Scalable Methods for Reliability Analysis in Digital Circuits using Physics-Based Device-Level Models

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Jianxin Fang

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Sachin S. Sapatnekar

October, 2012

Acknowledgements

First and foremost, I want to express my ultimate appreciation and gratitude to my academic advisor, Prof. Sachin Sapatnekar, for his great guidance and support throughout the past five years. He led me in to the field of EDA, directed me through tough challenges, and gave me continuous encouragement and support during the most difficult periods of my research. His superior vision, broad yet deep knowledge, and dedication to details and perfection, deeply impressed me and changed my way of thinking and working. This will be a lifetime treasure for me.

I am grateful to Prof. Chris Kim, Prof. Antonia Zhai, and Prof. Marc Riedel, for serving on my PhD degree committee, reviewing my thesis, and providing precious comments and suggestions.

I would also like to thank the Department of Electrical and Computer Engineering and the VEDA Lab of University of Minnesota that have provided an excellent environment for my research. I spent many enjoyable hours with my colleagues discussing work and life, and received plenty of valuable ideas and help from them.

Without the unwavering love and support from my family, this would have been a hard and lonely journey. I would like to thank my parents and my wonderful wife, who always stands beside me and steadily supports me.

Dedication

To my parents, my wife, and my lovely daughter.

Abstract

As technology has scaled aggressively, device reliability issues have become a growing concern in digital CMOS very large scale integrated (VLSI) circuits. There are three major effects that result in degradation of device reliability over time, namely, time-dependent dielectric breakdown (TDDB), bias-temperature instability (BTI), and hot carrier (HC) effects. Over the past several years, considerable success has been achieved at the level of individual devices to develop new models that accurately reconcile the empirical behavior of a device with the physics of reliability failure. However, there is a tremendous gulf between these achievements at the device level and the more primitive models that are actually used by circuit designers to drive the analysis and optimization of large systems. By and large, the latter models are decades old and fail to capture the intricacies of the major advances that have been made in understanding the physics of failure; hence, they cannot provide satisfactory accuracy. The few approaches that can be easily extended to handle new device models are primarily based on simulation at the transistor level, and are prohibitively computational for large circuits.

This thesis addresses the circuit-level analysis of these reliability issues from a new perspective. The overall goal of this body of work is to attempt to bridge the gap between device-level physics-based models and circuit analysis and optimization for digital logic circuits. This is achieved by assimilating updated device-level models into these approaches by developing appropriate algorithms and methodologies that admit scalability, resulting in the ability to handle large circuits. A common thread that flows through many of the analysis approaches involves performing accurate and computationally feasible cell-level modeling and characterization, once for each device technology, and then developing probabilistic techniques to utilize the properties of these characterized libraries to perform accurate analysis at the circuit level. Based on this philosophy, it is demonstrated that the proposed approaches for circuit reliability analysis can achieve accuracy, while simultaneously being scalable to handle large problem instances. The remainder of the abstract presents a list of specific contributions to addressing individual mechanisms at the circuit level.

Gate oxide TDDB is an effect that can result in circuit failure as devices carry unwanted and large amounts of current through the gate due to oxide breakdown. Realistically, this results in catastrophic failures in logic circuits, and a useful metric for circuit reliability under TDDB is the distribution of the failure probability. The first part of this thesis develops an analytic model to compute this failure probability, and differs from previous area-scaling based approaches that assumed that any device failure results in circuit failure. On the contrary, it is demonstrated that the location and circuit environment of a TDDB failure is critical in determining whether a circuit fails or not. Indeed, it is shown that a large number of device failures do not result in circuit failure due to the inherent resilience of logic circuits. The analysis begins by addressing the nominal case and extends this to analyze the effects of gate oxide TDDB in the more general case where process variations are taken into account. The result shows derivations that demonstrate that the circuit failure probability is a Weibull function of time in the nominal case, while has a lognormal distribution and at a specified time instant under process variations. This is then incorporated into a method that performs gate sizing to increase the robustness of a circuit to TDDB effect.

Unlike gate oxide TDDB, which results in catastrophic failures, both BTI and HC effects result in temporal increases in the transistor threshold voltages, causing a circuit to degrade over time, and eventually resulting in parametric failures as the circuit violates its timing specifications. Traditional analyses of the HC effects are based on the so-called lucky electron model (LEM), and all known circuit-level analysis tools build upon this model. The LEM predicts that as device geometries and supply voltages reduce to the level of today's technology nodes, the HC effects should disappear; however, this has clearly not been borne out by empirical observations on small-geometry devices. An alternative energy-based formulation to explain the HC effects has emerged from the device community: this thesis uses this formulation to develop a scalable methodology for hot carrier analysis at the circuit level. The approach is built upon an efficient one-time library characterization to determine the age gain associated with any transition at the input of a gate in the cell library. This information is then utilized for circuit-level analysis using a probabilistic method that captures the impact of HC effects over time, while incorporating the effect of process variations. This is combined with existing models for BTI, and simulation results show the combined impact of both BTI and HC effects on circuit delay degradation over time.

In the last year or two, the accepted models for BTI have also gone through a remarkable shift, and this is addressed in the last part of the thesis. The traditional approach to analyzing BTI, also used in earlier parts of this thesis, was based on the reaction-diffusion (R-D) model, but lately, the charge trapping (CT) model has gained a great deal of traction since it is capable of explaining some effects that R-D cannot; at the same time, there are some effects, notably the level of recovery, that are better explained by the R-D model. Device-level research has proposed that a combination of the two models can successfully explain BTI; however, most work on BTI has been carried out under the R-D model. One of the chief properties of the CT model is the high level of susceptibility of CT-based mechanisms to process variations: for example, it was shown that CT models can result in alarming variations of several orders of magnitude in device lifetime for small-geometry transistors. This work therefore develops a novel approach for BTI analysis that incorporates effect of the combined R-D and CT model, including variability effects, and determines whether the alarming level of variations at the device level are manifested in large logic circuits or not. The analysis techniques are embedded into a novel framework that uses library characterization and temporal statistical static timing analysis (T-SSTA) to capture process variations and variability correlations due to spatial or path correlations.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Under aggressive technology scaling, device reliability issues have become a growing concern in digital very-large-scale integrated (VLSI) circuits. As CMOS devices age, they are predominantly affected by three reliability mechanisms that degrade their performance:

- Time-dependent dielectric breakdown (TDDB) of gate oxides
- Bias-temperature instability (BTI)
- Hot carrier (HC) effects

The results of these mechanisms is that the circuit may fail *catastrophically* in being unable to achieve correct logic functionality, or *parametrically*, in being able to achieve correct logic functionality but not at the correct specifications (e.g., the timing specifications may be violated). Of the above effects, TDDB results in catastrophic failures in the gate oxide, potentially leading to catastrophic circuit failures in the behavior of the circuit. On the other hand, BTI and HC effects result in a gradual degradation in the transistor threshold voltage or mobility, causing the circuit to slow down over time, and eventually failing to meet its timing specifications (however, it may continue to operate correctly under a slower clock).

A great deal of work has been carried out at the device level to build accurate models. In fact, this area has been a hive of activity at the device level, with numerous innovative works being published within the last decade or so, explaining the physics of degradation and successfully matching theoretical advances with experimental measurements.

However, large systems can contain many billions of transistors, and considerable effort must be expended in taking these device-level results and performing system-level analyses that can predict the impact of aging effects at higher levels of abstraction. This is a formidable task that necessitates efforts at the logic block level, RTL level, and system level. This thesis makes a first start by addressing such issues at the logic block level, and lays the framework for similar analyses at higher levels of abstraction.

A review of current-day circuit-level approaches for analyzing aging and degradation effects shows that most commercial tools, and many academic efforts, continue to use old, and sometimes obsolete, models. While the circuits and design automation community has shown some interest in expanding this field, most of the existing work has only addressed BTI effects, and this too has merely skimmed the surface of the problem.

This thesis addresses the circuit-level analysis of the above three reliability issues from a new perspective. The overall goal of this body of work is to attempt to bridge the wide chasm between the device-level physics-based models, where tremendous advances have been made in the recent past, and the much more primitive models that are widely used for circuit analysis and optimization in digital logic circuits today. This goal is achieved by assimilating updated device-level models into these approaches by developing appropriate algorithms and methodologies that admit scalability, resulting in the ability to handle large circuits.

Further, we thoroughly investigate the impact of process variations, including correlation effects, on the impact of these circuit aging mechanism. We demonstrate that all aging mechanisms are, in some way, significantly affected by process variations, and develop methods that organically capture the effects of variations during our analyses.

A common thread that flows through many of the analysis approaches involves performing accurate and computationally feasible cell-level modeling and characterization, once for each device technology, and then developing probabilistic techniques to utilize the properties of these characterized libraries to perform accurate analysis at the circuit level. Based on this philosophy, it is demonstrated that the proposed approaches for circuit reliability analysis can achieve accuracy, while simultaneously being scalable to handle large problem instances. In fact, all of the proposed approaches in this thesis have a complexity that is linear in the number of gates in the circuit.

It is important to also state, at the outset, what this work does *not* do. This is a design automation effort that develops scalable solutions by building up on past work. To maintain focus, we do not fabricate and test our ideas on silicon, instead leaning on models that are provided by many other successful groups that work in this area, operating at the device level or building small test structures such as ring oscillators. This is a standard model in design automation, having been used successfully many times in the past.

In the remainder of this chapter, we expound on individual failure mechanisms and highlight the contributions of this thesis.

1.1 Time-Dependent Dielectric Breakdown

Time-dependent dielectric breakdown refers to the phenomenon where defects are generated in the SiO_2 gate oxide under the continued stress of normal operation over a long period. Eventually, the oxide becomes conductive when a critical defect density is reached at a certain location in the oxide. With device scaling, electric fields across the gate oxide have increased as supply voltages have scaled down more slowly than the oxide thickness, and transistors have become more susceptible to oxide breakdown.

At the device level, the mechanisms and modeling of oxide breakdown have been throughly studied, and various empirical or analytical models have been proposed for this phenomenon [4]. The time-to-breakdown characteristic for a MOS transistor is typically modeled as a Weibull random variable [5].

The effect of a breakdown is to provide a path for current to flow from the gate to the channel. The terms *hard breakdown* (HBD) and *soft breakdown* (SBD) are used to describe the severity of oxide breakdown occurrences. An HBD is a low-resistance breakdown that can cause significant current to flow through the gate, while an SBD has a higher resistance, and lower breakdown current through the gate [4]. Catastrophic functional failures, which are the focus of this work, can only be caused by HBDs (although, as we will show, not every HBD causes a functional failure).

At the circuit level, the traditional failure prediction method for large circuits has two weaknesses. First, every known prior method uses area-scaling, extrapolated from single-device characterization [4]. The idea is based on the weakest-link assumption, that the failure of any individual device will cause the failure of the whole chip. In this work, we show that this assumption are not always true since circuits have inherent resilience to some breakdown events, and some HBDs may not result in circuit failure. This implies that the traditional method is inaccurate. Second, many past works do not directly incorporate the effect of process variations, which significantly increase the probability of circuit failure under HBD. The methods that do incorporate process variations for circuit-level oxide reliability analysis [6, 7] do show reduced lifetimes, but are based on the simple notion of area-scaling, which is too pessimistic for circuit lifetime prediction. We demonstrate in our work that our method shows that lifetime predictions using existing approaches are excessively pessimistic by at least half an order of magnitude, as compared with methods that consider inherent circuit resilience.

Second, we explore the effects of process variations, and find that the predicted FP under nominal condition is significantly affected by variations. We extend the nominal case FP analysis to include the effect of process variations, and show that this still provide substantially better improvements in the predicted lifetime over the conventional area-scaling model. The circuit FP at a specified time instant is derived to have a lognormal distribution due to process variations.

We address this problem in Chapter 2 and present the following contributions. First, we develop a scalable method for analyzing the failure probability (FP) of large digital circuits, while realistically considering the circuit environment that leads to stress and oxide breakdown. To achieve this goal, at the *transistor level*, we introduce improved models for time-to-breakdown and post-breakdown behavior. At the *logic cell level*, we devise a procedure for performing precise FP analysis for standard cell based digital circuits, and present an effective library characterization scheme. At the *circuit level*, we derive a closed-form expression for the FP of large digital logic circuits considering the actual stress on devices and the probability of failure due to HBD. The cost of the analysis is linear in the number of gates in the circuit.

Based on the analytical result of circuit failure probability, we develop an optimization approach to mitigate the effect of gate oxide breakdown in Chapter 3. We formulate a problem that performs transistor sizing with the aim of increasing the time to circuit failure, while addressing conventional sizing goals such as power and delay. Experimental results show that circuit reliability can be improved by increasing the area, which runs counter to the prediction of the traditional area-scaling theory.

1.2 Hot Carrier Effect

Hot carrier effects in MOSFETs are caused by the acceleration of carriers (electrons or holes) under lateral electric fields in the channel, to the point where they gain high enough energy and momentum (and hence they are called *hot* carriers) to break the barriers of surrounding dielectric, such as the gate and sidewall oxides [8]. The presence of hot carriers triggers a series of physical processes that affects the device characteristics under normal circuit operation. These effects cumulatively build up over prolonged periods, causing the circuit to age with time, resulting in performance degradations that may eventually lead to circuit failure.

The rate of hot carrier generation increases with time t as $t^{1/2}$. Since the multiplicative constant for this proportionality is relatively small, in the short-term, HC effect is overshadowed by bias-temperature instability (BTI) effects, which increase as t^n, for $n \approx 0.1$–0.2, but with a larger constant multiplier. However, in the long term, the $t^{1/2}$ term dominates the t^n term, making HC effects particularly important for devices in the medium to long term, such as embedded/automotive applications and some computing applications. Figure 1.1 [1] shows that the impact of HC effects can be very substantial after long periods of operation.

Conventionally, HC effects were captured using the lucky electron model [9], which was valid in the period of high supply voltages. However, this model is inadequate in explaining HC effects in deeply-scaled CMOS with low supply voltages [10]. Recently, newer energy-driven theories [11–13] have been introduced to overcome the limitations of the lucky electron model, and to explain the mechanism of carriers-induced degradation for short-channel devices at low supply voltages. These theories have been experimentally validated on nanometer-scale technologies. The energy-driven framework includes the effects of electrons of various levels of energy, ranging from high-energy *channel hot carriers* (CHCs) to low-energy *channel cold carriers* (CCCs). Under this model, injection is not necessary for the device degradation, and carriers with enough energy

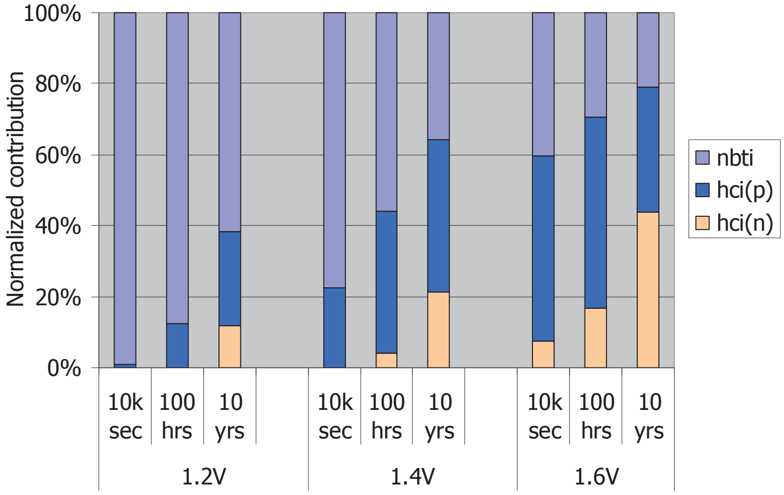

Figure 1.1: HCI and NBTI components for delay degradation of a ring oscillator as a function of stress time and stress voltage [1].

can affect the Si–SiO_2 interface directly. However, much of the published circuit-level work on HC effects is based on the lucky electron model, which is effectively obsolete.

Existing work on HC degradation analysis of digital circuits can be divided into to two categories. The first is based on device-level modeling/measurement tied to circuit-level analysis, including [14], commercial software such as Eldo using computationally-intensive simulations. While these methods are flexible enough to accept new models and mechanisms, they are not scalable for analyzing large circuits. Methods in the second category [15, 16] are based on a circuit-level perspective, using statistical information about device operation to estimate the circuit degradation. While these works are usually efficient and scalable to large digital circuits, they use over-simplistic models for device aging and cell characterization, and therefore cannot achieve the high accuracy provided by methods in the first category, especially for nanometer-scale technologies. Extending these methods to energy-driven models, including CHC and CCC, is highly nontrivial, and is certainly not a simple extension.

Beyond the issue of using better modeling techniques for analyzing the nominal case, it is also important to consider the effects of process variations, which significantly affect circuit timing in digital circuits [17] in current and future technologies. Since HC effects are closely dependent on the circuit operation and device stress conditions, they is also affected by process variations. The interaction between HC effects and process variations has gained increasing attentions in recent years. However, most of the published works only focus on device-level analysis [18, 19] or small-scale digital circuit [20], and the proposed methods are usually based on LEM model with HSPICE or Monte Carlo simulation, and are not scalable to large digital circuits.

Chapter 4 of this thesis provides a third path for CHC/CCC degradation analysis for large digital circuits, by using the newer multi-mode energy-driven degradation model [12, 13], performing cell-level characterization of transistor age gain per signal transition event, and utilizing signal statistics to perform circuit-level degradation analysis. Then the proposed approach is extended at the cell-level modeling and circuit-level analysis to incorporate process variations, and variation-aware circuit degradation analysis proposed based on the statistical static timing analysis (SSTA) framework. Moreover, the cumulative effect of HC and BTI is explored.

1.3 Bias-Temperature Instability

Bias-temperature instability causes the threshold voltage, V_{th}, of CMOS transistors to increase over time under voltage stress, resulting in a temporally-dependent degradation of digital logic circuit delay. The reaction-diffusion (R-D) model [21–24], based on dissociation of Si–H bonds at the Si/SiO$_2$ interface, has been the prevailing theory of BTI mechanism and has been widely used in research on circuit optimization and design automation. There have been considerable amount of work based on the R-D for circuit analysis [25–27], degradation monitoring [28, 29], and design mitigation techniques [30–37]. However, over the years, several limitations in the theory have been exposed and an alternative theory arose as the charge trapping and detrapping model [38–41], in which the defects in gate dielectrics can capture charged carriers, resulting in V_{th} degradations.

The major difference between the two models is the nature of the diffusing species and the medium that facilitates the diffusion. Based on published works, both R-D and charge trapping mechanisms exist in current semiconductor technologies, and the superposition of both models is shown to better match experimental device data [24].

In nanometer-scale technologies, variations in the BTI effect are gaining a great deal of attention under both R-D and charge trapping frameworks, due to the random nature of defect localization in smaller and smaller transistors; together, these result in increased variations in the number of defects in a transistor, leading to the variations in the BTI effect as predicted by both frameworks.

Most of the published circuit-level works incorporating BTI variations are based on the variability model of ΔN_{IT} randomness within the R-D framework, introduced by [42]. However, as explored in Chapter 5 of this thesis, for digital logic circuits, the ΔN_{IT} variation in the R-D based model has a relatively small impact on circuit timing variation. On the other hand, the variations of device-level BTI degradations under charge trapping has been discovered to be a significant issue for nanoscale transistors. Charge trapping and detrapping at each defect are random events that are characterized by the capture and emission time constants. This paradigm is intrinsically statistical and it captures not only the variations in the number of defects, but also the variations in ΔV_{th} induced by each defect [43–45]. Under this statistical model, the variation of device lifetime increases significantly.

However, the impact of BTI variations on circuit performance under charge trapping has not received much attention, with only limited works that explore this issue beyond the device level. In Chapter 5 of this thesis, we first introduce the notion of precharacterized *mean defect occupancy probability* for charge trapping to effectively reduce the complexity of circuit-level analysis and to make it possible to handle large-scale circuits. Then we incorporate variations under both the R-D and charge trapping into a novel temporal statistical static timing analysis (T-SSTA) framework, capturing randomness from both process variations and temporal BTI degradations. Our experimental results project the relative role of BTI charge trapping to circuit variability to increase significantly in the future technology nodes, but is less than the contribution of process variations.

Chapter 2

Circuit Failure Analysis due to Gate Oxide Breakdown

Gate oxide breakdown is an important reliability issue that has been widely studied at the individual transistor level, but has seen very little work at the circuit level. We first develop an analytic closed-form model for the failure probability of a large digital circuit due to this phenomenon. The new approach accounts for the fact that not every breakdown leads to circuit failure, and shows a 4.8–6.2× relaxation of the predicted lifetime with respect to the pessimistic area-scaling method for nominal process parameters. Next, we extend the failure analysis to include the effect of process variations, and derive that the circuit FP at a specified time instant has a lognormal distribution due to process variations. Circuits with variations show 19–24% lifetime degradation against nominal analysis and 4.7–5.9× lifetime relaxation against area-scaling method under variations. Both parts of our work are verified by extensive simulations and proved to be effective, accurate and scalable.

2.1 Introduction

Of late, reliability issues have become an increasingly important concern in CMOS VLSI circuits. Oxide breakdown refers to the phenomenon where defects are generated in the SiO_2 gate oxide under the continued stress of normal operation over a long period. Eventually, the oxide becomes conductive when a critical defect density is reached at a certain location in the oxide. With device scaling, as electric fields across the gate oxide have increased as supply voltages have scaled down more slowly than the oxide thickness, transistors have become more susceptible to oxide breakdown.

At the device level, the mechanisms and modeling of oxide breakdown have been studied for several decades, yielding a large number of publications, as surveyed in [4]. Various empirical and analytical models, including percolation models [46, 47] have been proposed for this phenomenon. The time-to-breakdown characteristic for a MOS transistor is typically modeled as a Weibull random variable, and characterized by accelerated experiments, in which MOS transistors or capacitors are subjected to high voltage stress at the gate terminal, with both the source and drain terminals grounded until breakdown is detected [5, 48].

The effect of a breakdown is to provide a path for current to flow from the gate to the channel. The terms *hard breakdown* and *soft breakdown* are widely used to describe the severity of oxide breakdown occurrences. Functional failures, which are the focus of this work, can only be caused by HBDs (although, as we will show, not every HBD causes a functional failure). Unlike in analog or memory circuits where SBDs can provoke circuit failure, SBDs in digital logic circuit can only cause parametric variations but not functional failures [4, 49, 50], therefore they are not considered in this work. Through the rest of this chapter, the term "circuit failure" implies a functional failure in digital logic circuits.

It is believed that there is no substantial difference between the physical origins of the HBD and SBD modes [51], and they are generally distinguished by the resistance of the breakdown path and the consequence to the devices. An HBD is a low-resistance breakdown that can cause significant current to flow through the gate, while an SBD has a higher resistance, and lower breakdown current through the gate [4]. A quantitative comparison of these two modes is presented in [2], and the concept of HBD and SBD has been verified for technologies down to 40nm [52].

At the circuit level, the traditional failure prediction method for a large circuit uses area-scaling, extrapolated from single device characterization [4]. The idea is based on the weakest-link assumption, that the failure of any individual device will cause the failure of the whole chip. Recently, new approaches have been proposed to improve the prediction accuracy by empirical calibration using real circuit test data [53], or by considering the variation of gate-oxide thickness [6]. The former is empirical and hard to generalize, while the latter does not consider the effect of breakdown location. Moreover, all existing methods circuit-level methods assume that (a) the transistors in the circuit are *always* under stress, and (b) any transistor breakdown *always* leads to a circuit failure. These assumptions are not always true, as discussed in Section 2.2.1.

Precise analysis or measured results on several small circuits have been published, based on the post-breakdown behavior models: for a 41-stage ring oscillator in [54], a 6T SRAM cell in [49], and current mirrors and RS latches in [50]. These methods, using either complex analysis models or are based on measurements, and cannot easily be extended to general large-scale digital circuits in a computationally scalable manner.

On the other hand, the probability of circuit failure is significantly affected by on-chip process variations. Recent work [6] proposed a statistical approach for full-chip oxide reliability analysis considering process variation of T_{ox}; however, this work did not present a path to determining the full distribution of the reliability function or statistics such as its variance. Subsequent work in [7] improved upon this by presenting a post-silicon analysis and mitigation method involving on-chip sensors and voltage tuning. The major drawback of these variation-aware approaches for circuit-level oxide reliability analysis is that they are all based on the simple notion of area-scaling, which is too pessimistic for circuit lifetime prediction.

The contribution of our work is twofold. First, we develop a scalable method for analyzing the failure probability of large digital circuits, while realistically considering the circuit environment that leads to stress and oxide breakdown. To achieve this goal, at the *transistor* level, we revise the Weibull time-to-breakdown model to incorporate the actual stress modes of transistors. We propose a new piecewise linear/log-linear resistor model for post-breakdown behavior of transistors as a function of the breakdown location within the transistor, in accordance with device-level experimental data in [2]. At the logic *cell* level, we devise a procedure for performing precise FP analysis for standard cell based digital circuits, and present an effective library characterization scheme. In particular, we demonstrate the circuits have inherent resilience to failure due to gate oxide breakdown, and we use this information to build a characterization methodology and analysis method that provides more correct FP computations than the area-scaling model. At the *circuit* level, we derive a closed-form expression for the FP of large digital logic circuits, based on the above characterization of the post-breakdown circuit operation. This analysis leads to the conclusion that area-scaling estimates are unduly pessimistic.

Second, we explore the effects of process variations, and find that the predicted FP under nominal condition is significantly affected by variations. We then extend the nominal case FP analysis to include the effect of process variations, and show that this still provide substantially better improvements in the predicted lifetime over the conventional area-scaling model. The transistor-level model and cell-level analysis are updated, and it is derived that the circuit FP at a specified time instant has a lognormal distribution due to process variations, and this distribution expands as the process variations and spatial correlation increase. Both parts of our work are verified by extensive simulations and results prove the proposed methods are effective, accurate and scalable.

We begin with an analysis of the nominal case. Section 2.2 presents an overview of transistor-level breakdown models, the post-breakdown behavior, and the value of the breakdown resistance, and introduces our empirical model. Next, Section 2.3 develops a method for cell-level FP computation. This is applied to circuit-level calculations in Section 2.4, where we derive a closed-form formula predicting the circuit-level FP. The theory for the nominal case is extended to variation-aware oxide reliability analysis in Section 2.5. Finally, Section 2.6 presents simulation results to validate the proposed methods, and we conclude in Section 2.7.

2.2 Transistor-Level Models

In this section, we discuss models for the time-to-breakdown and the post-breakdown behavior of a transistor. Sections 2.2.1 and 2.2.2 largely overview existing models, while Section 2.2.3 presents our new simple quantitative model for breakdown resistance that can be calibrated from experimental data.

Our discussion is guided by two observations:

- As shown in [4], only HBDs cause serious device degradations.
- As demonstrated in [48], the occurrence of HBD is very prevalent in NMOS transistors but relatively rare in PMOS devices.

Therefore, we only consider NMOS HBD in this work. However, the framework presented here can easily be extended to the cases where these two assumptions are relaxed.

Furthermore, we assume that a transistor will be affected by at most one HBD. This assumption is reasonable, and is similar in spirit to the single stuck-at fault assumption in the test arena: due to the statistical and infrequent nature of breakdown events, the probability of more than one independent breakdown striking the same transistor is very low.

2.2.1 Time-to-Breakdown

The transistor time-to-breakdown, T_{BD}, is widely modeled as a Weibull distribution with an area-scaling formula [5]. The breakdown probability of device i, with area a_i, at time t is

$$\mathrm{Pr}_{\mathrm{BD}}^{(i)}(t) = 1 - \exp\left(-\left(\frac{t}{\alpha}\right)^{\beta} a_i\right), \tag{2.1}$$

where α is the characteristic time corresponding to 63.2% of breakdown probability for the unit-size device with area $a_i = 1$, and β is the Weibull shape factor, also known as the Weibull slope. A common representation of a Weibull distribution is on the so-called *Weibull scale*, under the transform

$$W = \ln(-\ln(1 - \mathrm{Pr})) = \beta \ln(t/\alpha) + \ln(a_i) \tag{2.2}$$

In other words, if we plot W as a function of $\ln(t)$, the result is a straight line with slope β.

The Weibull parameters α and β in are usually characterized in experiments, as described in [2,5], where the gate oxide of the transistor is placed in inversion mode and subjected to a constant voltage stress. However, this experimental scenario is not an accurate representation of the way in which transistors function in real circuits, where the logic states at the transistor terminals change over time, with six possible static stress modes for a NMOS transistor, as shown in Fig. 2.1[1] .

An HBD occurs in the case of NMOS stressed in inversion, while an NMOS in accumulation almost always experiences SBD [48]. In Fig. 2.1, Mode A corresponds

[1] The other two combinations, with the gate at logic 1 and the source and drain at different voltages, are transient modes, not relevant for analyzing long-term stress.

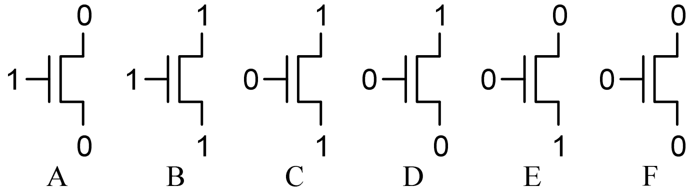

Figure 2.1: Stress modes for NMOS transistors.

to inversion, and Modes C, D and E to accumulation, while B and F do not impose a field that stresses the gate oxide. Thus, only the portion of time when the transistor is stressed in Mode A is effective in causing HBDs in a device, and potential circuit failure. We introduce the stress coefficient, γ_i, for device i to capture the proportion of this effective stress time, and reformulate (2.1) as

$$\mathrm{Pr}_{\mathrm{BD}}^{(i)}(t) = 1 - \exp\left(-\left(\frac{\gamma_i t}{\alpha}\right)^{\beta} a_i\right) \tag{2.3}$$

where $(\gamma_i t)$ represents the effective stress period after time t of circuit operation. The stress coefficient γ_i is the probability of Mode A, and can be represented by the joint probability mass function (jpmf) that the (gate, source, drain) terminals of transistor i have the logic pattern $(1, 0, 0)$. This can be calculated using the signal probability (SP) of each node, and maps on to a well-studied problem in CAD. These probabilities may be computed, for example, more approximately by using topological methods that assume independence [55], or using more computational methods that explicitly capture correlations, such as Monte Carlo approaches [56].

