and transistor levels [42]. Simulation techniques search for the minimum strength of a malignant transient noise (for instance, in the form of critical charge deposited by an SEU). Although accurate, the simulation results can only provide limited insight into the problem. A somewhat similar problem is encountered in the

characterization of combinational logic gates' susceptibility to transient errors [31][43][44]. Dynamic response of a gate is obtained by solving the simplified differential equations describing its dynamic behavior. However, because of the nonlinear cross-coupling between the nodal voltages of an SRAM cell, it is difficult to extend this literature to the problem at hand. Thus, a closed-form analytical formulation that reveals the relationship between the minimum strength of a malignant transient noise and the configuration parameters of the SRAM cell is sought. In addition to giving invaluable intuition about SRAM dynamic stability, it can be directly used for reliability, yield and optimization to avoid lengthy iterative SPICE simulations [45].

In this work, for the first time, a method for evaluating dynamic noise margins is developed analytically. The results allow predicting the transient error susceptibility of an SRAM cell using a closed-form expression. The key innovation involves using the methods of nonlinear system theory in developing this model. It is shown that when a transient noise of a given magnitude affects a sensitive node of a cell, the bi-stable, feedback-driven nature of the cell determine whether the noise will be suppressed or will evolve to eventually flip state. It is shown that for the flip to take place, the noise needs to exceed specific threshold amplitude, which is shown to be equal to the SNM, and be sustained longer than a minimum critical duration.

The inverters of the cross-coupled pair in an SRAM cell are described as operating in two modes: high gain and low gain, such that in each mode, the driving current of the inverter is a linear function of either the input voltage or the output voltage but not both [46]. This permits decoupling of the input and output voltage dependences. Using a piece-wise linear MOSFET model, rigorous analysis of the coupled nonlinear feedback system is performed. The specific formal and quantitative result is a closed-form expression that can be used to predict whether a cell flip will occur for a noise

signal with specific characteristics. For a given SRAM cell design, the expression evaluates the critical pulse width for various noise amplitudes.

1.4 NBTI UNDER DYNAMIC TEMPERATURE VARIATION

In the nanometer regime, negative bias temperature instability (NBTI) is one of the dominant factors of transistor aging [50][51]. NBTI manifests itself as a gradual increase in the magnitude of PMOS threshold voltage. Over time, this leads to an increased circuit delay. At the end of the lifetime of the chip, a timing error may occur. The time until such a failure occurs is defined as the circuit lifetime. Industrial practice reveals that NBTI has been significant since the 90nm technology node and will get worse with further scaling [52].

NBTI occurs when the gate of the PMOS is negatively biased (i.e. the gate voltage is 0 *Volt* and the source voltage is V_{dd}). The electrical field induced across the gate oxide generates a complex process of electro-chemical reaction that consists of several sub-processes and involves accumulation of interface traps and positive charge at the Si-SiO_2 interface leading to degradation of PMOS threshold voltage [53][54].

NBTI is highly sensitive to operating temperature. This is illustrated in Figure 1.2, where the NBTI-limited lifetime of an inverter chain is shown to decrease by 2.2X for every 10^{o}C increase in operating temperature. On-chip temperature in modern ICs exhibits significant variation that depends on the operating conditions of the chip. The precise spatial and temporal characteristics of temperature behavior are complex. They depend on the thermal properties of the substrate, the local current densities, and the input switching patterns. Spatially, differences of temperature of up to 50^{o}C are possible in high-end microprocessors [56]. Temporal temperature profile may change very rapidly. Figure 1.3 shows a trace of temperature in a circuit block [57] in which the temperature can vary by up to 50^{o}C within 3 seconds. Because of the above reasons, to accurately

predict the amount of NBTI-induced circuit performance degradation it is indispensable to have a model capable of handling arbitrary temperature variation in which devices actually operate. In this work, we focus on the impact of temporal variation of temperature, i.e., dynamic temperature variation. Without ambiguity, we use the term "temperature variation" to mean the temporal variation of temperature in the rest of the work.

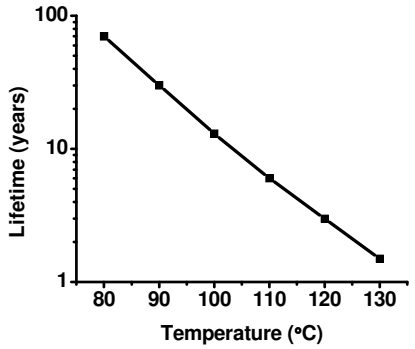


Figure 1.2: Lifetime of an inverter chain decreases by 2.2 X for every 10 ^{o}C increase in operating temperature due to NBTI. The temperature is held constant throughout the lifetime of the circuit. Degradation of PMOS threshold voltage is obtained from [51]. Circuit delay is obtained from SPICE simulation for 45nm technology node. Duty cycle of input signal is set to be 50%.

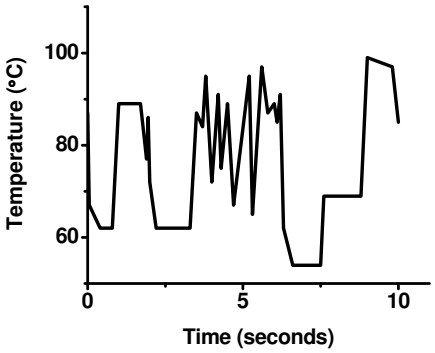


Figure 1.3: Operating temperature exhibits significant dynamic variation for workload C630 of the DRM block in a microprocessor design ([57]).

Several compact NBTI models suitable for circuit-level simulations have been proposed in the past. In [53][54][58][59], the models assume that both the temperature and the voltage stress are fixed at constant values. The impact of interrupting of voltage stress due to signal transition with a fixed temperature is studied in [51][60][61][62]. In [63], both the temperature and the voltage signal are assumed to change between only two levels. Moreover, the voltage signal at any time is assumed to be known. In the case of continuously-changing temperature that spans a large range, the assumption of such two-level temperature can result in significant error in the change of PMOS threshold voltage, if the rms value of the difference between the actual temperature and its two-level approximation is large.

In this work, a novel compact NBTI model that can handle arbitrary temperature variation is proposed. It takes into account the device-level characterization of the temperature dependence of the various sub-processes of NBTI [53][64][65][66] and is consistent with the general framework of reaction-diffusion (R-D) model [53][54]. When both temperature variation and signal transition are present, our model inherits the legacy of the existing models on the treatment of signal transition and also incorporates the impact of temperature variation at the same time.

The proposed model enables both optimal design of circuit for NBTI based on temperature-profiling and real-time adaptive control to mitigate the NBTI problem based on temperature-monitoring. For temperature-profiling, thermal simulators such as HotSpot [67] has been utilized at design-time to predict the spatial and temporal variation of temperature on the chip. Using the information of temperature variation can avoid over-design for NBTI mitigation. For real-time adaptive control, on-chip temperature sensors have been used to monitor the temperature variations during the usage of the

chip. The monitored temperature values are fed into a controller for real-time thermal management [68]. This scheme can be easily adapted for real-time control of NBTI if the controller is equipped with an NBTI estimator based on the monitored temperature values.

1.5 ONLINE CIRCUIT RELIABILITY MONITORING

In order to combat the degradation of devices, specifically NBTI, traditional design practice adds guard-band to the clock period to ensure that the circuit can properly function even if the circuit is stressed with the *worst-case* operating condition through its lifetime [52]. This pessimism inevitably incurs high cost in performance. To reduce the pessimism, online monitoring techniques have been employed to measure the *actual* degradation of the circuit [52][57][72]. The existing measurement-based techniques can be classified into three types. The first type measures the degradation of a simple structure (a PMOS transistor or a ring oscillator) and infers the degradation of a large circuit without considering the impact of its circuit topology [57]. The second type measures the delay degradation of the circuit directly, by sensors constantly monitoring part or all of the primary outputs of the circuit [52][72]. The major drawback of this technique is that the number of the sensors can be huge when there are a large number of primary outputs on the chip that need to be monitored. The third type involves occasional online full delay testing using test programs [73]. This approach requires expensive infrastructure for online delay testing and needs to interrupt the normal operation of the circuits.

To overcome the above limitations, we develop an effective reliability tracking framework that uses a circuit reliability macromodel and a hybrid network of on-chip sensors that consist of temperature sensors and ring oscillators. The key feature of our work, in contrast to the traditional tracking techniques that rely solely on direct

measurement [52][57][72][73], is an explicit reliability macromodel which takes real-time temperature measurements as input and maps them to the current circuit degradation (the increase of D_{max}). We choose to monitor temperature because among environmental factors affecting NBTI that include signal probabilities, supply voltage (V_{dd},) and temperature, the latest has the dominant impact on the rate of device degradation. We found through simulation, by assuming typical variations of the environmental factors, that variations of signal probabilities, V_{dd} and temperature can lead up to 14%, 16% and 56% differences in circuit delay increase (ΔD_{max}) respectively, compared to the nominal ΔD_{max} at the end of lifetime. In addition, cost-effective on-chip temperature sensors already exist and they have been used in commercial chips to monitor real-time chip temperature [74]. Another key aspect is the calibration scheme that allows model parameters refined via direct observation. This is necessary because of process variation and unknown physics [53][58][76].

1.6 DISSERTATION ORGANIZATION

The remainder of the dissertation is organized as follows. Chapter 2 presents a tool for fast analysis of soft error susceptibility for cell-based designs. Chapter 3 describes analytical modeling of SRAM dynamic stability. In Chapter 4, we present a model for NBTI under dynamic temperature variation. Chapter 5 describes a framework of online circuit reliability monitoring using a reliability macromodel and a network of sensors. Finally, Chapter 6 concludes this dissertation.

Chapter 2: FASER: Fast Analysis of Soft Error Susceptibility for Cell-Based Designs

Soft errors have emerged to be a major reliability concern for combinational logic circuits as technology scales down [1]. Estimating the soft error rate is essential for choosing the proper hardening strategy for the design.

In Chapter 1, we have described that soft errors are the result of high-energy particle strikes at the sensitive areas of MOS transistors. However, due to several masking mechanisms [1][12][21][22][23][24], namely, electrical masking, logic masking, and latching-window masking, the actual soft error rate (SER) is substantially lower than the raw particle strike rate. The raw particle strike rate is readily available [11], hence the key is to estimate the reduction of error rate likelihood due to these masking mechanisms.

The soft error de-rating due to the masking mechanisms for a combinational logic circuit is dependent upon the circuit's input vector. Existing work on estimating soft error rate for combinational logic circuit uses simulation techniques to explicitly enumerate the input vectors [12][21][22]. However, the simulation approach suffers the problem of poor scalability since the number of input vectors grows exponentially with the number of primary inputs.

In this work, we propose FASER, which stands for Fast Analysis of Soft ERror susceptibility, for cell-based signs. The innovation is that we use binary decision diagrams (BDD) to avoid explicit enumeration of input vectors, and we use SPICE to precharacterize the cells for error generation and propagation to ensure the accuracy of the tool.

This chapter is organized as follows. Section 2.1 describes the cell characterization procedure. In Sections 2.2, 2.3 and 2.4, we discuss the static analysis of SER. Section 2.5 presents the experimental results, and we summarize this chapter Section 2.6.

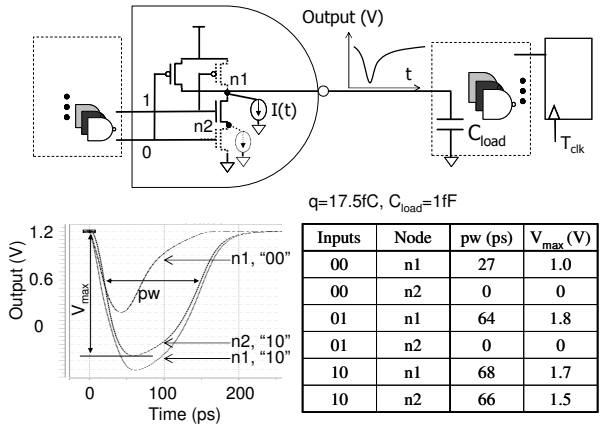
2.1 CELL LIBRARY CHARACTERIZATION

The proposed static SER analysis methodology FASER is targeted towards the use with the cell-based design methodology. Accurate library characterization is thus a key consideration. The two essential characterization steps are pulse generation and pulse attenuation (propagation).

A high-energy particle striking a node deposits charge which leads to a time-varying voltage pulse of a certain magnitude and shape. The characteristics of the pulse are dependent on the specific transistor network of each gate. Thus, the goal of the library characterization is to predict for every library gate the waveforms produced at the cell output for particle strikes at each vulnerable region. The current flow created by the charge deposited into the node is modeled as a single exponential for cosmic-ray related soft errors [1][22] (alternative models for alpha-particle related soft errors are also in existence [35][36][37]):

$$I(q,t) = \frac{2q}{\sqrt{\pi}T_s}\sqrt{\frac{t}{T_s}}e^{-\frac{t}{T_s}} \tag{2.1}$$

where q is the collected charge and T_s is the technology-dependent charge-collection time constant. Collected charge q depends on the particle energy, and follows an exponential distribution [1][22].



Inputs	Node	pw (ps)	V_{max} (V)
00	n1	27	1.0
00	n2	0	0
01	n1	64	1.8
01	n2	0	0
10	n1	68	1.7
10	n2	66	1.5

Figure 2.1: Pulse generation is characterized by circuit simulation with SPICE. The table shows the pulse produced at the output of NAND gate.

Figure 2.1 shows the SPICE simulation setup for characterization of pulse generation, where every vulnerable node is taken into account. The simulations are performed for a range of charge values (q) and load capacitances (C_{load}). The voltage pulses produced at the cell output by a specific charge deposited on an intra-cell node strongly depends on the biasing condition determined by the input vector. In the example circuit of Figure 2.1, both n1 and n2 are sensitive if the input vector is "10" (the pulse generated at n2 is slightly attenuated by the transistor above it), while only n1 is sensitive if the input vector is "01" or "00". When the input vector is "00", the pull-up network has the smallest resistance, resulting in the smallest falling pulse generated.

Existing tools [1][12][21][22][23][24] either ignore the effects of biasing conditions, or assume the worst case biasing conditions for every gate in the circuit. Experiments show a difference of 1.5X - 4X between the SPICE simulation result and the analysis performed under the worst-case assumptions. Clearly, accurate soft error analysis tool needs to consider different biasing conditions of the gates.

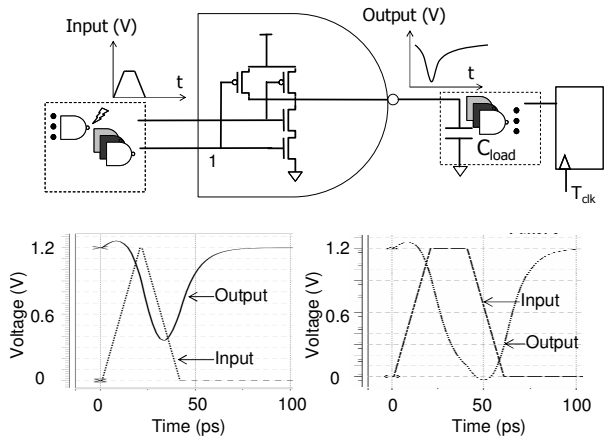


Figure 2.2: Pulse propagation is characterized by circuit simulation with SPICE. The curves show the low-pass characteristic of NAND gate.

The voltage pulse produced at the cell output is approximated by a trapezoidal waveform, and are captured by two parameters, pulse width (*pw*) measured at $0.5V_{DD}$ and maximum voltage value (V_{max}). The rise and fall times of the trapezoidal waveform are chosen to be typical values.

After a transient faulty pulse is generated, it propagates toward the primary outputs of the circuit. In the course of its propagation, the pulse's electrical properties, such as width and magnitude, evolve as a result of the low-pass characteristics of the gates it propagates through. Short pulses tend to be attenuated, while long pulses tend to maintain their original width and magnitude after passing through a combinational logic gate. Figure 2.2 shows the SPICE simulation setup for characterization of this dynamic transfer function, where *pw* and V_{max} of the output pulse is found as a function of the input pulse. While different input-pin to output paths may be characterized by somewhat different transfer characteristics, this is a secondary effect, which we have for now ignored.

Pulses may re-converge and overlap at a gate in a circuit if multiple paths exist between the particle-striking point (fault-site) and the gate. The interaction of two pulses

arriving simultaneously can be modeled. Currently, characterization captures only the first-order effect of pulse-overlapping with the output produced by simple superposition, followed by low-pass filtering by the gate's dynamic transfer function. The error of estimation for circuit SER due to this approximation appears to be minor compared with SPICE simulation for the benchmark circuits we have tested.

2.2 STATIC ANALYSIS OF FAULT EVENTS PROPAGATION

FASER is a static SER analysis methodology in that it relies on the implicit enumeration of the input vector space. The algorithm formally encodes and propagates the error pulses using binary decision diagrams. Binary decision diagrams are a powerful data structure proposed by Bryant [28] for efficient representation and manipulation of Boolean functions. By propagating the fault-encoding function to the primary outputs the algorithm can accurately predict output error probabilities. The error propagates only if the path from a fault-site to the output is sensitizable under the specific assignment of side inputs to the gates. The proposed BDD-based symbolic error manipulation algorithm succeeds in effectively capturing such logical masking [21][22]. However, the formation and propagation of the pulses symbolically is intrinsically linked in the algorithm with the accurate characterization of cell electrical properties, contained in the library. This guarantees accurate modeling of electrical masking [21].

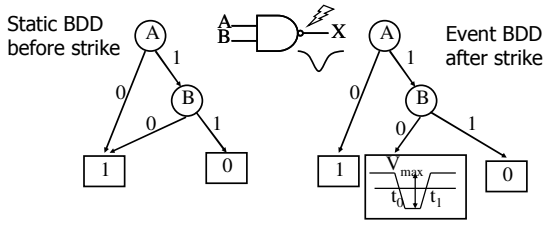


Figure 2.3: Fault-encoding with event BDD for the biasing condition "10". The strength of the pulse depends on the biasing condition and the strike location within the gate.