2.2.2 Post-Breakdown Behavior

Fig. 2.2(a) shows a two-dimensional schematic that displays the idea of oxide breakdown in a MOS transistor. The channel length is denoted by L, and the source/drain extensions are of length L_{ext}. The distance from the source is denoted by x, and the breakdown is assumed to be located at x_{BD}.

Various modeling approaches for post-breakdown analysis at the transistor- or cell-level have been proposed in the literature. Several approaches have proposed models for SBDs, e.g., [57,58], but these result in parametric failures rather than the functional

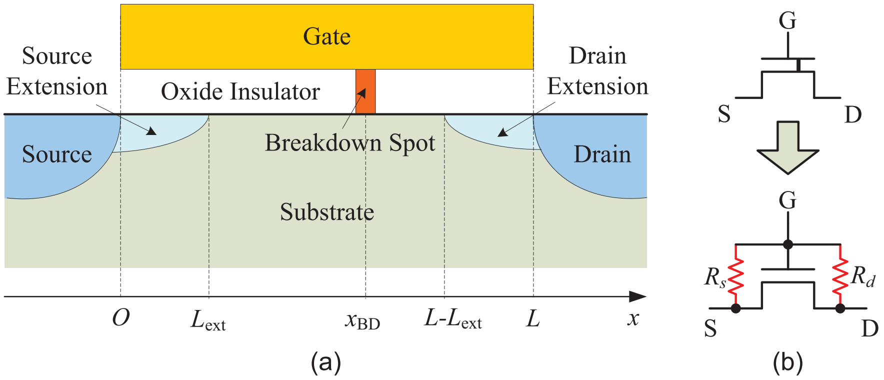

Figure 2.2: (a) Schematic of oxide breakdown in a transistor. (b) Resistor model for post-breakdown behavior.

failures that this work studies. The work in [59] suggests a complex physical model that reduces to a simple resistor model when the breakdown location is near the source or drain. As summarized in [4], independent experiments have reported that HBDs show a roughly linear (ohmic) I-V characteristic. Based on this, we use a simpler linear resistor model, similar to that in [60,61], for post-breakdown behavior analysis. A MOS transistor that has undergone oxide breakdown is replaced with a healthy clone and two resistors, R_s and R_d, as shown in Fig. 2.2(b). The values of these two resistors are dependent on the breakdown location, x_{BD}.

In characterizing the values of these resistances, it is important to lay down some requirements that they must fulfill. Fig. 2.3(a) shows the experimental measurement value of the effective breakdown resistance, R_{BD}, for HBDs as a function of x_{BD}, where both the source and drain nodes of the transistor are grounded, and R_{BD} is measured between the gate node and the ground [2]. The data points in this figure correspond to measurements, while the solid line is based on a detailed device simulation. Further experimental data in [2] (not shown here), demonstrate that over a range of channel lengths, the nature of the variation of R_{BD} with x_{BD} shows the same trend as in the figure. Specifically, the observations drawn from [2] are that:

- R_{BD} is smaller when the HBD occurs in the source or drain extension regions, and larger for x_{BD} in the channel.

- R_{BD} decreases exponentially (note the log scale on the y-axis) when x_{BD} approaches either end of the channel, while it does not vary significantly with x_{BD} in the center of the channel.

- The statistics of the breakdown location, x_{BD}, show a uniform distribution over the length of the channel.

In advanced high-k technology, [62] indicated that breakdowns are more likely to happen in the grain boundary (GB) sites, which also have uniform distribution in the dielectric layer.

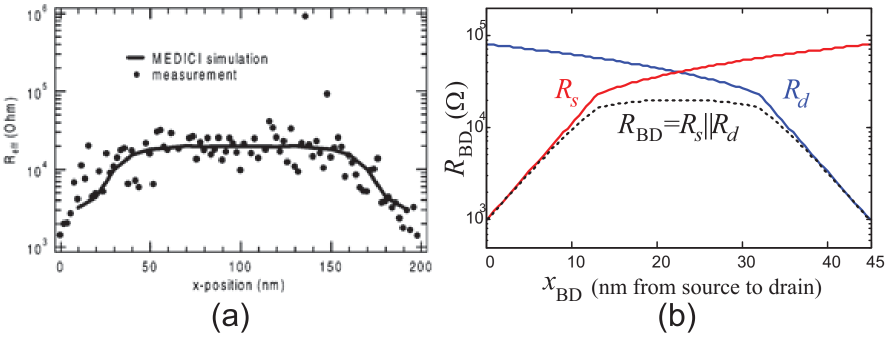

Figure 2.3: (a) The effective breakdown resistance as a function of the breakdown location [2]. (b) Modeling of the breakdown resistors.

2.2.3 Modeling the Breakdown Resistors

While the structure of the breakdown resistor model using R_s and R_d in Fig. 2.2(b) is not fundamentally new, there has been less work on deriving a model that relates the breakdown resistance with x_{BD}. The only known work is an equivalent circuit model in [59], but it requires a complex characterization process; moreover, the nonlinearity of the model makes its evaluation in a circuit simulator more time-consuming. We derive a much simpler model based on the idea of fitting the result from experiments and simulation which requires very few measurements for characterization.

The form of the model is guided by the breakdown resistance vs. x_{BD} curve in Fig. 2.3(a). We propose to capture the variation of the breakdown resistance with x_{BD} through a piecewise linear/log-linear model, where R_s $[R_d]$ varies exponentially with

x_{BD} in the source [drain] extension region, and linearly in the remainder of the channel:

$$R_s(x) = \begin{cases} ae^{bx}, & 0 \le x \le L_{\mathrm{ext}} \\ kx, & L_{\mathrm{ext}} \le x \le L \end{cases} \tag{2.4}$$

Due to source-drain symmetry, we obtain $R_d(x) = R_s(L - x)$. When both the source and drain nodes are grounded, $R_{\mathrm{BD}}(x) = R_s(x) \parallel R_d(x)$. The value of R_{BD} is at its minimum, $R_{\mathrm{BD\,min}}$, at $x = 0$ and $x = L$, and by symmetry, at its maximum, $R_{\mathrm{BD\,max}}$ at $x = L/2$. This discussion about R_{BD} is purely for illustration purposes: in our work, we do not directly use R_{BD}, but work with the R_s and R_d models in conjunction with MOS transistor models.

The constants k, a and b are obtained from experiment measurements in [2] by matching a set of boundary conditions. At $x = 0$, the value of R_s dominates the value of R_d, so that $R_{\mathrm{BD}} \simeq R_s(0) = R_{\mathrm{BD\,min}}$. Thus, $a = R_{\mathrm{BD\,min}}$.

At $x = L/2$, by symmetry, $R_s = R_d$, implying that $R_{\mathrm{BD}} = R_s(L/2)/2 = R_{\mathrm{BD\,max}}$. Therefore, $k = 4R_{\mathrm{BD\,max}}/L$.

Finally, to ensure the continuity between the linear and log-linear pieces of the piecewise model, we must ensure that $\lim_{x \to L_{\mathrm{ext}}^-} R_s(x) = \lim_{x \to L_{\mathrm{ext}}^+} R_s(x)$, i.e., $kL_{\mathrm{ext}} = ae^{bL_{\mathrm{ext}}}$. So,

$$b = \frac{1}{L_{\mathrm{ext}}} \ln \left(\frac{4R_{\mathrm{BD\,max}} L_{\mathrm{ext}}}{R_{\mathrm{BD\,min}} L} \right) \tag{2.5}$$

Four parameters are required to characterize this model: L, L_{ext}, $R_{\mathrm{BD\,min}}$ and $R_{\mathrm{BD\,max}}$. Fig. 2.3(b) shows an example plot for the parallel combination of R_s and R_d using this model, with the parameters $L = 45$nm, $L_{\mathrm{ext}} = 13$nm, $R_{\mathrm{BD\,max}} = 20\mathrm{k}\Omega$, and $R_{\mathrm{BD\,min}} = 1\mathrm{k}\Omega$[2] . It is easy to see that the results here are well matched to the trend of experimental measurements in Fig. 2.3(a).

2.3 Cell-Level Failure Analysis

Our entire technique for digital circuit failure analysis due to gate oxide breakdown is summarized in Figure 2.4. At the transistor level, the process parameters L and L_{ext},

[2] The values of $R_{\mathrm{BD\,max}}$ and $R_{\mathrm{BD\,min}}$ are input parameters and independent of the analysis approaches. Based on projections from the published literature, their values are taken to be 20kΩ and 1kΩ, respectively.

and R_{BD} measurements $R_{\mathrm{BD\,max}}$ and $R_{\mathrm{BD\,min}}$ are inputs to the method and utilized to characterize our post-breakdown $R_{\mathrm{BD}}(x_{\mathrm{BD}})$ model using the method discussed above. The values of α and β for the Weibull distribution that characterizes transistor-level failure are also input parameters. At the cell level, the driver and load I-V curves of each logic cell in the input cell library are precharacterized and stored into LUTs using Algorithm 1, which will be described in this section. The calculation of cell FP is performed in a circuit-specific context with Algorithm 2, also described later. Finally the circuit FP analysis is performed using the proposed method, using the result (2.16) presented in Theorem 1.

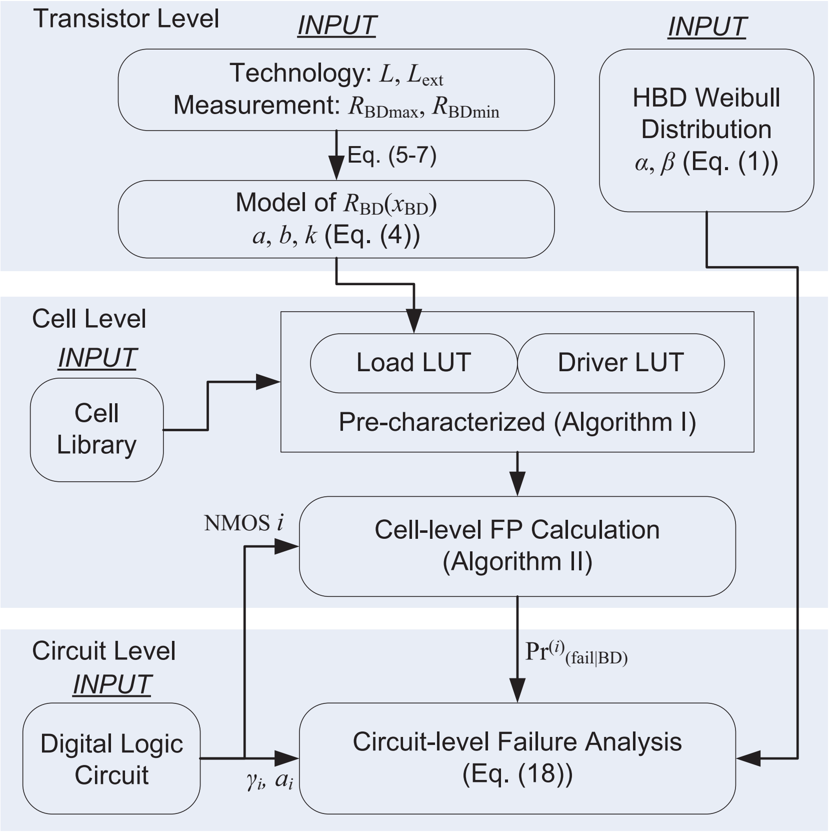
Transistor Level
INPUT
INPUT
Technology: L, L_{ext}
Measurement: R_{BDmax}, R_{BDmin}
Eq. (5-7)
Model of $R_{\mathrm{BD}}(x_{\mathrm{BD}})$
a, b, k (Eq. (4))
HBD Weibull Distribution
α, β (Eq. (1))
Cell Level
INPUT
Cell Library
Load LUT
Driver LUT
Pre-characterized (Algorithm I)
NMOS i
Cell-level FP Calculation (Algorithm II)
Circuit Level
INPUT
$\mathrm{Pr}^{(i)}{}_{(\mathrm{fail|BD})}$
Digital Logic Circuit
γ_i, a_i
Circuit-level Failure Analysis (Eq. (18))

Figure 2.4: Flow chart of digital circuit oxide reliability analysis.

This section focuses on analyzing the effects of oxide breakdown at the logic cell level. A formula for the FP for each breakdown case is developed, and a library characterization scheme is proposed for standard cell based digital circuits.

2.3.1 Breakdown Case Analysis

The effect of the gate oxide breakdown in an NMOS transistor is to create current paths from the gate node of the transistor to its source and drain nodes. In CMOS circuits, the gate node of a device is typically connected to the output of another logic cell or latching element, while the source/drain nodes are, by definition, connected to transistors within the same logic cell (or more generally, the same channel-connected component). This implies that while analyzing breakdown at the gate node of a transistor, it is necessary to consider both the logic cell that it belongs to and the preceding logic cell that drives the gate node of the transistor.

Consider a cell n that contains a transistor with oxide breakdown. Let k be the pin of cell n connected to the gate of this transistor, and let m be the logic cell that drives pin k of cell n. Then for any broken down NMOS transistor, we can find the corresponding case index (m, n, k). Fig. 2.5(a) shows an example of such a breakdown case, using a NAND2 as cell m, a NOR2 as cell n, and $k = 1$. Here we call cell m as the *driver cell* and cell n as the *load cell* of this case.

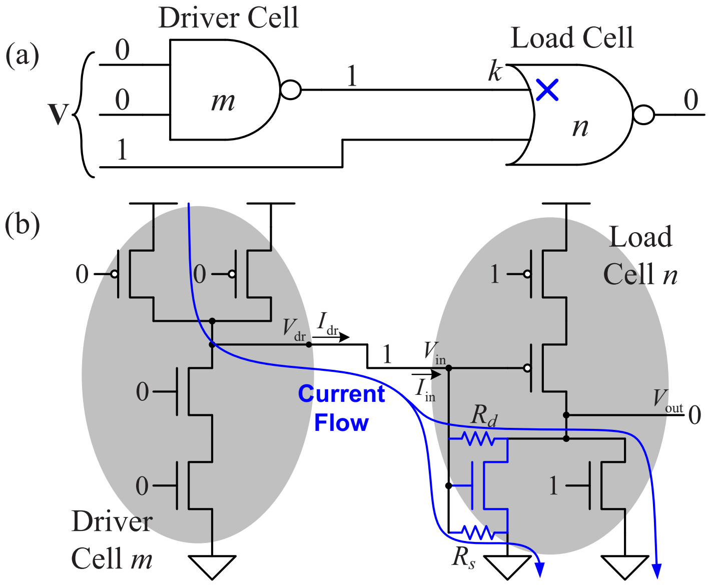

Figure 2.5: Cell-level analysis of the breakdown case.

To analyze each breakdown case (m, n, k), we must specify the input vector $\mathbf{V}$ for the free pins of the two cells. The input vector $\mathbf{V}$ is a Boolean vector of dimension

$q(m,n) = (\text{Fanin}(m) + \text{Fanin}(n) - 1)$, i.e., $\mathbf{V} \in \mathbb{B}^{q(m,n)}$, where $\text{Fanin}(i), i \in \{m,n\}$ represents the number of input pins of cell i; in Fig. 2.5(a), q = 3, and we consider the assignment $\mathbf{V} = (0,0,1)$. We refer to a breakdown case for a specific input vector as $(m,n,k,\mathbf{V})$. Any given $(m,n,k,\mathbf{V})$ combination can be analyzed based on the post-breakdown behavior model discussed in Section 2.2. The transistor-level circuit, using the resistor model, is shown in Fig. 2.5(b), with the current flow path due to oxide breakdown indicated. The worst case, over all input vectors (it should be noted that q is a small number) for this two-cell structure defines the FP, as quantified in the next subsection.

Essentially, Fig. 2.5(b) shows that the current lost due to the breakdown event has the potential to alter the logic value at the output of cell m or n or both; whether it actually does so or not depends on the strength of the opposing transistor that attempts to preserve the logic value.

2.3.2 Calculation of Failure Probabilities

The breakdown case in Fig. 2.5 is analyzed using SPICE DC sweep of x_{BD} with 45nm PTM model [63] and $V_{\text{dd}} = 1.2$V. The output voltages of driver cell m and load cell n, denoted by V_{dr} and V_{out}, as functions of x_{BD}, are shown in Fig. 2.6. This figure indicates that when breakdown occurs near the source or drain and the breakdown resistor, R_s or R_d, is small, the output voltages of cells m and n may shift away from their nominal values of V_{dd} and 0, respectively. Beyond certain limits, the logic could flip and result in circuit failure.

Note that the results for driver and load cells are asymmetric for the input excitation in Fig. 2.5, in that the driver cell m shows a failure when the defect lies at either end of the channel, while the failure for the load cell n appears only when the defect lies at the drain end. The difference lies in the case that x_{BD} is small where R_s is very small and R_d is large. In this case the other NMOS in cell n is on and the output voltage is relatively unaffected even in the presence of a breakdown.

We introduce two thresholds, V_H and V_L (in the figure, $V_H = 0.7V_{\text{dd}}, V_L = 0.3V_{\text{dd}}$), so that if the voltage surpasses these thresholds, a failure is deemed to occur. It can be shown that since the variation of the resistance with x_{BD} is monotonic near the drain [source], and since MOS transistors typically have monotonically increasing I-V curves, the output voltages of the impacted logic cells will also change monotonically with x_{BD} near the drain [source]. In other words, the *failure region* on either side of the channel is a continuous interval[3] , which is determined by the corresponding crossover point. We define the crossover points to be $x^{\text{dr}}_{\text{fail-s}}, x^{\text{ld}}_{\text{fail-s}}, x^{\text{dr}}_{\text{fail-d}}$, and $x^{\text{ld}}_{\text{fail-d}}$, which refer to the breakdown locations where the corresponding cell output voltages cross the threshold, as illustrated in Fig. 2.6[4] .

This result is not surprising: the R_{BD} is large in the channel and small in the source/drain extension regions, so that HBDs in the latter regions are liable to cause logic failures.

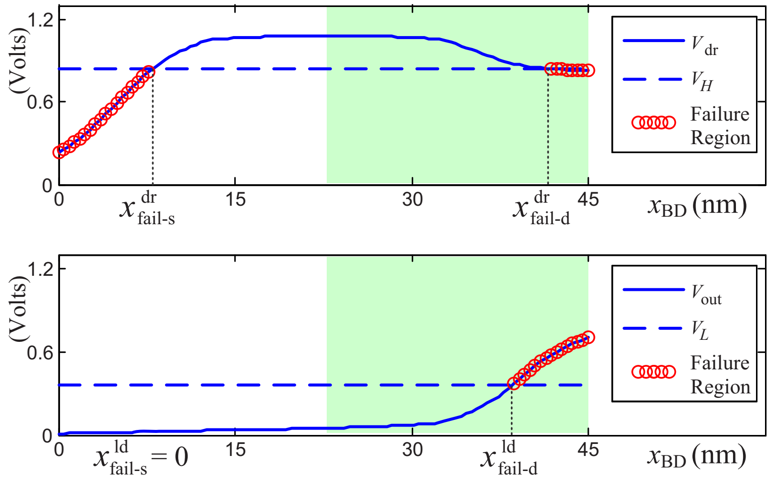

Figure 2.6: Cell output voltages under breakdown.

We can then obtain the source-side and drain-side *failure probability* (FP) separately for this specific breakdown case and input vector by evaluating the probability of x_{BD} falling within the corresponding failure region. According to [2, 62], the breakdown position is uniformly distributed in the channel, i.e., $x_{\text{BD}} \sim \text{U}[0, L]$. Therefore, these FPs are given by:

$$\begin{aligned} \Pr^{(m,n,k,\mathbf{V})}_{(\text{fail-s}|\text{BD})} &= \max\left(p^{\text{dr}}_s, p^{\text{ld}}_s\right) \\ \Pr^{(m,n,k,\mathbf{V})}_{(\text{fail-d}|\text{BD})} &= \max\left(p^{\text{dr}}_d, p^{\text{ld}}_d\right) \end{aligned} \tag{2.6}$$

[3] If the output voltage does not cross the threshold, the failure region may be an empty set, as in the left part of the lower graph of Fig. 2.6.

[4] If no crossing point exists, the value of the parameter is set to zero at the source end or L at the drain end.

where, for a given breakdown case $(m, n, k, \mathbf{V})$, the FP components are

$$\begin{aligned} p_s^{\text{dr}} &= \frac{x_{\text{fail-s}}^{\text{dr}}}{L}, \qquad p_d^{\text{dr}} = 1 - \frac{x_{\text{fail-d}}^{\text{dr}}}{L} \\ p_s^{\text{ld}} &= \frac{x_{\text{fail-s}}^{\text{ld}}}{L}, \qquad p_d^{\text{ld}} = 1 - \frac{x_{\text{fail-d}}^{\text{ld}}}{L} \end{aligned} \tag{2.7}$$

A transistor breakdown with case index (m, n, k) corresponds to a logic failure if such a failure is seen under any input vector $\mathbf{V} \in \mathbb{B}^{q(m,n)}$. This is because once the device-level failure occurs, the circuit is considered to functionally fail if it fails under *any* input vector. Therefore the FP of either side for case (m, n, k) is the worst over all input vectors $\mathbf{V} \in \mathbb{B}^{q(m,n)}$, i.e., the maximum probability among all input vectors. Under the assumption of at most one HBD per transistor, the events of source-side failure and drain-side failure are mutually exclusive, therefore the total FP for case (m, n, k) is the sum of the two sides:

$$\Pr{}_{(\text{fail}|\text{BD})}^{(m,n,k)} = \max_{\mathbf{V} \in \mathbb{B}^q} \Pr{}_{(\text{fail-s}|\text{BD})}^{(m,n,k,\mathbf{V})} + \max_{\mathbf{V} \in \mathbb{B}^q} \Pr{}_{(\text{fail-d}|\text{BD})}^{(m,n,k,\mathbf{V})} \tag{2.8}$$

Since the logic cells come from a common cell library, $\mathbb{C}$, it is possible to characterize a library over all breakdown cases as a precomputation. For circuit-level failure analysis, as described in Section 2.4, the precomputed FP results can be retrieved from the characterized library in $O(1)$ time.

2.3.3 Cell Library Characterization

The principles behind our cell-level failure analysis procedure have been outlined in the previous two subsections. However, the implementation of this approach involves the analysis of cases $(m, n, k, \mathbf{V})$, and a simple precharacterization would involves a quadratic-complexity enumeration of both driver and load cells from the library. Specifically, the number of SPICE simulations required for this precharacterization, N_{enum}, is computed as:

$$N_{\text{enum}} = N_{\text{cell}}^2 \cdot N_{\text{pin}} \cdot 2^{2N_{\text{pin}}-1} \tag{2.9}$$

Here, N_{cell} stands for the number of cells in the library, and N_{pin} is a bound on the number of fan-ins for a cell; practically, this is a small constant (and this is substantiated on a Nangate library in our experimental results). The number of enumerations, N_{enum}, is the total possible combinations of $(m, n, k, \mathbf{V})$, with $m, n \leq N_{\text{cell}}$, $k \leq N_{\text{pin}}$, and Boolean vector $\mathbf{V}$ has $2^{2N_{\text{pin}}-1}$ combinations. With N_{pin} well bounded ($N_{\text{pin}} \leq 6$ in the Nangate library we used), the case amount $N_{\text{enum}} \propto N_{\text{cell}}^2$ has quadratic complexity with library size, which presents a problem for the cell library characterization process, especially for libraries with a larger number of cells. For example, experiments on a 55-cell Nangate library show that about 1.7 million such enumerations are necessary: clearly, this is a very high cost, even for a one-time precharacterization step.

To overcome this cost without any significant sacrifice in accuracy, we propose a method that improves the scalability of our failure analysis approach. The essence of the idea is that instead of precharacterizing and storing all quadratic combinations, we precharacterize the I-V curves for the library cells and then solve the breakdown cases on the fly. The number of precharacterizations is linear in the number of cells, and the solution can be performed in constant time. Specifically, our library characterization and cell-level FP calculation scheme consists of two stages:

- In the first stage (*precharacterization*), we consider the possibility that each library cell may feature as a driver for another load gate and a load for another driver gate. Accordingly, each cell is characterized to obtain its driver I-V curve (when it acts a driver cell) and its load I-V-x_{BD} curve (when it acts as a load cell) separately, the curves are stored numerically in look up tables (LUTs).

- In the second stage (*FP calculation*), which is performed during the analysis of a specific circuit, the precharacterized curves are used to compute the FP of a specified $(m, n, k, \mathbf{V})$ case from the I-V curves of the driver cell and the load cell using the LUT data.

Fig. 2.7 shows an example that demonstrates our improved scheme. For the example shown in Fig. 2.5, Fig. 2.7(a) plots the precharacterized $I_{\text{dr}}(V_{\text{dr}})$ curve for the driver cell and the precharacterized family of $I_{\text{in}}(V_{\text{in}}, x_{\text{BD}})$ curves (indexed by x_{BD}) for the load cell, and these capture the interaction between the driver and the load cell at the output of the driver. The effect on the output voltage of the load cell is captured by Fig. 2.7(b), which shows the precharacterized family of curves for $V_{\text{out}}(V_{\text{in}}, x_{\text{BD}})$, indexed by the value of x_{BD}. Note that the load curves are shown for $x_{\text{BD}} \in [L/2, L]$, i.e., the drain side, and in this range, I_{in} and V_{out} are monotonic function of x_{BD}. As we

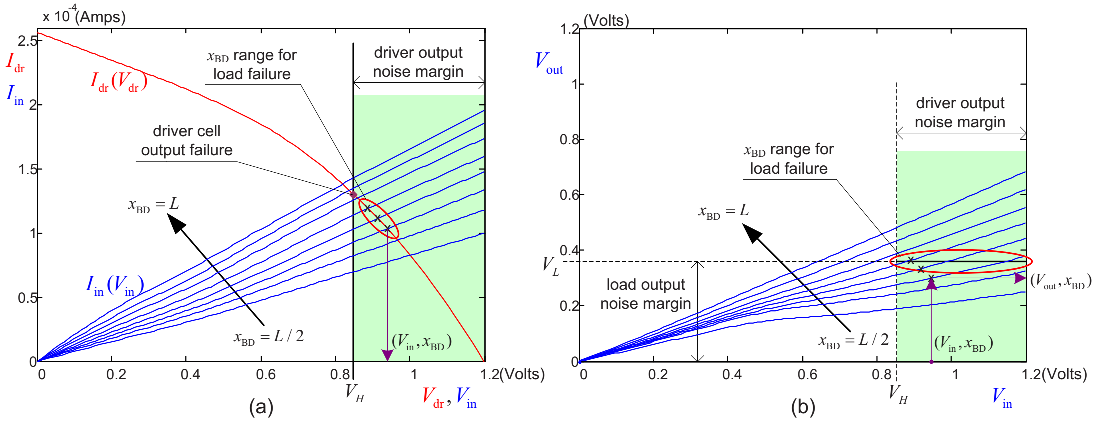

Figure 2.7: Demonstration of solving the cell-level FP using I-V curves of the driver and load cells.

will show later, these curves are adequate to capture the interaction between the driver and the load in any circuit.

Algorithm 1 presents the precharacterization procedure that precomputes these curves. Note that this precharacterization is performed off-line, like standard cell characterization, and must be carried out just once for a given technology. The complexity of this algorithm is linear in the size of the cell library, and the notations used within the algorithm are as follows: for a cell i, I_{dr} and V_{dr} stand for the current and voltage when the output pin of the cell acts as a driver; I_{in} and V_{in} stand for the input current and voltage when the input pin k of the cell acts as a load; and V_{out} stands for the voltage of the output pin of cell i when it acts as a load.

As mentioned earlier, each cell i in the library is characterized separately in its role as a driver and as a load. For the driver characterization, the $I_{\text{dr}}(V_{\text{dr}})$ curve is calculated with sampled values for V_{dr}, for all possible input combinations. Therefore the total number of driver I-V LUTs is $N_{\text{cell}} \cdot 2^{N_{\text{pin}}}$. The load characterization is performed similarly but with an additional enumeration that samples the breakdown location, x_{BD}. The total number of $I_{\text{in}}(V_{\text{in}}, x_{\text{BD}})$ and $V_{\text{out}}(V_{\text{in}}, x_{\text{BD}})$ LUTs corresponding to this is $2 \cdot N_{\text{cell}} \cdot N_{\text{pin}} \cdot 2^{N_{\text{pin}}}$. The storage overhead associated with all driver and load LUTs in the entire library is given by

$$\begin{aligned} \text{Driver} &: N_{\text{cell}} \cdot 2^{N_{\text{pin}}} \cdot N_V, \\ \text{Load} &: 2 \cdot N_{\text{cell}} \cdot N_{\text{pin}} \cdot 2^{N_{\text{pin}}} \cdot N_{x_{\text{BD}}} \cdot N_V \end{aligned} \tag{2.10}$$

where N_V stands for the number of V_{dr} and V_{in} samples, and $N_{x_{\text{BD}}}$ stands for the number of x_{BD} samples. This implies that the storage is linear in N_{cell} since the other terms in this expression are bounded by moderate constants in practice.

Algorithm 1 The characterization of cell library for FP calculation.

```
{Driver characterization}
for each cell $i$ in the library do
    for each input vector $\mathbf{V}$ of cell $i$ do
        Calculate $I_{\text{dr}}(V_{\text{dr}})$ for samples of $V_{\text{dr}}$
        Store $I_{\text{dr}}(V_{\text{dr}})$ in driver LUT for cell $i$ input $\mathbf{V}$
    end for
end for
{Load characterization}
for each cell $i$ in the library do
    for each input pin $k$ of cell $i$ do
        for each input vector $\mathbf{V}$ of cell $i$ do
            Calculate $I_{\text{in}}(V_{\text{in}}, x_{\text{BD}})$ and $V_{\text{out}}(V_{\text{in}}, x_{\text{BD}})$ for samples of $V_{\text{in}}$ and $x_{\text{BD}}$
            Store $I_{\text{in}}(V_{\text{in}}, x_{\text{BD}})$ and $V_{\text{out}}(V_{\text{in}}, x_{\text{BD}})$ in load LUT for cell $i$ pin $k$ input $\mathbf{V}$
        end for
    end for
end for
```

Using these precharacterized curves, the second stage, FP calculation, is applied in a circuit-specific context. Given a driver cell and a load cell, the FP calculation step must compute the unknown voltages at the output of the driver and the load. We now demonstrate this calculation for the scenario in Fig. 2.5, where the correct outputs of the driver and load cell correspond to logic 1 and 0, respectively. For this scenario, the following circuit equations must be solved to determine the unknown voltages:

$$\begin{aligned} I_{\text{dr}}(V_{\text{dr}}) &= I_{\text{in}}(V_{\text{in}}, x_{\text{BD}}) \\ V_{\text{dr}} &= V_{\text{in}} \\ V_{\text{out}} &= V_{\text{out}}(V_{\text{in}}, x_{\text{BD}}) \end{aligned} \tag{2.11}$$

Consider the problem of solving this for a HBD on the drain side, $x_{\mathrm{BD}} \in [L/2, L]$, affecting the voltage at the driver output, V_{dr}, as illustrated by the failure region on the right of the V_{dr} curve in Fig. 2.6. From (2.11), $I_{\mathrm{in}}(V_H, x_{\mathrm{BD}}) = I_{\mathrm{dr}}(V_H)$, corresponding to the intersection of two plots and the $V_{\mathrm{dr}} = V_{\mathrm{in}} = V_H$ line in Fig. 2.7(a). Therefore, for a specific value of V_H, the RHS of this equation can be obtained from the lookup table for the driver side gate. Finding the x_{BD} that solves the equation is then a matter of a reverse lookup on the lookup table for the load side gate.

At any value of V_{in}, since the family of $I_{\mathrm{in}}(V_{\mathrm{in}}, x_{\mathrm{BD}})$ curves increases monotonically with x_{BD}, a failure at $x_{\mathrm{BD}} = x_1$ implies a failure for all $x_{\mathrm{BD}} \geq x_1$, and this solution corresponds to the edge of the failure region, $x^{\mathrm{dr}}_{\text{fail-d}}$, shown in Fig. 2.6.

Now consider a failure at the load output, V_{out}. Since our goal is to sum up a set of disjoint probabilities, it is important only to consider load output failures that *do not* cause a driver output failure. The procedure consists of two steps:

(1) we consider all intersections in Fig. 2.7(a) between the I_{dr} and the family of I_{in} curves in the region $V_H \leq V_{\mathrm{in}} \leq V_{\mathrm{dd}}$, and for each of these, we determine the $(V_{\mathrm{in}}, x_{\mathrm{BD}})$ value, and

(2) we use the traced $(V_{\mathrm{in}}, x_{\mathrm{BD}})$ values in Fig. 2.7(b), using the V_{out} LUTs to determine the corresponding value of V_{out}: if this exceeds the threshold, V_L, then we have a failure.

In principle, a drain-side failure that occurs anywhere in the interval $[L/2, L]$ could cause a load output failure. However, we narrow down this range further. The idea is based on the observation that the $V_{\mathrm{out}}(V_{\mathrm{in}}, x_{\mathrm{BD}})$ curves in Fig. 2.7(b) cross V_L in the interval $V_H \leq V_{\mathrm{in}} \leq V_{\mathrm{dd}}$ only for a specific, typically small, range of x_{BD}. We exploit this idea to improve the efficiency of this procedure, restricting the search in the previous paragraph to this interval of x_{BD}: this is seen to yield considerable computational savings in practice.

To be general, the above idea must be extended to several cases, corresponding to breakdowns at the output of the driver and the load at both possible logic values, due to failures at the drain side and the source side. Thus, we must consider:

$$V_{\text{dr}} = V_{\text{TH}}^{\text{dr}},\ x_{\text{BD}} \in [0, L/2] \text{ for } x_{\text{fail-s}}^{\text{dr}};\ \text{or} \tag{2.12}$$

$$V_{\text{dr}} = V_{\text{TH}}^{\text{dr}},\ x_{\text{BD}} \in [L/2, L] \text{ for } x_{\text{fail-d}}^{\text{dr}};\ \text{or} \tag{2.13}$$

$$V_{\text{out}} = V_{\text{TH}}^{\text{out}},\ x_{\text{BD}} \in [0, L/2] \text{ for } x_{\text{fail-s}}^{\text{ld}};\ \text{or} \tag{2.14}$$

$$V_{\text{out}} = V_{\text{TH}}^{\text{out}},\ x_{\text{BD}} \in [L/2, L] \text{ for } x_{\text{fail-d}}^{\text{ld}}. \tag{2.15}$$

Here, $V_{\text{TH}}^{\text{dr/out}}$ stands for the corresponding threshold voltage (V_H or V_L) of $V_{\text{dr/out}}$.

Algorithm 2 lists the entire procedure for cell-level FP calculation including all four components. The cell-level case index (m, n, k) is determined for NMOS transistor i by finding out the driver cell m, the load cell n and the input pin k.

Algorithm 2 The calculation of cell-level FP using driver and load LUTs. Equation solving uses piecewise-linear approximation based on the LUT data. Failure criteria $V_{\text{TH}}^{\text{dr/out}} = V_H$ or V_L, depends on the nominal values of V_{dr} and V_{out}.

```
1:  for each NMOS transistor $i$ in the circuit do
2:      Determine the case index $(m, n, k)$ from $i$
3:      for each input vector $\mathbf{V}$ of this case do
4:          Determine input vectors for driver and load cells: $\mathbf{V}_{\text{dr}}, \mathbf{V}_{\text{ld}}$
5:          {Driver cell $m$ output failure:}
6:          For $x_{\text{BD}} \in [0, L/2]$, obtain $x_{\text{fail-s}}^{\text{dr}}$ as follows
            (if failed, set $x_{\text{fail-s}}^{\text{dr}} = 0$):
            a. Get $I_{\text{TH}} = I_{\text{dr}}(V_{\text{TH}}^{\text{dr}})$ using driver LUT;
            b. Get $x_{\text{BD}}$ by reverse lookup $I_{\text{in}}(V_{\text{TH}}^{\text{dr}}, x_{\text{BD}}) = I_{\text{TH}}$ using load LUT.
7:          Repeat 6 with $x_{\text{BD}} \in [L/2, L]$, obtain $x_{\text{fail-d}}^{\text{dr}}$
            (If failed, set $x_{\text{fail-d}}^{\text{dr}} = L$).
8:          {Load cell $n$ output failure:}
9:          For $x_{\text{BD}} \in [0, L/2]$, obtain $x_{\text{fail-s}}^{\text{ld}}$ as follows
            (If failed, set $x_{\text{fail-s}}^{\text{ld}} = 0$):
            a. Get subset of $x_{\text{BD}}$ samples, $\mathbf{X}$, satisfying
            $V_{\text{out}}(V_{\text{TH}}^{\text{dr}}, x_{\text{BD}}) \le V_{\text{TH}}^{\text{out}}$ and $V_{\text{out}}(V_{\text{nom}}^{\text{dr}}, x_{\text{BD}}) \ge V_{\text{TH}}^{\text{out}}$;
            b. For each $x_{\text{BD}} \in \mathbf{X}$, solve (2.11) for $V_{\text{out}}$, obtain new LUT $V_{\text{out}}(x_{\text{BD}})$;
            c. Solve $x_{\text{BD}}$ by reverse lookup $V_{\text{out}}(x_{\text{BD}}) = V_{\text{TH}}^{\text{out}}$ using the new LUT.
10:         Repeat 9 with $x_{\text{BD}} \in [L/2, L]$, obtain $x_{\text{fail-d}}^{\text{ld}}$
            (If failed, set $x_{\text{fail-d}}^{\text{ld}} = L$).
11:     end for
12:     Calculate $\Pr_{(\text{fail}|\text{BD})}^{(i)}$ using (2.6), (2.7) and (2.8).
13: end for
```

Since the number of V_{in} samples and x_{BD} samples is well bounded, the complexity of solving individual cases is bounded and can be considered as $O(1)$. The calculation of the entire circuit has a linear complexity to the circuit size. In practice, the cost of this is not large, as shown in our simulation results.

In summary, as compared to the direct calculation of cell-level FP which has quadratic complexity as (2.9), the proposed two-stage scheme effectively reduces both time and space complexity to linear in the library characterization stage, while introducing only a linear-complexity overhead to the circuit analysis stage. This scheme helps keep our entire analysis framework scalable for circuits as well as cell libraries.

2.4 Circuit-Level Failure Analysis

Oxide-breakdown-induced logic failure is a weakest-link problem, because failure of any individual logic cell causes the failure of the entire circuit[5] . As shown earlier, prior approaches considered both HBDs and SBDs, and did not adequately differentiate between breakdown events that cause failure and those that do not: in fact, SBD events do not cause functional failures in digital logic circuits [50]. As shown in Section 2.3, some, but not all, HBDs result in circuit failure. Our approach is predicated on identifying the probabilities of HBDs that can cause the circuit to become nonfunctional, and using this information to find the probability of circuit failure with time.

Our novel result on circuit-level FP analysis is stated below, and derives the probability density function of circuit FP based on the parameters of the transistor FP. Specifically, our new result shows that the probability distribution of the time-to-failure for an *entire circuit* is a Weibull distribution. Further, we will see that this implies that the conventional area-scaling based method for circuit FP estimation provides only a loose bound on the time-to-failure. The proof of the result is detailed in Appendix A.

Theorem 1 *The probability distribution $W(t)$, of the time-to-failure, t, for a logic circuit is given by the following distribution:*

$$W(t) = \beta \ln\left(\frac{t}{\alpha}\right) + \ln \sum_{i \in \text{NMOS}} \Pr^{(i)}_{(\text{fail}|\text{BD})} \gamma_i^{\beta} a_i. \tag{2.16}$$

[5] Some such failures may lie on false paths and be masked out, but we make the reasonable assumption that the probability that a cell lies on a false path is low, and this scenario can be neglected.

where α and β are the Weibull parameters for an unit-size device, and $\Pr^{(i)}_{(\text{fail}|\text{BD})}$, γ_i, and a_i are as previously defined.

This result leads to two important observations.
Observation 1: The time-to-breakdown PDF for a *circuit*, given by (2.16) is a Weibull distribution. Moreover:

- This distribution has the same Weibull slope, β, as the individual unit-sized device.
- The circuit-level distribution is shifted from that for a unit-sized device. The circuit FP curve is therefore parallel to the transistor FP curve, but is shifted vertically upwards by the *Weibull shift*, defined as:

  $$W_{\text{shift}} = \ln \sum_{i \in \text{NMOS}} \Pr^{(i)}_{(\text{fail}|\text{BD})} \gamma_i^{\beta} a_i. \tag{2.17}$$

  Alternatively, the shift along the horizontal axis shows the logarithm of the lifetime shifted to the left by an amount $\left(-\frac{1}{\beta} \ln \sum \Pr^{(i)}_{(\text{fail}|\text{BD})} \gamma_i^{\beta} a_i\right)$.
- The magnitude of this shift is determined by areas, stress coefficients and cell-level FP of transistors in the circuit.

Observation 2: Our method is more realistic than, and less pessimistic than, the traditional area-scaling-based method for predicting the FP distribution. Specifically, the area-scaling method yields the following Weibull distribution: [4]:

$$W' = \beta \ln \left(\frac{t'}{\alpha}\right) + \ln \sum_{i \in \text{NMOS}} a_i. \tag{2.18}$$

From (2.16) and (2.18), we can obtain that for the same circuit failure $W = W'$,

$$\frac{t}{t'} = \left(\frac{\sum a_i}{\sum \Pr^{(i)}_{(\text{fail}|\text{BD})} \gamma_i^{\beta} a_i}\right)^{\frac{1}{\beta}}. \tag{2.19}$$

This means our new method shows a relaxation of the circuit lifetime prediction against the traditional area-scaling by a multiplicative factor as given in (2.19). Since $\Pr^{(i)}_{(\text{fail}|\text{BD})}$ and γ_i are smaller than one, our new method always yields a longer lifetime prediction than the area-scaling approach.

Observation 2 can be interpreted as follows. Unlike the area-scaling-based traditional formula, our result can be considered to use a weighted sum of all areas, or the *effective area*, with the weighting term being $\Pr^{(i)}_{(\text{fail}|\text{BD})}\gamma_i^\beta$ for transistor i. This result complies with the intuition that (a) breakdown is slowed by a factor of γ_i, which is equivalent to the area shrinking by γ_i^β, (b) for each transistor only breakdowns in certain regions (near source or drain) lead to failure, so the effective area is further decreased by $\Pr^{(i)}_{(\text{fail}|\text{BD})}$ which is actually the worst-case proportion of the failure region.

2.5 Variation-Aware Oxide Reliability Analysis

While the analysis for the nominal case provides a clear framework for computing the FP, we find (as shown by our results in Section 2.6) that the effects of variation on the FP are significant. Therefore, in this section, we extend the proposed circuit failure analysis approach to include process variations and spatial correlation. First, we introduce the model for process variations. Next, the transistor-level model and cell-level analysis are updated to capture the effects of variation, and finally, the distribution of circuit failure probability under process variations is derived.

2.5.1 Modeling Process Variations

It is widely accepted that process parameter variations can be classified as lot-to-lot, die-to-die (D2D), and within-die (WID) variations, according to their scope; they can also be categorized as systematic and random variations by their causes and predictability. WID variations exhibit spatial dependence knows as spatial correlation, which must be considered for accurate circuit analysis.

We employ a widely-used variational model: a process parameter X is modeled as a random variable about its mean, X_0, as

$$\begin{aligned} X &= X_0 + X_g + X_s + X_r \qquad (2.20) \\ \sigma_X^2 &= \sigma_{X_g}^2 + \sigma_{X_s}^2 + \sigma_{X_r}^2 \end{aligned}$$

Here, X_g, X_s, and X_r stand for the global part (from lot-to-lot or D2D variations), the spatially correlated part (from WID variation), and the residual random part, respectively. Under this model, all devices on the same die have the same global part X_g. The spatially correlated part is modeled using a method similar as [17], where the entire chip is divided into grids. All devices within the same grid have the same spatially correlated part X_s, and devices in different grids are correlated, with the correlation falling off with the distance. The random part X_r is unique to each device in the system.

In this chapter we consider the variations in the transistor width (W), the channel length (L), and the oxide thickness (T_{ox}), and assume Gaussian-distributed parameters. The spatial correlation can be extracted as a correlation matrix [64], and processed using principal components analysis (PCA). The process parameter value in each grid is expressed as a linear combination of the independent principal components, with potentially reduced dimension. For a circuit with n transistors, with the three global parts for W, L and T_{ox}, the spatially correlated part and the n random parts, all the process parameters and their linear functions can be expressed in the random space with basis $\mathbf{e} = [\mathbf{e}_g, \mathbf{e}_s, \epsilon]^{\mathbf{T}}$ as

$$\begin{aligned} X &= X_0 + \Delta X = X_0 + \mathbf{k}_X^{\mathbf{T}}\mathbf{e} \qquad (2.21) \\ &= X_0 + \mathbf{k}_{Xg}^{\mathbf{T}}\mathbf{e}_g + \mathbf{k}_{Xs}^{\mathbf{T}}\mathbf{e}_s + k_\epsilon \epsilon \\ \sigma_X^2 &= \mathbf{k}_X^{\mathbf{T}}\mathbf{k}_X, \quad \mathrm{cov}(X_i, X_j) = \mathbf{k}_{Xi}^{\mathbf{T}}\mathbf{k}_{Xj} - k_{\epsilon_i}k_{\epsilon_j} \end{aligned}$$

Here, $\mathbf{e}_g = [e_{Wg}, e_{Lg}, e_{Tg}]^{\mathbf{T}}$ is the basis for global part, $\mathbf{e}_s = [e_1, ..., e_t]^{\mathbf{T}}$ is the basis of principal components for the spatial part, and $\epsilon \sim N(0,1)$ is the independent random part for each parameter.

2.5.2 Transistor-Level Models under Variations

For transistors with process variations, the Weibull slope β of the time-to-breakdown distribution is a linear function of oxide thickness [5, 65]:

$$\beta_i = \beta_0 + c\ \Delta T_{ox}^{(i)} = \beta_0 + c\ \mathbf{k}_{Ti}^{\mathbf{T}}\mathbf{e} \qquad (2.22)$$

where β_i stands for the Weibull slope for transistor i and β_0 denotes the nominal value. The T_{BD} distribution of i^{th} NMOS transistor under process variation has the same form as (2.3), with β replaced by β_i. Its area, $a_i = W_i L_i$, is a product of two correlated Gaussians.

The post-breakdown behavior model is also updated to capture the natural randomness of the breakdown resistance, as indicated in Fig. 2.3(a). The variational models of breakdown resistors, R_s and R_d, are modified to include the variations as follows,

$$\begin{aligned} R_s(x) = R_d(L-x) &= \begin{cases} ae^{bx}(1+\lambda_r\epsilon_r), & 0 \le x \le L_{\text{ext}} \\ kx(1+\lambda_r\epsilon_r), & L_{\text{ext}} \le x \le L \end{cases} \\ \epsilon_r &\sim N(0,1) \end{aligned}$$

This model is consistent with the variations shown in [2].

2.5.3 Cell-Level Analysis under Variations

Under process variations, the cell-level FP due to a NMOS HBD (taking the breakdown case in Fig. 2.5 for example) depends on the breakdown resistor and parameters of all transistors in involved driver cell m and load cell n. This dependence is modeled as a linear function of related parameters, using first-order Taylor Expansion. Thus the FP components defined in (2.7) are updated as

$$\begin{aligned} p &= p_0 + d_r^0\lambda_r\epsilon_r + \sum_j \left(d_{W_j}^0 \Delta W_j + d_{L_j}^0 \Delta L_j + d_{T_j}^0 \Delta T_j \right), \qquad (2.23) \\ p &\in \{p_s^{\text{dr}}, p_s^{\text{ld}}, p_d^{\text{dr}}, p_d^{\text{ld}}\} \end{aligned}$$

Here, d_x^0 is the first-order Taylor coefficients on parameter x. These coefficients are obtained using sensitivity analysis for the cell-level FP characterization, and ΔW_j, ΔL_j and ΔT_j are random variables that can be expressed in the form in (2.21). Since the FP component p is a linear combination of these process parameters and ϵ_r, it can also be expressed with vector $\mathbf{e}$,

$$\begin{aligned} p &= p_0 + \mathbf{k}_p^{\mathbf{T}}\mathbf{e} + d_r^0\lambda_r\epsilon_r, \qquad (2.24) \\ p &\in \{p_s^{\text{dr}}, p_s^{\text{ld}}, p_d^{\text{dr}}, p_d^{\text{ld}}\} \end{aligned}$$

Note that ϵ_r is the Gaussian representing the randomness of R_{BD}, and is independent of the elements in $\mathbf{e}$.

Using (2.6), (2.8), and (2.24) we can obtain the source-side and drain-side failure probabilities using analytical methods. This involves applying the max operation on correlated Gaussian variables. The work in [66] provided a solution for this max function and approximated the result as a Gaussian in the same random space $\mathbf{e}$. Using such an approach, the final FP for case (m,n,k) is calculated by (2.8) as the sum of two Gaussian variables, and has the form of

$$\mathrm{Pr}^{(i)}_{(\mathrm{fail|BD})} = \mathrm{Pr}^{(m,n,k)}_{(\mathrm{fail|BD})} = \mathrm{Pr}^{(i)}_{0} + \mathbf{k}^{\mathbf{T}}_{\mathrm{Pr}^{(i)}}\mathbf{e} + d_i\epsilon_{r_i} \tag{2.25}$$

The details of the calculation of failure sensitivities d^0_x's in (2.23) are given in Appendix B. The characterization and calculation process still maintains linear complexity to the size of library and circuit.

2.5.4 Circuit-Level Analysis under Variations

Based on the nominal analysis result (2.16) of circuit FP, we can derive the following under a statistical model:

$$\exp(W) \quad = \sum_{i\in\mathrm{NMOS}} \left(\frac{\gamma_i t}{\alpha}\right)^{\beta_i} \mathrm{Pr}^{(i)}_{(\mathrm{fail|BD})} a_i \tag{2.26}$$

Note that $(\frac{t}{\alpha})^{\beta_i}$ is no longer a common factor of the RHS expression due to the device-dependent β_i. Next we define y_i for each NMOS device i as following

$$\begin{aligned} \exp(W) &= \sum_i \exp(y_i) && (2.27)\\ \text{where} \;\; y_i &= \beta_i \ln\left(\frac{\gamma_i t}{\alpha}\right) + \ln\left(\mathrm{Pr}^{(i)}_{(\mathrm{fail|BD})} a_i\right) && (2.28)\\ &= \beta_i \ln\left(\frac{\gamma_i t}{\alpha}\right) + \ln \mathrm{Pr}^{(i)}_{(\mathrm{fail|BD})} + \ln W_i + \ln L_i \end{aligned}$$

Under process variations, for the i^{th} NMOS transistor, β_i is a Gaussian in random space $\mathbf{e}$ as shown in (2.22); $\mathrm{Pr}^{(i)}_{(\mathrm{fail|BD})}$ is a Gaussian in space $\mathbf{e} \cup \epsilon_{r_i}$ as in (2.25); W_i and L_i are also Gaussians in space $\mathbf{e}$ as assumed in Section 2.5.1.

We use two approximations to compute the FP. First, the above logarithms are approximated Gaussians using moment-matching (see Appendix C). As shown in our simulation results section, that approximation does not hurt the final result. Since $\mathrm{Pr}^{(i)}_{(\mathrm{fail|BD})}$ contains an additional random basis ϵ_{r_i} for breakdown resistor variation, the sum of the logarithms S_i will contain both $\mathbf{e}$ and ϵ_{r_i}. Denoting $\mathbf{k}_{S_i}$ and q_i as the coefficients for these two parts, and μ_{S_i} as the mean of S_i,

$$S_i = \ln \mathrm{Pr}^{(i)}_{(\mathrm{fail|BD})} + \ln W_i + \ln L_i = \mu_{S_i} + \mathbf{k}^{\mathbf{T}}_{S_i}\mathbf{e} + q_i\epsilon_{r_i} \tag{2.29}$$

Therefore y_i can be expressed as a Gaussian using $\mathbf{e}$ and ϵ_{r_i}. Denoting $F_i = \ln(\gamma_i t/\alpha)$ and substituting (2.28) with (2.29),

$$\begin{aligned} y_i &= \beta_i \ln\left(\frac{\gamma_i t}{\alpha}\right) + S_i \\ &= \beta_{i0} F_i + \mu_{S_i} + (cF_i \mathbf{k}_{T_i} + \mathbf{k}_{S_i})^{\mathbf{T}} \mathbf{e} + q_i \epsilon_{r_i} \qquad (2.30) \end{aligned}$$

which means that y_i is also a Gaussian expressed in terms of $\mathbf{e}$ and ϵ_{r_i}, and $\exp(y_i)$ will have a lognormal distribution. Note that y_i is the Weibull-scale FP corresponding to the HBD of i^{th} NMOS transistor.

From (2.27), $\exp(W)$ is the sum of correlated lognormal RVs. In the second approximation, we model this sum as a lognormal using Wilkinson's method [67], and its first two moments, u_1 and u_2, are[6]

$$\begin{aligned} u_1 &= \sum_i \exp\left(\mu_{y_i} + \sigma_{y_i}^2/2\right) \qquad (2.31) \\ u_2 &= \sum_i \exp\left(2\mu_{y_i} + 2\sigma_{y_i}^2\right) + 2\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} e^{\mu_{y_i}+\mu_{y_j}} e^{\frac{1}{2}(\sigma_{y_i}^2+\sigma_{y_j}^2+2r_{ij}\sigma_{y_i}\sigma_{y_j})} \end{aligned}$$

When $\exp(W)$ is small enough, using a first-order Taylor expansion, we find from (A.5) that

$$\begin{aligned} \mathrm{Pr}_{\mathrm{fail}}^{(\mathrm{ckt})} &= 1 - \exp\left(-\exp(W)\right) \qquad (2.32) \\ &\approx 1 - (1 - \exp(W)) = \exp(W). \qquad (2.33) \end{aligned}$$

This result indicates that, when the circuit FP $\mathrm{Pr}_{\mathrm{fail}}^{(\mathrm{ckt})}$ is small (which is actually the case we are interested in, since a circuit with a very large number of HBDs is unlikely to be functional), it can be approximated with $\exp(W)$, which has lognormal distribution with the first two moments given in (2.31). When $\mathrm{Pr}_{\mathrm{fail}}^{(\mathrm{ckt})}$ is large, its distribution is unknown, but the mean and variance still can be calculated using a numerical method based on (2.32). Using the distribution function, it is possible to predict the circuit FP at given time t with any specific confidence (e.g. 99%).

[6] The calculation of u_2 requires the covariance of y_i and y_j. When the HBD case for NMOS i also involves NMOS j (i.e., j belongs to cell m or n) or vice versa, the random parts ϵ of y_i and y_j are actually correlated since they contain process parameters from the same transistor(s). This kind of case is fairly rare (about $2/N$ for a circuit with N logic cells), hence the correlations of the random parts are omitted to simplify the computation.

The result also shows that the circuit-level mean-time-to-failure under process variation is no longer a strict Weibull distribution, since the $\sigma_{y_i}^2$ in (2.31) brings second order term $\ln^2 t$. Although this observation is based on approximations, it is confirmed by simulation results.

Due to the process variations, the mean value of circuit FP is increased by the $\sigma_{y_i}^2$ terms in (2.31). The variance $(u_2 - u_1^2)$ also increases with larger $\sigma_{y_i}^2$. This verifies that process variations exaggerate the likelihood of circuit failure. Moreover, u_2 contains the term r_{ij} which depends positively on the spatial correlation. This means higher spatial correlation will increase the variance of FP, thus elevating the reliability issue.

The calculation of y_i in (2.30) has $O(1)$ complexity due to the limited number of involved devices and principal components. Using the recursive technique proposed in [68], the sum operation over N lognormal variables in (2.27) can be computed as $N-1$ sum operations on two lognormal variables, keeping the computational complexity at $O(N)$.

2.6 Experimental Results

The proposed methods for circuit oxide reliability analysis were applied to the ISCAS85 and ITC99 benchmark circuits for testing. The circuits were synthesized by ABC [69] using the Nangate 45nm open cell library [70], and then placement was carried out using a simulated annealing algorithm. The cell-level library characterization was performed using HSPICE simulation and 45nm PTM model [63]. The circuit-level analysis was implemented in C++ and tested on a Linux PC with 3GHz CPU and 2GB RAM. The parameters for unit-size device the Weibull distribution are $\alpha = 10000$ (arbitrary unit) and $\beta = 1.2$ [5].

2.6.1 Results for Nominal Failure Analysis

Three methods for calculating the circuit FP are implemented using a C++ program: (a) Method 1 (M1) performing device-by-device calculation (Equation (A.1)); (b) Method 2 (M2) using our closed-form formula (Equation (A.4)); and (c) Monte Carlo (MC) simulation. The implementations of M1 and M2 assume signal independence when computing the stress coefficients, while this is factored into the MC simulation. The

MC simulation, performed for each of the time samples, consists of two parts: one, in which the jpmf (see Section 2.2.1) for each transistor stressed in mode A is computed, using 10000 randomized input vectors, and a second, where the breakdown transistors and x_{BD} are randomly generated for 5000 sample circuits, and the probability of circuit failure is computed. For computational efficiency, a biased Monte Carlo technique is utilized to help the verification for very low circuit FP situations.

Table 2.1: Runtime and error comparison for different methods and different benchmarks, as well as the lifetime relaxations.