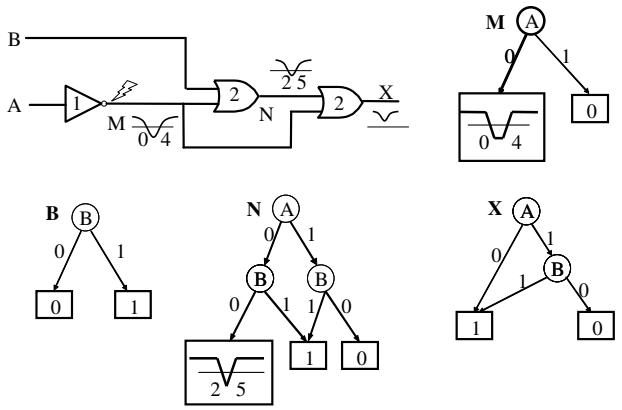


Figure 2.4: Pulse propagation in a simple circuit. Numbers inside the gates are their propagation delays. Terminals of event BDDs contain pulse propagation delays and durations. Pulses encoded with event BDDs of N and M cancel each other, resulting in no pulse at X.

The BDD describing the Boolean function at a given node in an error-free environment is termed *static* BDD. It is constructed using the classic rules of [28]. A particle strike at a node creates a transient pulse that can be represented by modifying the static BDD. Such a data structure is referred to as *event BDD*. In the event BDD, the terminal vertices encode both the error pulses and the original static logic values. The event BDD encoding will contain the arrival time (*AT*), the maximum voltage (V_{max}), and the width of the pulse (*pw*). Figure 2.3 shows fault-encoding with event BDD for biasing condition "10". If the pins of the gates are not primary inputs, the BDD describing the Boolean function of the biasing condition is found first, which is then modified to contain a pulse at one of its terminal vertex to become an event BDD.

Constructing the output event BDD for an operation on two input event BDDs is a recursive process similar to that of constructing the static BDD, which utilizes the standard BDD operations [28]. The operations are different only in how the terminal vertices are processed. Specifically, when the terminal vertex of one operand is reached, we check if the state of the output can be determined. If it can, a terminal vertex for the

output event BDD is generated. Otherwise, a non-terminal vertex for the output is generated, and the event BDD of the other operand is searched one level deeper. Determining the state of the output is through logic operation and table look-up from the library. Logic operation is performed, for example, if one operand has a controlling value and has no pulse, in which case, the output value is determined regardless of the state of another operand (logic masking). Table lookup is performed when the analog characteristics of the output pulse is to be determined (electrical masking).

As noted in the previous section, different biasing conditions may result in very distinct output pulse, sensitive area and hence latching error probability due to particle strikes at a given gate. In order to achieve near SPICE-level accuracy, this dependence needs to be taken into account during the analysis, which requires enumerating all gate biasing conditions and all intra-gate nodes. This clearly increases the computational burden on the algorithm since for each biasing condition and each intra-gate node an event BDD is now generated at the fault site and propagated to the latches. We have found that this is crucial for accuracy improvement and that the penalty is affordable in most cases. Indeed, assuming that the average cell fan-in is k, the increase in complexity due to this enumeration is $O(k2^k)$. Since k is typically between 1 and 3, the cost is manageable.

Propagating the fault events *statically* is equivalent to constructing the event BDDs for the circuit nodes in the fan-out cone of the fault-site where the particle-strike occurs. The event BDD of a circuit node is simply its static BDD if it is outside the fan-out-cone of the fault site, since no error-pulses will occur at the node.

To illustrate fully the working of the algorithm, consider a small circuit example of Figure 2.4. To simplify the discussion, the pulse magnitude is ignored and only the pulse width is taken into account. Given the collected charge, an event BDD is generated

for each biasing condition and intra-gate node of the fault site (node M) is constructed. Due to electrical masking, the pulse width changes along the propagation. The error-pulse is logically-masked when B=1. The event BDD at node X is the same as its static BDD because the pulse at node X is too small due to re-convergence to reach the gate threshold voltage.

2.3 ALGORITHM FLOW AND LATCHING PROBABILITY COMPUTATION

Ultimately, the static analysis of FASER is based on computing the probability of an error at the latch due to the totality of pulses propagating towards primary outputs. First, it is assumed that a particle can strike every node (diffusion region) in the circuit with the probability given by the ratio of the node area to total area. Second, the primary inputs remain stable. The validity of this assumption for SER analysis was demonstrated in [12]. Third, the equilibrium probabilities of the primary inputs are known and independent of each other. This last assumption has been successfully applied for power estimation and circuit reliability assessment [29]. Soft error estimation for circuits with strongly correlated primary inputs will be our future work.

The core of the algorithm is to find the conditional latching error probability $P(q, bc, i, j, k)$, given the collected charge q, biasing condition bc of the victim gate (fault site), intra-gate node j of gate i, and latch k. Calculating $P(q, bc, i, j, k)$ is discussed in the next paragraph. The contribution to the bit error rate (BER) of latch k by gate i is,

$$PL = \frac{pw - w}{T_{clk}} \; BER(i,k) = \sum_{j} \sum_{bc} \sum_{q} P(q, bc, i, j, k)(R\ (q, i, j)\Delta q) \tag{2.2}$$

where $R\ (q, i, j)\Delta q$ is the strike rate for collected charge in the range of q and $q + \Delta q$, which is proportional to the area of node j of gate i. We use the average BER of all output latches by particle strikes on all gates as a merit of a circuit's soft error susceptibility.

However, other criteria, such as the largest BER of the latches, can be used as well, depending on the application.

Propagating an event BDD to the primary output gives us a reliable measure of the occurrence probabilities and strengths of the pulses that will appear at the latch inputs. However, the latching error probability is linked to another masking mechanism, known as latching window masking. Latching window masking occurs due to the *temporal randomness* of the particle strike time [1][12][21][22], and the realization that the pulse arrival time at the latch has to be within the latching window for the error to occur. Assuming a uniform strike-time probability, the actual latching probability for a pulse is:

$$PL = \frac{pw - w}{T_{clk}} \tag{2.3}$$

where *PL* is the latching probability, *pw* is the width of the faulty pulse present at the input of the latch, *w* is latching-window size of the latch, and T_{clk} is the clock period. Given an event described by the set of parameters (*q, bc ,i, j, k*), the event BDD at the primary output *k* is constructed first (Figure 2.5). With the assumption that the primary inputs are independent of each other, each edge of the event BDD is assigned a probability, based on the primary input the edge corresponds to. By recursively traversing the event BDD, the probability for an event contained in a terminal v to occur, $p(v)$ can be calculated [10]. The conditional latching error probability *P*(*q, bc ,i, j, k*), is then

$$P(q, bc, i, j, k) = \sum_{v} p(v) PL(v) \tag{2.4}$$

where $PL(v)$ is the latching error probability of the pulse contained in terminal v of the event BDD, determined by (2.3).

Generate static BDD for every circuit node;
BER ← 0;
for all combinations of (*q, bc, i, j, k*)

1. Retrieve the generated pulse shape for (*q, bc, i, j, k*), and generate an event BDD at the fault site *i*.
2. Propagate the event BDD in the fan-out cone of fault site *i*.
3. The event BDD at the primary output *k* (input of latch *k*) is traversed to find *P*(*q, bc, i, j, k*).
4. $BER \mathrel{+}= P(q,bc,i,j,k)(R\ (q,i,j)\Delta q)$

end
BER ← *BER / num_of_latches*;

Figure 2.5: Pseudo-code of FASER flow.

2.4 CIRCUIT PARTITIONING FOR SPEED-UP

It is well known that the worst-case complexity of the BDD encodings of logic functions is exponential in the number of variables [28]. To make manipulation of BDDs efficient, a partitioning heuristic is adopted. This is a common practice in CAD techniques that use BDDs [33][34]. The use of partitioning allows a significant speed-up without a noticeable loss of prediction accuracy.

The circuit is partitioned into smaller domains as shown in Figure 2.6. Some nodes are designated to be pseudo primary inputs, and serve as the boundary between the partitions. Signal correlations are only considered within the domains. To estimate the latching error probability due to a particle-strike on a particular gate, an event BDD is generated at the fault site and propagated to the boundary nodes, where the pulse occurrence probabilities are estimated by traversing the event BDDs. Next, these pulses are treated as being generated at the boundary nodes and assumed to be independent of each other. In principle, these secondary pulses can be independently propagated further

to the latches. The latching error probability due to particle strike at the fault site is approximated by the sum of the latching probabilities of the secondary pulses weighted by their respective occurrence probabilities. In practice, if we process the fault sites backward starting from the gates closest to the latches, the latching probabilities of the secondary pulses can be directly estimated from the latching error probabilities due to particle-strike at the boundary nodes, without further propagation of the secondary pulses.

We define the partition size as the maximum number of primary/pseudo primary inputs of each domain. A larger partition size allows a more global account of signal correlations but at the cost of a rapidly growing run time and memory usage. The tradeoff between speed and accuracy is performed by adjusting the partition size. We believe that the most significant impact of signal correlation on pulse propagation occurs in the neighborhood of the fault site. Therefore, the improvement of accuracy beyond a certain partition size (typically, 15-20) is expected to be minimal, and this is confirmed by the experimental results.

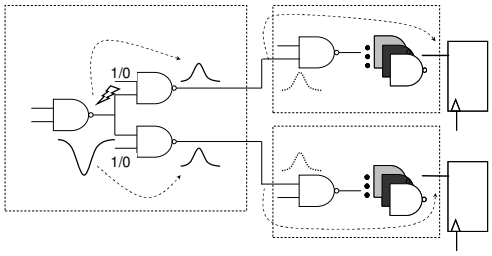


Figure 2.6: Partitioning circuits for speed-up leads to the loss of correlations between pulses of different domains. The loss of accuracy is minimal for partition size beyond a certain point.

2.5 EXPERIMENTAL RESULTS

The static SER analysis tool FASER was implemented in C++. The experiments investigated the accuracy of the static technique, the runtime of the algorithm, and the speed-accuracy trade-off using the partition heuristic. FASER takes a technology-mapped netlist, equilibrium probability of the primary inputs, clock period, and flux rate of the high-energy neutrons, and gives BER of the latches at the primary outputs. The widely utilized flux rates for New York City were used for analysis [11]. The SPICE technology files were based on the Berkeley Predictive Technology Model (BPTM) for the 100nm technology [30].

In order to verify the validity of the FASER, a Monte-Carlo experiment based on SPICE simulation was utilized. Since Monte-Carlo SPICE simulation is very time-consuming, we were only able to perform the tests on small artificially constructed benchmark circuits, with the largest circuit containing 35 gates.

The Monte-Carlo simulation is designed to measure the latching error probability given a particle strike with a random data set of (collected charge q, strike time t, gate i, node j, input vector V). Collected charge q follows an exponential distribution [1][22]. Strike time t is uniformly distributed between 0 and T_{clk}. The probability for a strike to occur in node j of gate i is proportional to node j's area. We assume that all input vectors have equal probability

The experiments were conducted as follows. For a random data set of (q, t, i, j, V), a current pulse with magnitude corresponding to the collected charge q and polarity corresponding to node j's diffusion type is injected to node j of gate i at time t, with input vector V. Voltage samples are taken at the latch output at T_{clk} and $2T_{clk}$. If either value did not match the correct one, an error is declared. Under this set-up, the conditional latching error probability is equal to the number of errors divided by the total number of

simulations. The run time of the Monte-Carlo tests ranges from 15 minutes to 45 minutes for the artificial benchmark circuits. FASER takes into account 5 different pulse strengths and its run-time for every test circuit is less than 0.01 seconds, giving a speed-up of over 90,000X. All experiments are conducted on a Dell GX260 workstation running Redhat Linux. The latching error probabilities of the benchmark circuits are compared in Figure 2.7. The average error between the two sets of data is 12% and can be well attributed to the simplified 2-parameter modeling of the error pulse in cell library characterization. Figure 2.8 shows the latching error probability due to each gate in circuit C1.

FASER with circuit partitioning heuristic is validated on the ISCAS'85 benchmark circuits. Partition size is expressed in terms of the maximum number of primary/pseudo primary inputs of the pulse propagation domains. Experimental results in Table 2.1 show that good accuracy can be achieved with relatively small partition size. The run-time varies between 22 seconds and 638 seconds for partition size 15 (Table 2.2). The runtime increases rapidly with partition size as shown in Table 2.1. However, improvement in accuracy is very small for partition sizes beyond 15. The estimated BER for 100nm CMOS technology is on the order of 10^{-5} FIT, where 1 FIT is defined as 1 failure in 10^{9} hours. It is to be noted that the partition size is not the sole factor that affects the run time. The circuit structure and the choice of pseudo primary inputs in partitioning also can greatly affect the BDD size, which is a well known property of BDD.

As a proxy of the memory usage, the maximum BDD size in terms of the number of vertices of the BDD is measured and shown Table 2.2. The general trend is that the BDD size increases drastically with partition size. The increase of BDD size with respect to the circuit size under the same partition size is due to the increased complexity of the nodal functions.

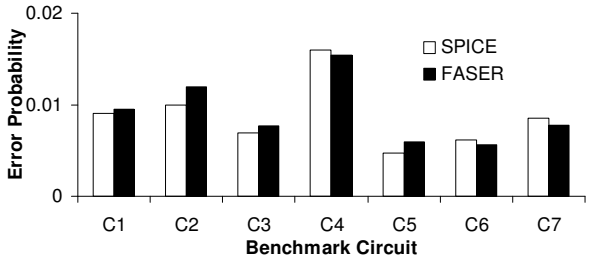


Figure 2.7: Error probabilities by FASER and SPICE simulation. The average error is 12%.

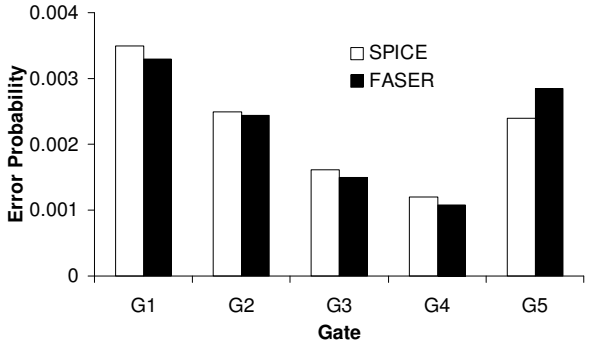


Figure 2.8: Latching error probability due to each gate in circuit C1.

Table 2.1: Bit Error Rates for ISCAS' 85 benchmark circuits with different partition sizes (Np)

Circuits	Bit Error Rates (10^{-5}FIT)		
	Np=15	Np=20	Np=30
C432	3.0	3.2	3.1
C499	2.0	2.0	2.0
C1908	2.2	2.1	1.9
C1355	2.0	2.0	2.0
C3540	2.6	2.6	2.4
C5315	1.1	1.1	1.1
C7552	1.9	1.8	1.8

Table 2.2: Run-time and the maximum BDD size for the ISCAS' 85 benchmark circuits.

Circuit	Run-time(s)			Max BDD Size		
	Np=15	Np=20	Np=30	Np=15	Np=20	Np=30
C432	22	76	465	99	1223	86083
C499	39	63	129	101	145	404
C1908	66	86	1050	169	187	26393
C1355	40	62	119	101	145	406
C3540	149	195	5400	1028	1353	17861
C5315	278	546	1515	1372	6115	12100
C7552	638	780	7200	1813	6702	11602

2.6 SUMMARY

In this chapter, we proposed a fast static soft error analysis tool FASER. Accurate models are based on STA-like precharacterization methods, and logical masking is computed via binary decision diagrams with circuit partitioning. Experimental results indicate that the FASER achieves good accuracy compared to the SPICE-based simulation method. The average error across the benchmark circuits is 12% at over 90,000X speed up.

Chapter 3: Analytical Modeling of SRAM Dynamic Stability

SRAM-based memory arrays are a major component of integrated circuits, and verifying the stability of SRAM cells is essential to ensure the quality of the integrated system. Evaluating the stability of an SRAM cell using static noise margin (SNM) static noise margin (SNM) [38] is straightforward, but it is at the cost of reduced design margins. On the other hand, checking stability using dynamic noise margins (DNM) requires complex time-dependent stability analysis but permits a more aggressive flexible design, without jeopardizing reliability. DNM take into account spectral and time-dependent properties of the specific noise patterns. In this work, we specifically study the DNM of an SRAM cell under the impact of single event upsets (SEU) [11][7][8][9].

Our contribution is this work is: 1) we propose a criterion for DNM of an SRAM cell; 2) we obtain an analytical form of DNM, under the approximation of piece-wise linearization. This chapter is organized as follows. In Section 3.1, we set up differential equations to describe the nonlinear dynamics of the SRAM cell. Based on the mathematical formulation of Section 3.1, we conduct the analytical study on the dynamic stability of the SRAM cell in Sections 3.2 and 3.3. We present the experimental results in Section 3.4. Finally we summarize this chapter in Section 3.5.