Circuit	Size	MC Runtime		Method 1 (M1)		Method 2 (M2)		Lifetime
Name	(#Cells)	jpmf	Breakdown	Runtime	$\mathrm{Err}_{\text{M1-MC}}$	Runtime	$\mathrm{Err}_{\text{M2-M1}}$	Relaxation
c432	221	0.39s	9.11s	0.21s	2.37%	10ms	7.51e-5	5.48×
c880	384	0.74s	18.7s	0.34s	2.30%	10ms	2.87e-5	5.50×
c1355	596	1.02s	31.3s	0.29s	2.22%	10ms	2.62e-5	5.34×
c2670	759	1.41s	36.2s	0.83s	3.08%	30ms	2.70e-5	6.16×
c3540	1033	2.55s	67.2s	1.43s	2.21%	60ms	1.27e-5	5.58×
c5315	1699	3.45s	93.9s	1.17s	1.37%	40ms	8.93e-6	5.48×
c6288	3560	17.6s	398s	3.52s	1.74%	130ms	2.93e-6	5.40×
c7552	2316	6.12s	127s	1.69s	1.49%	60ms	5.07e-6	5.29×
b14	4996	35.5s	985s	6.40s	2.81%	250ms	2.09e-6	5.30×
b15	6548	53.3s	2251s	8.53s	1.93%	340ms	2.31e-6	4.83×
b17	20407	209s	8011s	26.6s	3.01%	1060ms	4.56e-7	4.83×
b20	11033	106s	3218s	13.3s	2.06%	530ms	8.01e-7	5.09×
b21	10873	103s	3126s	12.4s	1.69%	490ms	7.78e-7	5.01×
b22	14974	148s	4968s	16.3s	1.16%	650ms	6.34e-7	4.99×

Table 2.1 presents the detailed runtime and error comparisons for these methods and benchmarks, and shows the lifetime prediction of our method against that of the area-scaling method, as determined by (2.19). Here, $\mathrm{Err}_{\text{M1-MC}}$ is the error between methods M1 and MC, and $\mathrm{Err}_{\text{M2-M1}}$ is the error between methods M2 and M1. Both errors are measured as the average relative error of FP over a number of time samples. The comparison of M1 with MC shows the effectiveness of the proposed method and demonstrates that the signal independence assumption is appropriate for our benchmarks. The comparison between M2 and M1 validates the approximations made in the proof of Theorem 1. Runtime comparisons (circuit read-in time is not counted in) indicate that the proposed method reduces the runtime by 3 to 4 orders of magnitude, compared with MC. In summary, our new method M2 for circuit failure analysis in (A.4) is fast and accurate, and it gives a 4.8–6.2× relaxation in the predicted circuit lifetime, as against the traditional area-scaling method.

Fig. 2.8 visualizes the FP curves for benchmark c7552 which has 2316 cells, as well as the curves using traditional area-scaling and the curve for a unit-size device. The three methods, M1, M2 and MC yield very close results, and all degradation curves share the same Weibull slope. We show a significant relaxation in the circuit lifetime against traditional area-scaling.

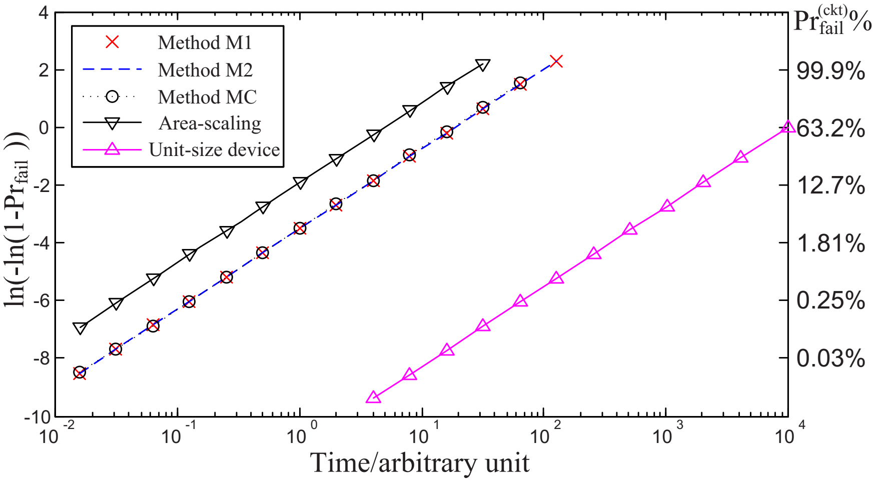

Figure 2.8: Result of benchmark circuit c7552 and comparison with traditional area-scaling method and unit-size device.

2.6.2 Results for Variation-Aware Failure Analysis

The process variation of T_{ox} is chosen so that its 3σ point is 4% of its mean [7], and is split into 20% of global variation, 20% of spatially correlated variation and 60% of random variation. The variation of W and L sets the 3σ point to 12% of the mean [71], and is split to 40% of global variation, 40% of spatially correlated variation and 20% of random variation. The correlation matrix uses the distance based method in [64]. The number of grids grows with the circuit size.

For each benchmark circuit, the mean and standard deviation of the failure probability are calculated at the time when the nominal circuit has a failure probability of 1%, using the proposed method and Monte Carlo (MC) simulation, separately. The MC simulation randomly generates 5000 circuit instances with different process parameters according to their distribution and correlation models: for each sample, we evaluate the FP by using the random value of the process parameters, and performing the nominal analysis described in Sections 2.2 to 2.4.

Table 2.2: Comparisons of the mean μ and σ of circuit failure.

Circuit	Size		Failure probability		Error to MC		Runtime		3σ lifetime	
Name	#Cells	#Grids	μ	$\frac{\sigma}{\mu}$	μ	σ	Proposed	MC	Nominal	AreaScaling
c432	221	4	1.018%	8.73%	0.18%	0.93%	1.32s	130s	-18.9%	5.23×
c880	384	9	1.024%	8.82%	0.87%	1.52%	1.88s	203s	-19.5%	5.26×
c1355	596	9	1.022%	8.97%	0.11%	0.69%	2.54s	207s	-19.6%	5.16×
c2670	759	16	1.023%	9.10%	0.64%	0.91%	5.70s	532s	-19.9%	5.94×
c3540	1033	16	1.023%	9.34%	0.41%	1.99%	8.16s	842s	-20.3%	5.36×
c5315	1699	25	1.025%	9.49%	0.79%	0.73%	7.75s	743s	-20.6%	5.25×
c6288	3560	64	1.028%	10.4%	0.81%	0.36%	22.7s	2210s	-22.2%	5.23×
c7552	2316	36	1.026%	9.75%	0.73%	0.88%	11.1s	1075s	-21.1%	5.07×
b14	4996	81	1.028%	10.1%	0.56%	1.14%	36.5s	3875s	-21.8%	5.16×
b15	6548	100	1.027%	10.2%	0.52%	0.61%	53.5s	5285s	-21.9%	4.76×
b17	20407	361	1.033%	11.3%	1.13%	0.76%	181s	16634s	-23.8%	4.71×
b20	11033	169	1.031%	10.7%	1.01%	2.75%	80.3s	8100s	-22.8%	4.93×
b21	10873	169	1.031%	10.6%	0.95%	1.32%	73.9s	7593s	-22.7%	4.87×
b22	14974	225	1.032%	10.9%	1.40%	2.65%	104s	10290s	-23.2%	4.84×

Table 2.2 presents the statistics of the circuit failure probability using the proposed method. The first three columns represent the circuit name and its characteristics. Information about the mean and standard deviation of the FP using our approach are presented in the next two columns, and the corresponding relative errors to MC in the following two. It can be seen that our approach closely matches MC, with average errors of 0.72% for the mean and 1.23% for the standard deviation. The value of the mean is very close to the nominal FP of 1%, but the standard deviation is considerable. The last two columns compare the circuit lifetime at FP=1% for our approach (using μ+3σ FP) with the nominal approach (using nominal FP) and the area-scaling method under variations (using μ+3σ FP), respectively. We see that the circuit lifetime decrease 19–23% due to process variation, and the proposed approach shows 4.7–5.9× lifetime relaxation against the pessimistic area-scaling method.

Fig. 2.9 plots the probability density function (PDF) and cumulative density function (CDF) of benchmark c7552 at the nominal failure probability of 1%. The dotted curves show results of MC simulation, while the solid curves show lognormal distribution obtained using proposed method. The nearly perfect match of these two methods validates the approximations made during the analysis, and demonstrates that the circuit FP has a lognormal distribution in the region of interest.

The proposed method is also tested with other process parameter variance and correlation data besides the condition assumed above. Table 2.3 shows the μ+3σ value

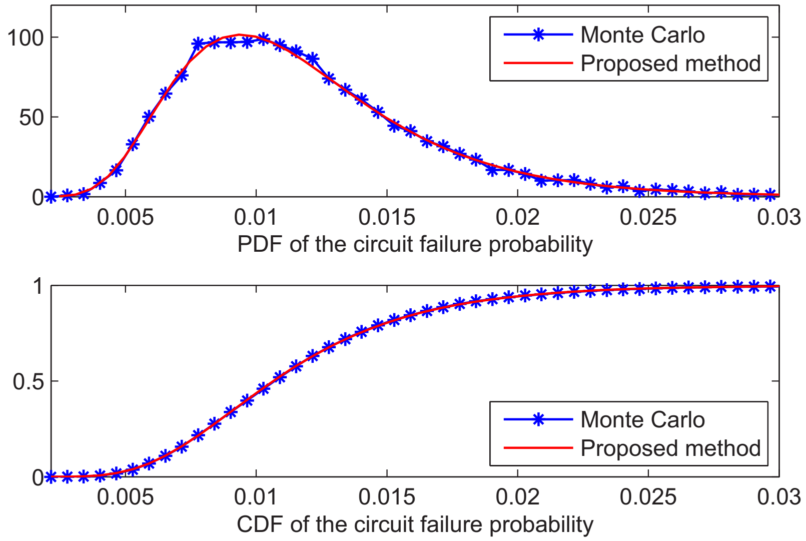

Figure 2.9: Comparison of the PDF and CDF of circuit failure.

of circuit failure when nominal circuit FP is 1%, and its relative error against MC simulation for benchmark c7552, under several process variation and spatial correlation conditions:

Table 2.3: Circuit failure of c7552 under different test conditions.

Process Variation	Less correlation $g/s/r$=10/10/80%		Medium correlation $g/s/r$=30/40/30%		More correlation $g/s/r$=50/40/10%	
W, L, T_{ox}	μ+3σ	Error	μ+3σ	Error	μ+3σ	Error
σ/μ=1%	1.13%	0.29%	1.23%	0.08%	1.27%	0.09%
σ/μ=2%	1.27%	0.37%	1.47%	0.06%	1.56%	1.60%
σ/μ=5%	1.89%	1.10%	2.48%	0.83%	2.75%	2.58%
σ/μ=10%	4.32%	1.23%	6.57%	3.07%	7.72%	6.23%

The labels g, s, r in the table stand for the global part, the spatially correlated part and the random part of the parameter variations. The results indicate that the relative error to MC simulation is small under all the test conditions, indicating the proposed method is accurate and robust to different conditions of process variations. Moreover, we observe that as the μ+3σ value of the FP increases when the process variation increases, or when the correlation increases. This verifies again that the process variations and spatial correlation elevate the reliability issues due to oxide breakdown.

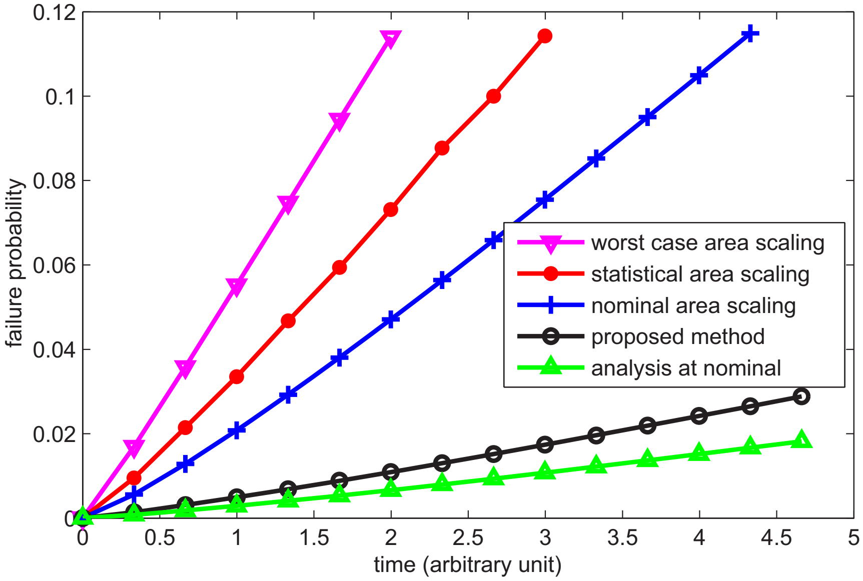

Figure 2.10: Comparison of circuit failure, as predicted by various methods, for c7552.

Finally, Fig. 2.10 shows a comparison of FP vs. time for benchmark c7552 using (a) area scaling with worst-case T_{ox}, (b) area scaling with the T_{ox} variation model in [6], (c) area scaling with nominal T_{ox}, (d) the variation-aware approach proposed in Section 2.5, (e) the analysis method using nominal process parameters as in Section 2.4. The μ+3σ FP value is used for (b) and (d). The figure leads to several important conclusions. First, it is clear that there are significant differences between area-scaling based methods and our approaches, and that the area-scaling methods are generally too pessimistic. Therefore, to accurately predict circuit reliability, it is essential to account for the inherent circuit resilience and process variations simultaneously. Second, it demonstrates that under either model, the nominal case provides optimistic estimates of the lifetime, and that it is essential to incorporate the effects of variations in order to obtain more accurate lifetime estimates.

2.7 Conclusion

The chapter has focused on the reliability issues caused by gate oxide breakdown in CMOS digital circuits, with the consideration of the inherent resilience in digital circuits that prevents every breakdown from causing circuit failure. The proposed approach takes account for the effective stress for HBD generation and the probability of circuit failure after HBD occurrences. The FP for large digital logic circuits is derived in closed form, and it is demonstrated that the circuit-level time-to-failure also follows Weibull distribution and shares the same Weibull slope with the unit-size device. Then the proposed failure analysis approach is extended to include the effect of process variations. The circuit FP at specified time instant is derived to be a lognormal distribution due the process variations, and this distribution expands as the process variations and spatial correlation increase. Experimental results show the proposed approaches are effective and accurate compared with Monte Carlo simulation, and give significant better lifetime predictions than the pessimistic area-scaling method.

Chapter 3

Optimization of Circuit Oxide Lifetime using Gate Sizing

In this chapter we use the analysis results in Chapter 2 to develop an optimization approach to mitigate the effect of gate oxide breakdown. We demonstrate that by appropriately sizing the devices, the circuit can be made more resilient, so that it performs correctly even in the presence of oxide breakdown events. We formulate a problem that performs transistor sizing with the aim of increasing the time to circuit failure, while addressing conventional sizing goals such as power and delay. Experimental results show that circuit reliability can be improved by increasing the area, which runs counter to the prediction of the traditional area-scaling theory.

3.1 Introduction

The circuit level failure analysis in Section 2.4 shows that for a circuit designed in a given technology, the FP is affected by the Weibull shift, W_{shift}, given by Equation (2.17).

We define the *lifetime* of a circuit as the time corresponding to a specified failure probability, W. In other words, this is the time at which the right hand side of Equation (2.16) evaluates to W. It is easy to show that under this failure probability, if the Weibull shift for a circuit is reduced from $W_{\text{shift}}^{(0)}$ to $W_{\text{shift}}^{(1)}$, then the impact on the circuit lifetime is given by the following exponential relationship:

$$\frac{t_1}{t_0} = \exp\left(\frac{1}{\beta}\left(W_{\text{shift}}^{(0)} - W_{\text{shift}}^{(1)}\right)\right) \tag{3.1}$$

Therefore by reducing the Weibull shift, it is possible to lower the FP and prolong the lifetime of the circuit.

We achieve this through gate sizing, a widely-used circuit optimization method that traditionally explores area/delay/power tradeoffs by sizing the logic cells in the circuit [72, 73]. The conventional gate sizing problem is commonly formulated as:

$$\begin{aligned} \text{minimize} \quad & \mathit{Area}(\mathbf{s}) \\ \text{subject to} \quad & \mathit{Delay}(\mathbf{s}) \leq D_{\max} \\ & \mathit{Power}(\mathbf{s}) \leq P_{\max} \end{aligned} \tag{3.2}$$

Here the optimization variable $\mathbf{s} = \{s_1, s_2, ..., s_j\}$ is the array of sizing factors for each logic cell in the circuit.

We will next demonstrate how the Weibull shift is affected by the sizes of the logic cells in the circuit in Section 3.2, and use it to build a framework for reliability-driven gate sizing in Section 3.3. Experimental results prove the effectiveness of the proposed model and method in Section 3.4, and conclusions are given in Section 3.5.

3.2 Modeling of the Weibull Shift

From Section 2.3, the cell-level FP, $\Pr_{(\text{fail}|\text{BD})}^{(i)}$, is obtained by analyzing the breakdown case of cells m and n (Figure 2.5). Therefore it depends on the sizes of these cells and can be represented as:

$$\Pr_{(\text{fail}|\text{BD})}^{(i)} = f(s_m, s_n), \tag{3.3}$$

where s_m and s_n are the sizing factors for cells m and n, i.e., the multiples of their sizes with respect to their nominal sizes. Clearly, the area of an NMOS transistor i, $a_i = s_n a_{i(\text{nominal})}$, and this depends on s_n. Therefore, we define a set of new functions $Q^{(i)}$ to include all the sizing-dependent elements in Equation (2.17):

$$Q^{(i)}(s_m, s_n) = \Pr_{(\text{fail}|\text{BD})}^{(i)} s_n = s_n f(s_m, s_n). \tag{3.4}$$

The Weibull shift of the circuit can be rewritten as

$$W_{\text{shift}} = \ln \sum_{i \in \text{NMOS}} Q^{(i)}(s_{m(i)}, s_{n(i)}) \gamma_i^{\beta} a_{i(\text{nominal})}, \tag{3.5}$$

where $n(i)$ $[m(i)]$ refers to the logic cell that contains [drives] the i^{th} NMOS transistor.

The computation of the $Q^{(i)}$ functions requires the calculation of FP $\Pr^{(i)}_{(\text{fail}|\text{BD})}$, which does not admit a simple closed form. Therefore, to find the $Q^{(i)}(s_m, s_n)$ function for each breakdown case, we perform SPICE-based analysis as a numerical alternative. For each case $i \to (m, n, k)$, the $Q^{(i)}(s_m, s_n)$ function is computed with a set of sampled s_m and s_n values and stored in a look-up table during library characterization. The plot in Figure 3.1 shows an example of the Q function corresponding to the breakdown case shown in Figure 2.5, obtained using a SPICE simulation.

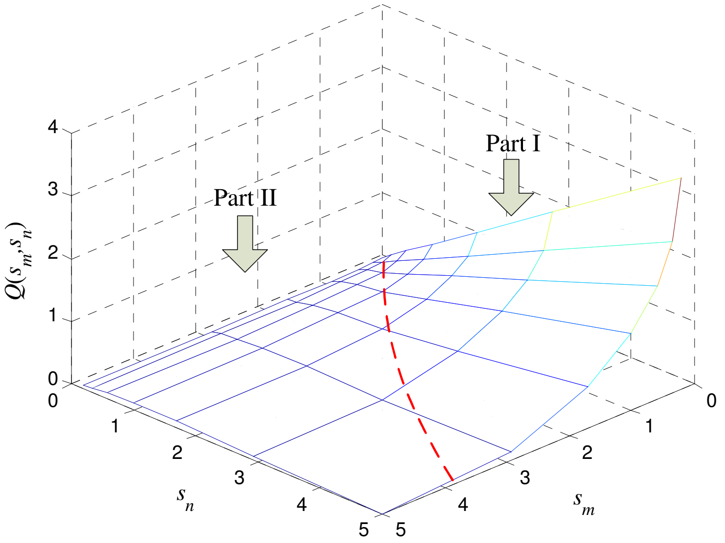

Figure 3.1: A plot of a representative Q function.

The plot shows that the Q function can be divided into two parts, as labeled in the figure. Within Part I, Q increases significantly as s_m decreases and s_n increases, while in Part II, the Q function is "flat" and has a small value. To achieve lower Q values, and hence lower Weibull shifts, the sizing point should be kept in the Part II. On closer examination, it can be verified that the Q function is not convex; however, we find that it can be approximated very well by a generalized posynomial function, and we will exploit this idea.

3.3 Reliability-Driven Gate Sizing

In order to take circuit failure into consideration, we can add a new constraint for the Weibull shift to the sizing problem, to limit the shift in the Weibull curve, $W_{\text{shift}} \leq W_{\max}$, where $W_{\max}$ denotes the maximum acceptable Weibull shift under a circuit lifetime spec.

The conventional gate sizing problem is usually solved using geometric programming (GP) [74], in which the objective and constraints are modeled using posynomials, and the problem is then transformed to a convex optimization problem and solved by standard solvers. However the Weibull shift function, a weighted sum of Q functions of all transistors, cannot be directly represented as a posynomial of the sizing factors. To address this problem and adapt the Weibull shift constraint into the GP framework, an empirical generalized posynomial fit for the Q functions is proposed:

$$\begin{aligned} & Q_{fit} = \max(Q_{f1}, Q_{f2}) - q, \\ \text{where} \quad & Q_{f1} = c_1 \left(\frac{s_n}{s_m}\right)^{b_1}; \; Q_{f2} = c_2 \left(\frac{1}{s_m}\right)^{b_2} + d; \\ & b_1, b_2, c_1, c_2, d, q \geq 0. \end{aligned} \tag{3.6}$$

Here Q_{fit} is the maximum of two posynomial functions, Q_{f1} for the higher side (Part I in Figure 3.1), and Q_{f2} for the lower side (Part II in Figure 3.1). Experimental results show a 5.82% average relative error of fitting for the tested library in Section 2.6.1. Since all fitting parameters are non-negative, $Q_{fit} + q$ is a generalized posynomial.

Based on the proposed model, we define intermediate variables $Q_m = \max(Q_{f1}, Q_{f2})$ to ensure the posynomial property, and use Q_{fit} to replace Q in Equation (3.5) to obtain

$$\exp(W_{\text{shift}}) = \sum_{i \in \text{NMOS}} Q_m^{(i)} \gamma_i^{\beta} a_{i(\text{nominal})} - \sum_{i \in \text{NMOS}} q_i \gamma_i^{\beta} a_{i(\text{nominal})}. \tag{3.7}$$

The constraint $W_{\text{shift}} \leq W_{\max}$ can now be rewritten as

$$\begin{aligned} \sum_{i \in \text{NMOS}} Q_m^{(i)} \gamma_i^{\beta} a_{i(\text{nominal})} & \leq \exp(W_{\max}) + \sum_{i \in \text{NMOS}} q_i \gamma_i^{\beta} a_{i(\text{nominal})}, \\ Q_{f1}^{(i)}/Q_m^{(i)} & \leq 1, \quad i \in \text{NMOS}, \\ Q_{f2}^{(i)}/Q_m^{(i)} & \leq 1, \quad i \in \text{NMOS}. \end{aligned} \tag{3.8}$$

Note that all right hand sides above are constants, and these constraints are in posynomial form and can directly be applied to the conventional sizing problem in Equation (3.2). The new problem, containing the Weibull shift constraints, can be solved by traditional GP solvers.

Due to the nonconvex property of the original Weibull shift function, it is difficult to find the global optimum of the sizing problem. The newly proposed posynomial fit for Q functions adjusts the search space to a convex set, with minimal loss in accuracy. Thus the global optimum of the modified problem can be regarded as a close approximation for the solution of the original problem.

3.4 Experimental Results

For reliability-driven gate sizing, we work with a library that is characterized by calculating the Q function with sampled sizing factors for all breakdown cases, and then fitting the Q functions using Matlab for each case. For the total of 119 cases, there is a 5.82% average relative error of fitting (RMSE divided by the maximum of Q function, then averaged for all cases).

The benchmark circuits were initially mapped to the library consisting of the nominal-size logic cells. Then we use transistor area and delay models consistent with [73], and Mosek [75] Optimization Toolbox for Matlab as the GP solver.

To verify the usefulness of gate sizing for reliability, each of the benchmark circuits is first optimized for delay without a W_{shift} constraint, to obtain the minimum delay d_0 and corresponding area a_0. The corresponding unoptimized value of W_{shift} at this minimum delay point is shown in the second column of Table 3.1. The circuit is then optimized to minimize W_{shift}, subject to a delay constraint of 1.1× d_0 and an area constraint of a_0. The solution, listed in the third column, shows the W_{shift} improvement at the cost of 10% more delay. The fourth column lists the corresponding lifetime improvement calculated using Equation (3.1). For the fifth column, the area constraint is loosened to $2a_0$, for further improvement of W_{shift}, and the corresponding lifetime improvement is provided in the last column. Over all tested benchmarks, the results show 1.1–1.5× lifetime improvement when the delay constraint is relaxed to 1.1× of the minimum delay, and another 1.2–1.9× improvement when an additional 2× area is allowed.

Table 3.1: Lifetime improvement by gate sizing.

Circuit Name	W_{shift} at min delay d_0, a_0 (I)	min W_{shift} $D \leq 1.1d_0$ $A \leq a_0$ (II)	Lifetime Improve II vs. I	min W_{shift} $D \leq 1.1d_0$ $A \leq 2a_0$ (III)	Lifetime Improve III vs. II
c432	6.01	5.52	1.50×	5.12	1.40×
c880	5.98	5.83	1.13×	5.27	1.59×
c2670	7.02	6.80	1.20×	6.45	1.33×
c3540	7.47	7.14	1.31×	6.76	1.37×
c5315	7.91	7.66	1.23×	7.40	1.24×
c6288	7.95	7.67	1.26×	6.92	1.87×
c7552	8.23	8.06	1.16×	7.81	1.23×

As a typical gate sizing example, Figure 3.2 presents the area vs. Weibull shift trade-off curves under different delay constraints for benchmark circuits c880, c2670, and c3540. The triangle points in the plot indicate the area a_0 and W_{shift} at minimum delay for each circuit. Two curves under different delay constraints are plotted for each circuit. The x-axis shows both W_{shift} and the absolute lifetime when circuit FP = 5%. The figure shows that the circuits sized for minimum delay generally have the worst lifetime values, i.e., the triangles are to the right of the curves, and by loosening the delay and/or area constraints, the lifetime can be improved.

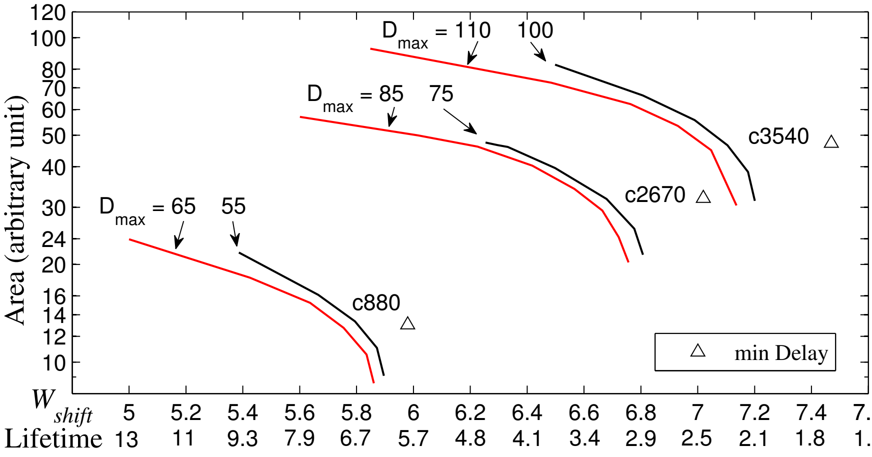

Figure 3.2: The area vs. Weibull shift trade-off curves.

We have shown that *circuit reliability can be improved by increasing the area, which runs counter to the prediction of the traditional area-scaling theory* of Equation (2.18), which claims higher FP for larger circuit size. This apparent contradiction can be explained by seeing that larger sizes make the gates more resilient and prevent logic failures even in the presence of breakdown current. This causes the failure regions in Figure 2.6 to shrink, counteracting the tendency of larger areas to be susceptible to more failures.

3.5 Conclusion

This chapter applies the failure probability analysis result in Section 2.4 to the development of an optimization approach to improve the reliability against oxide breakdown. A novel Weibull shift model is proposed and applied to the conventional GP-form gate sizing problem, and it is proved effective by experiments. It is shown that circuit reliability can be improved by increasing the sizes of transistors, which runs counter to the prediction of the traditional area-scaling theory.

Chapter 4

Incorporating Hot Carrier Injection Effects into Timing Analysis

This chapter focuses on hot carrier (HC) effects in modern CMOS technologies and proposes a scalable method for analyzing circuit-level delay degradations in large digital circuits, using methods that take abstractions up from the transistor level to the circuit level. We begin with an exposition of our approach for the nominal case. At the transistor level, a multi-mode energy-driven model for nanometer technologies is employed. At the logic cell level, a methodology that captures the aging of a device as a sum of device age gains per signal transition is described, and the age gain is characterized using SPICE simulation. At the circuit level, the cell-level characterizations are used in conjunction with probabilistic methods to perform fast degradation analysis. Next, we extend the nominal-case analysis to include the effect of process variations. Finally, we show the composite effect of these approaches in the presence of other aging variations, notably bias-temperature instability (BTI), and study the relative impact of each component of aging on the temporal trends of circuit delay degradations. The analysis approaches for nominal and variational cases are both validated by Monte Carlo simulation on various benchmark circuits, and are proven to be accurate, efficient and scalable.

4.1 Introduction

Hot carrier (HC) effects in MOSFETs are caused by the acceleration of carriers (electrons or holes) under lateral electric fields in the channel, to the point where they gain high enough energy and momentum (and hence they are called *hot* carriers) to break the barriers of surrounding dielectric, such as the gate and sidewall oxides [8]. The presence of hot carriers triggers a series of physical processes that affects the device characteristics under normal circuit operation. These effects cumulatively build up over prolonged periods, causing the circuit to age with time, resulting in performance degradations that may eventually lead to circuit failure.

The phenomenon of HC effects is not new: it was a significant reliability issue from the 1970s to the 1990s, when circuits operated under high supply voltages (2.5–5V), which led to a high lateral electric field in the MOSFET channel. The effects of HCs were mitigated by the introduction of special process techniques such as lightly doped drains (LDDs). The traditional theory of HC mechanisms was based on a field-driven model, in which the peak energy of carriers (electrons or holes) is determined by the lateral field of the channel [9]. This was based on the theory of the so-called lucky electron model, capturing the confluence of events due to which an electron is "lucky" enough to do damage – to gain energy from the channel field, to be redirected towards the silicon/oxide interface, and to avoid energy-robbing collisions along the way.