3.1 System Modeling Setup

In this work, we focus on the stability of a 6T SRAM cell with respect to an SEU. An SEU is a rare and random event. At the moment when it occurs, the SRAM cell is most likely in the standby mode. Therefore, we will only evaluate the SRAM stability in the standby mode. For other noise sources, the most interested modes might be the read or write mode. However, the framework built in this work can be easily extended to those

cases. A 6T SRAM cell consists of two identical cross-coupled inverters and two access transistors. For notational simplicity, the two access transistors are not included in the analysis. Without loss of generality, the noise source in this work is modeled as a current pulse injected into the internal node where the initial voltage is zero, Figure 3.1. If the noise source flows in the opposite direction, i.e. drawing positive charge from the internal node where the initial voltage is $V_{dd,}$ the following discussion still applies, but with the roles of the NMOS and PMOS being switched. Capacitance, *C,* is the total lumped capacitance at the output node of each inverter. We describe the *state* of the cell by the *state vector* $\vec{V} = (V_1, V_2)$, where V_1 and V_2 are the two nodal voltages of the cell. For a given state vector value, the currents flowing through the four transistors can be determined. The set of $\vec{V}$ of all possible values forms the *state space*. The transient noise current is modeled as a square pulse with amplitude, I_n, and pulse width, pw:

$$i_n(t) = \begin{cases} I_n & (0 < t < pw) \\ 0 & (t \le 0 \text{ or } t \ge pw) \end{cases} \tag{3.1}$$

The question of interest is whether under the impact of the noise signal the cell will undergo a *state-flip*, that is, will change the state. In SNM analysis [38], the voltage transfer curves are used to identify the minimum voltage/current noise amplitude that will lead to a state-flip. Figure 3.2 shows the time behavior of the two nodal voltages under the impact of two separate transient noises with the same amplitude but different

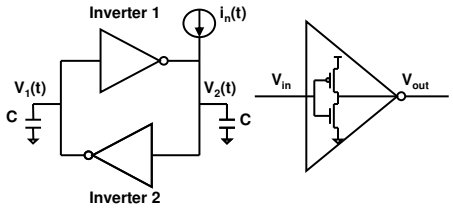
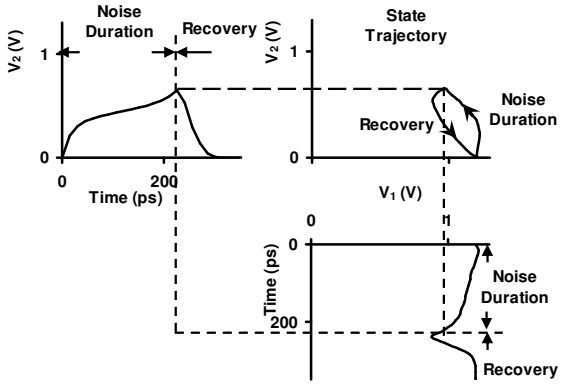
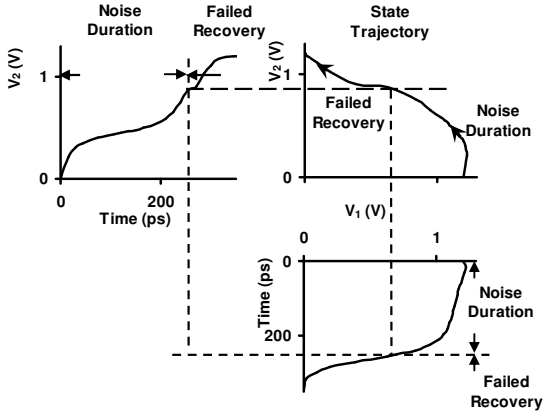


Figure 3.1: A 6T SRAM cell with a transient current noise being injected (access transistors not shown).



(a)



(b)

Figure 3.2: State space and time-domain plots of SRAM nodal voltages with injected noise. Depending on noise duration (a) cell dynamically recovers, or (b) state-flip occurs.

pulse widths. In Figure 3.2 (a), the SRAM cell successfully recovers to its original state after the short noise vanishes, while in Figure 3.2 (b), the SRAM cell fails to recover due to longer noise duration and undergoes a state-flip. Figure 3.2 also shows the *state trajectories* for both cases. The state trajectories are obtained by plotting the state vectors

at various times in the state space. It can be seen that the state trajectory is a closed loop if no state-flip occurs and an open curve otherwise.

We start by setting up a general framework. It can be later shown how SNM analysis can be derived from it. The dynamic stability analysis is based on a system of coupled differential equations (3.2) and (3.3) that describe the dynamics of the SRAM cell - the evolution of the state vector driven by the charging/discharging of the nodal capacitances:

$$dV_1(t)/dt = -I_{inv}(V_2(t), V_1(t))/C \tag{3.2}$$

$$dV_2(t)/dt = [-I_{inv}(V_1(t), V_2(t)) + i_n(t)]/C \tag{3.3}$$

In these equations, $I_{inv}(V_{in}, V_{out})$ is the *driving current* of an inverter, i.e. the current flowing into the inverter through its output node. The positive current direction is chosen to be discharging any nodal capacitor. Since I_{inv} is a nonlinear function of the inverter's input and output voltages, equations (3.2) and (3.3) describe a *nonlinear system*. The driving current I_{inv} is the sum of the drain-to-source currents of the NMOS and PMOS transistors, I_{nmos} and I_{pmos}:

$$I_{inv}(V_{in}, V_{out}) = I_{nmos}(V_{in}, V_{out}) + I_{pmos}(V_{in}, V_{out}) \tag{3.4}$$

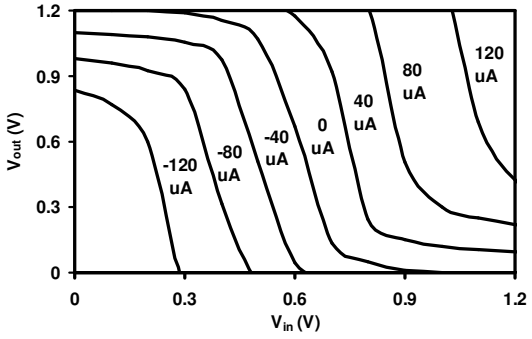


Figure 3.3: Equi-current drive curves are used to capture I-V characteristics of an inverter.

For both transistors, the driving currents are completely characterized by their gate-to-source voltage (V_{gs}) and drain-to-source voltage (V_{ds}). In particular, for NMOS, $V_{gs} = V_{in}$ and $V_{ds} = V_{out}$.

The essence of the inverter's driving current is best captured through the *drive curves* [46]. The drive curves are a contour map of the inverter's driving current versus its input and output voltages, Figure 3.3. The drive curves can be characterized by SPICE simulations. The characterization procedure is similar to that of the dc transfer curve, except that the output node of the inverter is loaded with a constant current source which the driving current is equal to. The dc voltage transfer curve is just one in the set of the drive curves, with:

$$I_{inv}(V_{in}, V_{out}) = 0 \tag{3.5}$$

Drive curves are described by $I_{inv}(V_{in}, V_{out}) = \kappa$, where κ is a constant. Each curve divides the state space into two regions, with one region $I_{inv}(V_{in}, V_{out}) > \kappa$ and the other $I_{inv}(V_{in}, V_{out}) < \kappa$.

Analytical study of the dynamic stability requires analytical modeling of the dependence of the driving currents on circuit parameters. Solvability of a nonlinear system problem is critical because known mathematical methods do not provide powerful enough means to analytically solve many nonlinear system problems [47]. Therefore, in addition to accuracy, the model of driving current must allow us to solve the SRAM equations (3.2) and (3.3) *analytically*.

The above requirements mean that a *linear* and *separable* model of driving current needs to be employed. We adopt the linear gate model of [46]. It is based on functional variable decoupling and piece-wise linearization. Specifically, an inverter is modeled as operating in two modes - a high gain and a low gain mode. In each mode, the

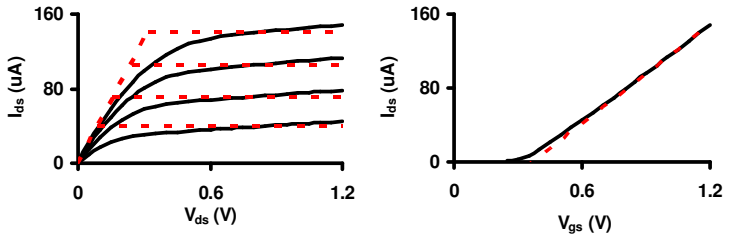
driving current of the inverter is either a linear function of the input voltage or the output voltage, but never both.

Additionally, the model assumes that the short-circuit current is negligibly small [48]. This means that the NMOS and PMOS transistors never conduct simultaneously. (This assumption holds if the rise/fall time of the output voltage of the inverter is longer or comparable to that of the input voltage [48]. The assumption is verified and appears justified in this work). When the NMOS transistor conducts, the driving current of the inverter, I_{inv}, can be written as:

$$I_{inv}(V_{in}, V_{out}) = \begin{cases} 0 \text{ (cutoff)} \\ g_{mn}(V_{in} - V_{thn}) \text{ (saturation)} \\ V_{out} / R_n \text{ (linear)} \end{cases} \tag{3.6}$$

where g_{mn}, V_{thn}, R_n are respectively the transconductance, threshold voltage, and linear-region resistance of the NMOS transistor. We refer to (3.6) as the "linear gate model". The key to the linear gate model is that these parameters are independent of the input and output voltages. When the PMOS transistor conducts, I_{inv} can be modeled similarly. With this approximation, the number of piece-wise linear regions that need to be considered is greatly reduced. This enables us to obtain a concise analytical solution without significant loss of accuracy.

Characterizing these parameters is based on the I-V characteristics of the conducting MOS transistor, Figure 3.4. MOS transistors in the deep sub-micron domain follow the α-power law [49]. As transistor feature size scales down, α approaches one (α = 1.27 for NMOS and 1.46 for PMOS in 100nm technology). As a result, the transistor's drain-to-source current, I_{ds}, in saturation mode will be closer to a linear function of V_{gs} [49]. Thus, we can expect a better match between the piece-wise linear model and the actual I-V characteristic in future technologies.



(a) (b)

Figure 3.4: To enable analytical solution a linear decoupled MOSFET model is used. I-V plots of a 100nm NMOS show that the approximation (dashed lines) is reasonable: (a) I_{ds}-V_{ds}, and (b) I_{ds}-V_{gs} for V_{ds} = 1.2V.

3.2 DYNAMIC STATE SPACE ANALYSIS

According to the SRAM equations (3.2) and (3.3), a dynamic process of charging/discharging of the two capacitors that causes the evolution of the state vector $\vec{V}$ will continue until it reaches the *steady state* ($d\vec{V} / dt = 0$).The values of state vectors over time form the state trajectory, a curve linking the initial and steady states. If the currents are finite, the state trajectory is continuous. In the language of nonlinear system theory [47], any state that satisfies the condition $d\vec{V} / dt = 0$ is an *equilibrium state*. Thus, the steady state of a dynamic process is an equilibrium state. When there are multiple equilibrium states, the steady state is also determined by the initial state of the dynamic process. Evaluating (3.2) and (3.3) for $d\vec{V} / dt = 0$, it can be seen that the equilibrium states of an SRAM cell are given by the roots of the following equations:

$$-I_{inv}(V_2, V_1) / C = 0 \tag{3.7}$$

$$[-I_{inv}(V_1, V_2) + i_n(t)] / C = 0 \tag{3.8}$$

The above equations suggest that in the presence of transient noise, equilibrium states change over time. Under the noise model in (3.1), the change is not continuous and only occurs at $t = 0$ and $t = pw$. Specifically, equations (3.7) and (3.8) simplify to

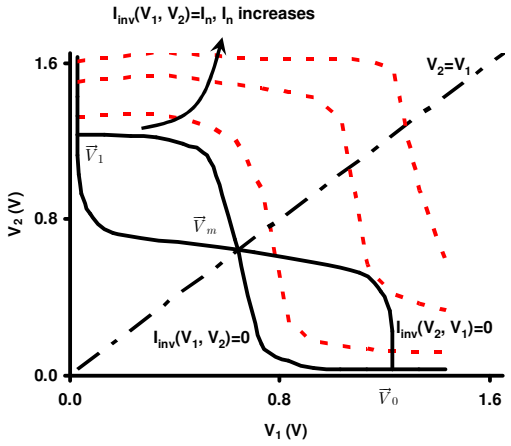


Figure 3.5: Stability analysis is done using superposition of mirrored transfer curves (drive curves). Injection of noise dynamically shifts the curves changing equilibrium conditions.

$$I_{inv}(V_2, V_1) = 0 \tag{3.9}$$

$$I_{inv}(V_1, V_2) = \begin{cases} I_n & (0 < t < pw) \\ 0 & (t \le 0 \text{ or } t \ge pw) \end{cases} \tag{3.10}$$

The traditional way of showing the roots of these equations is to plot the functions in (3.9) and (3.10) in the state space [38], Figure 3.5. The function described by $I_{inv}(V_1, V_2) = I_n$ is plotted for different I_n. These curves are in fact the drive curves of inverters discussed earlier.

The three crossing points between the drive curves $I_{inv}(V_2, V_1) = 0$ and $I_{inv}(V_1, V_2) = 0$ form the equilibrium states for $t \le 0$ or $t \ge pw$: $\vec{V}_0$, $\vec{V}_1$, and $\vec{V}_m$, referring to the states "0" , "1", and "meta-stable". Importantly, these are also the equilibrium states for the noise-free case. The crossing points between the curve $I_{inv}(V_2, V_1) = 0$ and a dashed line are the equilibrium states for a particular I_n during $0 < t < pw$. It can be seen from Figure 3.5 that as I_n increases, the number of equilibrium states for $0 < t < pw$ reduces from three to one.

It is also possible to relate the current discussion to the traditional SNM analysis. Several different criteria have been shown equivalent, including the one based on coincidence of equilibrium states. This criterion is clearly identical to the case when the number of equilibrium states is two, Figure 3.5. The noise amplitude that will result in this case is then precisely the SNM of the cell (I_{snm}). Any higher noise amplitude will result in there remaining a single equilibrium state, making the SRAM cell unable to store either "0" or "1". Specifically, if the remaining equilibrium state represents a "1", the charging/discharging process on the two capacitors will pull any initial state, even if it is a "0", to the only equilibrium state "1". Effectively, SNM analysis ignores the fact that the change of the equilibrium states is only temporary, due to the transient nature of the noise. After the transient noise vanishes $(t > pw)$, the equilibrium states will return to their original locations. If during the time when the noise is present ($0 \le t \le pw$) the state trajectory has not moved sufficiently far away from the initial state, the driving currents of the inverters are able to pull the state vector back to the correct equilibrium state (the initial state) after the noise vanishes.

After the transient noise vanishes, any initial state in this recovering stage, which is also the final state in the noise duration, Figure 3.2, will eventually be driven to one of the three equilibrium states as $t \to \infty$. Figure 3.6 shows three such state trajectories in the recovering stage. The equilibrium states behave as attractors, with each equilibrium state $\vec{V}_e$ having its *attraction region* $\mathbf{A}(\vec{V}_e)$ [47]. The attraction region is defined by the following: for any initial state vector in $\mathbf{A}(\vec{V}_e)$, its state trajectory will terminate in $\vec{V}_e$ for $t \to \infty$. Because of the assumed symmetry of the SRAM cell, the two cross-coupled inverters are identical. By symmetry, the boundary of attraction regions in the recovering stage is:

$$V_2 = V_1 \tag{3.11}$$

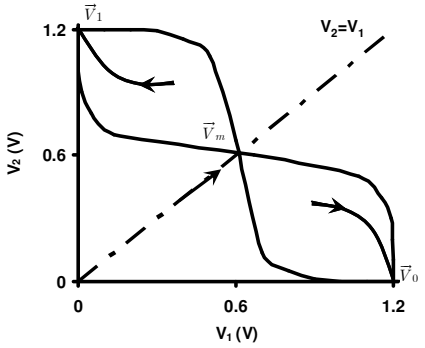


Figure 3.6: Without noise, the region of attraction for $\vec{V}_1$ ($\vec{V}_0$) is the entire region above (below) V_2=V_1. Any state trajectory starting from an initial state in a region of attraction of an equilibrium state will reach it eventually.

Any point in the state space that is below (above) the boundary belongs to the attraction region of $\vec{V}_0$ ($\vec{V}_1$). Any point on the line is attracted to $\vec{V}_m$. The transient noise modifies (moves) the equilibrium states of the SRAM cell and their attraction regions only during $0 < t < pw$. During this period, the attraction regions can be approximately identified from the locations of their associated equilibrium states in the state space. However, the exact location of the boundary of the attraction regions for $0 < t < pw$ is not required in this work.

3.3 TRANSIENT BEHAVIOR OF SRAM UNDER NOISE

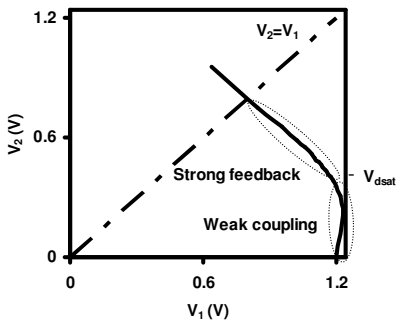
The effect of the transient noise on the stability of the SRAM cell is as follows. Suppose the cell is in the equilibrium state $\vec{V}_0 = (V_{dd}, 0)$ before the transient noise occurs:

$$\vec{V}(t \le 0) = \vec{V}_0 \tag{3.12}$$

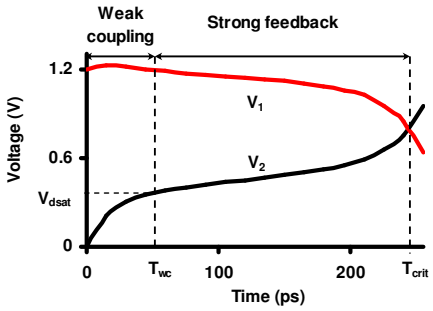
During $0 < t < pw$, the noise current and driving currents of the two inverters force the state vector of the cell to move away from $\vec{V}_0$. If by the time the transient noise vanishes ($t = pw$), the state vector is still within the attraction region of $\vec{V}_0$ (i.e., $V_2 < V_1$), the state trajectory will eventually terminate on $\vec{V}_0$, causing no error. Otherwise,

the final state would be $\vec{V}_1$ or $\vec{V}_m$ (unstable), and a state flip will occur. Therefore, the criterion for dynamic stability is whether the state trajectory can cross the boundary of attraction regions *by the time* the transient noise vanishes.

First, we analyze the dynamic stability of the SRAM cell for $I_n \leq I_{snm}$, i.e. the noise amplitude is smaller than the SNM. The appearance of noise at $t = 0$ relocates the equilibrium state originally at $\vec{V}_0$ to a higher position, Figure 3.5. From the symmetry of



(a)



(b)

Figure 3.7: For noise amplitudes higher than the SNM (I_n>I_{snm}) only one equilibrium state exists (above V_2=V_1): (a) state trajectory; (b) time-domain node voltage evolution.