Extrapolating this theory, it was expected that at today's supply voltages, HC effects would almost disappear as carriers cannot gain enough energy when the electric field is reduced to these levels. However, experimental evidence on nanoscale technologies shows that this is not true, and hot carrier degradation remains significant for MOSFETs at lower voltages [10]. Moreover, these issues are projected to worsen in future generations of devices.

The rate of hot carrier generation increases with time t as $t^{1/2}$. Since the multiplicative constant for this proportionality is relatively small, in the short-term, HC effect is overshadowed by bias-temperature instability (BTI) effects, which increase as t^n, for $n \approx 0.1$–0.2, but with a larger constant multiplier. However, in the long term, the $t^{1/2}$ term dominates the t^n term, making HC effects particularly important for devices in the medium to long term. It has been shown in [1], for example, that HC effects can contribute to 40-80% of all aging after 10 years of operation. Therefore, HC effects are a significant factor in the short term and are dominant in applications with longer lifetimes, such as embedded/automotive applications and some computing applications.

Recently, newer energy-driven theories [11–13] have been introduced to overcome the limitations of the lucky electron model, and to explain the mechanism of carriers-induced degradation for short-channel devices at low supply voltages. These theories have been experimentally validated on nanometer-scale technologies. The energy-driven framework includes the effects of electrons of various levels of energy, ranging from high-energy *channel hot carriers* (CHCs) to low-energy *channel cold carriers* (CCCs). Under this model, injection is not necessary for the device degradation, and carriers with enough energy can affect the Si–SiO_2 interface directly. However, much of the published circuit-level work on HC effects is based on the lucky electron model, which is effectively obsolete.

Existing work on HC degradation analysis of digital circuits can be divided into to two categories. The first is based on device-level modeling/measurement tied to circuit-level analysis, including [14], commercial software such as Eldo using computationally-intensive simulations, and [1], which predicts the lifetime of a ring oscillator using measured data. While these methods are flexible enough to accept new models and mechanisms, they are not scalable for analyzing large circuits.

Methods in the second category are based on a circuit-level perspective, using statistical information about device operation to estimate the circuit degradation. In [15], a hierarchical analysis method for delay degradation, based on a simple device-level HC model, was proposed. The work in [16] defined a duty factor to capture the effective stress time for HC effects, which assumes constant HC stress during signal transitions and models the duty factor to be proportional to the transition time. The characterization of HC degradation is performed in the device level and only considers the switching transistors, with other transistors in the stack ignored. Then signal probability (SP) and transition density (TD) is utilized for aging analysis. While these works are usually efficient and scalable to large digital circuits, they use over-simplistic models for device aging and cell characterization, and therefore cannot achieve the high accuracy provided by methods in the first category, especially for nanometer-scale technologies. Extending these methods to energy-driven models, including CHC and CCC, is highly nontrivial, and is certainly not a simple extension.

Beyond the issue of using better modeling techniques for analyzing the nominal case, it is also important to consider the effects of process variations, which significantly affect circuit timing in digital circuits [17] in current and future technologies. Since HC effects are closely dependent on the circuit operation and device stress conditions, they is also affected by process variations. The interaction between HC effects and process variations has gained increasing attentions in recent years. However, most of the published works only focus on device-level analysis [18, 19] or small-scale digital circuit [20], and the proposed methods are usually based on LEM model with HSPICE or Monte Carlo simulation, and are not scalable to large digital circuits.

This chapter provides a third path for CHC/CCC degradation analysis for large digital circuits, and makes the following contributions:

- Instead of using a simple empirical degradation model [15], or a device model assuming constant stress during transition [16], our method is built upon the newer multi-mode energy-driven degradation model [12, 13].

- It introduces the novel concept of age gain (AG) to capture the amount by which a transistor ages in each signal transition, and develops a quasistatic approach for accurate analysis of AG.

- It performs cell-level characterization of AG, in which the AGs of all transistors in a logic cell corresponding to a signal transition event is computed simultaneously, instead of only considering switching transistors [16].

- It utilizes signal statistics, leveraged from techniques for power estimation, to perform circuit-level degradation analysis. The multilevel hierarchy of modeling and analysis enables both high accuracy and great scalability of the proposed approach.

- It demonstrates that the circuit delay degradation has a slight but negligible deceleration effect due to the degradation of signal transition, in contrast to the significant acceleration effect predicted in [16]. The work in [16] assumes HC aging to be proportional to the transition time, which increases with aging; however, this is not entirely accurate since slower transition times excite fewer energetic carriers and actually cause *less* damage.

- The proposed approach for circuit degradation analysis using the energy-driven model is extended at the cell-level modeling and circuit-level analysis to incorporate the impact of process variations on both device aging and circuit timing. The variation-aware circuit degradation analysis is fitted into the statistical static timing analysis (SSTA) framework, with good accuracy and scalability.

Our work bridges the wide chasm between the tremendous advances at the device level with the much simpler models that are currently used at the circuit level. Our approach maintains accuracy and scalability at all levels of design, employing accurate modeling and characterization at the device and cell levels, and a scalable algorithm at the circuit level. We begin with an approach for analyzing the nominal case, neglecting the effects of variations. At the *transistor level*, we employ the energy-driven model for device aging [12, 13], as outlined in Section 4.2. At the *logic cell level*, we characterize (offline) the device age gain per signal transition for cells within a library using SPICE simulations, as described in Section 4.3. At the *circuit level*, the signal probability and activity factor are utilized to perform fast degradation analysis, based on the cell-level characterization, as explained in Section 4.4. Next, we extend the engines developed above to include the impact process variations on both HC aging and circuit timing, as discussed in details in Section 4.5. The proposed analysis approaches for both nominal and variational cases are validated by Monte Carlo simulation on various benchmark circuits, and is demonstrated in Section 4.6 to be accurate, efficient and scalable. This chapter ends by presenting concluding remarks in Section 4.7.

As in other work considering hot and cold carriers, we refer to the CHC/CCC problem under all energy modes as "hot carrier"/"HC" degradation, but it is implicit that the CCC case is also included.

4.2 CHC/CCC Aging: Device Models

4.2.1 Traditional Mechanisms

The traditional lucky electron model for HC degradation was based on *direct electron excitation* (DEE), i.e., the theory of impact ionization and interface trap generation due to broken Si–H bonds [8], based on a set of chemical reactions. Let us denote the silicon-hydrogen bonds at the surface as $\equiv \mathrm{Si}_s\mathrm{H}$, where the subscript s denotes the surface, i.e., the oxide-substrate interface, with three other bonds ("$\equiv$") connected to other silicon atoms in the substrate, One of the reactions that causes HC injection involves trap generation by electrons (e^-) that breaks the silicon-hydrogen bond, i.e.,

$$\equiv \mathrm{Si}_s\mathrm{H} + e^- \quad \rightarrow \quad \mathrm{Si}^* + \mathrm{H} \tag{4.1}$$

Another is related to trap generation by electrons and holes (h^+) as they interact with hydrogen atoms (H) and molecules (H_2), i.e.,

$$\begin{aligned} e^- + h^+ + \mathrm{H}_2 \quad &\rightarrow \quad 2\mathrm{H} \\ \equiv \mathrm{Si}_s\mathrm{H} + \mathrm{H} \quad &\rightarrow \quad \mathrm{Si}^* + \mathrm{H}_2 \end{aligned} \tag{4.2}$$

It is also possible for holes to break the $\equiv \mathrm{Si}_s - \mathrm{H}$ bound.

4.2.2 Energy-driven Mechanisms

From an energy perspective, hot electrons change the distribution of the electron energy distribution function (EEDF). The expression

$$\text{Rate } = \int f(E)S(E)dE \tag{4.3}$$

describes the hot carrier rate, where f, the EEDF, and S, the interaction cross section or scattering rate, are functions of energy E. It has been shown that the dominant energies associated with this integrand are at a set of "knee" points in either f or S. There are four major mechanisms that affect the above rate [11, 12]:

- In the field-driven paradigm of the lucky electron model, f has no significant knee, and the dominant energies are driven by the S function. This is the first mechanism, and its effect is decreasing in nanometer-scale technologies.

- In addition, there are knees in the EEDF beyond the range of the lucky electron model. It has been shown that the EEDF has a significant knee at the point at which there is a steep potential drop at the drain, corresponding to the potential from the drain to the channel pinch-off point, V_{EFF}, and a second knee is seen at about $2V_{EFF}$ due to *electron-electron scattering* (EES).
- The third mechanism, linked to high-energy carriers, is through *single vibrational excitation* (SVE) due to energy transfer to the phonon system, adding to energy from lattice vibrations.
- Finally, there is evidence that the bonds may be broken by *channel cold carrier* (CCC) effects, through a fourth mechanism corresponding to *multiple vibrational excitation* (MVE). This corresponds to direct excitation of the vibrational modes of the bond by multiple carrier impacts, each of which individually have low energy, but which can cumulatively break the bond [76]. MVE degradation is strongly correlated to the current, i.e, the number of electrons “hitting” the bond per second.

The energy-driven theory for HC generation [12] uses quantum-mechanical models to explain the process of carriers gaining energy, through three different mechanisms: (1) *High-energy channel hot carriers* based on direct electron excitation (DEE), consistent with the Lucky Electron Model (LEM), and on the SVE mechanism, (2) *Medium energy electrons* based on the EES mechanism, and (3) *Channel cold carriers* based on the MVE mechanism, which creates lower-energy carriers that cause degradations.

4.2.3 Device Aging Model

The degradations of the saturation drain current, $\Delta I_{\text{on}}/I_{\text{on}}$, of a transistor due to HC effects follow a power model [13]:

$$(\Delta I_{\text{on}}/I_{\text{on}})_j = A(\text{age}_j)^n \tag{4.4}$$

The exponent n is widely accepted to be 0.5 over a range of processes. The value of A can be obtained from device-level experiments, e.g., from the plots in [13]. The age function of a MOSFET is given by

$$\text{age} = t/\tau = R_{\text{it}}t \tag{4.5}$$

where R_{it} can be interpreted as the *rate of aging* over time, and corresponds to the rate of interface trap generation. The quantity τ is its inverse and is referred to as the device lifetime. Over the years, considerable effort has been expended in characterizing R_{it} at the device level. Under the classical field-driven LEM scenario, this has the form:

$$R_{\mathrm{it(LEM)}} = \frac{1}{\tau} = K \left(\frac{I_{ds}}{W}\right) \left(\frac{I_{bs}}{I_{ds}}\right)^{m} \tag{4.6}$$

The more accurate multi-mode energy-driven model for HC degradation for fine-geometry CMOS devices changes this to [13]:

$$\begin{aligned} R_{\mathrm{it}} = \frac{1}{\tau} \quad &= \quad C_1 \left(\frac{I_{ds}}{W}\right) \left(\frac{I_{bs}}{I_{ds}}\right)^{m} + C_2 \left(\frac{I_{ds}}{W}\right)^{a_2} \left(\frac{I_{bs}}{I_{ds}}\right)^{m} \\ &\quad + C_3\, V_{ds}^{\frac{a_3}{2}} \left(\frac{I_{ds}}{W}\right)^{a_3} \exp\left(\frac{-E_{emi}}{k_B T}\right) \end{aligned} \tag{4.7}$$

The three terms in the expression correspond to degradation in the high-energy mode (corresponding to LEM), the medium-energy mode, and through channel cold carriers, respectively.

The relation between R_{it} and I_{ds}/W in Equation (4.6) is linear, and experimental data [12, 13] show that this is grossly incorrect. The nonlinear model in Equation (4.7) shows excellent fits to experimental measurements, and therefore our analysis is based on this model.

HC degradation has positive dependence on temperature, and a corner-based approach with worst-case temperature is used in this work. If more information about thermal characteristics is available, this model can easily be extended.

4.3 Cell-level Characterization

The device-level models outlined in the previous section provide a means for computing the aging due to CHC/CCC effects. To determine their effects on the circuit, our approach begins by building a cell-level characterization technique for the standard cell library that computes the delay drift over time. The remainder of this section describes the precharacterization method: we begin by determining the aging effect on each transistor of a library cell, and then compute its effect on the cell delay.

4.3.1 Transistor Age Gain Per Transition

For most of the time during the operation of a digital circuit, the MOS transistors in the circuit are in off or triode state, where there is minimal HC degradation. The period during which there is a sufficient number of carriers in the channel, with various levels of energy, corresponds to only the active (switching) state, and it is sufficient for only this state to be considered in analyzing HC degradation at the transistor level.

Therefore, HC aging does not occur over all time, and the defect generation rate function in Equation (4.7) becomes time-varying, and can be written as $R_{\text{it}}(t)$. Fig. 4.1 shows the $R_{\text{it}}(t)$ of the NMOS transistor in an inverter with a rising input signal: notice that the value is zero outside the transition, and contains non-zero components from medium energy and cold carriers over the period of transition. The medium energy component shows two peaks in the beginning and end of transition due to the peaks of I_{bs}/I_{ds}, while the cold carriers component has one peak near the end of transition due to the peak of I_{ds}. The active state of a logic cell can be characterized using the input signal transition time and output load capacitance. For example, a faster transition results in higher-energy carriers, while a slower transition to a larger load may result in a larger volume of lower-energy carriers.

It is important to note that, unlike NBTI, HC effects do not experience recovery effects, and the application of HC stress results in monotone aging. We introduce a term, called the age gain (AG) per transition of a MOSFET, to capture the effect of degradation due to HC aging as a result of each transition that the MOSFET undergoes. The age function, which increases monotonically over the life of a device, is the sum of AGs of all transitions:

$$\text{age} = \sum_{\text{all transitions}} \text{AG} \tag{4.8}$$

We compute the AG using a a quasistatic approach: such methods have been accepted for HC analysis [77]. With this approach, the device AG over each transition period with time-dependent aging rate is computed as the integral of the aging rate function $R_{\text{it}}(t)$, as shown below,

$$\text{AG} = \int_{\text{tran}} R_{\text{it}}(t)\text{dt} \tag{4.9}$$

Here, tran stands for the interval of a specific transition, and $R_{\text{it}}(t)$ is defined in Equation (4.7) with time-dependent operation voltages and currents. The integral computes

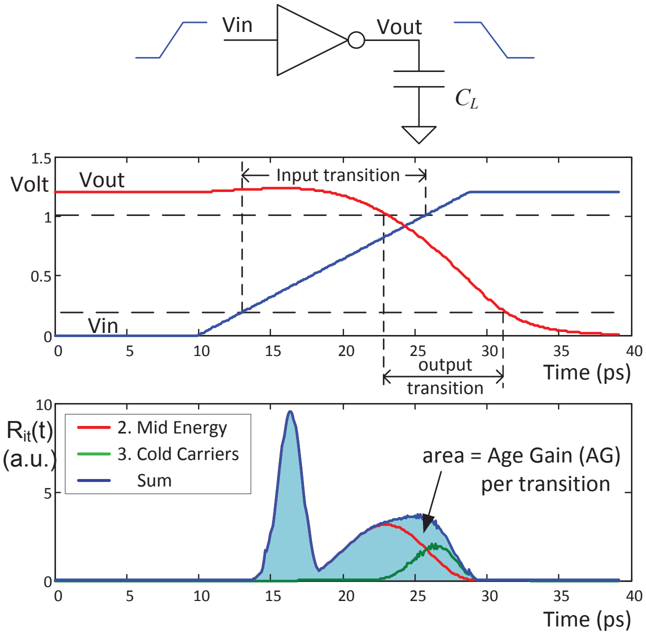

Figure 4.1: An example that shows the age function, $R_{\text{it}}(t)$, during a signal transition of an inverter.

the age gain associated with one specific transition and uses the quasistatic approach to approximate the integral as a finite sum.

4.3.2 Library Characterization

For a digital circuit, the AG calculations can be characterized at the cell-level as a part of library characterization. Under a specified input signal type (rise or fall), a transition time, and an output load capacitance, the time-varying voltages and currents of all MOS transistors inside the logic cell can be computed using SPICE transient analysis. The AG of each transistor is computed by the numerical integration of $R_{\text{it}}(t)$ in Equation (4.7), as given by (4.9).

Examining the procedure outlined above, it is easy to see that for library-based digital circuits, where all logic cells are from a cell library, the degradation of HC effect can be precharacterized for cells in the library and stored in a look-up table (LUT). Fig. 4.2 illustrates how a NAND2 cell may be characterized, by enumerating the signal input pin, the signal type, the transition time denoted as tr, and the output load denoted as C_L. Note that a transistor can experience age gain even if there is no transition on its gate input: for example, for a two-transistor NMOS pulldown in the NAND2 cell, a transition that turns on the upper input, while the lower input is already on, can cause an increase in AG for the lower transistor. We capture such effects in our model. For example, for each case shown in the figure with specified tr and C_L, the AGs of all four transistors in the NAND2 cell are computed simultaneously.

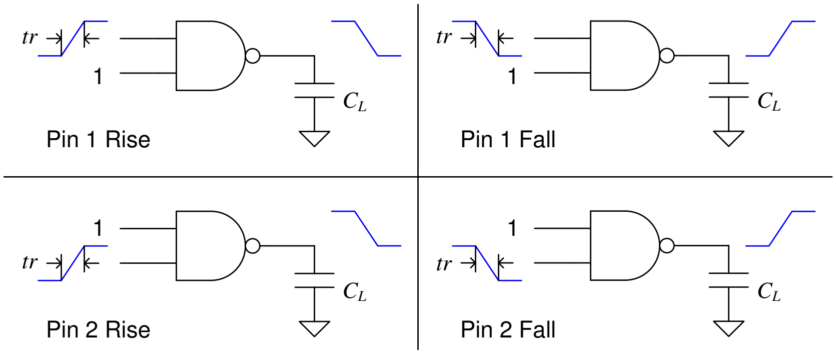

Figure 4.2: Characterization of a NAND2 cell. The number of simulations required for characterization is identical to those of timing characterization.

The LUT of each cell i outputs the AGs of all transistors j inside the cell, and has five input parameters as expressed in Equation (4.10), where k stands for the input pin with signal transition[1] , r/f for the transition type (rise or fall), inp for the input vector of the remaining input pins (if more than one input vector can make pin k critical), tr_k for the input transition time on pin k and C_L for the load capacitance.

$$\left\{\mathrm{AG}_{j,k}^{r/f}\right\}_{j\in\text{cell } i} = \mathrm{LUT}_{\mathrm{AG}}(k, r/f, inp, tr_k, C_L) \tag{4.10}$$

Characterization cost: The number of simulations required to characterize AG is *identical* to that required for timing characterization: in fact, the same simulations

[1] As in static timing analysis, we operate under the single input switching (SIS) assumption, i.e., the signal transition for a logic cell is triggered by one signal. This can be extended to the multiple input switching (MIS) scenario, where more than one signal arrives during the transition using methods similar to those used for timing characterization. However, given that the age function is computed cumulatively over long periods of time and that the probability of MIS is typically much lower than that of SIS, the SIS assumption gives an adequate level of accuracy. Further improvements in accuracy here are likely to be overshadowed by modeling errors at the device level.

are used, but additional currents/voltages are monitored, followed by a post-processing phase in which mathematical operations (such as numerically integration) are performed on this data to compute AG. Therefore, the number of simulations is $O(N_{\text{cell}})$, where N_{cell} is the number of cells, and so is the storage complexity of the LUT.

The effect of aging on a transistor is to alter its saturation drain current I_{on}. This in turn affects key performance parameters such as the propagation delay and output signal transition time of a logic cell that the transistor lies in. Given that the aging perturbations are small, we use first-order models for these relationships, as is done in other variational methods [78]. The propagation delay d_i and signal transition tr_i of cell i are modeled using the following linear relationship with the $\Delta I_{\text{on}}/I_{\text{on}}$ of transistors j inside cell i:

$$\begin{aligned} d_i &= d_{i0} + \sum_{j \in \text{cell } i} S_{ij}^{d} (\Delta I_{\text{on}}/I_{\text{on}})_j && (4.11) \\ tr_i &= tr_{i0} + \sum_{j \in \text{cell } i} S_{ij}^{tr} (\Delta I_{\text{on}}/I_{\text{on}})_j && (4.12) \end{aligned}$$

The propagation delay d_i, signal transition time tr_i, and their sensitivities S_{ij}^{d} and S_{ij}^{tr} to the transistor $\Delta I_{\text{on}}/I_{\text{on}}$ values are calculated using standard techniques. The approximation that mobility degradation $\Delta\mu/\mu = \Delta I_{\text{on}}/I_{\text{on}}$ is used for device model in SPICE analysis. As pointed out earlier, these correspond to the same SPICE simulations that are used for AG characterization, although different results are extracted from the simulations. The results are stored in LUTs, expressed as follows:

$$\begin{aligned} d_i &= \text{LUT}_d(k, r/f, inp, tr_k, C_L) && (4.13) \\ \{S_{ij}^{d}\}_j &= \text{LUT}_{S_d}(k, r/f, inp, tr_k, C_L) && (4.14) \\ tr_i &= \text{LUT}_{tr}(k, r/f, inp, tr_k, C_L) && (4.15) \\ \{S_{ij}^{d}\}_j &= \text{LUT}_{S_{tr}}(k, r/f, inp, tr_k, C_L) && (4.16) \end{aligned}$$

As stated earlier, the computations of these LUTs has similar complexity as that of AG characterization. Moreover, there are established methods for computing each one of these, as they are used in variational/statistical analysis.

4.4 Circuit-Level Analysis of HC Degradation

Given a set of precharacterized cells, our task at the circuit level is to efficiently use this information to perform scalable circuit-level analysis using these accurate models. Our analysis consists of four steps, described in this section: first, finding the distribution of the signal transition time at each node; second, calculating the AG for all gates in a circuit, considering their context in the circuit; third, using this information to analyze device aging; and fourth, analyzing the delay degradation of the circuit.

4.4.1 Distribution of Signal Transition Time

Due to the discrete nature and finite (but potentially large) number of signal paths in digital circuit, the signal transition time, $tr(q)$, at a certain node q has a discrete probability distribution, $\Pr(tr(q))$, which is nonzero at all values of $tr(q) \in \mathbf{Tr^{(q)}}$, where $\mathbf{Tr^{(q)}}$ stands for the set of all possible tr values of node q.

We assume that the signal transition times of the primary inputs is known (and assumed to be constant). The signal transition distribution of all internal nodes can be calculated either using Monte Carlo simulation, or using a probabilistic method. Here, we introduce a transition propagation (TP) method to calculate the transition time distribution (rise and fall separately) at each node, which is similar in spirit as static timing analysis (STA), but calculates the complete distribution information of transition time using signal probability (SP) and activity factor (AF), instead of just solving for the longest delay and transition time, as in conventional STA.

As each gate q is processed in topological order, given the distribution of transition times at each input pin of the gate, we use the LUT_{tr} in Equation (4.15) to compute the distribution of $tr(q)$ at the output. A single transition at the output of q can be triggered under a number of different logical input conditions. We enumerate these conditions for each gate, which correspond to enumerating, for each input pin k, the set of noncontrolling inputs that allow a transition at k to become a transition at q. Under each condition, we compute $tr(q)$ using LUT_{tr}, and $\Pr(tr(q))$ using the activity factor (AF) of the corresponding input transition and the signal probability (SP) of the nontransitioning inputs.

The enumeration over all patterns on a gate is not expensive for a gate with a reasonable number of inputs; however, we must also perform an enumeration over all transition times. In principle, this could lead to state explosion as the number of possible elements of $\mathbf{Tr^{(q)}}$ are enumerated. To control this, we use data binning to reduce the number of data points that represent the distribution by approximating it with a discrete distribution over a smaller number of points, denoted as $\mathbf{Tr_s}$. We find that the error due to this approximation is negligible.

Theoretically, it is necessary to analyze the distribution of tr at each circuit node and to use this result for AG calculation. However, as will be shown in Section 4.6.1, the error of using a single value of tr from STA result is very small compared with using this full tr distribution, since this is already a second-order effect; moreover, the actual distribution of tr tends to have a small standard deviation and the tr from STA result gives a close approximation. In the following work (e.g., the variation-aware analysis in Section 4.5.4) the value of tr from STA can be used safely to reduce complexity.

4.4.2 Mean AG Calculation in Digital Circuits

As discussed in Section 4.3.1, device aging in a library cell is modeled using age gain per transition $\mathrm{AG}_{j,k}^{r/f}$ and characterized using a quasistatic approach at the cell-level. At the circuit level, since each input pin k of a logic cell i has different probability distribution of transition time $tr_{r/f}$ (r/f for rise and fall), computed using the results of the method in Section 4.4.1, the age gain from each rise or fall signal on pin k also has a unique distribution.

Unlike the case of static timing analysis (STA) for delay analysis, where the focus of the analysis is to determine the slowest path, the aging analysis must consider the average operational conditions (and then find the slowest path in the circuit at various points in time). Therefore the mean value of the age gain distribution is calculated as shown in Equation (4.17), where the new term $\mathrm{AG}_{k,j}$ is defined as the mean age gain of transistor j per input signal cycle (including one rise and one fall signal) on pin k.

$$\begin{aligned} \mathrm{AG}_{k,j} &= \mathrm{AG}_{k,j}^{r} + \mathrm{AG}_{k,j}^{f} \\ \text{where} \quad \mathrm{AG}_{k,j}^{r} &= \sum_{tr_r \in \mathbf{Tr_s}} \mathrm{AG}_{k,tr_r,j}^{r} \cdot \Pr(tr_r) \\ \mathrm{AG}_{k,j}^{f} &= \sum_{tr_f \in \mathbf{Tr_s}} \mathrm{AG}_{k,tr_f,j}^{f} \cdot \Pr(tr_f) \end{aligned} \tag{4.17}$$

where $\mathbf{Tr_s}$ is the approximate discretized version of $\mathbf{Tr}$. Here $\mathrm{AG}_{k,j}$ is calculated as the sum of the mean age gain per rise signal, $\mathrm{AG}^r_{k,j}$, and mean age gain per fall signal $\mathrm{AG}^f_{k,j}$, which are computed separately using age gain per transition under specific transition time tr_r and tr_f from the cell-level AG LUT in Equation (4.10), and the signal transition time distribution in Section 4.4.1.

4.4.3 Analysis of Device Aging

The circuit-level device aging analysis is performed based on analysis of the device age gain per signal cycle in the above section, and the statistical estimation of signal cycles in a given period of circuit operations. All signal paths (instead of only critical ones in STA) are considered in the device aging analysis, because all signal propagations affect the device aging. If a circuit is V_{dd}-gated or power-gated, the device aging model incorporates this effect using signal statistics, as shown below.

During a time period t of circuit operation, the age of a transistor j in a digital circuit is the accumulation of age gains (AGs) due to signal cycles on its input pins that occurred from time 0 to t. Since we have already obtained the mean AG per signal cycle in Equation (4.17), the device age function can be written as the number of signal cycles on each pin k times AG per cycle of k, summed for all input pins of cell i (where transistor j belongs), as follows:

$$\mathrm{age}_j(t) \quad = \quad \sum_{k \in \mathrm{pin}_i} N_k \cdot \mathrm{AG}_{k,j} = \sum_{k \in \mathrm{pin}_i} \eta_k \cdot t \cdot \mathrm{AG}_{k,j} \tag{4.18}$$

Here $\mathrm{AG}_{k,j}$ stands for the mean age gain of transistor j per cycle of input signal on pin k; N_k stands for the number of signal cycles on pin k during time period of t, and $\eta_k = N_k/t$ is defined as the rate of effective signal cycle on input pin k, that causes cell switching. The value of η_k can be obtained using the statistical information of signal probability (SP) and activity factor (AF) as

$$\eta_k = f_{\mathrm{ref}} \cdot \mathrm{AF}_k \cdot \mathrm{Pr}_{k\ \mathrm{critical}} \tag{4.19}$$

Here, f_{ref} is the frequency of reference clock, AF_k is the activity factor of the k^{th} input pin of cell i, i.e., the average number of signal transition cycles in a reference clock cycle [79], and $\mathrm{Pr}_{k\ \mathrm{critical}}$ is the critical probability of pin k, i.e., the probability that the cell output is dependent on the input logic of pin k:

$$\text{Pr}_{k\ \text{critical}} = \text{Prob}(\text{output}(\text{pin}_k = 0) \neq \text{output}(\text{pin}_k = 1)) \tag{4.20}$$

This can easily be calculated using the joint signal probability of the input pins (computed from the Monte Carlo-based SP simulations described in Section 4.6) and the truth table of the logic cell.

4.4.4 Analysis of Circuit Delay Degradation

The circuit delay degradation analysis is performed based on the models discussed in the previous sections, and static timing analysis (STA) is performed using a PERT-like traversal [17] to calculate the delay of the fresh and the aged circuits.

Since the HC aging is dependent on the signal transition as modeled in Section 4.3.1, an initial STA of the fresh circuit is necessary for calculating the HC aging, based on which the circuit delay degradation after a period of operation can be computed by doing STA again with aged device parameters.

HC effects can slow down the signal transition during the aging process, which in turn reduces the age gain per transition, further slowing down HC-based circuit aging. Therefore, in principle, the circuit delay degradation is generally a *decelerating* process, as will be pointed out in Section 4.6, and it may need iterations for accurate analysis that recalculate the slowdown in signal transition times in multiple steps and update the age gains. Our experimental results in Section 4.6 explore a *one-step* method (where the signal transition times at $t = 0$ are used throughout the simulation) with a *N-step* iterative method (where the transition times are updated N times through the life of the circuit). The experimental results in Section 4.6.1 demonstrate that in practice this deceleration effect of aging is quite insignificant and can be safely ignored, so that the degradation analysis can be performed efficiently without iterations[2] .

The degraded critical path delay D in a digital circuit is given by

$$D = \sum_{i \in \text{path}} d_{i0} + \sum_{i \in \text{path}} \Delta d_i = D_0 + \Delta D \tag{4.21}$$

[2] Other authors [16] have found nontrivial *acceleration* effects of HC degradation, mainly due to the inaccuracy of their model assumption of constant HC stress during signal transitions (in contrast to our time-varying model illustrated in Fig. 4.1).

The cell-level delay degradation Δd_i, which is modeled as a linear function of all transistor degradation $\Delta I_{\text{on}}/I_{\text{on}}$ in Equation (4.11), can be derived as following using the models of $\Delta I_{\text{on}}/I_{\text{on}}$ and $\text{AG}_{k,j}$ in Equation (4.4) and (4.17).

$$\begin{aligned} \Delta d_i &= \sum_{j \in \text{cell } i} S^d_{ij} \cdot A \cdot (\text{AR}^{(i)}_j \cdot t)^n \\ \text{where} \quad \text{AR}^{(i)}_j &= \sum_{k \in \text{pin}_i} \eta^{(i)}_k \text{AG}^{(i)}_{k,j} \end{aligned} \tag{4.22}$$

Therefore the critical path delay degradation is

$$\Delta D = At^n \sum_{i \in \text{path}} \sum_{j \in \text{cell } i} S^d_{ij} \cdot (\text{AR}^{(i)}_j)^n \tag{4.23}$$

Equation (4.23) indicates that the path delay degradation of digital circuits has a power function versus time, with the same exponent n as the power model of device degradation in Equation (4.4). However, since devices on different paths have different rate of aging, the longest-delay path may change after a period of degradation, as will be shown in the experimental results in Section 4.6.1.

4.5 Variation-aware Hot Carrier Aging Analysis

In Sections 4.3 and hc:circuitana, we had developed machinery for analyzing circuit delay degradation due to HC effects for the nominal case. In this section, we extend the proposed HC aging and circuit degradation analysis approach to incorporate the effects o process variations. We begin by presenting the underlying models used to represent process variations. Next, we modify the previously described transistor-level model and cell-level characterization approaches for HC aging to incorporate the effects of variations. Finally, we devise an efficient method for performing circuit-level delay degradation analysis due to HC effects under process variations. As will be demonstrated in our results in Section 4.6.2, the impact of process variations on aging are significant.

4.5.1 Modeling Process Variations

The variations of the process parameter created at the fabrication time can be classified as lot-to-lot, die-to-die (D2D), and within-die (WID) variations, according to their scope; they can also be divided as systematic and random variations by the cause and predictability. Usually WID variations exhibit spatial dependence knows as spatial correlation, which need be considered for accurate analysis.

This chapter employs a widely-used model for process variations: a process parameter X is modeled as a random variable about its mean, X_0, as [80]:

$$\begin{aligned} X &= X_0 + X_g + X_s + X_r \\ \sigma_X^2 &= \sigma_{X_g}^2 + \sigma_{X_s}^2 + \sigma_{X_r}^2 \end{aligned} \tag{4.24}$$

Here, X_g, X_s, and X_r stand for the global part (from lot-to-lot or D2D variations), the spatially correlated part (from WID variation), and the residual random part, respectively. This model assumes all the devices on the same die have the same global part X_g. The spatially correlated part is captured by a grid-based model similar to [17]. The entire circuit is divided into equally-sized grids by its geometry layout. All transistors within the same grid have the same spatially correlated part X_s, and the transistor parameters in different grids are correlated, with the correlation coefficient falling off with the distance increasing. The random part X_r is unique to each transistor in the circuit.

In this chapter we consider the variations in the transistor width (W), the channel length (L), the oxide thickness (T_{ox}), as well as shifts in the threshold voltage V_{th} due to random dopant fluctuations (RDFs). In other words, for each device, X represents elements of the set $\{W, L, T_{ox}, V_{th}\}$. As in the large body of work on SSTA, we assume Gaussian-distributed parameters for each of these process parameters, with W and L exhibiting spatial correlation, and T_{ox} and V_{th} being uncorrelated from one device to the next. The essential idea can be extended to incorporate other types of variations into the formulation. The spatial correlation can be extracted as a correlation matrix using model proposed in [64], and then processed using principal components analysis (PCA) to reduce the data dimension. The value of the process parameter in each grid is expressed as a linear combination of the independent principal components. Notationally, each process parameter X is expressed as a vector in a random space, with basis $\mathbf{e} = [\mathbf{e}_g, \mathbf{e}_s, \mathbf{e}_r, \epsilon]^{\mathbf{T}}$, as

$$\begin{aligned} X &= X_0 + \Delta X = X_0 + \mathbf{k}_X^{\mathbf{T}}\mathbf{e} \qquad (4.25) \\ &= X_0 + \mathbf{k}_{Xg}^{\mathbf{T}}\mathbf{e}_g + \mathbf{k}_{Xs}^{\mathbf{T}}\mathbf{e}_s + \mathbf{k}_{Xr}^{\mathbf{T}}\mathbf{e}_r + k_\epsilon \epsilon \\ \sigma_X^2 &= \mathbf{k}_X^{\mathbf{T}}\mathbf{k}_X, \quad \mathrm{cov}(X_i, X_j) = \mathbf{k}_{Xi}^{\mathbf{T}}\mathbf{k}_{Xj} - k_{\epsilon_i}k_{\epsilon_j} \end{aligned}$$

Here, $\mathbf{e}_g = [e_{Wg}, e_{Lg}]^{\mathbf{T}}$ is the basis for global part (T_{ox} variation is considered purely random hence does not have global and spatial parts), $\mathbf{e}_s = [e_1, ..., e_t]^{\mathbf{T}}$ is the basis of principal components for the spatially correlated part, where t is the number of dimensions after the PCA processing, and $\mathbf{e}_r = [\epsilon_1, ..., \epsilon_m]^{\mathbf{T}}$ is the basis of random part. Its dimension, m, will depend on the implementation of the SSTA algorithm, and can vary from constant to linear (of circuit size), as will be discussed later in this chapter. The random part vector $\mathbf{k}_r$ can be implemented using a sparse data structure. The Gaussian variable $\epsilon \sim N(0, 1)$ is a separate independent random part for use in circuit-level timing analysis.

4.5.2 Transistor-Level Aging Model under Variations

The variations in the process parameters corresponding to the transistor width, W, length, L, and oxide thickness, T_{ox}, originate in the fabrication process, and are “baked in” (i.e., remain constant) for each manufactured circuit through its lifetime. Under these process variations, the *rate of aging* defined in (4.7) will be affected by the fluctuation of the process parameters as well as the input/output conditions, and can be updated as

$$R_{\mathrm{it}}^{\mathrm{var}} = R_{\mathrm{it}} + f(\mathbf{\Delta X}) + g(\mathbf{\Delta Y}) \qquad (4.26)$$

Here $\mathbf{\Delta X} = \{\Delta W_j, \Delta L_j, \Delta T_{ox,j}\}, j \in$ cell i is the fluctuation vector of process parameters of all transistors within the logic cell i, and $\mathbf{\Delta Y}$ is the fluctuation vector of circuit-specific conditions, including input signal transition time Δtr and load capacitance ΔC_L, which come from the process variations of transistors in the fanin/fanout cells in the circuit.

Using the quasistatic approach in (4.9), the transistor age gain per signal transition under process variation is

$$\begin{aligned} \mathrm{AG}^{\mathrm{var}} &= \int_{\mathrm{tran}} R_{\mathrm{it}}^{\mathrm{var}}(t)\mathrm{dt} \\ &= \mathrm{AG}_{\mathrm{nom}} + F(\mathbf{\Delta X}) + G(\mathbf{\Delta Y}) \end{aligned} \tag{4.27}$$

Here $F(\mathbf{\Delta X})$ and $G(\mathbf{\Delta Y})$ are the integral of $f(\mathbf{\Delta X})$ and $g(\mathbf{\Delta Y})$, respectively, and they stand for the variation of *age gain* per signal transition due to the process parameters $\mathbf{\Delta X}$ and circuit-specific conditions $\mathbf{\Delta Y}$, respectively.

4.5.3 Cell-Level Model and Characterization under Variations

Our cell-level modeling and characterization under process variations consists of two parts: transistor age gain characterization and cell timing characterization.

Following the quasistatic approach (4.27), we use first-order Taylor expansion to model the device age gain of one signal transition event under process variations as follows:

$$\mathrm{AG}^{\mathrm{var}} = \mathrm{AG}_{\mathrm{nom}} + F(\mathbf{\Delta X}) + G(\mathbf{\Delta Y}), \tag{4.28}$$

$$\text{where } F(\mathbf{\Delta X}) = \sum_{j \in \text{cell } i} \left(\frac{\partial \mathrm{AG}}{\partial W_j} \Delta W_j + \frac{\partial \mathrm{AG}}{\partial L_j} \Delta L_j + \frac{\partial \mathrm{AG}}{\partial T_{ox,j}} \Delta T_{ox,j} \right) \tag{4.29}$$

$$G(\mathbf{\Delta Y}) = \frac{\partial \mathrm{AG}}{\partial tr} \Delta tr + \frac{\partial \mathrm{AG}}{\partial C_L} \Delta C_L \tag{4.30}$$

The nominal value $\mathrm{AG}_{\mathrm{nom}}$ and its sensitivities $\partial \mathrm{AG}/\partial(\cdot)$ to the variational parameters are all characterized by HSPICE simulation in a manner similar to Section 4.3.2, and the results are stored in LUTs for the use of circuit-level analysis. The variational parameters ΔW_j, ΔL_j, $\Delta T_{ox,j}$ and Δtr are Gaussian random variables in vector form in the random variable space $\mathbf{e}$, as defined in (4.25). The load capacitance C_L has following relationship with the process parameters of the fanout transistors

$$C_L = K \sum_{k \in \mathrm{Fanout}(i)} \frac{W_k L_k}{T_{ox,k}} \tag{4.31}$$

Under assumption of σ/μ being a relatively small value (e.g. $< 10\%$), which is true for reasonable processes, their product and quotient can be approximated as a Gaussian random variable in the same space e, using the linear approximation based on moment matching method discussed in Appendix G. In this manner, C_L is also expressed as a vector in the RV space $\mathbf{e}$. The input signal transition time Δtr will be modeled in the cell timing characterization below in (4.34). Based on these models, the transistor age gain under variations, which depends linearly on these parameters as (4.28), can also be presented as a Gaussian random variable in the space $\mathbf{e}$.

The transistor drain current degradation, modeled by (4.4), is a power function of age

$$(\Delta I_{\text{on}}/I_{\text{on}})_j = A(\text{age}_j)^n \tag{4.32}$$

with the exponent $n\approx 1/2$. Since the transistor age calculated by (4.18) is a Gaussian random variable in the space $\mathbf{e}$, its power function can also be approximated as Gaussian in the same RV space using methods proposed in Appendix F, using the mean and variance computed by Appendix D or E.

The cell timing characterization under process variations is performed based on following first-order model of propagation delay d_i and output signal transition time tr_i,

$$d_i = d_{i0} + \sum_{X\in P_i} \frac{\partial d_i}{\partial X}\Delta X \tag{4.33}$$

$$tr_i = tr_{i0} + \sum_{X\in P_i} \frac{\partial tr_i}{\partial X}\Delta X \tag{4.34}$$

Here, $P_i = \{W_j, L_j, T_{ox,j}, (\Delta I_{\text{on}}/I_{\text{on}})_j, V_{th,j}\}, j \in$ cell i are the process and aging parameters of the transistors in the cell that are considered for timing variation and degradation, and $\partial d_i/\partial(\cdot)$ are the corresponding sensitivities to these parameters. These sensitivity values are computed using HSPICE analysis and stored in a look-up table for the use of circuit level timing analysis.

4.5.4 Circuit-Level Analysis under Variations

Under process variations, the circuit-level delay and degradation analysis is based on the SSTA framework proposed in [17], which handles all cell delays, arrive times, and signal transitions as Gaussian random variables in the space $\mathbf{e}$, and performs sum and max operations in space $\mathbf{e}$ while performing a PERT-like traversal to calculate the total delay.

As in the nominal case discussed in Section 4.4.4, an initial SSTA of fresh circuit is performed to calculate the signal transition of all nodes, from which the device aging under process variations is calculated for a given period of operation time. Then SSTA is performed again with degraded device parameters to calculate the circuit delay. The results from the nominal case analysis already demonstrate the approximation of using signal transition tr from STA results, and using 1-step analysis are both very accurate. Meanwhile when process variations are considered, their impact on transition time tr is a second-order effect, and a small change due to variations will have a very diluted impact on delay degradation. Therefore in the variational case we keep using tr value from SSTA results instead of the full distribution, and using a 1-step analysis to effectively reduce runtime while maintaining accuracy.

In SSTA, the way of handling the random part $\mathbf{k}_{Xr}^{\mathbf{T}}\mathbf{e}_r$ could affect the runtime and accuracy of the results. Since the random parts come from the process parameters of different devices in the circuit and are independent with each other, when calculating the signal arrival time through a circuit traversal, the size of the random parts can grow significantly, resulting in quadratic complexity for runtime and storage. A simple remedy is to merge the random part $\mathbf{k}_{Xr}^{\mathbf{T}}\mathbf{e}_r$ of each random variable into a separate scalar term $k_\epsilon \epsilon$. This method greatly reduces the time and storage complexity, but the accuracy is slightly affected due to the path reconvergence in the circuit topology which introduces correlations to the random parts. In [81] a better trade-off between accuracy and complexity was proposed by removing smaller elements in the random vector $\mathbf{k}_{Xr}$ using preset threshold and merge them into the separate term k_ϵ. We use this method to handle the random part and the results in Section 4.6.2 verify that it is accurate and efficient.

4.6 Experimental Results

The proposed method for delay degradation analysis of digital circuit is applied to the ISCAS85 and ITC99 benchmark circuits for testing. The circuits are mapped to a subset of the Nangate 45nm open cell library [70] using ABC [69], with placement carried out using a simulated annealing algorithm. The cell-level library characterization was performed using HSPICE simulation and 45nm PTM model [63]. The circuit-level analysis was implemented in C++ and run on a Linux PC with 3GHz CPU and 2GB RAM. The parameters a_2, a_3 and m of the device-level HC model in Equation (4.7) is from [13]. The coefficients C_1, C_2 and C_3 for different energy-driven modes have arbitrary units (a.u.) and are selected empirically according to the τ vs I_{ds}/W plot in [13]. The parameter A in Equation (4.4) also has a.u..

4.6.1 Results of Nominal Degradation Analysis

The cell-level characterization of transistor age gain, as well as the degradation of cell delay and output transition is performed using HSPICE simulation with the enumeration of all signal input cases for each cell. In nominal case (where the process variations are not included), the characterization of the library which contains 55 logic cells takes 1 hour and 52 minutes of runtime and 8.4MB of hard drive storage (in ASCII format). This is 1.9× runtime and 5.9× storage overhead compared with timing characterization (Equations (4.13–4.16)), which is well within reasonable range.

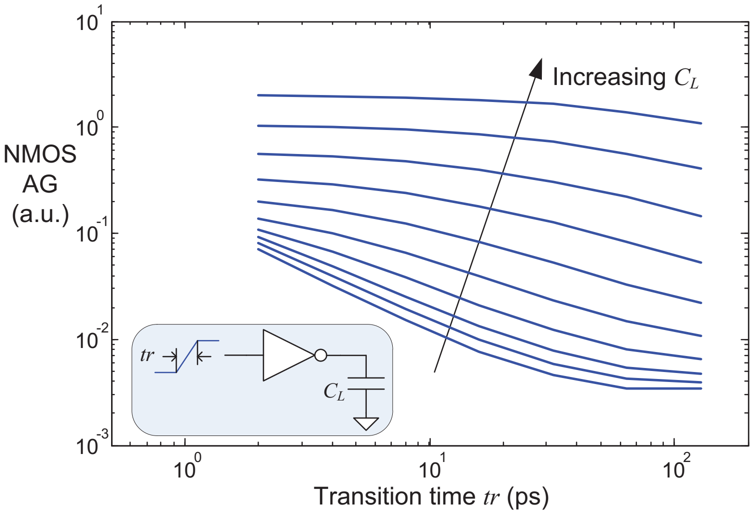

Figure 4.3: Age gain versus signal transition time and load capacitance.

Fig. 4.3 plots the curves of the NMOS transistor age gain (AG) versus the input signal transition tr of an inverter with a rising input signal, under different load capacitance C_L. The figure indicates that the input signal transition generally causes more damage to the transistor (larger AG) when the load C_L is large, or when the transition

tr is small. This is explained by the fact that HC degradations are caused by the charge carriers flowing through the channel, and larger load C_L requires more charge to be moved, while smaller transition tr makes the carriers moving faster, thus causing more damage. This result is consistent with the data presented in [14]. In other transition cases, with different cells and input signals, the AG vs. tr and C_L plots may be slightly different, but all have a trend similar to Fig. 4.3. Specifically, for the small range in which the transition time increases as a result of aging (<2%), the AG generally reduces slightly, i.e., aging slightly decelerates with time.

Fig. 4.4 shows a plot of the circuit degradation analysis result of the N-step method discuss in Section 4.4.4 with N set from 1 to 256. The plot indicates that the overall error between 1-step and N-step is very small ($< 1\%$), and that N=64 is an adequate step number for the balance between runtime and accuracy. Therefore in the following analysis N=64 will be used for the N-step method.

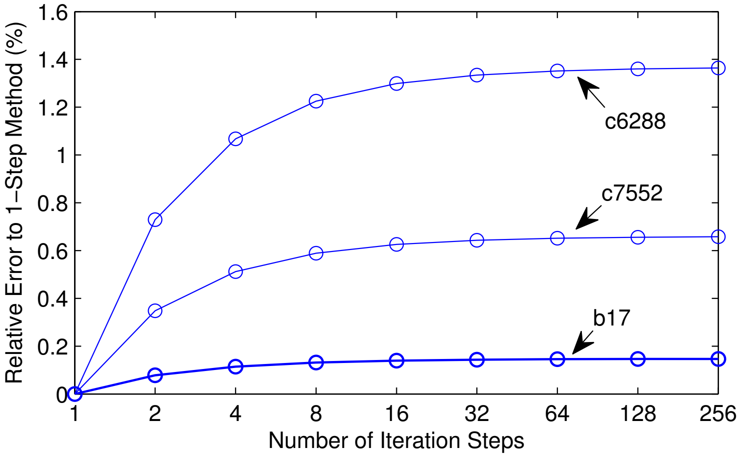

Figure 4.4: Circuit degradation analysis with different number of iterations.

The results of the proposed approach for circuit degradation analysis under HC effects are presented in Table 4.1 for different benchmark circuits. The sizes of the circuits range from 221 cells (c432) to 20407 cells (b17). Three methods are implemented and applied on each benchmark: the first is a Monte Carlo (MC) simulation to calculate the circuit degradation by stressing the circuit using 10000 random input signal transitions; the second is the proposed analysis approach using one-step approximation that ignores the deceleration of aging, and the third incorporates the deceleration process using N-step method, updating the aged tr at 64 time points over the life of the circuit (see Section 4.4.4 for details). The signal probability (SP) and activity factor (AF) data for the latter two methods is obtained using Monte Carlo method with 10000 random input transition samples. The circuit degradations are calculated at t=10000 (a.u.) with reference clock f_{ref}=1GHz, input SP=0.5 and AF=0.05.

Table 4.1: Runtime and degradation comparison of different methods for nominal case HC aging analysis

Circuit	Size	Fresh	SP/AF	1-step Analysis			64-step Analysis		MC		SimpleDF
Name	#Cells	Delay	T_{exe}	T_{exe}	ΔD	ΔD_{tr}	T_{exe}	ΔD	T_{exe}	ΔD	ΔD_{err}
c1908	442	835ps	1.5s	0.080s	135ps	136ps	4.6s	134ps	2.1s	135ps	13.6%
c2670	759	1228ps	2.1s	0.120s	134ps	134ps	7.3s	134ps	2.1s	138ps	24.9%
c3540	1033	1397ps	4.0s	0.230s	425ps	425ps	13.4s	424ps	6.4s	439ps	-22.9%
c5315	1699	1133ps	5.7s	0.260s	80ps	80ps	14.6s	80ps	8.3s	82ps	56.1%
c6288	3560	2579ps	30.8s	0.640s	489ps	488ps	37.4s	483ps	55.9s	500ps	42.7%
c7552	2361	1209ps	9.6s	0.350s	138ps	137ps	20.3s	137ps	17.9s	152ps	1.2%
b14	4996	2586ps	60.7s	1.010s	789ps	790ps	58.6s	789ps	74.7s	766ps	-44.4%
b15	6548	2628ps	90.4s	1.330s	574ps	574ps	77.5s	573ps	100.2s	609ps	-24.2%
b17	20407	3201ps	320.6s	4.120s	131ps	130ps	237.5s	130ps	343.6s	135ps	-28.9%
b20	11033	2586ps	171.2s	2.170s	502ps	501ps	124.6s	502ps	189.4s	495ps	-7.1%
b21	10873	2837ps	162.1s	2.000s	269ps	269ps	116.1s	269ps	179.8s	262ps	8.2%
b22	14794	2845ps	232.9s	2.740s	325ps	326ps	158.3s	325ps	254.8s	335ps	-36.6%
Average Error to MC					3.3%	3.4%		3.6%			

In Table 4.1, the first column lists the benchmark circuit name, the second and third columns list the number of cells and the fresh delay of the circuit, the fourth column lists the runtime of SP and AF calculation, and the remaining columns show the runtime and circuit delay degradation obtained using the three methods. The results show that the one-step analysis and 64-step analysis yield very close results (<1% relative error), and that the error between using the full tr distribution (ΔD) and using a single tr value from the STA result (ΔD_{tr}) is negligible, demonstrating that the effect of tr distribution and its dynamics on the circuit degradation is very small and can be safely ignored to reduce computation. The error between one-step analysis and MC is small (3.3% relative error) while the one-step method has much lower runtime, indicating the proposed analysis method is efficient and accurate compared with Monte Carlo simulation.

The last column shows a comparison with a simple duty-factor based scheme, similar to [16]. Note that in contrast with this method, our approach performs quasistatic analysis with newer energy-driven model, which captures the time-varying HC stress, and indicates that the transistor AG decreases when signal transition slows down (Fig. 4.3). In addition, [16] uses an empirical device HC model which only considers the switching transistors and ignores the other transistors in the stack which also experience current stress. Our approach perform the device degradation analysis in the cell level, and the AG of all transistors in a logic cell is computed simultaneously. The results in the last column assume constant HC stress through signal transitions, ignores non-switching transistor degradation, and uses worst-case transition time. Experimental results of all tested benchmarks show errors of −44% to +64% for this method. It is clear that the use of such simplifying assumptions, commonplace in all prior work on large-scale circuits, results in serious errors.

It is important to note that the SP and AF analysis take more time than the HC degradation calculation; however (a) this computation is a common overhead shared by other circuit analyses, such as power estimation, oxide reliability, BTI degradation, etc., and should not be counted solely towards the proposed approach, and (b) our implementation uses Monte Carlo simulation to generate these probabilities; faster graph traversal based methods may also be used.

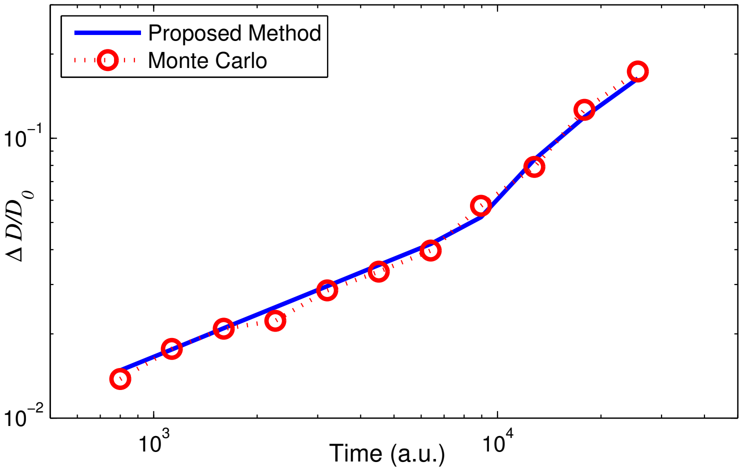

Figure 4.5: Circuit delay degradation versus time.

Fig. 4.5 shows the circuit delay degradation versus time on a logarithm scale for benchmark c1908 using both the proposed analysis method (one-step) and MC simulation. The results from these two approaches match well with each other, and the delay degradation is a power function of time with exponent 0.5 before t=10000 (a.u.). After that the delay degradation is no longer a power function and increases at a faster rate since the critical path may change, as discussed in Section 4.4.4.

An examination of the degradation of tr confirms that the effects of aging deceleration are negligible. Fig. 4.6 shows the histograms of the degradation $\Delta tr/tr_0$ of benchmark c7552, where tr_0 is the signal transition time of a node in the fresh circuit, and Δtr is the transition increment of each node at t=10000 (a.u.). The histograms of rise and fall signal transition degradation are plotted separately. We can see that although the circuit has nearly 13% delay increasing, the degradation of transition time is only around 5% in average, which causes very small impact on AG, according to the AG/tr curves in Fig. 4.3. This explains the fact that in Table 4.1, the results of ΔD using 1-step analysis and 64-step analysis are very close. In contrast, the simply duty factor model assumes constant HC stress in the off-to-on transition, leading to the result that the transition time degradation elevates the duty factor and accelerate circuit degradation, which is incorrect.

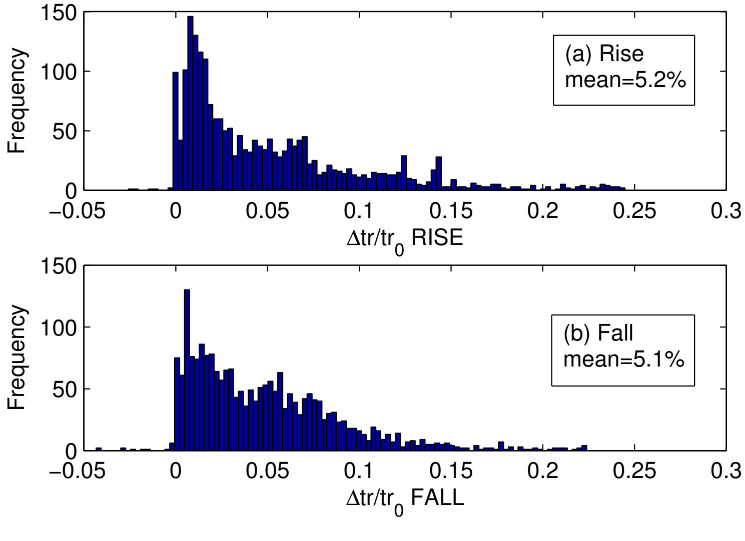

Figure 4.6: Histogram of transition time degradations.

We further examined the effect of tr distribution due to different signal paths. Section 4.4.1 has discussed that theoretically it is necessary to calculate full tr distribution at each circuit node for AG calculation. We captured the tr distribution of all primary output nodes in benchmark c7552 in MC simulation. Results show the actual tr distribution has a very small standard deviation (average σ_{tr}/μ_{tr}=1.34%), and its mean value is very close to the STA result (average μ_{tr}/tr_0=0.98). This explains the results in Table 4.1 that using tr approximation from STA (column ΔD_{tr}) yields very high accuracy.

4.6.2 Results for Variation-Aware Degradation Analysis

With the consideration of process variations, the library characterization of timing, aging and their sensitivities to process parameters using HSPICE takes 4 hour and 27 minutes of runtime and 21.3MB of hard drive storage (in ASCII format). This is 4.5× runtime and 15.1× the storage overhead compared with the basic timing characterization required for the nominal case (Equations (4.13–4.16)), which is reasonable.

The process variations in parameters W, L, and T_{ox} are set to 3σ=4% of their mean values [7]. The V_{th} variation due to RDF is dependent on the device size [82]. It has a Gaussian distribution with mean value $\mu = 0$, and standard variation

$$\sigma_{V_{th}} = \sigma_{V_{th0}} \sqrt{\frac{W_0 L_0}{WL}} \tag{4.35}$$

in which $\sigma_{V_{th0}}$ is the RDF-induce threshold standard deviation of a minimum-sized device (W_0 by L_0). The value of $\sigma_{V_{th0}}$ is dependent on process parameters and the doping profile. Here we assume $3\sigma_{V_{th0}} = 5\%$ of the nominal V_{th}. The parameter variations of W and L are split into 20% of global variation, 20% of spatially correlated variation and 60% of random variation, while the variations of T_{ox} and V_{th} are fully random. The grid-based spatial correlation matrix is generated using the distance based method in [64], with the number of grids growing with circuit size, as shown in Table 4.2. We have used the same assumptions for operation time, frequency, and input SP/AF as the nominal case in previous section.

Based on the conclusions from the experimental results of nominal case in previous section, we use the tr value from SSTA results as an valid alternative of full tr distribution, and 1-step method instead of multi-step iterative calculation of circuit aging, in order to reduce the runtime of the variation-aware analysis of circuit degradation under HC effect.

The proposed analytical approach is verified by Monte Carlo (MC) simulation. For the variational case, the MC simulation is performed by generating 5000 circuit samples with randomized process parameters, then for each circuit sample with the parameters determined, calculating the HC aging and delay degradation of using the proposed analytical method for nominal case. Using this simulation scheme is based on the consideration that the proposed analytical method for nominal case has already been verified by MC simulation in Section 4.6.1, and performing a full MC simulation with both randomized process parameters and input vectors will be too time-consuming and impractical.