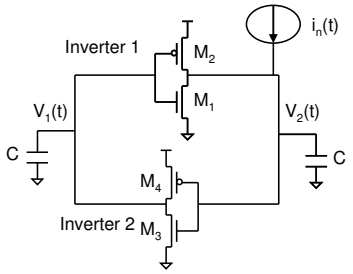


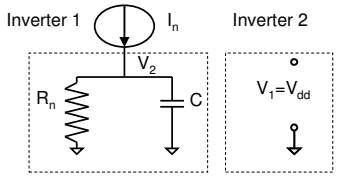
Figure 3.8: Schematics of the SRAM cell with a transient current noise being injected.

the system and the monotonicity of the drive curves, we can see that this new equilibrium state is always below the line $V_2 = V_1$. During $0 < t < pw$, the driving currents of the inverters and the noise current move the state vector in the direction of this new equilibrium state. When the transient noise vanishes, the state vector is guaranteed to be within the attraction region of $\vec{V}_0$, i.e. below the line $V_2 = V_1$, regardless of how large the noise pulse width pw is, and thus no state flip will occur. This is consistent with the conclusion of the SNM analysis. Second, we consider the case when $I_n > I_{snm}$. For $0 < t < pw$, according to the discussion in Section 3.2, there exists only one equilibrium state, which is above the line $V_2 = V_1$. Figure 3.7(a) shows that under the attraction of this equilibrium state, the state trajectory moves toward the line $V_2 = V_1$ and may cross the line if the noise pulse width pw exceeds the *critical pulse width*, T_{crit}, which is the time the trajectory arrives at the line. T_{crit} is a function of the pulse amplitude I_n. Figure 3.7(b) shows the corresponding SPICE-generated plot of the time evolution of the state vector. Thus, the criterion of dynamic stability becomes whether the pulse width pw is greater than T_{crit}. To determine T_{crit}, we only need to analyze the transient behavior before the line $V_2 = V_1$ is reached. Once the line is crossed, a flip will occur. During this time, the noise current is being drained through NMOS transistor M1, and the output voltage of inverter 2 switches low through the NMOS transistor too, Figure 3.8. The majority of the

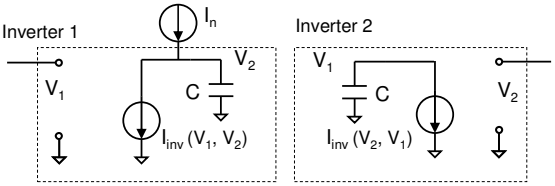
driving currents of both inverters are carried by the NMOS transistors. For inverter 2, since $V_2 < V_1$, M3 is either in cutoff mode or saturation mode, making the driving current primarily controlled by V_2 through transconductance. It is useful to divide the operation of the SRAM cell, Figure 3.7, into two regions: *weak coupling* and *strong feedback mode*. The boundary between the two regions is at the saturation voltage of M1, that is when $V_2 = V_{dsat}$ of M1 with $V_{gs} = V_{dd}$.

The state trajectory starts in the weak coupling mode, in which M1 is in the linear region and M3 is mostly in cut-off if M3's threshold voltage V_{thn} does not differ from V_{dsat} too much. The state trajectory enters the strong feedback mode when M1 becomes saturated. In the weak coupling mode, V_1 nearly remains at V_{dd} and V_2 increases from 0 to V_{dsat}. Figure 3.7(b) shows that the increase of V_2 gets slower with time in the weak-coupling mode, an asymptotic behavior that is common for RC circuits. This feature will be used later for developing the model. In the strong feedback mode ($V_2 > V_{dsat}$), the driving currents of both inverters are primarily controlled by transconductance, i.e., the change of V_1 is primarily determined by V_2 and vice versa, which results in positive feedback. The positive feedback makes both nodal voltages change in a nearly exponential fashion, as shown in Figure 3.7(b).

The above regions of operation can be modeled using the linear gate model of (3.6). Specifically, in the weak coupling region, Figure 3.9(a), inverter 1 behaves like a resistor from V_2 to the ground, with resistance being the linear-region resistance of M1, R_n. Inverter 2 is approximated as not switching. This approximation is only true if V_1 nearly remains at V_{dd} in this mode. We found that for a cell designed in the 100nm technology V_1 can differ from V_{dd} by at most 0.05V for $V_2 \le V_{dsat}$, indicating that the approximation is reasonable. From the above analysis, the time behavior of the SRAM cell in the weak coupling mode is governed by the following equations:



(a)



(b)

Figure 3.9: Using piece-wise linear gate model, SRAM operation is divided into regions of (a) weak coupling and (b) strong feedback.

$$dV_2 / dt = -V_2 / (R_n C) + I_n / C \tag{3.13}$$

$$V_1 = V_{dd} \tag{3.14}$$

with initial state being $(V_{dd}, 0)$, and final state being (V_{dd}, V_{dsat}), which is the point where the state trajectory exits the weak coupling mode. From (3.13), V_2 increases asymptotically toward the steady state value $I_n R_n$ until the working mode is switched.

Solving (3.13) with the initial and final conditions, we find that the time until the trajectory of the state vector leaves the weak coupling mode, T_{wc}, is given by:

$$T_{wc} = -R_n C \ln[1 - V_{dsat} / (I_n R_n)] \tag{3.15}$$

It is clear from (3.15) that to reach the strong feedback mode, the noise amplitude I_n must be larger than V_{dsat} / R_n. Otherwise, positive feedback will not occur and no state-flip will result. Therefore, the SNM I_{snm} can be approximated by V_{dsat} / R_n.

In the strong feedback mode, Figure 3.9(b), both inverters are modeled as voltage controlled current sources between the inverters' output nodes to the ground. By applying

the linear gate model (3.6) in the saturation mode for both inverters, we obtain two cross-coupled linear equations:

$$\begin{bmatrix} dV_1(t)/dt \\ dV_2(t)/dt \end{bmatrix} = -\begin{bmatrix} 0 & g_{mn}/C \\ g_{mn}/C & 0 \end{bmatrix}\begin{bmatrix} V_1(t) \\ V_2(t) \end{bmatrix} + \begin{bmatrix} g_{mn}V_{thn}/C \\ (g_{mn}V_{thn} + I_n)/C \end{bmatrix} \tag{3.16}$$

with initial conditions

$$V_1(t = T_{wc}) = V_{dd}, \quad V_2(t = T_{wc}) = V_{dsat} \tag{3.17}$$

Equation (3.16) describes a linear system. The linear system theory [47] can then be used to solve for the state trajectory analytically. Specifically, linear transformation is first performed on equation (3.16) so that the matrix is diagonalized. In this way, we obtain two equations governing the dynamics of two characteristic functions which are not cross-coupled. The two transformed equations are similar to equation (3.13) and can be solved easily. By performing a reverse-transformation on the characteristic functions, we can obtain a closed-form solution for the state vector. Since our goal is to find the critical pulse width T_{crit}, we impose the final condition: $V_2 = V_1$. This gives:

$$\begin{aligned} T_{crit}(I_n) = {} & -R_n C \ln[1 - V_{dsat}/(I_n R_n)] - \\ & C/g_{mn} \ln[1 - g_{mn}(V_{dd} - V_{thn})/I_n] \end{aligned} \tag{3.18}$$

In (3.18), the first term is T_{wc}, the time spent in the weak-coupling mode, and the second term is the time spent in the strong feedback mode before the attraction boundary is reached. It can be proven that if the first term in (3.18) is finite, i.e., $I_n > V_{dsat}R_n$, and $V_{dsat} > V_{thn}$, the second term is finite too. In other words, if the noise amplitude exceeds the SNM, given sufficiently long (but still finite) duration for the noise pulse, state-flip can occur. The expression (3.18) for the critical pulse width gives the minimum pulse duration that a noise pulse must have in order to cause a state-flip.

It is instructive to check the asymptotic behavior of the critical pulse width when $I_n \to \infty$. In this case, we have a simple bound on the critical pulse width:

$$T_{crit}(I_n) = CV_{dd}/I_n \tag{3.19}$$

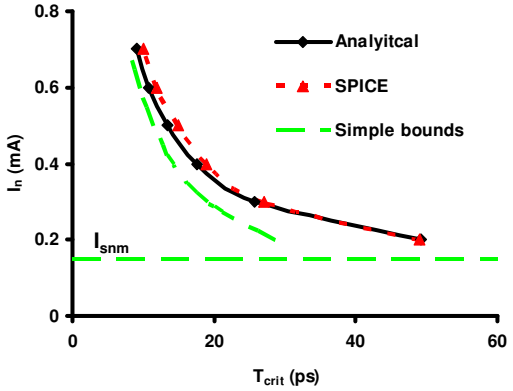


Figure 3.10: Noise amplitude vs. pulse width: region below is safe. Analytical prediction of dynamic margins matches a simulation well. A substantial improvement over the bounds based on the SNM and a simple bound is also verified.

This expression states that when the noise amplitude is sufficiently high, the SRAM cell will undergo a state-flip once the total charge deposited by the transient noise is equal to the charge stored on the node, CV_{dd}. However, if the noise amplitude is comparable to those of the driving currents, the critical pulse width given by (3.18) is greater than that given by (3.19), as confirmed by the experiments in the next section. This is because for smaller noise amplitude, it will take longer time to build enough charge on the victim node for a state-flip. During this time, however, M1 can drain more current from the transient noise, requiring more charge to be deposited by the noise source for a state-flip to occur.

3.4 EXPERIMENTAL RESULTS

In order to verify the validity of the developed theory, and specifically of the critical pulse width T_{crit} (Eq. (3.18)), we designed an SRAM cell using the PTM 100nm technology [30], with V_{dd}=1.2V, $(W/L)_{nmos}$= 2, $(W/L)_{pmos}$=3, and $C = 4.8fC$. We performed SPICE simulation on the SRAM cell as follows. The SRAM cell was put in standby mode. A transient noise having a square-pulse shape was injected into the node

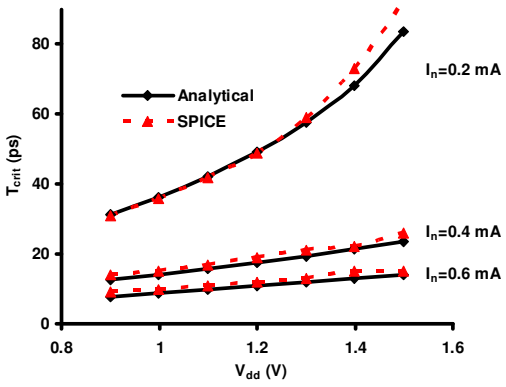


Figure 3.11: The developed analytical model allows SRAM design space exploration, e.g., studying critical pulse T_{crit} at different V_{dd} and noise amplitudes.

where the nodal voltage is zero. The critical pulse width for certain noise amplitude was taken as the smallest pulse width that will result in a state-flip. The SNM was measured as the minimum noise amplitude that can possibly result in a state-flip, with no limitation on pulse width. Figure 3.10 shows the comparison between the SPICE simulation result and the result calculated by (3.18). For the convenience of viewing, Figure 3.10 plots critical pulse width in the horizontal direction. This gives the maximum tolerable noise amplitude given a pulse width. Figure 3.10 also shows the SNM and the asymptotic behavior of the critical pulse width in (3.19) as the simple bounds. It can be seen that excellent match is achieved between the analytical solution and the SPICE simulation result. The error is only significant when the noise amplitude is very close to the SNM. This is expected since in this region, the system performance is very sensitive to the circuit parameters. The dependency of T_{crit} on V_{dd} is shown in Figure 3.11 for three transient noise amplitudes. When adjusting V_{dd}, there is a tradeoff between the amounts of stored charge and feedback delay. Larger V_{dd} leads to shorter feedback delay time, making it easier for a transient noise to be latched. However, larger V_{dd} also results in

more charge stored, making it more difficult for a transient noise to be latched. Figure 3.11 shows that this later effect dominates. It can be seen that as V_{dd} increases, the critical pulse width also increases, indicating that the SRAM cell is more resilient to transient noises for higher supply voltages.

In order to show the validity of the analytical solution for different circuit sizing, we fix the PMOS size at $(W/L)_{pmos}$=3 and vary the NMOS transistor size. Adjusting the transistor size faces similar tradeoff as adjusting V_{dd}: larger transistor size leads to shorter feedback delay time, making it easier for a transient noise to be latched. However, larger transistor size also results in more charges stored and more noise current drained, making it more difficult for a transient noise to be latched. Figure 3.12 shows that this later effects dominates.

An important application of the critical pulse width is to evaluate the impact of an SEU on the stability of an SRAM cell. An SEU induced by cosmic rays typically injects an "exponential" current pulse into an internal node of the cell [27]:

$$I(q,t) = \frac{2q}{\sqrt{\pi}T_s}\sqrt{\frac{t}{T_s}}e^{-\frac{t}{T_s}} \tag{3.20}$$

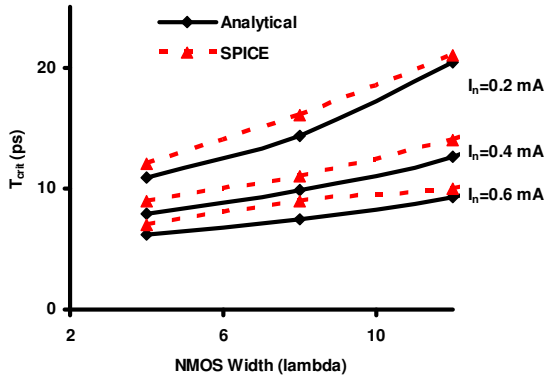


Figure 3.12: Dependency of T_{crit} on the NMOS size for different transient noise amplitudes. PMOS size is $(W/L)_{pmos}$=3.

where q is the collected charge, or total amount of charge deposited by the current pulse, and T_s is the technology-dependent charge-collection time constant It can bee seen from (3.20) that the pulse duration is technology-dependent, and the pulse amplitude is determined by q. The *critical charge* is defined as the minimum amount of collected charge that is able to flip a cell. To compute the critical charge of an SRAM cell, a mapping that needs to be enabled is between the noise profile in (3.20) and the square-pulse noise model on which the theory developed here relies. The mapping criteria we used are: (a) matching the total charge between the square pulse and the "exponential" pulse, and (b) injecting the square pulse and its "exponential" equivalent on an SRAM cell should generate similar effect, i.e. similar state trajectories. It was found through extensive experimentation that setting the pulse width to a value of $3T_s$ produces a reasonable match. The magnitude of the current pulse can then be easily found as $q/(3T_s)$, which can be shown to correspond to 69% of the maximum amplitude of the "exponential" pulse. Figure 3.13 shows a good match between the state trajectories induced by such equivalent pulses for two different collected charges. The procedure of computing the critical charge is as follows. First, we set the critical pulse width to be $3T_s$, and use (3.18) to compute the corresponding noise amplitude I_n, which is the maximum tolerable noise amplitude for this noise duration. Then, the critical charge is obtained as $3T_s I_n$. To verify the result, SPICE simulation was performed to iteratively search for the critical charge by injecting the "exponential" current pulse into the node where the nodal voltage is zero. The comparison in Figure 3.14 shows that good match is achieved and the average estimation error is 11%. Figure 3.14 indicates that higher V_{dd} makes the cell more soft-error resilient, since for higher V_{dd}, more charge is needed to flip a cell.

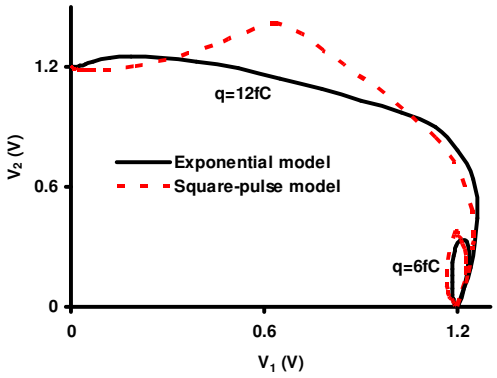


Figure 3.13: A square pulse model approximates here exponential current source by matching total charge. State-trajectories for pulses with two total charge values are shown (q=6 fC and q=12 fC). Trajectories are obtained via SPICE.

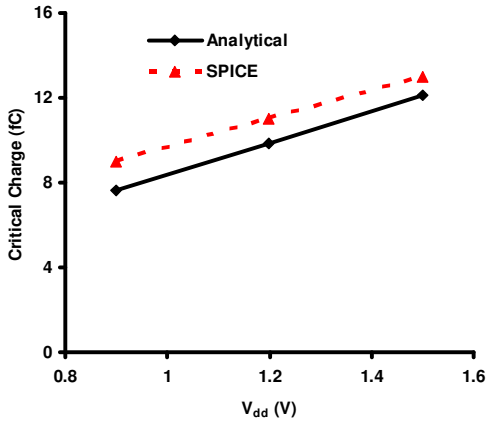


Figure 3.14: Critical charge that will cause a state-flip by a single event upset.

3.5 SUMMARY

In this chapter, for the first time, we present an analytical study of transient error susceptibility of SRAM cells. We propose a criterion for the SRAM dynamic stability.

We model the transient current noise as a square-pulse. By using piece-wise linear modeling of the nonlinear feedback behavior of an SRAM cell, we obtain analytical closed-form solution of critical pulse width that can flip the state of the SRAM cell. Experiments show that excellent match is achieved between the analytical prediction and the SPICE simulation results.

Chapter 4: Modeling of NBTI-Induced PMOS Degradation under Arbitrary Dynamic Temperature Variation

In Chapter 2 and 3, we have analyzed the impact of transient noise on both combinational logic and SRAM memory array. In Chapter 4 and 5, we will focus on the impact of permanent errors, specifically, the timing errors due to negative bias temperature instability (NBTI). The physical mechanism of NBTI is described in Chapter 1. NBTI causes the magnitude of PMOS threshold voltage to increase over time, which leads to the increase of circuit delay. Circuit designers must ensure that the increase of circuit delay shall not result in any timing errors within the lifetime of the integrated circuits. NBTI is highly sensitive to operating temperature. On the other hand, modern ICs exhibit significant variation in operating temperature, both spatially and temporally [56][57]. Therefore, in order to accurately predict the amount of NBTI-induced performance degradation, it is necessary to take into account the effect of temperature variation. In this work, we focus on the impact of temporal variation of temperature, i.e., dynamic temperature variation.

Most prior works on NBTI modeling assume that both the temperature and the voltage stress are fixed at constant values [53][54][58][59]. The only one that deals with dynamic temperature variation is [63], but it only handles two-level temperature variation and the model lacks validation.

In this work, we propose a compact NBTI model that handles arbitrary temperature variation. In addition, the effect of voltage signal toggling is also considered. The proposed NBTI model is consistent with the general framework of reaction-diffusion (R-D) model [53][54].

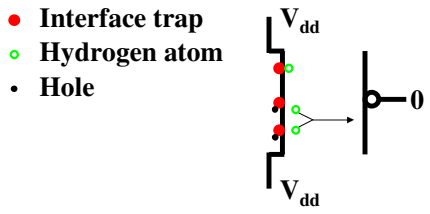


Figure 4.1: Illustration of the NBTI process. Holes break the Si-H bonds to generate an equal number of interface traps and free hydrogen atoms at the Si-SiO_2 interface. Part of the interface traps and free hydrogen atoms re-combine, while the rest of the hydrogen atoms subsequently form molecules and diffuse away, leaving behind un-annealed interface traps.