The proposed approach for variation-aware circuit degradation analysis under HC effects are presented in Table 4.2 for different benchmark circuits. Three methods for circuit degradation analysis are implemented and compared. The first is denoted as Variational Analysis 1 (VA1), which simply combines nominal case HC aging analysis and SSTA with process variations, without considering the impacts of process variations on HC aging (i.e., the $F(\mathbf{\Delta X}) + G(\mathbf{\Delta Y})$ term in (4.27)). The second, denoted as Variational Analysis 2 (VA2), is similar to VA1 but does include the impacts of process variations on HC aging (the $F(\mathbf{\Delta X}) + G(\mathbf{\Delta Y})$ term). The last one is a Monte Carlo (MC) simulation with the same assumptions as VA2. The runtime, mean and standard deviation (SD) value of the circuit delay degradation ΔD for these three methods are listed in Table 4.2, along with the fresh delay and delay degradation of the nominal case.

The VA2 results matched the MC simulation very well, with a 1.6% average error in $\mu_{\Delta D}$ and 4.1% average error of $\sigma_{\Delta D}$), and much lower runtime, indicating the proposed variation-aware degradation analysis approach VA2 is accurate and efficient. In comparison, the VA1 method, which ignores the interaction between process variations and HC aging, has much shorter runtime than VA2, has close $\mu_{\Delta D}$ results as compared to VA2 and MC (4.1% error to MC), but large errors on $\sigma_{\Delta D}$ (23.9% to MC).

Fig. 4.7 shows the probability distribution function (PDF) of circuit delay of benchmark b21 at t=10000 (a.u.). The visual match of VA2 and MC verifies that the circuit delay under HC aging and process variation effects follows Gaussian distribution with its mean and SD accurately predicted by proposed VA2 approach. The result from VA1 has noticeable error compared with VA2 and MC. These results indicate that VA2 has much better accuracy and should be used for degradation analysis in the variational case. However in the scenario of runtime in higher priority than accuracy, VA1 can be used instead.

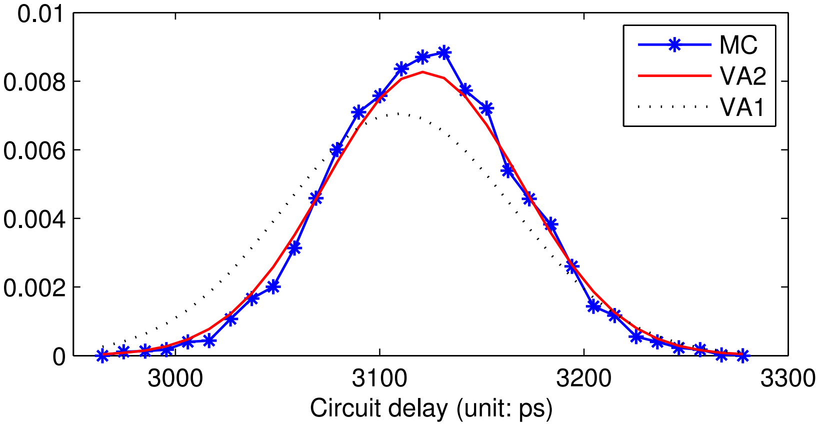

Figure 4.7: The circuit delay PDF of benchmark b21 under variations.

Table 4.2: Runtime and degradation comparison of variation-aware HC aging analysis methods (circuit delay unit: ps)

Circuit	Size		Nominal Delay		VA 1			VA 2			MC		
Name	#Cells	#Grids	Fresh	ΔD	T_{exe}	$\mu_{\Delta D}$	$\sigma_{\Delta D}$	T_{exe}	$\mu_{\Delta D}$	$\sigma_{\Delta D}$	T_{exe}	$\mu_{\Delta D}$	$\sigma_{\Delta D}$
c1908	442	9	835ps	136ps	2.2s	127ps	15.1ps	17.5s	140ps	12.4ps	233s	140ps	12.4ps
c2670	759	16	1228ps	134ps	3.8s	133ps	23.6ps	33.5s	141ps	21.2ps	401s	138ps	20.6ps
c3540	1033	16	1397ps	425ps	6.3s	410ps	27.6ps	56.9s	429ps	17.1ps	688s	428ps	16.1ps
c5315	1699	16	1133ps	80ps	8.3s	82ps	22.6ps	71.0s	86ps	20.9ps	817s	84ps	21.7ps
c6288	3560	64	2579ps	488ps	19.1s	444ps	44.4ps	171.4s	508ps	33.7ps	1926s	507ps	35.1ps
c7552	2361	36	1209ps	137ps	10.9s	145ps	25.5ps	93.7s	149ps	22.4ps	1098s	148ps	21.7ps
b14	4996	81	2586ps	790ps	29.1s	808ps	58.8ps	287.0s	810ps	42.9ps	3084s	808ps	44.9ps
b15	6548	100	2628ps	574ps	40.3s	584ps	53.6ps	400.7s	591ps	42.7ps	4272s	586ps	38.5ps
b17	20407	361	3201ps	130ps	128.9s	135ps	58.5ps	1317.6s	146ps	56.3ps	13286s	135ps	58.0ps
b20	11033	169	2586ps	501ps	64.5s	493ps	52.5ps	657.1s	512ps	45.1ps	6653s	507ps	43.1ps
b21	10873	169	2837ps	269ps	61.8s	269ps	53.8ps	598.6s	285ps	48.2ps	6353s	281ps	45.7ps
b22	14794	225	2845ps	326ps	84.6s	334ps	57.2ps	829.9s	334ps	48.0ps	8605s	328ps	47.3ps
Average Error to MC						3.6%	23.9%		1.6%	4.1%			

4.6.3 Delay degradation under both HC and BTI effects

To provide a holistic picture, we now show the impact of aging due to all major reliability issues. In addition to HC effects, the analysis of bias-temperature instability (BTI) effects is also added to determine circuit delay degradation. For BTI, we follow the SP/AF based modeling and analysis approach proposed in [79], and the assumption of the relative relationship between HC and BTI effect is based on the results of [1]. For simplicity the BTI degradation is considered for nominal case, with the time exponent assumed to be n=1/6.

Fig. 4.8 plots the circuit delay degradation as a function of time for nominal HC aging, nominal BTI degradation, HC aging under process variations using VA2 (results shown as $\mu + 3\sigma$ values), and the total delay degradation. The results indicate that BTI is the dominant aging effect in the early lifetime of digital circuit but grows slower with time, while HC effect begins low but grows faster, and surpasses BTI effect in the late lifetime. This is in consistent with the fact that HC effect has a larger time exponent (∼1/2) than the BTI effect (∼1/6).

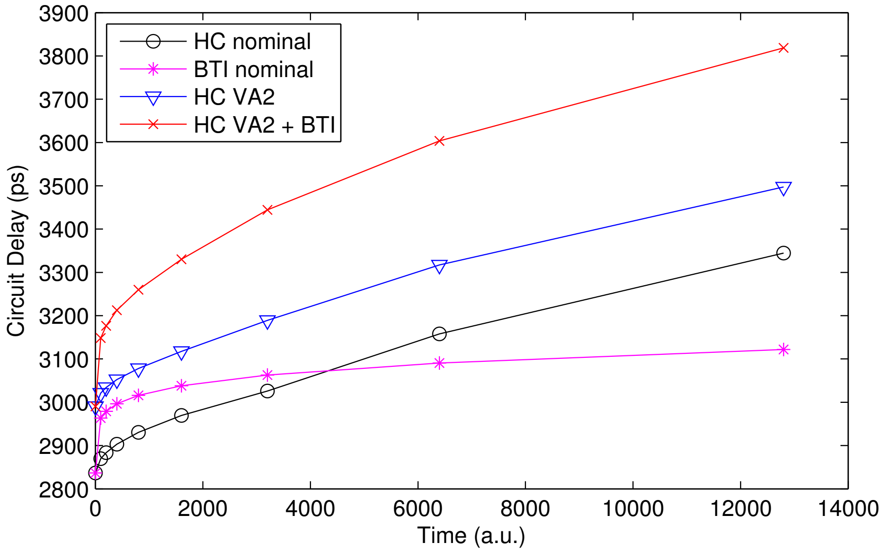

Figure 4.8: Delay degradation vs. time of b21 for different degradation components.

Fig. 4.9 shows the bar plot of the circuit delay degradations (normalized to fresh delay) of different benchmark, due to the effect of process variations (PV), BTI, and

HC, at time points of 100, 500, and 4000 (a.u.), representing the early, medium and late lifetime scenarios. The plot indicates that while effect of PV remain at the same level, the BTI effect dominates the early stage and the HC effect dominates the late stage of the circuit lifetime.

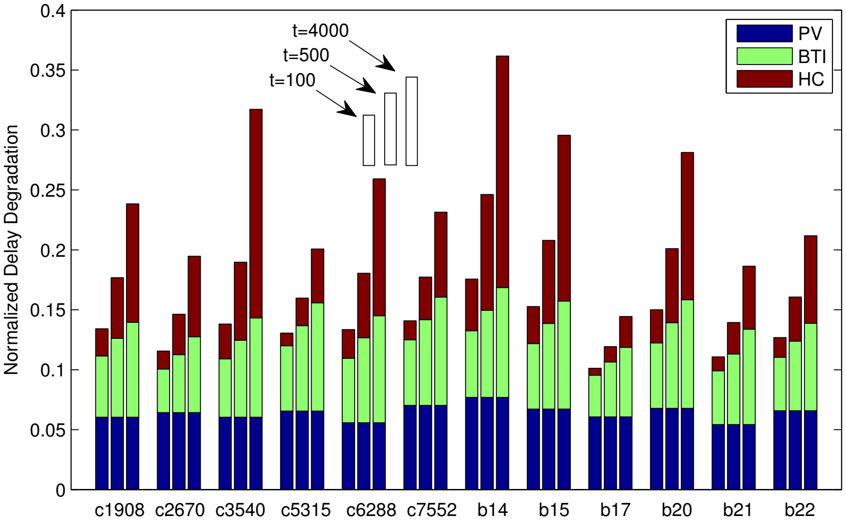

Figure 4.9: Circuit degradation components from process variations (PV), BTI effects, and HC effects.

4.7 Conclusion

This chapter focuses on the HC effect in large scale digital circuits, and proposes scalable approaches for analyzing CHC/CCC-induced delay degradation, with innovations in analysis at the transistor, cell, and circuit levels. The proposed approaches can handle nominal case and variation-aware case, both validated by Monte Carlo simulations on benchmark circuits and are shown to be efficient and accurate.

The interactions between process variations and HC effects are investigated and discovered to have nonnegligible impact to the circuit degradation, and need to be included in the analysis. The deceleration dynamics of HC aging, as well as its dependence on signal transition distributions are discovered to be negligible, thus can be approximated using simple alternatives so that the computational complexity could be effectively reduced.

Chapter 5

The Impact of BTI Variations on Timing in Digital Logic Circuits

A new framework for analyzing the impact of bias-temperature instability (BTI) variations on timing in large-scale digital logic circuits is proposed in this chapter. This approach incorporates both the reaction-diffusion model and the charge trapping model for BTI, and embeds these into a temporal statistical static timing analysis (T-SSTA) framework capturing process variations and path correlations. Experimental results on 32nm, 22nm and 16nm technology models, verified through Monte Carlo simulation, confirm that the proposed approach is fast, accurate and scalable, and indicate that BTI variations make a significant contribution to circuit-level timing variations.

5.1 Introduction

Reliability issues in very large scale integrated (VLSI) circuits have been a growing concern as technology trends in semiconductor technologies show progressive downscaling of feature sizes. One of the major reliability issues is bias-temperature instability (BTI), which causes the threshold voltage, V_{th}, of CMOS transistors to increase over time under voltage stress, resulting in a temporally-dependent degradation of digital logic circuit delay. Various optimizations have been proposed to cope with this degradation, such as slowing the operating frequency with time, adding delay guardbands, and using adaptive methods to recover from delay degradation.

The reaction-diffusion (R-D) model [21–24], based on dissociation of Si–H bonds at the Si/SiO$_2$ interface, has been the prevailing theory of BTI mechanism and has been widely used in research on circuit optimization and design automation. There have been considerable amount of work based on the R-D for circuit analysis [25–27], degradation monitoring [28, 29], and design mitigation techniques [30–37]. However, over the years, several limitations in the theory have been exposed. For instance, in R-D theory, the rate of recovery is determined by the diffusion of neutral hydrogen atoms, which is not affected by the gate bias. However the measured device recovery begins faster and lasts longer than the prediction of R-D theory, and shows strong dependence on the applied gate bias. An alternative mechanism for explaining BTI effects is the charge trapping and detrapping model [38–41], in which the defects in gate dielectrics can capture charged carriers, resulting in V_{th} degradations. The major difference between the two models is the nature of the diffusing species and the medium that facilitates the diffusion. Based on published works, both R-D and charge trapping mechanisms exist in current semiconductor technologies, and the superposition of both models is shown to better match experimental device data [24].

In nanometer-scale technologies, variations in the BTI effect are gaining a great deal of attention under both R-D and charge trapping frameworks, due to the random nature of defect localization in smaller and smaller transistors; together, these result in increased variations in the number of defects in a transistor. While there has been a great deal of research on timing variability due to process variations [17, 83, 84], and a few previous works have combined random variation effects from process variations with deterministic BTI degradations [85–87], the problem of BTI variations has not received much attention.

Most of the published circuit-level works incorporating BTI variations are based on the variability model of ΔN_{IT} randomness within the R-D framework, introduced by [42]. This model was applied to analytically determine the effect of BTI variations on SRAM and logic cells, and on circuit and pipeline performance using Monte Carlo simulations in [88, 89]. However, as explored in this chapter, for digital logic circuits, the ΔN_{IT} variation in this R-D based model has a relatively small impact on circuit timing variation, as compared with variations under the charge trapping model and process variations. Another model of the BTI-related variations was considered in [86], as caused by process perturbations. Since these small model perturbations lead to a relatively small change in the BTI-driven delay shift, the impact on circuit timing is a second-order perturbation that is relatively small.

On the other hand, the variations of device-level BTI degradations under the charge trapping model has been discovered to be a significant issue for nanoscale transistors. Charge trapping and detrapping at each defect are random events that are characterized by the capture and emission time constants. This paradigm is intrinsically statistical and it captures not only the variations in the number of defects, but also the variations in ΔV_{th} induced by each defect [43–45]. Under this statistical model, the variation of device lifetime increases significantly, especially for devices with a smaller number of defects N, as illustrated in Fig. 5.1.

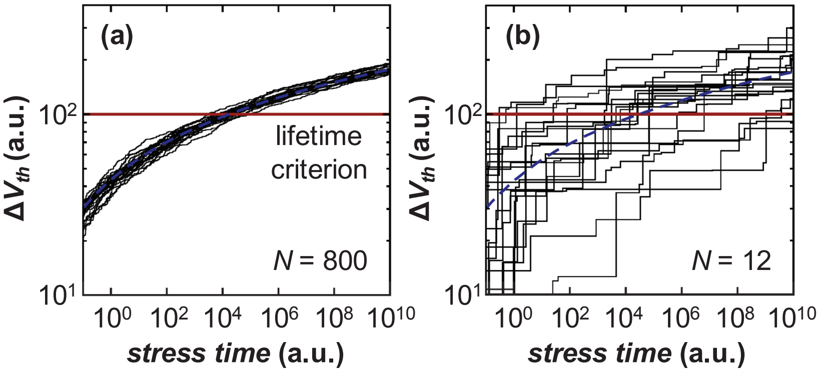

Figure 5.1: [3]: (a) Narrow distribution of lifetime in large devices where randomness averages out; (b) Large variation of lifetime in small devices where stochasticity predominates.

However, the impact of BTI variations under the charge trapping model on circuit performance has not received much attention, with only limited works that explore this issue beyond the device level. In [3], models and approaches were proposed for analyzing the impact of BTI variations on circuit performance; however the proposed SPICE-based atomistic approaches are time-consuming and not scalable to large-scale circuits. As we will show, the charge trapping model is the dominant component of BTI intrinsic variations, and makes significant contributions to circuit delay variation. Furthermore its impact grows rapidly as devices scale down, posing increasingly severe reliability issues to digital logic circuits.

In this chapter, we first introduce the notion of precharacterized *mean defect occupancy probability* for the charge trapping model to effectively reduce the complexity of circuit-level analysis and to make it possible to handle large-scale circuits. Then we incorporate variations under both the R-D model and the charge trapping model into a novel temporal statistical static timing analysis (T-SSTA) framework, capturing randomness from both process variations and temporal BTI degradations. We exercise this approach on large digital logic circuits and show simulation results for the 32nm, 22nm, and 16nm technology nodes. The correlation of process parameters due to path reconvergence is considered efficiently in modeling and analysis to guarantee both high accuracy and low complexity. To the best of our knowledge, this is the first circuit-level work that incorporates variations in BTI effects into SSTA under a scalable and computationally efficient procedure.

Our experimental results are based on simulations, and show that the proposed analysis approach has an accuracy that lies within 2.2% of Monte Carlo simulation while speeding up the calculation by 15×. Averaging over all benchmarks considered in our work, the fraction of the variance attributable to process variations, BTI R-D effects, and BTI charge trapping effects is, respectively, 81%, 3%, and 16% at the 32nm node, 70%, 4%, and 26% at the 22nm node, and 66%, 5%, and 29% at the 16nm node. Thus, under these models, the relative role of BTI charge trapping to circuit variability is projected to increase significantly in the future, but is less than the contribution of process variations.

5.2 Modeling Variations

This section introduces the models used to capture the effects of variations that affect BTI-induced aging. We begin by discussing BTI variations under both the R-D and charge trapping models. Next, we overview models for process variations, including spatial correlation effects. As in [24], the total threshold degradation ΔV_{th} of an MOS device is modeled by superposition as

$$\Delta V_{\text{th}} = \Delta V_{\text{th-RD}} + \Delta V_{\text{th-CT}} + \Delta V_{\text{th-RDF}}, \tag{5.1}$$

in which the BTI terms $\Delta V_{\text{th-RD}}$ and $\Delta V_{\text{th-CT}}$ are independent Gaussian random variables that will be given in (5.5) and (5.15), and $\Delta V_{\text{th-RDF}}$ is the variation component due to random dopant fluctuation (RDF) [90,91], which also follows Gaussian, and have no spatial correlations [82]. For each transistor, the sum, ΔV_{th}, of Gaussian variables is still a Gaussian, and this sum is an independent random variable for different MOS transistors.

5.2.1 BTI Variability under the R-D Model

Under the R-D framework, the mechanism of BTI in a MOS transistor is explained through the dissociation of Si–H bonds at the Si/SiO$_2$ interface and the diffusion of hydrogen into dielectric and gate. The number of generated interface traps, Si$^+$, is denoted as ΔN_{IT}, and absolute value of the induced threshold voltage shift, ΔV_{th}, is

$$\Delta V_{\text{th}} = \frac{q \Delta N_{\text{IT}}}{C_{\text{ox}}} \tag{5.2}$$

Under the R-D model, the long term threshold voltage shift ΔV_{th} under AC BTI stress is modeled in [88, 92] as

$$\Delta V_{\text{th}}^{(\text{nom})} = f_{\text{AC}}(SP) \cdot K_{\text{DC}} \cdot t^n \tag{5.3}$$

in which K_{DC} is a technology dependent constant for DC BTI degradation, and $f_{\text{AC}}(SP)$ is the coefficient that captures the AC degradation with signal probability SP (the probability of effective BTI stress). The function $f_{\text{AC}}(SP)$ can be precomputed numerically using method proposed in [23].

For deeply scaled technologies, the device size is small enough that ΔN_{IT} is a random variable, modeled as a Poisson distribution [42]:

$$\begin{aligned} \Delta N_{\text{IT}} &\sim \text{Poisson}(\lambda), \\ \text{where } \lambda &= \Delta N_{\text{IT}}^{(\text{nom})} = \Delta V_{\text{th}}^{(\text{nom})} \cdot C_{\text{ox}}/q \end{aligned} \tag{5.4}$$

Our reliability analysis focuses on late lifetime behavior, when the average numbers of interface traps λ in MOS transistors have relatively large values. For instance, the value of λ corresponding to $\Delta V_{\text{th}} = 0.1\text{V}$ for a device with $\frac{W}{L} = 2$ is about 49 for 32nm PTM [63] model, or 15 for 16nm PTM model, and it increases proportionally with the device size. It is well-known that for $\lambda > 10$, a Gaussian approximates the Poisson distribution well [93]. Therefore, to simplify our analysis without significant loss of accuracy, this Poisson distribution is approximated as a Gaussian distribution with the same mean and variance $\mu = \sigma^2 = \lambda$, hence $\Delta N_{\mathrm{IT}} \sim N(\lambda, \lambda)$. From (5.2), the threshold voltage degradation under R-D model has the distribution

$$\Delta V_{\text{th-RD}} \sim N\left(\frac{q\lambda}{C_{\mathrm{ox}}}, \frac{q^2\lambda}{C_{\mathrm{ox}}^2}\right) \tag{5.5}$$

As will be shown in Sec 5.4, this distribution approximation does not induce significant errors to the circuit level results.

5.2.2 BTI Variability under the Charge Trapping Model

Recent work [38] on the BTI effect of small-area devices reveals that the degradation and recovery of ΔV_{th} proceed in discrete steps, with variable heights, which could not be explained by the R-D model, but are fully consistent with charge trapping, which is also observed in random telegraph noise (RTN) and $1/f^2$ noise.

Based on these observations, a newer charge-trapping model was proposed for the BTI effect, in which each defect is characterized by parameters of the capture time τ_c and emission time τ_e, and each defect's contribution to the device threshold change, ΔV_t. These parameters are characterized using the time-dependent defect spectroscopy (TDDS) method [38,43], as a distribution shown in the form of a density map as Fig. 5.2, in which defects with similar time constants are binned together, and the total ΔV_t is shown in each grid.

If this characterization is performed on a large enough device, with the assumption that ΔV_t of all defects are independent and identically distributed (i.i.d.), the density map could be interpreted as the distribution of defects, in which each grid's value represents the probability of defects falling into that grid. The generation of this distribution is part of technology process characterization and independent of circuit structure.

Charge trapping (capture) and detrapping (emission) is a stochastic process. Following the models in [45], the capture time, τ_c, and the emission time, τ_e, are strongly dependent on bias voltage and temperature. In digital circuits there are only two non-transient voltage stages, logic 1 and logic 0, hence the bias condition can be simplified to two static modes of stress and relaxation. We capture the temperature dependence

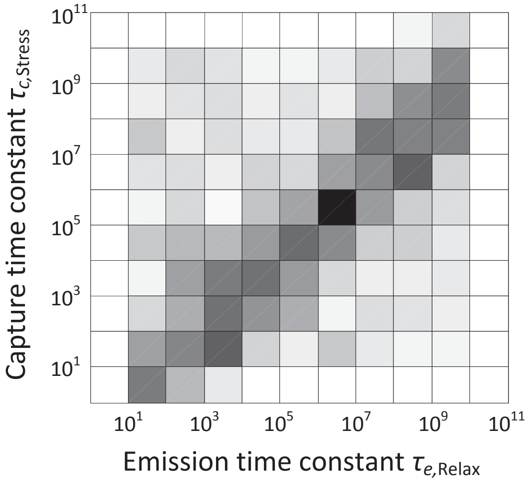

Figure 5.2: The distribution of defects according to their capture time constant τ_c and emission time constant τ_e.

effect by the use of a standard corner-based approach where the worst-case temperature corner is assumed. In this way each defect can be described by four time constants, denoted by the vector $\vec{\tau}$ as

$$\vec{\tau} = (\tau_{c,\text{Stress}}, \tau_{c,\text{Relax}}, \tau_{e,\text{Stress}}, \tau_{e,\text{Relax}}). \tag{5.6}$$

The defect occupancy probability (i.e., the probability of charge trapping) of a single defect with time constants $\vec{\tau}$ under AC stress of duty factor DF and time span t is derived in [45] to be:

$$P_c(DF, t, \vec{\tau}) = \frac{\tau_e^*}{\tau_c^* + \tau_e^*}\left(1 - \exp\left(-\left(\frac{1}{\tau_c^*} + \frac{1}{\tau_e^*}\right)t\right)\right), \tag{5.7}$$

Here the duty factor DF of a device under AC stress is defined as the probability of the transistor in accumulation mode that is effective for BTI stress (in some papers, DF is also referred to as the signal probability SP). The parameters τ_c^* and τ_e^* are defined as the effective capture and emission time constants under AC stress, which account for the duty factor effect:

$$\frac{1}{\tau_c^*} = \frac{DF}{\tau_{c,\text{Stress}}} + \frac{1-DF}{\tau_{c,\text{Relax}}} \tag{5.8}$$

$$\frac{1}{\tau_e^*} = \frac{DF}{\tau_{e,\text{Stress}}} + \frac{1-DF}{\tau_{e,\text{Relax}}} \tag{5.9}$$

Fig. 5.3 shows an example plot of the occupancy probability function, $P_c(DF, t, \vec{\tau})$, of a single defect as defined in (5.7), with the values of the time constants shown in the figure. The plot indicates that the occupancy probability P_c increases gradually with DF, but rises rapidly with time at the range of 10^5 to 10^6 a.u..

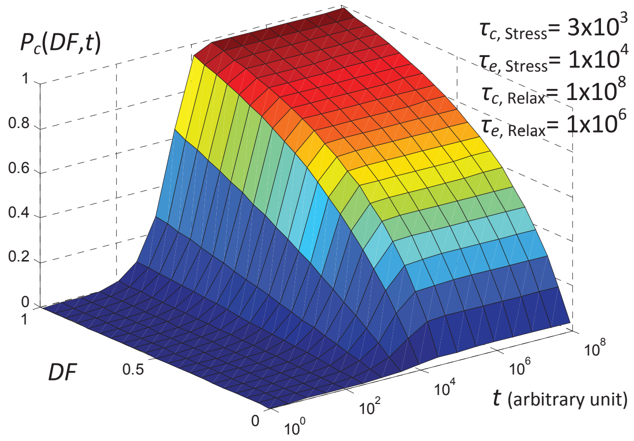

Figure 5.3: An example plot of defect occupancy probability function $P_c(DF, t, \vec{\tau})$ of a single defect.

Since the defect precursors (Si$-$Si bond in the SiO$_2$ dielectric according to [94]) are created in the fabrication process and uniformly distributed in the dielectric layer, the statistical distribution of capture/emission time constants associated with each defect is i.i.d. For each defect, the four components of $\vec{\tau}$ are correlated [38], and their joint distribution can be characterized for a specific technology. In this chapter, we follow the assumptions in [3] to generate the distributions of time constants. Fig. 5.2 shows an example 2-D histogram of the joint distribution of $\tau_{c,\text{Stress}}$ and $\tau_{e,\text{Relax}}$, which are the dominant components of $\vec{\tau}$. The proposed approaches in this chapter are independent of the distribution of $\vec{\tau}$.

We introduce the concept of the *mean defect occupancy probability*, $\bar{P}_c(DF, t)$, which captures the expected value of the probability of a defect charged with carriers (i.e., captured), based on the single defect model of (5.7), and $f(\vec{\tau})$, the joint pdf of $\vec{\tau}$:

$$\bar{P}_c(DF, t) = \int P_c(DF, t, \vec{\tau}) f(\vec{\tau}) d\vec{\tau} \tag{5.10}$$

Fig. 5.4 shows an example of $\bar{P}_c(DF, t)$ function corresponding to the assumed $f(\vec{\tau})$

plotted in Fig. 5.2. This plot indicates that the mean occupancy probability is a monotonically increasing function of both DF and time. Due to averaging effects over large number of defects with different $\vec{\tau}$, $\bar{P}_c(DF, t)$ changes more gradually with time, compared with P_c of a single defect in Fig. 5.3.

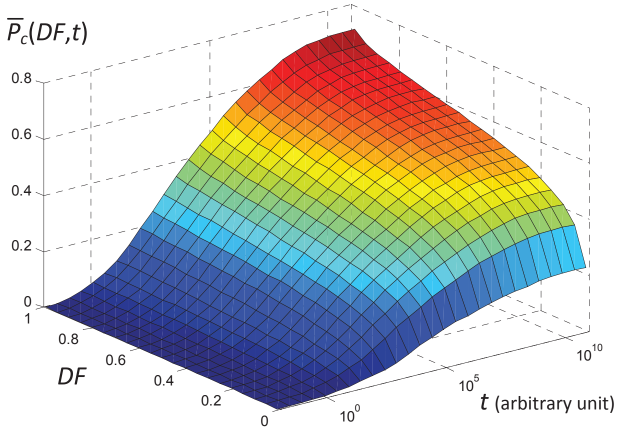

Figure 5.4: The plot of mean defect occupancy probability function $\bar{P}_c(DF, t)$.

Since $\bar{P}_c(DF, t)$ is only determined by the distribution $f(\vec{\tau})$ and is independent of the circuit structure, it can be pre-characterized numerically using (5.10) and stored in a look-up table (LUT) for use in the circuit analysis.

For small-geometry devices, the number of defects in a MOS transistor is a relatively small number with relatively large variation [43]. For a transistor of length L and width W, the total number of oxide defects n is empirically modeled as a Poisson distribution [42]:

$$\begin{aligned} n &\sim \text{Poisson}(N), \\ \text{where} \quad N &= N_{ot} W L. \end{aligned} \tag{5.11}$$

Here N_{ot} is the density of defects in the dielectric, and N is the total number of defects in the MOS transistor. Note that the Poisson distribution in (5.11) has similar form as the R-D model (5.4), but they are from different underlying mechanisms: R-D is modeled with interface traps (Si$-$H bond), while T-D is modeled with bulk oxide traps (missing oxygen atom in Si$-$O$-$Si bond [94]). Both kinds of traps are modeled as Poisson distributions due to the random location of the traps in small devices, however these two distributions are not correlated in nature.