This chapter is organized as follows. In Section 4.1, we study NBTI under dynamic temperature variation and constant voltage stress. In Section 4.2, the joint impact of temperature variation and voltage signal transition is investigated. We present the experimental results in Section 4.3 and summarize this chapter in Section 4.4.

4.1 Model of NBTI under Dynamic Temperature Variation and Constant Voltage Stress

The origin of NBTI can be traced back to the intrinsic structural mismatch at the Si-SiO_2 interface of a MOS transistor, which results in dangling Si- bonds acting as interface traps. Figure 4.2 ([55]) illustrates 3 types of interface traps, Pb, Pb0 and Pb1, which differ by electrical activity and density. When the PMOS is in inversion, the interface traps become positively charged, which leads to the increase of the magnitude of the threshold voltage [55].

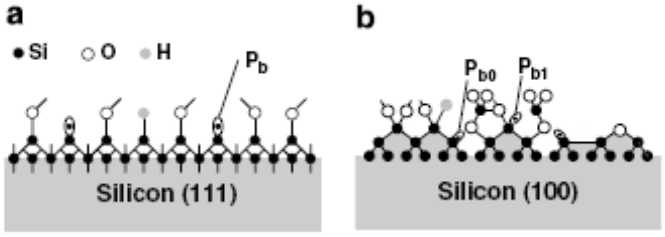


Figure 4.2: Interface traps at the (a): (111) Si surface and (b): (100) Si surface ([55])

The interface traps are detrimental to device performance. Therefore, in IC fabrication, hydrogen passivation is introduced to anneal these traps, forming electrically in-active Si-H bonds [69]. NBTI happens during the usage of a PMOS device in which the Si-H bonds can be broken by holes accumulated in the conducting channel when the gate of the PMOS device is negatively biased (i.e., the gate voltage is 0 *Volt* and the source voltage is V_{dd}), generating interface traps [53][54]. NBTI consists of the following sub-processes having distinct temperature-dependences [53][59][64]:

1. Accumulation of holes: negative voltage bias of the PMOS gate induces the electrical field across the gate, leading to the accumulation of holes in the conducting channel of the PMOS device.

2. Forward reaction: holes break the Si-H bonds to generate an equal number of interface traps and free hydrogen atoms at the Si-SiO_2 interface.

3. Backward reaction: one interface trap and one free hydrogen atom re-combine to form a Si-H bond at the Si-SiO_2 interface.

4. Pairs of hydrogen atoms that are not involved in backward reaction form hydrogen molecules: $\mathrm{H}+\mathrm{H} \rightarrow \mathrm{H}_2$.

5. The hydrogen molecules diffuse into the gate oxide and then the poly-silicon. It is the diffusion of the hydrogen molecules that leaves a portion of the interface traps permanently un-annealed. The un-annealed interface traps manifest themselves in increased magnitude of the PMOS threshold voltage, $\left|V_{th}\right|$.

These processes are illustrated in Figure 4.1. The dynamic combination of the above sub-processes can be described by the reaction–diffusion (R-D) model [59].

$$\begin{aligned}\frac{dN_{IT}(t)}{dt} &= k_F\left[V,T(t)\right]\left[N_0 - N_{IT}(t)\right] \\ &\quad -k_R\left[V,T(t)\right]N_{IT}(t)\sqrt{N_{H_2}(x,t)\big|_{x=0}}\end{aligned} \tag{4.1}$$

$$\frac{\partial N_{H_2}(x,t)}{\partial t} = D_{H_2}[T(t)]\frac{\partial^2 N_{H_2}(x,t)}{\partial x^2} \qquad (x \geq 0) \tag{4.2}$$

subject to the boundary condition,

$$-D_{H_2}[T(t)]\frac{\partial N_{H_2}(x,t)}{\partial x}\bigg|_{x=0} = \frac{1}{2}\frac{dN_{IT}(t)}{dt} \tag{4.3}$$

and the initial condition,

$$N_{H_2}(\forall x \geq 0, t=0) = 0, \quad N_{IT}(t=0) = 0 \tag{4.4}$$

where k_F and k_R are the forward and backward reaction rates respectively; N_0, N_{IT}, and N_{H_2} are the initial Si-H bond density, interface trap density, and H_2 density respectively; D_{H_2} is the diffusivity of H_2; V and T are voltage stress and temperature respectively. The interface is at $x = 0$. Note that the sub-processes 1 and 2 have been combined to be the forward-reaction term, i.e. the first term on the right-hand side of (4.1). The sub-process 3 is included as the second term on the right-hand side of (4.1), where $\sqrt{N_{H_2}(x,t)|_{x=0}}$ is a result of the following fact: when the sub-process 4 reaches equilibrium, $N_H(x,t)|_{x=0} \propto \sqrt{N_{H_2}(x,t)|_{x=0}}$ [59]. The diffusion of H_2 is described in (4.2). The boundary condition (4.3) describes the fact that when two un-annealed interface traps are generated, one H_2 will diffuse into the gate oxide. It is to be noted that in the case of *temperature variation* and constant voltage stress, *the parameters k_F, k_R, and D_{H_2} can vary with time if they have a non-zero temperature activation energy*, in contrast to the assumption generally made in the literature which treats these parameters as constant throughout the lifetime of the PMOS [53][54][59][64].

The reaction rates k_F, k_R, and D_{H_2} are related to V and T through (4.5)-(4.6) [53]:

$$\begin{aligned}\frac{k_F}{k_R} &\propto E_{ox} e^{E_{ox}/E_0} e^{[-E_a(k_F)+E_a(k_R)]/kT} \\ &\equiv k_{FR}(V) e^{[-E_a(k_F)+E_a(k_R)]/kT}\end{aligned} \tag{4.5}$$

$$D_{H_2} \propto e^{-Ea(D_{H_2})/kT} \tag{4.6}$$

with

$$E_{ox} = (|V| - |V_{th}|) / t_{ox} \tag{4.7}$$

where E_{ox}, t_{ox}, E_0, and k are the oxide field, the oxide thickness, the field acceleration constant, and the Boltzman constant, respectively; and $E_a(k_F)$, $E_a(k_R)$, $E_a(D_{H_2})$ are the temperature activation energies of k_F, k_R, and D_{H_2} respectively. $k_{FR}(V)$ is the voltage dependence of k_F / k_R. Note that D_{H2} is independent of V since H_2 is electrically neutral.

Although NBTI can cause the shift of several device parameters of PMOS including the threshold voltage, carrier mobility, and series resistance, the degradation of carrier mobility and series resistance is negligible compared to the degradation of V_{th} [65]. The change of threshold voltage of the PMOS device after stress time of t, $\Delta|V_{th}(t)|$, can be found as [53]:

$$\Delta|V_{th}(t)| = qN_{IT}(t) / C_{ox} \tag{4.8}$$

where q and C_{ox} are electronic charge and gate capacitance per unit area, respectively.

We first make the following observations:

1. Experimentally, it was observed that [53]:

$$E_a(k_F) \approx E_a(k_R) \tag{4.9}$$

2. Diffusion is much slower than reaction so that reaction can be approximated to be in quasi-equilibrium [53][54][59], i.e. $dN_{IT}(t) / dt \approx 0$ at any time t in (4.1). Based on this fact, together with (4.5) and (4.9), (4.1) can be simplified to

$$\frac{N_{IT}(t)\sqrt{N_{H_2}(x,t)\big|_{x=0}}}{N_0 - N_{IT}(t)} = k_{FR}(V) \tag{4.1a}$$

Note that the right-hand side of (4.1a) is independent of the temperature. We then make (4.2) and (4.3) "time-invariant" by transforming the variable t into "equivalent time" τ ($t \to \tau$):

$$\begin{aligned}\tau &= \frac{1}{D_{H_2}(T_{ref})}\int_0^t D_{H_2}[T(t^\dagger)]dt^\dagger \\ &= \int_0^t e^{[E_a(D_{H_2})/k]\bullet[1/T_{ref}-1/T(t^\dagger)]}dt^\dagger\end{aligned} \tag{4.10}$$

where T_{ref} is a constant reference temperature. We set $T_{ref} = 125^\circ C$ throughout this work. From (4.10) we have

$$dt = \frac{D_{H_2}[T_{ref}]d\tau}{D_{H_2}[T(t)]} \tag{4.11}$$

Substituting (4.10) and (4.11) into (4.2) and (4.3), and including (4.1a) for completeness, we have,

$$\frac{N_{IT}(\tau)\sqrt{N_{H_2}(x,\tau)\Big|_{x=0}}}{N_0 - N_{IT}(\tau)} = k_{FR}(V) \tag{4.1a}$$

$$\frac{\partial N_{H_2}(x,\tau)}{\partial \tau} = D_{H_2}(T_{ref})\frac{\partial^2 N_{H_2}(x,\tau)}{\partial x^2} \quad (x \geq 0) \tag{4.2a}$$

$$-D_{H_2}(T_{ref})\frac{\partial N_{H_2}(x,\tau)}{\partial x}\bigg|_{x=0} = \frac{1}{2}\frac{dN_{IT}(\tau)}{d\tau} \tag{4.3a}$$

Note that the diffusivity in (4.2a) and (4.3a) is now a *time-independent* constant, $D_{H_2}(T_{ref})$. The dependence of the diffusion process, (4.2a), on temperature variation is *implicitly* captured through the conversion between t and τ in (4.10).

In (4.1a)-(4.3a), *explicit* dependence of the reaction and diffusion processes on temperature variation is eliminated. Indeed, (4.1a)-(4.3a) describe NBTI under constant temperature T_{ref} in the τ domain. The problem of NBTI under dynamic temperature variation $T(t)$, (4.1)-(4.3), is thus converted to the problem of NBTI under constant temperature, T_{ref}, (4.1a)-(4.3a), and we have

$$N_{IT}[t, T(t)] = N_{IT}[\tau(t), T = T_{ref}] \tag{4.12}$$

From (4.1a)-(4.3a), $N_{IT}(\tau)$ can be rigorously solved for $N_{IT} \ll N_0$ by applying Laplace transform [54], Appendix. The solution is given in (4.13), where N_{IT} has been converted to $\Delta \mid V_{th} \mid$ using (4.8):

$$\begin{aligned} \Delta \mid V_{th}(\tau) \mid = \left(q / C_{ox}\right)\left[2\Gamma\left(2/3\right) / \Gamma\left(7/6\right)\right]^{1/3} \times \\ \left[k_{FR}(V) N_0\right]^{2/3} \left[D_{H_2}(T_{ref})\right]^{1/6} \tau^{1/6} \end{aligned} \tag{4.13}$$

Combining (4.10), (4.12), and (4.13), we obtain

$$\begin{aligned} \Delta \mid V_{th}[t, T(t)] \mid = \left(q / C_{ox}\right)\left[2\Gamma\left(2/3\right) / \Gamma\left(7/6\right)\right]^{1/3} \times \\ \left[k_{FR}(V) N_0\right]^{2/3} \left[D_{H_2}(T_{ref})\right]^{1/6} \times \\ \{\int_0^t e^{[E_a(D_{H_2})/k] \bullet [1/T_{ref} - 1/T(t^\dagger)]} dt^\dagger\}^{1/6} \end{aligned} \tag{4.14}$$

where $\Gamma()$ is the gamma function. We can re-write (4.14) as

$$\begin{aligned} \Delta \mid V_{th}[t, T(t)] \mid = \Delta \mid V_{th}(t, T = T_{ref}) \mid \times \\ \left[(1/t)\int_0^t e^{[E_a(D_{H_2})/k] \bullet [1/T_{ref} - 1/T(t^\dagger)]} dt^\dagger\right]^{1/6} \end{aligned} \tag{4.15}$$

It is shown in (4.15) that $\Delta \mid V_{th} \mid$ under temperature variation can be obtained from $\Delta \mid V_{th} \mid$ under constant temperature by scaling the stress time based on the history of temperature variation. The equation (4.14) forms our proposed model for predicting NBTI under temperature variation. Alternatively, (4.15) can be used if $\Delta \mid V_{th}(t, T = T_{ref}) \mid$ is easier to characterize than k_{FR} and D_{H_2}.

In our model, the computational challenge is to find the equivalent stress time τ, which is the integral in (4.14). The following two algorithms can be used for computing τ:

1. If the sequence of temperatures is obtained sequentially from an on-line temperature monitor, two variables can be used to store the temperature T and the equivalent stress time τ for the current time. They are updated as follows:

$$\begin{aligned} T &\leftarrow \text{current temperature} \\ \tau &\leftarrow \tau + e^{[E_a(D_{H_2})/k] \bullet (1/T_{ref} - 1/T)} \Delta t \end{aligned} \tag{4.16}$$

2. If the distribution of the temperature profile can be obtained through thermal modeling or on-line temperature monitoring, we have

$$\begin{aligned}\tau &= t\int_{T_{\min}}^{T_{\max}} e^{[E_a(D_{H_2})/k]\bullet[1/T_{ref}-1/T]}p(T)dT \\ &\approx t\sum_{T=T_{\min}}^{T_{\max}} e^{[E_a(D_{H_2})/k]\bullet[1/T_{ref}-1/T]}p(T)\Delta T\end{aligned} \tag{4.17}$$

where $p(T)$ is the pdf of the temperature distribution over time. Using (4.17), we can avoid lengthy integration over time, since the integral variable is now the temperature and the number of discretized temperatures that need to be considered is much less than that of the time samples.

4.2 JOINT IMPACT OF TEMPERATURE VARIATION AND VOLTAGE SIGNAL TRANSITION

Under *constant temperature*, degradation due to NBTI with voltage stress changing between V_{dd} and 0 *Volt* as a result of signal transition has been well studied [51][60]-[63]. In these studies, instead of directly solving the partial differential equations (4.1)-(4.4), an approximate approach based on the analysis of the *diffusion front* [53], which is the longest distance that the hydrogen molecules diffuse away from the Si-SiO2 interface, is employed. In the stress stage (stress voltage is V_{dd}) $\Delta|V_{th}(t)|$ increases with time, while in the recovery stage (stress voltage is 0 *Volt*) $\Delta|V_{th}(t)|$ decreases with time, but never returns to 0. When stress-recovery occurs *periodically*, the diffusion front analysis shows that in the long term, i.e., $t \gg T_{clk}$ [51]

$$\Delta|V_{th}(t)| \le \left[\frac{(\Delta|V_{th}(T_{clk})|)^3}{1-\beta_t^3}\right]^{1/3} \tag{4.18}$$

where T_{clk} is the clock period; $\Delta|V_{th}(T_{clk})|$ is the shifting of V_{th} under *constant voltage stress* for a duration of T_{clk}; and β_t is determined by the duty cycle, α, which is the percentage of the time spent on voltage stress

$$\beta_t = 1-\sqrt{0.5(1-\alpha)T_{clk}/t} \tag{4.19}$$

In (4.19), the thickness effect of the gate oxide has been ignored for simplicity of the description. The dependence of $\Delta|V_{th}(t)|$ on T_{clk} is weak for frequencies lower than 10MHZ [61].

In [61], Monte Carlo simulations show that $\Delta|V_{th}(t)|$ due to *non-periodical* voltage switches is approximately the same as that in the case of *periodical* signals with the same duty cycle.

When temperature does vary over time, we first make the transformation $t \to \tau$, as we did in Section 4.1, so that in the τ domain, the *explicit* dependence of NBTI on dynamic temperature variation is eliminated. Then, in the τ domain, we apply the established solution for constant temperature, (4.18), to compute $\Delta|V_{th}(\tau)|$, and $\Delta|V_{th}(t)| = \Delta|V_{th}[\tau(t)]|$.

In Section 4.1, we have shown that when the voltage stress is held at a constant value, the $t \to \tau$ transformation can be performed by (4.10). In fact, we can similarly show that if the voltage stress changes between V_{dd} and 0, the transformation (4.10) can also eliminate the explicit dependence of NBTI on temperature variation. However, since the transformation scales the time dynamically according to the temperature variation, a period of T_{clk} in the t domain may correspond to different equivalent periods, $\hat{T}_{clk}$, in the τ domain at different temperature values, we take the average value of the equivalent periods as an approximation:

$$\hat{T}_{clk} \approx [\tau(t)/t]T_{clk} \tag{4.20}$$

Similarly, the equivalent duty cycle in the τ domain can be approximated as:

$$\hat{\alpha} = [1/\tau(t)]\int_0^t S(t^\dagger)e^{[E_a(D_{H_2})/k]\bullet[1/T_{ref}-1/T(t^\dagger)]}dt^\dagger \tag{4.21}$$

where $S(t^\dagger)$ is equal to 1 if there is voltage stress at $t^\dagger$ and 0 otherwise. When $S(t)$ is independent of the temperature $T(t)$, we have:

$$\hat{\alpha} = (1/t)\int_0^t S(t^\dagger)dt^\dagger \equiv \alpha \tag{4.22}$$

To summarize, when a PMOS is subject to both temperature variation and voltage signal transition, the shift of PMOS threshold voltage is given by (4.18) with $t \to \tau(t)$, $T_{clk} \to \hat{T}_{clk}$, and $\alpha \to \hat{\alpha}$.

4.3 MODEL VALIDATION

We first validate our model with temperature variation and constant voltage stress, (4.14). Although the impact of temperature variation on NBTI is crucial to the operation of integrated circuits, measurements of NBTI with temporal variation of temperature have not been reported in the literature, to the best knowledge of the authors. Therefore, we validate our model in two stages. First, we implement a simulator that solves the partial differential equations (4.1)-(4.4) numerically. The parameters we use in the simulator are either directly from the literature or extracted from the published data [53][64][66]. We validate the simulator by comparing its prediction with the measurements at fixed temperature values [66], Figure 4.3(a). Then we validate our model by comparing its prediction with the result of the simulator under temperature variation. The specific temperature profile we use for validation is from thermal cycling that is of significant concern for IC reliability [70]. The temperature profile can be modeled as a sinusoidal wave:

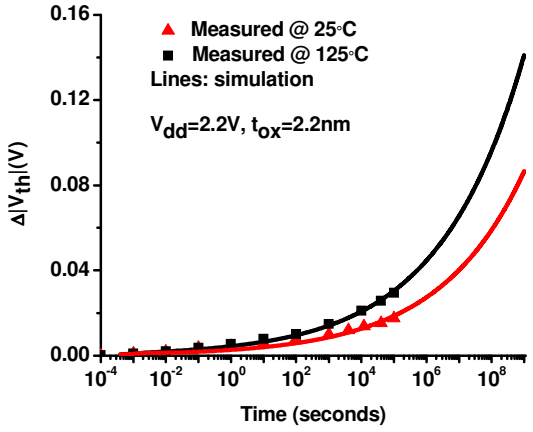
$$T(t) = (1/2)\times(T_{\min} + T_{\max}) + (1/2)\times(T_{\max} - T_{\min})\sin(2\pi f t) \tag{4.23}$$

where $T_{\min}$ and $T_{\max}$ are the minimum and maximum temperature values respectively and f is the thermal frequency. By applying our model, we can show that for $t \gg 1/f$:

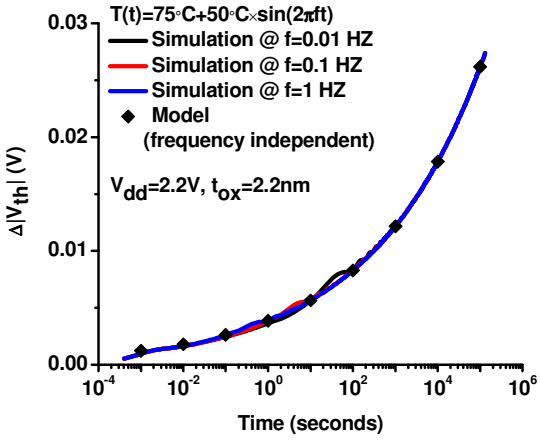
$$\Delta \mid V_{th}[t, T(t)] \mid = ct^{1/6} \tag{4.24}$$

where c is a time-independent constant, which depends on $T_{\min}$ and $T_{\max}$ but not f. This leads to an interesting observation as follows:

The shift of PMOS threshold voltage at t is independent of the frequency of thermal cycling with sinusoidal temperature waveform f, given $t \gg 1/f$.



(a)



(b)

Figure 4.3: Two-stage validation of the NBTI model under dynamic temperature variation and constant voltage stress. (a): a simulator based on solving the partial differential equations (4.1)-(4.4) is implemented and validated. Measurement data is from [66]. (b): the simulator is used to validate the analytical model (4.14) under thermal cycling in which the temperature varies as a sinusoidal wave of different frequencies.

Validation of the model under thermal cycling is shown in Figure 4.3(b). It can be seen that very good match has been achieved between the result of the model and that of the simulation. A special case of the temperature variation is that $T(t)$ can only change between two levels as in [63]. We can show that in this case, our model will lead to the same result as in [63].

We also validate our model subject to both temperature variation and voltage signal transition. The temperature variation profile is generated as follows. First, we generate 10 temporally-correlated temperature waveforms similar to Figure 1.3 by Gaussian processes. For each waveform, we artificially set the mean value and the standard deviation, so that these waveforms correspond to different workloads. These waveforms are then randomly selected with equal probabilities and concatenated to form a sequence of temperature variation. The histogram of the temperature is shown in Figure 4.4. Voltage stress is randomly applied, independent of temperature, with 50% duty cycle. Clock period is set to be 100 seconds and correlation time of the temperature

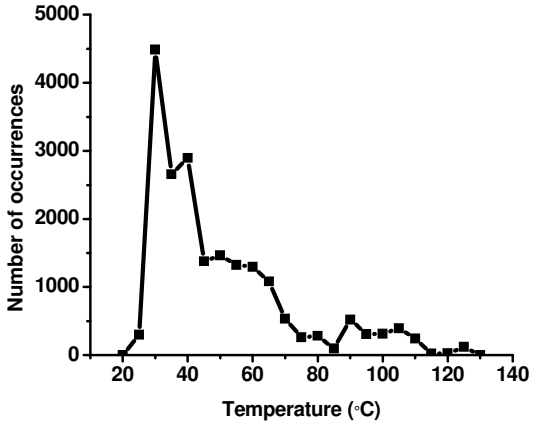


Figure 4.4: Histogram of the dynamic temperature variation. Worst-case temperature is 123 ^{o}C.

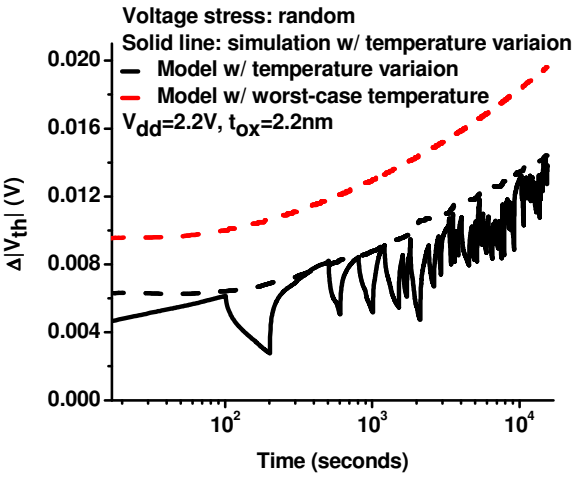


Figure 4.5: NBTI with both temperature variation and random transition of voltage signal. Duty cycle of the voltage signal is 50%. The proposed model accounting for temperature variation provides a significantly tighter bound for the simulation than the model that ignores the temperature variation and assumes a constant (worst-case) temperature, 123°C.

waveforms is set to be 30 seconds so that the simulator can compute $\Delta|V_{th}|$ after long enough time, while it is still able to capture the details of the transient behavior of the voltage and temperature. It is to be noted that the analytical model is much less constrained by this limit on clock period and correlation time of the temperature waveforms.

We assume that an on-chip temperature sensor is available and it samples the temperature every 1 second. For comparison, we also estimate $\Delta|V_{th}|$ based on worst-case assumption of temperature. The results from the simulation and the models are shown in Figure 4.5. It can be seen that the proposed model accounting for temperature variation provides a significantly tighter bound for the simulation than that from the model that ignores the temperature variation and assumes a constant (worst-case) temperature. In this experiment, the amount of degradation predicted by the proposed

dynamic temperature model is on average 46% less conservative compared to the model based on the worst-case temperature. Note that the errors of the models are with respect to the simulation result

4.4 SUMMARY

In this work, we have proposed a novel model of NBTI under arbitrary dynamic temperature variation based on the reaction- diffusion model. An analytical model for NBTI under temperature variation is developed and validated. The analytical model is especially useful for on-line reliability monitoring and reliable circuit design when thermal-profiling or modeling are available.

Appendix

In this appendix, we show the derivation for obtaining (4.13) from (4.1a)-(4.3a). We repeat (4.1a)-(4.3a) here, under the condition $N_{IT} \ll N_0$,

$$\frac{N_{IT}(\tau)\sqrt{N_{H_2}(x,\tau)\Big|_{x=0}}}{N_0} = k_{FR}(V) \tag{A1}$$

$$\frac{\partial N_{H_2}(x,\tau)}{\partial \tau} = D_{H_2}(T_{ref})\frac{\partial^2 N_{H_2}(x,\tau)}{\partial x^2} \quad (x \geq 0) \tag{A2}$$

$$-D_{H_2}(T_{ref})\frac{\partial N_{H_2}(x,\tau)}{\partial x}\bigg|_{x=0} = \frac{1}{2}\frac{dN_{IT}(\tau)}{d\tau} \tag{A3}$$

We perform Laplace transform on (A2), so that $N_{H_2}(x,\tau) \rightarrow \widetilde{N}_{H_2}(x,s)$:

$$s\widetilde{N}_{H_2}(x,s) = D_{H_2}(T_{ref})\frac{\partial^2 \widetilde{N}_{H_2}(x,s)}{\partial x^2} \quad (x \geq 0) \tag{A4}$$

Treating s as a "constant", the general solution of (A4) can be obtained:

$$\begin{aligned}\widetilde{N}_{H_2}(x,s) = {} & C_1(s)\exp[-\sqrt{s / D_{H_2}(T_{ref})}x] \\ & + C_2(s)\exp[+\sqrt{s / D_{H_2}(T_{ref})}x] \\ & (x \geq 0)\end{aligned} \tag{A5}$$

where $C_1(s)$ and $C_2(s)$ are independent of x and need to be determined by boundary and initial conditions. Since $\widetilde{N}_{H_2}(x,s) \rightarrow 0$ for $x \rightarrow \infty$, $C_2(s)$ must be 0. Therefore,

$$\begin{aligned}\widetilde{N}_{H_2}(x,s) = {} & C_1(s)\exp[-\sqrt{s / D_{H_2}(T_{ref})}x] \\ & (x \geq 0)\end{aligned} \tag{A6}$$

The Laplace transform of (A3) is

$$-D_{H_2}(T_{ref})\frac{\partial \widetilde{N}_{H_2}(x,s)}{\partial x}\bigg|_{x=0} = \frac{1}{2}s\widetilde{N}_{IT}(s) \tag{A7}$$

Substituting (A6) into (A7), we have

$$\sqrt{sD_{H_2}(T_{ref})}\,C_1(s)\exp[-\sqrt{s / D_{H_2}(T_{ref})}x\Big|_{x=0} = \frac{1}{2}s\widetilde{N}_{IT}(s) \tag{A8}$$

Hence,

$$\sqrt{sD_{H_2}(T_{ref})}\widetilde{N}_{H_2}(x,s)\,|_{x=0} = \frac{1}{2}s\widetilde{N}_{IT}(s) \tag{A9}$$

We have not used (A1) so far. To take advantage of it, assume that $N_{IT}(\tau)$ can be expressed as

$$N_{IT}(\tau) = R\tau^n \tag{A10}$$

where R and n are time-independent constants that need to be determined. From (A1), we have

$$N_{H_2}(x,\tau)\Big|_{x=0} = [k_{FR}(V)N_0 \ / \ R\tau^n]^2 \tag{A11}$$

We make Laplace transform on (A10) and (A11),

$$\widetilde{N}_{IT}(s) = R\Gamma(n+1) \ / \ s^{n+1} \tag{A12}$$

$$\widetilde{N}_{H_2}(x,s)\Big|_{x=0} = [k_{FR}(V)N_0 \ / \ R]^2\Gamma(1-2n) \ / \ s^{1-2n} \tag{A13}$$

Substituting (A12) and (A13) into (A9), we have

$$R^3 = 2\Gamma(1-2n) \ / \ \Gamma(n+1)[k_{FR}(V)N_0]^2\sqrt{D_{H_2}(T_{ref})}s^{-1/2+3n} \tag{A14}$$

Since R is a time-independent constant, R should not be a function of s. Therefore,

$$-1 \ / \ 2 + 3n = 0 \tag{A15}$$

Hence,

$$n = 1 \ / \ 6 \tag{A16}$$

and

$$R = [2\Gamma(2 \ / \ 3) \ / \ \Gamma(7 \ / \ 6)]^{1/3}[k_{FR}(V)N_0]^{2/3}D_{H_2}(T_{ref})^{1/6} \tag{A17}$$

Using (A10), we have

$$N_{IT}(\tau) = [2\Gamma(2 \ / \ 3) \ / \ \Gamma(7 \ / \ 6)]^{1/3}[k_{FR}(V)N_0]^{2/3}D_{H_2}(T_{ref})^{1/6}\tau^{1/6} \tag{A18}$$

Given $\Delta\left|V_{th}(\tau)\right| = qN_{IT}(\tau) \ / \ C_{ox}$, we can obtain (4.13) from (A18).

Chapter 5: Online Circuit Reliability Monitoring

In Chapter 1 and 4, we have shown that NBTI-induced circuit degradation can vary significantly with the chip temperature. In traditional design practice, worst-case based guard-band is added to the clock period to prevent timing errors within the circuit's lifetime [52]. This pessimism inevitably incurs high cost in performance. One approach to overcome this problem is to use online monitoring to measure the *actual* degradation of the circuit [52][57][72]. The existing measurement techniques are either inaccurate, since they measure only the degradation of a single structure (a PMOS transistor or a ring oscillator) and infers the degradation of a large circuit without considering the impact of its circuit topology [57], or expensive, since they measure the circuit delay by sensors constantly monitoring the part or all of the primary outputs of the circuit[52][72][73].

In this work we propose an online reliability tracking framework that utilizes a hybrid network of on-chip temperature and delay sensors together with a circuit reliability macromodel. The key feature of our work is an explicit macromodel which maps operating temperature to circuit degradation. The macromodel allows for cost-effective reliability tracking. The accuracy of the model is improved by online calibration of model parameters via monitoring the delay degradation of ring oscillators. The number of model parameters is relatively small. For example, in ISCAS'85 benchmark circuits, at most 21 parameters are required for the macromodel. The prediction of circuit degradation using our online monitoring strategy can be up to 20% less conservative compared to the worst-case reliability prediction.

This chapter is organized as follows. In Section 5.1, we develop the macromodel for NBTI-induced circuit degradation. In Section 5.2, we describe the online calibration

using ring oscillators. We present experimental results in Section 5.3, and summarize this chapter in Section 5.4.

5.1 CIRCUIT-LEVEL RELIABILITY MACROMODEL

The circuit reliability macromodel is built upon the existing device-level reliability models [53][75][77][78]. We first give a brief review of these models.

5.1.1 Device-Level NBTI Modeling

NBTI is caused by the accumulation of interface traps at the SiO_2-Si interface when the gate of a PMOS is negatively biased with respect to its drain or source voltages [51][53][54]. Under constant temperature, T, and fixed (DC) voltage stress, V_{dd}, the shift of the PMOS threshold voltage, ΔV_{th-DC}, at any time, t, is [51][53][58][59]:

$$\Delta V_{th-DC}(t \mid T) = c(V_{dd}, V_{th0}, T)t^{1/6} \tag{5.1}$$

where c is a function that depends on V_{dd}, T, and the initial threshold voltage V_{th0}. It has been shown that in [75][76], at constant temperature, if the gate voltage toggles between V_{dd} and 0 with a signal probability of *sp*, the shift of PMOS threshold voltage under this AC voltage stress, ΔV_{th-AC}, is proportional to ΔV_{th-DC} in the long term,

$$\Delta V_{th-AC}(t \mid T, sp) = g(sp)\Delta V_{th-DC}(t \mid T) \tag{5.2}$$

with the proportionality constant g determined by *sp*. The function $g(sp)$ can be precharacterized by device-level simulation [75] or measurement [76]. Figure 5.1 shows $g(sp)$ for a 65nm PMOS transistor [75].

In [77], it is shown that, in the case of varying temperature, $T(\tau)$ ($\forall \tau, 0 \le \tau \le t$), the forms of (5.1) and (5.2) can be well preserved, with t replaced by the temperature-dependent equivalent stress-time, t_{eq}, defined with respect to a constant reference temperature T_{ref}:

$$\Delta V_{th-DC}[t \mid T(0 \le \tau \le t)] = c(V_{dd}, V_{th0}, T_{ref}){t_{eq}}^{1/6} \tag{5.3}$$

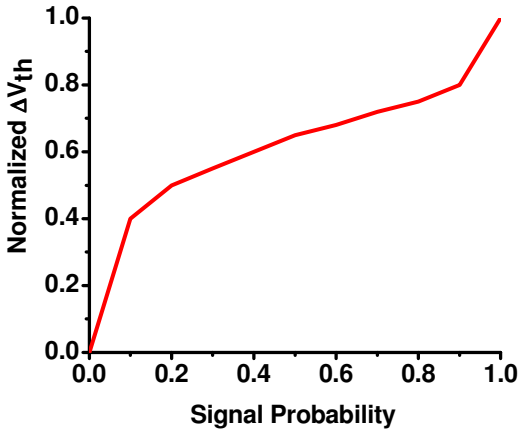


Figure 5.1: The proportionality function relating degradation model at AC voltage stress to that at DC voltage stress (signal probability=1) is extracted through simulation [75].

$$\begin{aligned} &\Delta V_{th-AC}[t \mid T(0 \le \tau \le t), sp] \\ &= g(sp)\Delta V_{th-DC}[t \mid T(0 \le \tau \le t)] \end{aligned} \tag{5.4}$$

where

$$t_{eq} = \int_0^t e^{[E_a/k]\bullet[1/T_{ref}-1/T(\tau)]} d\tau \tag{5.5}$$

In (5.5), E_a is the temperature activation energy and k is the Boltzman constant. We define the *device degradation function*, f, to be

$$f[c, E_a, T(0), \ldots, T(t)] \equiv \Delta V_{th-DC}[t \mid T(0 \le \tau \le t)] \tag{5.6}$$

Then

$$\Delta V_{th-AC}[t \mid T(0 \le \tau \le t), sp] = g(sp) f[c, E_a, T(0), \ldots, T(t)] \tag{5.7}$$

To summarize, the shift of PMOS threshold voltage under typical stress condition is proportional to the device degradation function f in the long term. The function f depends on the temperature history, as well as device-level parameters.

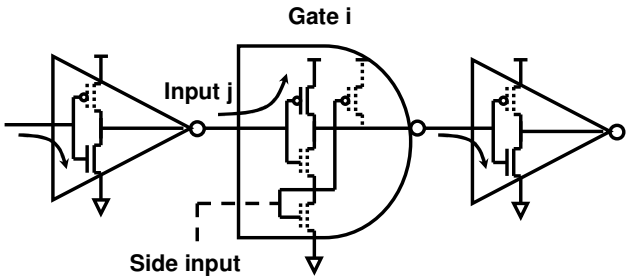


Figure 5.2: Illustration of a portion of a path and its switching direction. The transistors on the path are shown in solid lines.

5.1.2 Development of the Reliability Macromodel

Figure 5.2 illustrates a portion of a path and the signal switching direction. Under the impact of NBTI, the path delay *p* increases over time and can be described by:

$$p(t) = p\_up(t) + p\_down \tag{5.8}$$

where $p\_up(t)$ and p_down are the sum of pull-up and pull-down delays along the path, respectively. Note that p_down does not depend on time, because NMOS transistors are not affected by NBTI. $p\_up(t)$ can be further written as

$$p\_up(t) = p\_up^0 + \sum_i \Delta d_i(t) \qquad i \in \{\text{pull-up gates}\} \tag{5.9}$$

where p_up^0 is p_up at $t = 0$, and $\{\Delta d_i\}$ are the delay increase of the pull-up gates on the path, due to V_{th} shift of the PMOS transistors. For any gate i, suppose that input j is on the path being considered, we have [78]

$$\Delta d_i(t) = \lambda_{ij} \Delta V_{th,ij}(t) \tag{5.10}$$

where $\Delta V_{th,ij}(t)$ is the V_{th} shift of the PMOS (i, j) that is associated with input j of gate i, and λ_{ij} is a proportionality constant that can be characterized by SPICE simulation. From (5.7),

$$\Delta V_{th,ij}(t) = g(sp_{ij}) f[c, E_a, T(0), ..., T(t)] \tag{5.11}$$

where sp_{ij} is the input signal probability of PMOS (i, j). For PMOS transistors in series, sp_{ij} is the effective signal probability [75]. Putting (5.8)-(5.11) together and assuming that the gates of a combinational logic circuit block are compactly placed together so that they share the same temperature, we have:

$$p(t) = p^0 + p^* f[c, E_a, T(0), ..., T(t)] \tag{5.12}$$

where p^0 is the path delay at $t = 0$:

$$p^0 = p\_up^0 + p\_down \tag{5.13}$$

and

$$p^* = \sum_i \lambda_{ij} g(sp_{ij}) \tag{5.14}$$

In (5.12), we express the path delay at any time t, $p(t)$, *compactly* using the fresh path delay p^0, the device-level NBTI degradation $f[c, E_a, T(0), ..., T(t)]$, and the path-dependent degradation rate p^*. For a general combinational circuit with multiple paths, the circuit delay $D_{\max}$ is the maximum among all path delays:

$$\begin{aligned} D_{\max}(t) &= \max(p_k(t), \forall k) \\ &= \max(p_k^0 + p_k^* f[c, E_a, T(0), ..., T(t)], \forall k) \end{aligned} \tag{5.15}$$

where *k* is the index variable of the paths. In (5.15), the time dependence of D_{max} is through *f*(.) only. Therefore, we can treat D_{max} as a function of *f*:

$$D_{\max}(f) = \max(p_k^0 + p_k^* f, \forall k) \tag{5.16}$$

In (5.16), the function D_{max} (*f*) describes the relationship between the device degradation function *f* and the degraded circuit delay. It can be piece-wisely linearized for the sake of model compactness, as illustrated in Figure 5.3. In other words, if we re-write (5.16) as:

$$D_{\max}(t) = D_{\max}^0 + D_{\max}^* f \tag{5.17}$$

we need multiple pairs of $(D^0_{\max}, D^*_{\max})$ values for the macromodel. This effect is due to critical-path re-ranking. The number of these pairs is upper-bounded by κ/γ, where κ is the maximum percentage increase of $D_{max}(t)$ throughout the lifetime, and γ is the

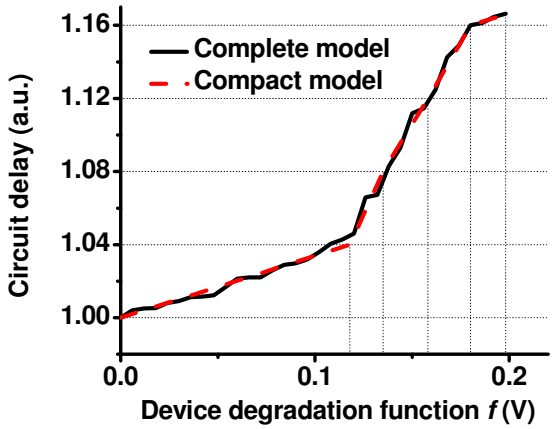


Figure 5.3: Illustration that the compact model of the circuit delay with respect to the device degradation function *f* can be obtained by piece-wise linearizing the complete model

percentage precision requirement. Determining the function $D_{max}(f)$ is enabled through a one-time precharacterization procedure using gate level static timing analysis (STA) tool that contains NBTI-aware gate models. STA eliminates the need of enumerating all paths, while achieving the same result as in (5.16). The STA tool we developed is similar to the one in [78]. It assumes that the signal probabilities of the primary inputs of a circuit are fixed. The signal probabilities of the primary inputs are then propagated to every circuit node. ΔV_{th} for each PMOS can be computed from *f* and the signal probabilities using (5.11). From ΔV_{th} of each PMOS, the increase of each gate delay can be found by (5.10). Finally, timing analysis is performed using the updated gate delays. For each combinational logic block, the tool generates delay as a function of *f*, similar to Figure 5.3. A lookup table can be constructed to hold the extracted $(D^0_{\max}, D^*_{\max})$ values, indexed by the interval of *f*. The lookup table can be retrieved at run-time.

The macromodel (5.17) enables online monitoring of the circuit delay degradation through tracking the device degradation function *f*. Tracking *f* is performed by computing

the accumulated equivalent stress time t_{eq} with the monitored temperature, Eq. (5.5), and then using t_{eq} to find $\Delta V_{th\text{-}DC}$, Eq. (5.3).

5.2 ONLINE MODEL CALIBRATION

The device degradation model of Eqs. (5.3), (5.5) and (5.6) contains two process-dependent model parameters, c and E_a. A straightforward approach is to use the nominal values of c and E_a to compute f. However, the actual values of c and E_a can be significantly different from their nominal values due to process variation. In order to account for this inaccuracy, we perform real-time updating of the model parameters by measuring the frequencies of the ring oscillators under normal operating conditions. The ring oscillators are placed close to the circuits being monitored to capture the systematic intra-die process variation, in addition to inter-die process variation. One of the most important sources of variations we need to consider is the variation of initial threshold voltage V_{th0} as a result of L_{gate} variation [80]. Intra-die L_{gate} variation shows significant systematic feature on a chip. At the same time,

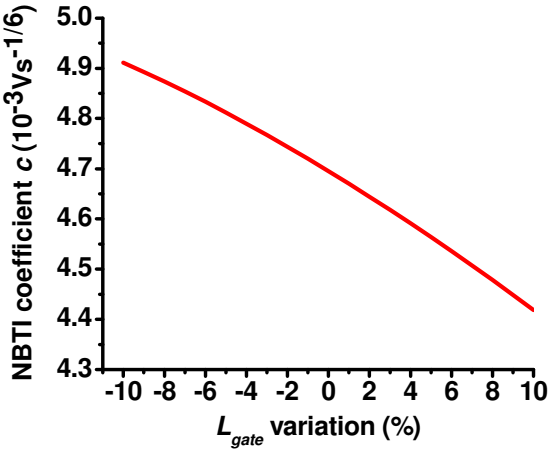


Figure 5.4: NBTI model parameter c is affected by L_{gate} variation.

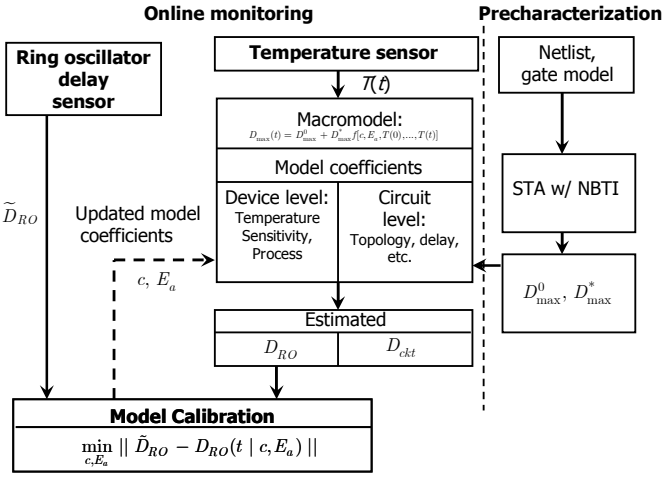


Figure 5.5: The flow chart of the online reliability monitoring scheme

$$\Delta V_{th0} \sim \exp(a\Delta L_{gate}) \tag{5.18}$$

where a is a constant. Because the model parameter c is a function of V_{th0}, Eq. (5.1), the variation of L_{gate} affects c, Figure 5.4.

The calibration process works as follows. The actual degraded path delay $\widetilde{D}_{RO}$ of a ring oscillator can be measured directly. On the other hand, an estimated path delay D_{RO} can be obtained from the macromodel for given c and E_a. We perform minimum mean square error (MMSE) fitting [81] between the time series of $\widetilde{D}_{RO}$ and D_{RO} to find the optimal values of c and E_a. MMSE also reduces the random noise in measurement. The updated values of c and E_a are then fed back to the macromodel to estimate the circuit delay of the function blocks. Figure 5.5 shows the flow chart of the online reliability monitoring scheme.

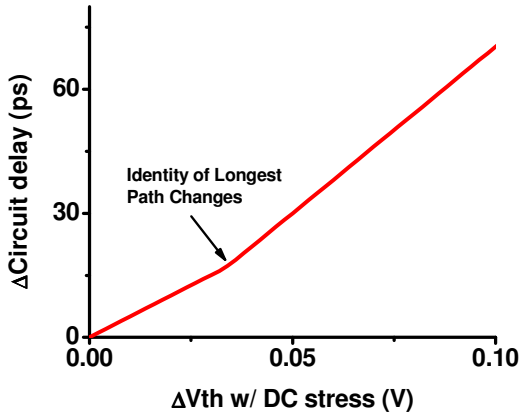


Figure 5.6: ΔCircuit delay (ΔD_{max}) as a function of device degradation functions f for benchmark circuit c1355.

5.3 EXPERIMENTAL RESULTS

We consider a ring oscillator and combinational logic block sitting in close proximity on a chip so that they share the same intra-die systematic process variation (ΔL_{gate}). However, they may experience different local temperatures. We demonstrate the accuracy of the macromodel by comparing the circuit delay degradation estimated by the macromodel to that of *direct* NBTI-aware STA. The latter is used as a proxy of the actual circuit delay degradation. Random temperature variation is assumed in all cases.

We first perform precharacterization on the circuits to extract the macromodel parameters (D^0_{max}, D^*_{max}) as described in Section 5.1. The precharacterization starts with generating the function $\Delta D_{max}(f)$ or $D_{max}(f)$, followed by piece-wise linearizing the function to obtain the parameters for each piece-wise linearization region. Figure 5.6 shows ΔD_{max} versus f for ISCAS'85 benchmark circuit c1355. Our precharacterization on the ISCAS'85 benchmark demonstrates that the macromodel can be compact. In Figure 5.7, we show that the most complex macromodel requires 7 sets of (D^0_{max}, D^*_{max}) values. This corresponds to 21 parameters in total, since each set consists of two

parameters and an additional parameter is required for each set to denote the end of the piece-wise linearization region.

We then demonstrate the necessity of using multi-set of macromodel parameters (D^0_{max}, D^*_{max}). In Figure 5.8, the estimated ΔD_{max} from the macromodel is compared with that generated by NBTI-aware STA. It can be seen that using a single set of macromodel

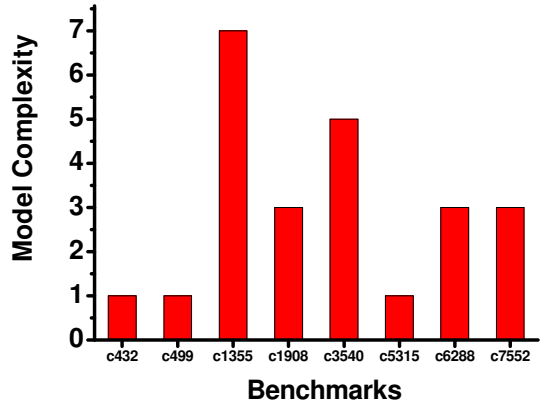


Figure 5.7: The complexity of the macromodel in terms of the number of parameter sets required for ISCAS'85 benchmark circuits.

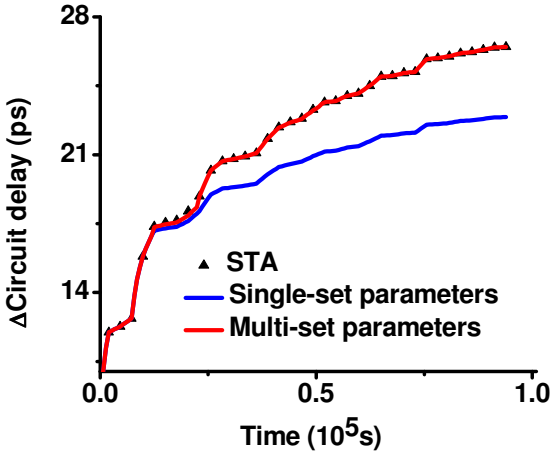


Figure 5.8: The necessity of using multi-set of macromodel parameters. The results are for benchmark circuit c1355.

parameters may estimate the circuit delay well at the beginning, but the error quickly grows with time due to re-ranking of the paths. The error can be as large as 15%. On the other hand, using 7 sets of macromodel parameters generates good estimations throughout the time.

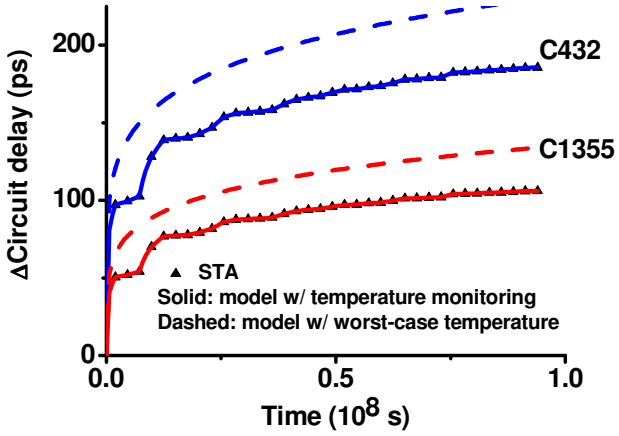


Figure 5.9: The effect of temperature monitoring on estimation of circuit delay degradation.

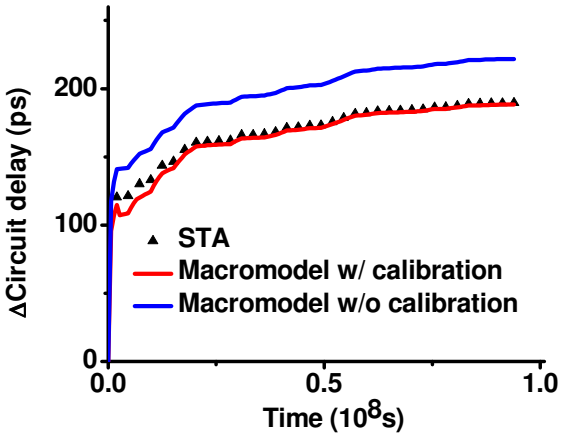


Figure 5.10: Calibrating model parameters to correct estimation errors due to process variation.

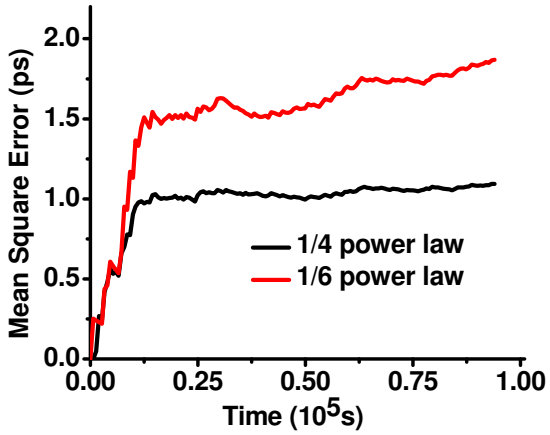


Figure 5.11: Model identification through online calibration of model coefficients. The MSE between the macromodel estimations and the STA result is an indicator of the model fitness for model selection.

The effect of temperature monitoring on estimation of ΔD_{max} is shown in Figure 5.9. To isolate this effect, process variation is assumed to be not present and the device level parameters are known exactly. The macromodel with monitored temperature estimates ΔD_{max} that closely follows the STA result. On the other hand, the estimation error using the macromodel but assuming worst-case temperature can be as large as 20%.

We demonstrate the effect of calibrating the model parameters in Figure 5.10. The device parameters c and E_a have nominal values in the macromodel, but assume non-nominal values in the STA due to process variation. Using the nominal model parameters may result in error as large as 18%. Online calibration using direct measurements of the ring oscillators can correct this estimation error.

Finally, we demonstrate that the proposed framework is capable of model identification. It is generally known that there are many competing device-level NBTI models existing [51][53][76]. Two popular ones are the 1/4 and 1/6 power laws [51][53]:

$$\Delta V_{th}(t) = \begin{cases} c_{1/4} t^{1/4} & (1/4 \text{ law}) \\ c_{1/6} t^{1/6} & (1/6 \text{ law}) \end{cases} \tag{5.19}$$

In the case of temperature variation, t in (5.19) is replaced with t_{eq}, Eq. (5.5). Without knowing which one is the actual model, we develop the macromodels for both device-level models. In the online calibration stage, MMSE method is used to fit the estimated results of the macromodels to the measured data for the ring oscillator. The mean square error (MSE) of the fitting can be used as a measure of model fitness. In this experiment, we assume that the "actual" device-level model is the 1/4 power law and the fitting is performed between macromodel estimations and the STA result. Figure 5.11 shows the MSE for macromodels based on both the 1/4 and 1/6 power laws. It can be seen that the MSE for the 1/4 power law model is smaller and converges. Therefore, the 1/4 power law model is selected for estimating D_{max} of the circuit.

5.4 SUMMARY

In this chapter, we developed a compact macromodel for online monitoring of NBTI-induced circuit degradation. The model parameters that depend on circuit topology are precharacterized by an NBTI-aware STA tool. Real time temperature measured by temperature sensors is used in the macromodel to estimate the circuit degradation. Process variation and inaccuracy of the device level model can be accounted for by calibrating the model coefficients using on-chip ring oscillators. Compared to the existing monitoring techniques that only measure the degradation of a simple structure (a PMOS or a ring oscillator), our approach is more accurate, because we take into account the effect of circuit topology. Compared to other types of existing monitoring techniques, which directly measure the circuit delay, our approach is more cost-effective, because it does not need a large number of sensors.

Chapter 6: Conclusions

A bottleneck of the performance-driven scaling of transistor feature size is the device reliability. As technology scales down, error rates due to both soft errors and permanent errors increase significantly. Traditional device-level and worst-case based reliability analysis may result in over-conservatism, which sacrifices the performance gained by downscaling. The objective of this dissertation is to develop efficient algorithms to take into account the effect of the error masking mechanisms of the circuit and the actual operating condition, so that the designers can more accurately evaluate the actual impact of the failure mechanisms. This dissertation has investigated: 1) fast analysis of soft error susceptibility for cell-based designs; 2) analytical modeling of SRAM dynamic stability; 3) modeling of NBTI-induced PMOS degradation under arbitrary dynamic temperature variation, and 4) online circuit reliability monitoring.

Soft errors in combinational logic circuit are becoming a serious concern for commercial electronics. Several error masking mechanisms can make the actual soft error rates significantly lower than the raw particle strike rate. Different input vectors of the circuit can result in different error masking capability. This dissertation proposes an algorithm that explicitly enumerates the input vectors using binary decision diagram and precharacterizes the cell library for error generation and propagation. Compared to the traditional simulation-based techniques, the proposed algorithm is computationally fast, while still maintaining estimation accuracy.

Verifying the stability of SRAM-based memory arrays is an essential design task. Traditional static noise margin is overly conservative for soft errors. In this dissertation, we propose a criterion of dynamic noise margin for SRAM cells under the impact of single event upsets. In addition, we obtain a close form of the dynamic noise margin

under the approximation of piece-wise linearization. The dynamic noise margin not only considers the amplitude of the transient noise, but also takes into account the temporal properties of the transient noise. We use nonlinear system theory to show that for the state stored in the SRAM cell to flip, the noise needs to exceed specific threshold amplitude, which is shown to be the static noise margin, and be sustained longer than a minimum critical period.

While soft errors cause only temporal failures of the circuit, permanent errors such as NBTI can result in permanent damage to the circuit. NBTI can cause timing errors for combinational logic circuits. NBTI is highly sensitive to operating temperature, and modern chips exhibit significant temperature variation both spatially and temporally. In this work, we have exploited recent experimental findings and built an NBTI model that can handle arbitrary dynamic temperature variation. The proposed model is consistent with the reaction-diffusion model and is validated by simulation.

The online circuit reliability monitoring framework utilizes the NBTI model under dynamic temperature variation. It is based on a hybrid network of on-chip sensors, consisting of temperature sensors and ring oscillators. The key feature of our work, in contrast to the traditional tracking techniques that rely solely on direct measurement of the increase of threshold voltage or circuit delay, is an explicit macromodel which maps operating temperature to the circuit degradation. The macromodel allows for cost-effective tracking reliability using temperature sensors. The macromodel is precharacterized by an STA-like NBTI-aware analysis tool. The incorporation of ring oscillators allows for online model calibration.

In the future technology generations, reliability will remain a major problem. Failure mechanisms that have been traditionally considered minor may emerge to be primary lifetime limitors. Thus, analysis of the impact of these failure mechanisms will

be indispensable for circuit designers. Increasingly, such analysis will have to be combined with models at circuit and even architectural levels to account for the effect of actual workload.

Bibliography

[1] P. Shivakumar, *et al*, "Modeling the Effect of Technology Trends on the Soft Error Rate of Combinational Logic," *Proc. DSTN*, pp. 389-398, 2002.

[2] D. J. Frank *et al*, "Device Scaling Limits of Si MOSFETs and Their Application Dependencies", *Proc. IEEE*, vol. 89, pp. 259-288, Mar, 2001.

[3] J. P. Spratt et al, "Effectiveness of IC shielded packages against space radiation," *IEEE Trans. Nucl. Sci*., vol. 44, pp. 2018 - 2025, Dec, 1997.

[4] J. Srinivasan, S. V. Adve, P. Bose and J. A. Rivers, "The case for lifetime reliability-aware microprocessors," *Proc. Intl. Syms. Comp. Arch*., 2004, pp. 276-287.

[5] T. J. O'Gorman et al, "Field testing for cosmic ray soft errors in semiconductor memories," *IBM J. Res. Dev*., vol. 40, pp. 41–49, Jan. 1996.

[6] T. C. May and M. H. Woods, "Alpha-particle-induced soft error in dynamic memories," *IEEE Trans. Elec. Dev.*, vol. 26, pp. 2–9, Jan. 1979.

[7] H. H. Chen and J. S. Neely, "Interconnect and circuit modeling techniques for full-chip power supply noise analysis," *IEEE Trans. Components, Packaging, and Manufacturing Technology*-Part B, 1998, pp. 209-215.

[8] R. Gharpurey and R. G. Meyer, "Modeling and analysis of substrate coupling in IC's," *IEEE Custom-Integrated Circuit Conf*., 1995, pp. 125–128.

[9] H.-R. Cha and O.-K. Kwon, "An analytical model of simultaneous switching noise in CMOS systems," *IEEE Trans. Advanced Packaging*, 2000, pp. 62-68.

[10] H. Su, S. S. Sapatnekar, S. R. Nassif, "Optimal decoupling capacitor sizing and placement for standard-cell layout designs," *IEEE Trans. Computer Aided Design*, vol. 22, pp. 428- 436, Apr, 2003.

[11] P. E. Dodd, *et al*, "Basic mechanisms and modeling of single-event upset in digital microelectronics," *IEEE Trans. Nucl. Sci.* Vol. 50, pp. 583–602, 2003.

[12] H. Cha, *et al*, "A Gate-Level Simulation Environment for Alpha-Particle-Induced," *IEEE Trans. Computers*, Vol. 45, pp. 1248-1256, 1996.

[13] D. Bossen, "CMOS soft errors and server design," *IRPS* tutorial, 2002.

[14] R. Degraeve, “New insights in the relation between electron trap generation and the statistical properties of oxide breakdown,” *IEEE Trans. Elec. Devices*, vol. 45, pp. 904-910, Apr, 1998.

[15] M. A. Alam, B. E. Weir and P. J. Silverman, “A study of soft and hard breakdown—part II: principles of area, thickness, and voltage scaling,” *IEEE Trans. Elec. Devices*, vol. 49, pp. 239-246, Feb, 2002.

[16] M. A. Alam, R. K. Smith, B. E. Weir and P. J. Silverman , “Uncorrelated breakdown of integrated circuits,” *Nature*, vol. 420, pp. 378, Nov, 2002.

[17] K. N. Tu, “Recent advances on electromigration in very-large-scale-integration of interconnects,” *J. Appl. Phys*., vol. 94, pp. 5451-5473, Nov, 2003.

[18] J. Tao, N. W. Cheung, C. Hu, “Electromigration characteristics of copper interconnects,” *IEEE Electr. Device Lett.*, vol. 14, pp. 249-251, May, 1993.

[19] C. Hu et al, “Hot-electron induced MOSFET degradation-model, monitor and improvement,” *IEEE Trans. Electr. Devices*, vol. ED-32, pp. 375–385, 1985.

[20] G. Chen et al, “Dynamic NBTI of PMOS transistors and its impact on device lifetime,” *Proc. IRPS*, 2002, pp. 196–202.

[21] K. Mohanram, *et al*, “Cost-Effective Approach for Reducing Soft Error Failure Rate in Logic Circuits,” *Proc. ITC*, pp. 893-901, 2003.

[22] M. Zhang, *et al*, “A Soft Error Rate Analysis (SERA) Methodology,” *Proc. ICCAD*, pp.111-118, 2004.

[23] Q. Zhou, *et a*l, "Transistor Sizing for Radiation Hardening," *Proc. IRPS,* pp. 310-315, 2004.

[24] Q. Zhou, *et al*, "Cost-Effective Radiation Hardening Technique for Combinational Logic," *Proc. ICCAD,* pp. 100-106, 2004.

[25] N. Miskov-Zivanov and D. Marculescu, “MARS-C: modeling and reduction of soft errors in combinational circuits,” *Proc. DAC*, 2006, pp. 767 – 772

[26] P. C. Murley, *et al*, “Soft-error Monte Carlo modeling program, SEMM,” *IBM J. Res. Develop*., Vol. 40, pp. 109-118, 1996.

[27] P. Hazucha, *et al*, “Impact of CMOS Technology Scaling on the Atmospheric Neutron Soft Error Rate,” *IEEE Trans. Nucl. Sc*i., Vol. 47, pp. 2586–2594, 2000.

[28] R. Bryant, “Graph-based algorithms for Boolean function manipulation,” *IEEE Trans. Computers*., Vol. 35, pp.677-691, 1986.

[29] F. Najm, “Transition density, a stochastic measure of activity in digital circuits,” *Proc. DAC*, pp.644-649, 1991.

[30] BPTM, http://www-device.eecs.berkeley.edu/~ptm/.

[31] K. Mohanram, “Closed-form simulation and robustness models for SEU tolerant design,” *Proc. VLSI Test Symposium*, pp. 327–333, 2005.

[32] K. Mohanram, "Simulation of transients caused by single-event upsets in combinational logic," *Proc. ITC,* 2005.

[33] J. Jain, *et al*. “Functional partitioning for verification and related problems,” *Brown/MIT VLSI Conference*, 1992.

[34] D. Sahoo*, et al,* “A Partitioning Methodology for BDD-based Verification,” *Proc. FMCAD,* 2004.

[35] Dharchoudhury, *et al*, “A switch-level algorithm for simulation of transients in combinational logic,” *Proc. Int. Fault-Tolerant Computing Symp.*, pp. 207-216, 1995

[36] G. C. Messenger, “Collection of charge on junction nodes from ion tracks,” *IEEE Trans. Nucl. Sci.*, Vol. 29, pp.2024-2031, 1982

[37] G. R. Srinivasan, *et al*, “Accurate predictive modeling of soft error rate due to cosmic rays and chip alpha radiation,” *Proc. Intl. Reliability Phys. Symp*., pp. 12-16, 1994.

[38] E. Seevinck, F. J. List and J. Lohstroh, "Static-noise margin analysis of MOS SRAM cells," *JSSC*, 1987, pp. 748-754.

[39] J. Lohstroh, E. Seevinck and J. D. Groot, “Worst-case static noise margin criteria for logic circuits and their mathematical equivalence,” *JSSC*, 1983, pp. 803-807.

[40] D. C. Pham, *et. al*., “Overview of the architecture, circuit design, and physical implementation of a first-generation cell processor,” *JSSC*, 2006, pp. 179-196.

[41] J. S. Fu, C. L. Axness and H. T. Weaver, “Two-dimensional simulation of single event induced bipolar current in CMOS structures,” *IEEE Trans. Nucl. Sci*., 1984, pp. 1155–1159.

[42] K. Mayaram, J.-H. Chern and P. Yang, “Algorithms for transient three-dimensional mixed-level circuit and device simulation,” *IEEE Trans. Computer-Aided Design*, 1993, pp. 1726–1733.

[43] A. Dharchoudhury, et al., "Fast timing simulation of transient faults in digital circuits," *ICCAD*, 1994, pp. 719-726.

[44] M. Omana, et al., "A model for transient fault propagation in combinatorial logic," *Intl. On-line Testing Symp*., 2003, pp. 111-115.

[45] R. Heald and P. Wang, "Variability in sub-100nm SRAM designs," *ICCAD*, 2004, pp. 347-352.

[46] M. Horowitz, "Timing models for MOS circuits," Ph. D. Dissertation, Stanford University, 1984.

[47] Z. Vukic, *et. al*., Nonlinear Control Systems, Marcel Dekker Inc., 2003.

[48] N. Hedenstierna and K. O. Jeppson, " CMOS circuit speed and buffer optimization, " *IEEE Trans. Computer-Aided Design*, 1987, pp. 270-281.

[49] T. Sakurai and A. R. Newton, "Alpha-power law MOSFET model and its applications to CMOS inverter," *JSSC*, 1990, pp. 584-594.

[50] N. Kimizuka *et al*., "The impact of bias temperature instability for direct-tunneling ultra-thin gate oxide on MOSFET scaling," *VLSI Symp. on Tech*., 1999, pp. 73-74.

[51] S. Bhardwaj "Predictive modeling of the NBTI effect for reliable design," CICC, 2006, pp. 189-192.

[52] M. Agarwal *et al*., "Circuit Failure Prediction and Its Application to Transistor Aging", *VLSI Test Symp*., 2007, pp. 277 – 286.

[53] M. A. Alam and S. Mahapatra, "A comprehensive model of PMOS NBTI degradation," *Microelectro. Reliability*, 45, pp. 71-81, 2005.

[54] K. O. Jeppson and C. M. Svensson, "Negative bias stress of MOS devices at high electric fields and degradation of MNOS," *J. of Appl. Phys*., 48, pp. 2004-2014, May 1977.

[55] D. K. Schroder, "Negative bias temperature instability: What do we understand?"*Microelectronics Reliability* 47 (2007), pp. 841-852.

[56] S. Borkar, "Microarchitecture and Design Challenges for Gigascale Integration," *Intl. Symp. Microarchitecture*., 2004, pp. 3-3.

[57] E. Karl *et al*., "Reliability modeling and management in dynamic microprocessor-based systems," *DAC*, 2006, pp. 1057-1060.

[58] S. Chakravarthi *et al*., "A Comprehensive Framework For Predictive Modeling of Negative Bias Temperature Instability," *IRPS*, 2004, pp. 273-282.

[59] A. Krishnan *et al*., "Negative bias temperature instability mechanism: The role of molecular hydrogen," *Appl. Phys. Lett*. 88, 153518, 2006.

[60] S. Kumar *et al*., "Impact of NBTI on SRAM read stability and design for reliability," *ISQED*, 2006, pp. 210 – 218.

[61] S. Kumar *et al*., "An analytical model for negative bias temperature instability," *ICCAD*, 2006, pp. 493-496.

[62] W. Wang *et al*., "The Impact of NBTI on the Performance of Combinational and Sequential Circuits," *DAC*, 2007, pp. 364-369.

[63] H. Luo *et al*., "Modeling of PMOS NBTI effect considering temperature variation," *ISQED*, 2007, pp. 139-144.

[64] A. E. Islam, H. Kufluoglu, D. Varghese and M. A. Alam, "Critical analysis of short-term negative bias temperature instability measurements: explaining the effect of time-zero delay for on-the-fly measurements," *Appl. Phys. Lett*. 90, 083505, Feb. 2007.

[65] A. Krishnan *et al*., "Material dependence of hydrogen diffusion: implications for NBTI degradation," *IEDM*, 2005, pp.688-691.

[66] H. Reisinger *et al*., "Analysis of NBTI degradation- and recovery- behavior based on ultra fast VT-measurements," *IRPS*, 2006, pp. 448-453.

[67] W. Huang *et al*., "An improved block-based thermal model in HotSpot 4.0 with granularity considerations," *WDDD*, 2007.

[68] D. Brooks and M. Martonosi, "Dynamic Thermal Management for High-Performance Microprocessors," *HPCA*, 2001, pp.171-182.

[69] B. Sopori *et al*., "Silicon device processing in H-ambients: H-diffusion mechanisms and influence on electronic properties," *J. Electro. Materials*, pp. 1616-1627, Dec. 2001.

[70] R. C. Blish, II, "Thermal cycling and thermal shock failure rate modeling," *IRPS*., 1997, pp. 110-117.

[71] S. Borkar., "Designing reliable systems from unreliable components: the challenges of transistor variability and degradation," *IEEE Micro*, v. 25, n. 6, pp.10-16, Nov. 2005.

[72] J. Blome *et al*., "Self-calibrating online wearout detection," *Micro*, 2007, pp. 109-122.

[73] Y. Li, S. Makar and S. Mitra, "CASP: concurrent autonomous chip self-test using stored test patterns," presented in *SELSE workshop*, Mar. 2008, Austin, TX.

[74] M. Royd *et al*., "System power management support in the IBM POWER6 microprocessor," *IBM J. Res. & Dev.*, v. 51, n. 6, pp. 733-746, Nov. 2007.

[75] S. Kumar, C. Kim, and S. Sapatnekar, "NBTI-Aware synthesis of digital circuits," *DAC*, 2007, pp. 370-375.

[76] V. Huard et al., "New characterization and modeling approach for NBTI degradation from transistor to product level," *IEDM*, 2007, pp. 797-800.

[77] B. Zhang and M. Orshansky, "Modeling of NBTI-induced PMOS degradation under arbitrary dynamic temperature variation," *ISQED*, 2008, pp. 774-779.

[78] B. Paul *et al*., "Impact of NBTI on the temporal performance degradation of digital circuits," *IEEE Electron Device Lett*., v. 26, n. 8, pp. 560-562, Aug. 2005.

[79] E. Grochowski, D. Ayers, and V. Tiwari, "Microarchitectural simulation and control of di/dt-induced power supply voltage variation," *HPCA*, 2002, pp. 2-12.

[80] A. Srivastava *et al*., "Modeling and analysis of leakage power considering within-Die process variations," *ISLPED*, 2002, pp. 64-67.

[81] D. Montgomery and G. Runger, *Applied Statistics an d probability for engineers*, 2nd ed., John Wiley & Sons, 1999.

Vita

Bin Zhang received his B. S. in Applied Physics from Shanghai Jiao Tong University, China, in 1994. After obtaining his M. S. in Physics from Peking University, China, in 1997. He went on to study in the United States, where he received his M. S. in Physics from Texas A& M University in 1999 and M. S. in Electrical & Computer Engineering from University of Texas at Austin in 2001. He worked as an intern in Broadwing Communications Inc. and Cisco Systems, and a software test engineer in Austin Test Inc., all in Austin, TX, during the period between 2000 and 2002. He entered the PhD program in the Department of Electrical and Computer Engineering, the University of Texas at Austin in 2003. His paper, co-authored with Wei-shen Wang and Michael Orshansky, "FASER: Fast analysis of soft error susceptibility for cell-based designs", received the Best Paper Award from *International Symposium on Quality Electronic Design* in 2006. His another paper, co-authored with Michael Orshansky, "Modeling of NBTI-induced PMOS degradation under arbitrary dynamic temperature variation", received the IBM Research Award at the 2008 Graduate and Industry Networking Conference (GAIN 2008), University of Texas at Austin, for "the best paper & presentation at the GAIN 2008 conference in materials science and Nano, Micro, Bio & MEMS engineering".

Permanent address: N1-5-1 Nanxiaoqu, Suining, Sichuan, 629000, China

This dissertation was typed by the author.