Similarly, the number of *occupied* defects, n_c, in a transistor also has a Poisson distribution[1] , with its mean value N_c calculated as follows.

$$\begin{aligned} n_c &\sim \text{Poisson}(N_c), \\ \text{where } N_c &= N \cdot \bar{P}_c(DF, t) \end{aligned} \tag{5.12}$$

Observed in [44], the BTI-induced threshold degradation corresponding to each single defect follows an exponential distribution. Each defect $k = 1, ..., n$, contributes a threshold degradation of:

$$\begin{aligned} \Delta V_{\text{th}}^{(k)} &\sim \text{Exp}(\eta), \\ \text{where } \eta &= \eta_0/(WL). \end{aligned} \tag{5.13}$$

Like N_{ot}, η_0 is a technology-specific constant.

The total threshold voltage degradation, ΔV_{th}, of a transistor is the sum of contributions $\Delta V_{\text{th}}^{(k)}$ from all *occupied* defects k in the transistor, i.e.,

$$\Delta V_{\text{th}} = \sum_{k=1}^{n_c} \Delta V_{\text{th}}^{(k)}. \tag{5.14}$$

A closed form of this sum is derived in [44], and the PDF of ΔV_{th} turns out be to a complex distribution with mean $\mu = N_c\eta$ and variance $\sigma^2 = 2N_c\eta^2$. The mean value corresponds to the nominal case (i.e., each of N_c defects results in a threshold degradation of η). In [44], the probit plot of ΔV_{th} indicates that for an adequate number of defects (e.g., $N_c \geq 10$), the transistor ΔV_{th} distribution can be approximated as a Gaussian by matching the mean and variance, resulting in the distribution:

$$\Delta V_{\text{th-CT}} \sim N(N_c\eta, 2N_c\eta^2) \tag{5.15}$$

When the number of occupied defects, N_c, is sufficiently large, this Gaussian approximation is justified by central limit theorem (CLT), using the fact that the total

[1] The number of occupied defects in a device follows a Poisson distribution by definition because (a) each occupied defect has the same occurrence rate $N_c/(WL)$ within the device area of W by L, and (b) the occurrence of all occupied defects are independent with each other. This is similar to the number of all defects which follows $n \sim \text{Poisson}(N)$, and is verified by experimental results in Sec 5.4.

threshold degradation is the sum of ΔV_{th} from each defect, which are i.i.d. exponential. For smaller devices with lower values of N_c, this Gaussian approximation is not necessarily accurate for individual devices, but the circuit level timing analysis results still have good accuracy compared with Monte Carlo simulation, which can be justified by the central limit theorem (CLT) because the circuit delay is the sum of the cell delays along the critical paths and approaches a Gaussian distribution. A more detailed discussion about this Gaussian approximation model is available in Sec 5.4.

5.2.3 Process Variations and Spatial Correlation

Variations in the process parameters also contribute to BTI variability. Process variations are typically classified as lot-to-lot, die-to-die (D2D), and within-die (WID) variations, according to their scope; they can also be categorized, based on their causes and predictability, as systematic or random variations. Some (but not all) WID variations exhibit spatial dependence knows as spatial correlation, which must be considered for accurate circuit analysis.

We employ a widely-used variational paradigm, where a process parameter X is modeled as a random variable about its mean, X_0, as [80]:

$$\begin{aligned} X &= X_0 + \Delta X \\ \Delta X &= X_g + X_s + X_r \\ \sigma_X^2 &= \sigma_{X_g}^2 + \sigma_{X_s}^2 + \sigma_{X_r}^2 \end{aligned} \tag{5.16}$$

Here, X_g, X_s, and X_r stand for, respectively, the global component (from lot-to-lot or D2D variations), the spatially correlated component (from WID variation), and the residual random component of process variations. Under this model, all devices on the same die have the same global part X_g. The spatially correlated part is modeled using a widely-used grid-based method [17] for the parameters that exhibit this property, and is zero for those that are spatially uncorrelated. Under the spatial correlation model, the entire chip is divided into grids. All devices within the same grid have the same spatially correlated part X_s; the X_s parameters for devices in different grids are correlated, with the correlation falling off with the distance. The random part X_r is unique to each device in the system.

In this chapter we consider the variations in the transistor width (W), the channel length (L), the oxide thickness (T_{ox}), as well as shifts in the threshold voltage V_{th} due to random dopant fluctuations (RDFs). In other words, for each device, X represents elements of the set $\{W, L, T_{\mathrm{ox}}, V_{\mathrm{th}}\}$. As in the large body of work on SSTA, we assume Gaussian-distributed parameters for each of these process parameters, with W and L exhibiting spatial correlation, and T_{ox} and V_{th} being uncorrelated from one device to the next. The spatial correlation structure is extracted as a correlation matrix [64], and processed using principal components analysis (PCA) to facilitate fast timing analysis [17]. The process parameter value in each grid is expressed as a linear combination of the independent principal components, with potentially reduced dimension.

Notationally, we express each process parameter X as a vector in a random space, with basis $\mathbf{e} = [\mathbf{e}_g, \mathbf{e}_s, \mathbf{e}_r, \epsilon]^{\mathbf{T}}$, as

$$\begin{aligned} X &= X_0 + \Delta X = X_0 + \mathbf{k}_X^{\mathbf{T}}\mathbf{e} \\ &= X_0 + \mathbf{k}_{Xg}^{\mathbf{T}}\mathbf{e}_g + \mathbf{k}_{Xs}^{\mathbf{T}}\mathbf{e}_s + \mathbf{k}_r^{\mathbf{T}}\mathbf{e}_r + k_\epsilon \epsilon \qquad (5.17) \\ \sigma_X^2 &= \mathbf{k}_X^{\mathbf{T}}\mathbf{k}_X, \quad \mathrm{cov}(X_i, X_j) = \mathbf{k}_{Xi}^{\mathbf{T}}\mathbf{k}_{Xj} - k_{\epsilon_i}k_{\epsilon_j} \qquad (5.18) \end{aligned}$$

Here, $\mathbf{e}_g = [e_{Wg}, e_{Lg}]^{\mathbf{T}}$ is the basis for the global part (T_{ox} variation and RDF effect are purely random hence do not have a global part), $\mathbf{e}_s = [e_1, ..., e_t]^{\mathbf{T}}$ is the basis of principal components for the spatially correlated part, in which t is the number of dimensions after the PCA processing of correlated part, and $\mathbf{e}_r = [\epsilon_1, ..., \epsilon_m]^{\mathbf{T}}$ is the basis of random part. The dimension of random part, m, will depend on the implementation of the SSTA algorithm, and can vary from constant to linear (of circuit size), as will be shown later in this chapter. The random basis $\mathbf{e}_r$ and its coefficient vector $\mathbf{k}_r$ are implemented using a sparse data structure. The Gaussian variable $\epsilon \sim N(0, 1)$ is a separate independent random part for use in circuit-level timing analysis.

5.2.4 Consideration of Process Variations and BTI Interaction

Process variations are created at manufacture time, while BTI degradation occurs during the circuit operation. Therefore the effect of process variations is independent from BTI, but the BTI effect will be impacted by process variations, i.e., the BTI degradation is dependent on the actual W, L and T_{ox} of a transistor. This chapter assumes the process variations and BTI effects (both R-D and charge trapping model including variabilities) to be independent and uses a superposition model to calculate the total effect on circuit-level degradations. This is based on the following facts and considerations.

- The impact of process variations on BTI degradation is a second order effect that is relatively small in nature.

- For W and L variations, [95] indicates the NBTI effect is more pronounced in narrow and long transistors. However the transistors on the critical paths are normally sized larger (wider) for timing performance, hence less affected by the W and L variations.

- For T_{ox} variation, a smaller T_{ox} causes elevated BTI degradation speed, but also gives smaller initial V_{th} and propagation delay. Therefore the interaction effect actually cancels out with each other to some degree, and ignoring it yields pessimistic and safe approximations.

- The independent assumption simplifies the modeling and analysis and helps achieve linear computational complexity and good scalability (Section 5.3.3).

5.3 Timing Analysis under Variations

This section introduces the logic cell delay model under BTI variations and process variations. Based on this model, a scalable approach for statistical timing analysis of large digital logic circuits is outlined.

5.3.1 Cell Timing Model and Characterization

We use a cell delay degradation model that is similar to [78]. The delay d_i of cell i is modeled using a first-order Taylor approximation, as a linear function of process parameters W_j, L_j and $T_{\text{ox-}j}$ of each transistor j in cell i, and BTI degradation $\Delta V_{\text{th}}^{(j)}$ of each transistor j:

$$d_i = d_{i0} + \Delta d_i = d_{i0} + \sum_{X \in \mathbf{P}_i} \frac{\partial d_i}{\partial X} \Delta X$$

Here $\mathbf{P}_i = \{W_j, L_j, T_{\text{ox-}j}, V_{\text{th}}^{(j)}\}, j \in$ cell i is the set of variational parameters. The nominal propagation delay d_{i0} and its sensitivity $\partial d_i / \partial X$ to parameter $X \in \mathbf{P}_i$ are computed using standard techniques through SPICE simulations. This part of the calculation is circuit-independent and performed as part of library characterization.

Since all variational parameters $X \in \mathbf{P}_i$ are expressed as vectors in the random variable space $\mathbf{e}$ in Section 5.2.3, d_i, which is a linear combination of these parameters, is also a vector in space $\mathbf{e}$:

$$\begin{aligned} d_i &= d_{i0} + \left(\sum_{X \in \mathbf{P}_i} \frac{\partial d_i}{\partial X} \mathbf{k}_X \right)^{\mathbf{T}} \mathbf{e} \\ &= d_{i0} + \mathbf{k}_{d_g}^{\mathbf{T}} \mathbf{e}_g + \mathbf{k}_{d_s}^{\mathbf{T}} \mathbf{e}_s + \mathbf{k}_{d_r}^{\mathbf{T}} \mathbf{e}_r \end{aligned} \tag{5.19}$$

Here the random part $\mathbf{e}_r = \{\epsilon_X\}_{X \in \mathbf{P}_i}$ is extended to include the random parts from all variational parameters $X \in \mathbf{P}_i$ in cell i.

5.3.2 Circuit Level Timing Analysis

At the circuit level, timing analysis is performed using a PERT-like traversal [17] at a fixed time point, where the contributions of the temporal BTI variations can be characterized using the models described in Sections 5.2.1 and 5.2.2. The V_{th} degradation due to these two models are uncorrelated, and are found to substantially affect the circuit level delay.

In our initial implementation of algorithm, as in [84], the random part $\mathbf{k}_r^T \mathbf{e}_r$ of arrival time is merged into the separate independent term $k_\epsilon \epsilon$ that is the product of scalars to reduce the computational complexity. The temporal statistical static timing analysis (T-SSTA) result of this method is denoted as T-SSTA1. Table 5.1 shows the mean and standard deviation of circuit delay degradation on benchmark c3540 under a 16nm technology model at t=2000 (a.u.), splitting the contribution of the mean and variance into those attributable to BTI R-D, BTI charge trapping (CT), process variations (PV), and finally presenting the combined values (ALL). The results of mean and standard deviation calculated by Monte Carlo (MC) simulation are listed as reference, and the results indicate that the mean value of delay degradation is mainly contributed by BTI RD and BTI CT, while the standard deviation is mainly contributed by BTI CT and PV effects. By comparing results for "ALL" from T-SSTA1 method with MC, we can see T-SSTA1 overestimates the mean value and underestimates the standard deviation, where the errors are mainly coming from the BTI CT part.

Table 5.1: T-SSTA results under variations (time unit: ps)

c3540 16nm	MC		T-SSTA1		T-SSTA2		T-SSTA3	
D_0=582.3	$\mu_{\Delta D}$	$\sigma_{\Delta D}$	$\mu_{\Delta D}$	$\sigma_{\Delta D}$	$\mu_{\Delta D}$	$\sigma_{\Delta D}$	$\mu_{\Delta D}$	$\sigma_{\Delta D}$
BTI RD	23.8	3.6	23.8	3.7	23.8	3.7	23.8	3.7
BTI CT	29.8	8.9	29.7	8.8	29.6	8.8	29.6	8.8
PV	0.5	14.8	4.2	14.1	0.3	14.6	1.3	14.6
ALL	53.9	17.2	57.1	17.0	54.0	17.5	54.8	17.4

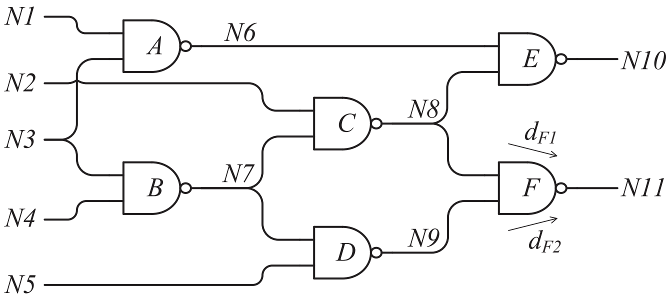

Figure 5.5: An example circuit showing path reconvergence.

Investigating this further, we find that the error between conventional method (T-SSTA1) and Monte Carlo (MC) simulation can be attributed to the correlations that arise due to path reconvergence, which are neglected in T-SSTA1. The BTI CT part of V_{th} degradation contains a significant amount of independent random component in the form (5.17), hence generates large errors due to path reconvergence. We illustrate the path reconvergence effect through Fig. 5.5, which shows an example circuit where the arrival time AT of node $N11$, denoted as AT_{N11}, is calculated as follows

$$AT_{N11} = \max\left(AT_{N8} + d_{F1}, AT_{N9} + d_{F2}\right) \tag{5.20}$$

where AT_{N8} and AT_{N9} stand for the arrival times of node $N8$ and $N9$, while d_{F1} and d_{F2} stand for the delays from the first and second input to the output of cell F. The arrival times and cell delays are modeled as vectors in random variable space $\mathbf{e}$. Note that since cell C and D have a common fanin of cell B, AT_{N8} and AT_{N9} are both dependent on the random component of the parameters of cell B, corresponding to the impact of X_r in Equation (5.16). As a result, the independent components in the expression for AT_{N8} and AT_{N9} are not independent of each other, but are correlated. However the conventional SSTA method, using an separate independent term $k_\epsilon \epsilon$ to replace $\mathbf{k}_r\mathbf{e}_r$, does not capture this path correlation and introduces errors. The same situation occurs when calculating the total delay using the maximum of AT_{N10} and AT_{N11}, which are correlated because the paths from node $N8$ reconverge.

One natural way to resolve this problem is to preserve the entire random part $\mathbf{k}_r\mathbf{e}_r$ when calculating the arrival times, by which the path correlation is captured. This method is denoted as T-SSTA2 in Table 5.1, and the results indicate this method is much more accurate than T-SSTA1. However, the cost paid for this accuracy is in the increased computation time associated with the growing size of the random part (e.g., AT_{N11} in the example contains components from cells B, C, D, and F). The computational complexity is discussed with more details in Section 5.3.3 and the experimental runtime and storage comparison will be given in Section 5.4.

We employ a third method, denoted as T-SSTA3 in Table 5.1, taken from [81], to provide a trade off between the accuracy and complexity. This removes only the smaller elements in the random vector $\mathbf{k}_r$ using preset threshold and merges them into the separate term k_ϵ. Results in Table 5.1 show that this method achieves good accuracy (within 2% error compared with T-SSTA2 and Monte Carlo) with low computation. We will expand on this in Section 5.4.

5.3.3 Computational Complexity

To calculate the circuit delay, the SSTA algorithm does a topological traversal through the digital circuit. For each node (logic cell), the timing analysis performs k sum-of-two and $k-1$ max-of-two operations, where k is the number of fan-in of the cell. In random space $\mathbf{e}$ with dimension d, the numbers of total sum and max operations for SSTA are

$$N_{sum} = n \cdot k \cdot d \tag{5.21}$$

$$N_{max} = n \cdot (k-1) \cdot d \tag{5.22}$$

Here $d = d_g + d_s + d_r$, in which d_g, d_s and d_r are the dimension of global component $\mathbf{e_g}$, spatial component $\mathbf{e_s}$ and random component $\mathbf{e_r}$, respectively. The values of d_g and d_s are well bounded by PCA algorithm therefore can be regarded as constant. The values of d_r depends on how the random part is handled as discussed in Section 5.3.2. For methods T-SSTA1 and T-SSTA3 d_r is bounded by a constant, while for T-SSTA2 d_r can grow significantly depending on the circuit size and structure. For simplicity it can be roughly approximated as $d_r \propto \sqrt{n}$, which corresponds to the depth of the circuit (number of cells on the critical paths). Therefore the computational complexity is $O(n)$ for T-SSTA1 and T-SSTA3, and $O(n^{1.5})$ for T-SSTA2. This result indicates the proposed T-SSTA3 method has good scalability to handle large scale circuits.

5.4 Experimental Results

Our approach for timing analysis under BTI variations and process variations is applied to ISCAS85 and ITC99 benchmarks. The benchmark circuits are mapped to a subset of the Nangate cell library [70] using ABC [69], with placement carried out using a simulated annealing algorithm. The benchmark circuits are scaled down to 32nm, 22nm and 16nm for comparisons under different technology models. The characterization of cell delay and of its sensitivities to variational parameters is performed using HSPICE simulation under PTM models [63]. Both the proposed analytical method and the Monte Carlo method (for verification) are implemented in C++ and run on a Linux PC (3GHz CPU, 2GB RAM).

The process variations in W, L, and T_{ox} are set to $3\sigma = 4\%$ of their mean values [7]. The V_{th} variation due to RDF is dependent on the device size [82]. It has a Gaussian distribution with mean value $\mu = 0$, and standard variation

$$\sigma_{V_{\text{th}}} = \sigma_{V_{\text{th0}}} \sqrt{\frac{W_0 L_0}{WL}} \tag{5.23}$$

in which $\sigma_{V_{\text{th0}}}$ is the RDF-induce threshold standard deviation of a minimum-sized device (W_0 by L_0). The value of $\sigma_{V_{\text{th0}}}$ is dependent on process parameters T_{ox} and N_a, as well as the doping profile of the channel [82]. Here we assume $3\sigma_{V_{\text{th0}}} = 5\%$ of the nominal V_{th}. The parameter variations of W and L are split into 20% of global variation, 20% of spatially correlated variation and 60% of random variation, while the variations of T_{ox} and V_{th} are fully random. The grid-based spatial correlation matrix is generated using the distance based method in [64], with the number of grids growing with circuit size, as shown in the Table 5.3.

The Monte Carlo simulation framework for verification of the proposed approach is set up as follows: the simulation program randomly generates 5000 circuit instances (we found it a good trade-off of accuracy and runtime). For each circuit instance, the ΔV_{th}

Table 5.2: Mean and SD of circuit delay using different methods (time unit: ps, V_{th} unit: mV, average error shown for $\mu_{\Delta D}$ and σ_D)

Circuit &		Initial	T-SSTA1		T-SSTA2		T-SSTA3		MC		$V_{\text{th-RD}}$		$V_{\text{th-CT}}$	
Technology		D_0	μ_D	σ_D	μ_D	σ_D	μ_D	σ_D	μ_D	σ_D	μ	σ	μ	σ
32 nm	c2670	734	802	13.3	795	14.6	796	14.4	794	14.7	17.8	3.3	16.9	6.1
	c3540	812	910	15.6	903	16.7	904	16.7	903	16.5	17.7	3.1	16.6	5.7
	c5315	666	736	12.9	735	13.1	735	13.1	735	12.9	18.1	3.2	17.5	6.0
	c6288	1416	1580	23.5	1574	24.1	1576	24.0	1574	23.6	17.3	2.9	16.5	5.3
	c7552	650	714	11.6	709	12.5	710	12.5	710	12.4	18.2	3.3	17.6	6.2
	b15	1416	1580	24.2	1571	26.2	1574	25.7	1572	25.9	17.0	3.1	16.5	5.6
	b17	1634	1770	27.0	1750	29.1	1757	28.5	1752	28.7	16.6	3.1	16.1	5.5
	b20	1432	1566	24.1	1554	26.8	1555	26.7	1555	26.3	17.8	3.2	17.3	6.0
	b21	1598	1765	27.0	1757	29.9	1758	29.7	1758	30.5	17.8	3.3	17.4	6.1
	b22	1520	1655	23.4	1645	25.3	1646	25.1	1645	25.3	17.4	3.2	17.0	6.0
Avg Err %			7.24	6.19	0.46	1.36	1.22	1.36						
22 nm	c2670	617	669	12.6	663	13.8	664	13.6	663	13.9	18.1	4.6	17.4	9.0
	c3540	671	760	15.1	754	16.5	755	16.4	754	16.7	17.6	4.3	16.7	8.1
	c5315	557	625	11.9	624	12.2	624	12.1	624	12.3	17.7	4.4	17.0	8.4
	c6288	1211	1366	22.8	1359	23.5	1362	23.3	1359	23.6	17.3	4.0	16.6	7.7
	c7552	549	607	10.6	601	12.3	602	12.2	601	12.3	17.9	4.5	17.3	8.8
	b15	1151	1287	23.9	1274	27.4	1275	27.0	1275	26.4	16.8	4.3	16.2	8.0
	b17	1312	1444	23.7	1437	24.4	1443	23.4	1440	25.4	16.5	4.3	16.0	7.9
	b20	1144	1293	23.4	1273	27.8	1274	27.6	1274	28.0	18.1	4.5	17.6	8.9
	b21	1251	1392	25.0	1388	26.9	1388	26.8	1388	26.6	18.0	4.5	17.6	8.9
	b22	1252	1379	23.0	1371	24.7	1372	24.6	1372	24.9	17.7	4.5	17.4	8.8
Avg Err %			7.48	8.45	0.75	1.50	1.30	2.16						
16 nm	c2670	537	592	13.9	582	15.7	584	15.4	582	15.9	18.5	6.2	17.9	12.7
	c3540	582	657	17.5	652	18.1	653	18.0	652	18.9	17.0	5.4	16.1	10.7
	c5315	489	570	14.2	567	14.6	568	14.5	568	15.1	18.2	5.8	17.7	12.0
	c6288	1092	1253	26.2	1247	26.7	1251	26.5	1247	26.6	17.3	5.3	16.6	10.5
	c7552	480	559	12.9	545	16.8	548	16.6	546	16.9	18.2	6.0	17.6	12.4
	b15	986	1143	26.0	1121	31.9	1124	31.2	1125	33.1	16.8	5.6	16.3	11.0
	b17	1100	1318	28.6	1293	33.2	1299	32.8	1293	33.6	16.7	5.6	16.2	11.0
	b20	941	1083	23.6	1075	23.5	1080	23.0	1078	25.2	17.9	5.9	17.4	12.1
	b21	1038	1158	26.7	1148	30.5	1149	29.8	1149	31.1	17.2	5.8	16.7	11.6
	b22	1072	1219	25.8	1206	29.0	1207	28.7	1208	29.1	17.9	6.0	17.5	12.2
Avg Err %			9.97	11.95	0.97	2.39	1.53	3.64						
Total Avg Err			8.23	8.86	0.73	1.75	1.35	2.39						

of each MOS transistor is calculated as the sum of the following three components:

(a) $\Delta V_{\text{th-RD}}$, which is set by (5.5) and is randomly generated based on the distribution of ΔN_{IT} as specified in (5.4),

(b) $\Delta V_{\text{th-CT}}$, which is set as the sum of ΔV_{th} of all defects that are randomly generated using distributions (5.11) and (5.13), and

(c) $\Delta V_{\text{th-RDF}}$, which is due to RDF effects and set by (5.23).

The contributions of $\Delta V_{\text{th-RD}}$ and $\Delta V_{\text{th-CT}}$ vary with different technologies [24]. In the experiments it is assumed that these two have comparable mean values, so that their contributions to circuit-level variations can be easily visualized. The process parameters W, L, and T_{ox} of each MOS transistor are also generated according to their distributions and correlation models. Then the propagation delay of each cell is calculated using (5.19) and pre-characterized cell delay and sensitivity data. Based on these values and a PERT-like traversal, the total delay of the circuit instances is evaluated using statistical static timing analysis (SSTA).

For each benchmark circuit, the mean and standard deviation of the circuit delay are calculated at time t=2000 (a.u.), using the proposed analytical method and Monte Carlo (MC) simulation. The three methods of handling random parts discussed in Section 5.3.2 are implemented separately. As before,

- T-SSTA1 merges the random part into one variable,
- T-SSTA2 preserves full random part, and
- T-SSTA3 partially lumps the random part.

Table 5.2 shows the nominal delay D_0 of each benchmark circuit, as well as the mean μ and standard deviation (SD), σ, of the circuit delay using three analytical methods and the MC simulation, at 32nm, 22nm, and 16nm. The last row shows the relative error of $\mu_{\Delta D}$ and σ_D of each analytical method, compared with MC.

The mean and SD of the ΔV_{th} contributions (averaged over all devices in the circuit) from the R-D model and from the charge trapping model are also listed in Table 5.2. Note that the simulation is based on the assumption that the ΔV_{th} contributions (mean value) from R-D model and charge trapping model are comparable. This assumption is made to give a general insight that the charge trapping model predicts significantly larger BTI variability than R-D model. The proposed approaches for circuit degradation analysis is actually independent with this assumption and can handle different cases of the BTI degradation model. In general cases of application, both R-D and charge trapping model of BTI effects can be characterized for given technology and used for analyzing the circuit timing degradations.

It is also worth noting that under certain cases (especially at 16nm, under the charge trapping model), the value of 3σ can be larger than μ, indicating Gaussian distribution may not be an accurate approximation of $\Delta V_{\rm th}$ since $\Delta V_{\rm th}$ from BTI effects should always be positive. However this inaccuracy of the Gaussian approximation is averaged out by the sum of delay along the critical path, and the circuit level delay, calculated by sum and max operations in SSTA, and approaches a Gaussian distribution according to the central limit theorem (CLT), which does not require the transistor $\Delta V_{\rm th}$ to be Gaussian. Therefore the proposed method appears to be robust even under this model inaccuracy, as verified by the good accuracy indicated in Table 5.2, and the visually-verified match between distribution functions plotted in Fig. 5.6, which shows an example of the circuit delay distribution for c3540 at 16nm at t=2000 (a.u.). The T-SSTA3 and MC methods match well, verifying the validity of our assumptions; T-SSTA1 is significantly different, due to the omission of path correlations.

Table 5.3: Comparison of computational complexities, where T_{exe}= runtime, [Cells] = average number of correlated cells.

Circuit	Size		T-SSTA1	T-SSTA2		T-SSTA3		MC
Name	#cells	#grids	T_{exe}	T_{exe}	[Cells]	T_{exe}	[Cells]	T_{exe}
c2670	759	16	3.4s	5.8s	26.2	6.7s	3.0	108s
c3540	1033	16	5.7s	13.5s	109.0	12.8s	2.7	201s
c5315	1699	16	7.2s	14.1s	40.8	15.2s	2.9	261s
c6288	3560	64	17.1s	137.8s	473.6	38.9s	2.8	627s
c7552	2361	36	9.8s	21.1s	53.7	20.3s	3.2	352s
b15	6548	100	34.8s	352.4s	512.1	89.6s	3.0	1181s
b17	20407	361	109.3s	1513s	421.6	306.0s	3.5	3645s
b20	11033	169	55.2s	482.1s	362.3	139.0s	3.4	1926s
b21	10873	169	52.9s	439.1s	351.4	133.5s	2.8	1845s
b22	14794	225	72.4s	671.2s	304.9	188.1s	3.1	2507s

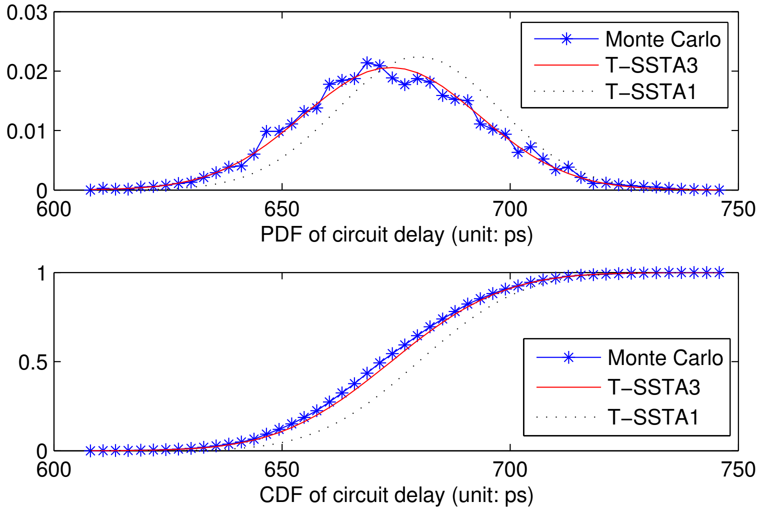

Figure 5.6: The delay PDF and CDF of c3540 with 16nm model.

Table 5.3 compares the runtime and storage complexity (in terms of the average number of correlated cells, denoted as [Cells]) of the analytical methods and MC. Fig. 5.7 shows the runtime vs. circuit size (number of logic cells) for the different methods.

The results indicate that the runtime of partially lumping random part (T-SSTA3) method grows linearly with circuit size increasing, indicating good scalability. It has an overall error of about 2% to MC, and is 15× faster on average. Furthermore, it reduces runtime by 60% and storage by 98% on average compared with T-SSTA2, with similar accuracy. The conventional method (T-SSTA1) has the shortest runtime, but has nearly 9% errors with respect to MC. The results also verify that the Gaussian approximations for ΔV_{th} in BTI R-D and charge trapping models are valid; the method is accurate, efficient, and scalable. Moreover, the standard deviation of circuit delay σ_D increases with technology downscaling, indicating that random timing variation attributable to BTI is a growing issue.

Fig. 5.8 shows the variance of circuit delay that originates from process variations, R-D BTI variations, and charge trapping BTI variations separately, for different benchmarks under the 32nm, 22nm, and 16nm technology models. For better presentation of data, the variances are normalized to the total variance of 32nm model for each benchmark. These results indicate that the charge trapping model is the dominant

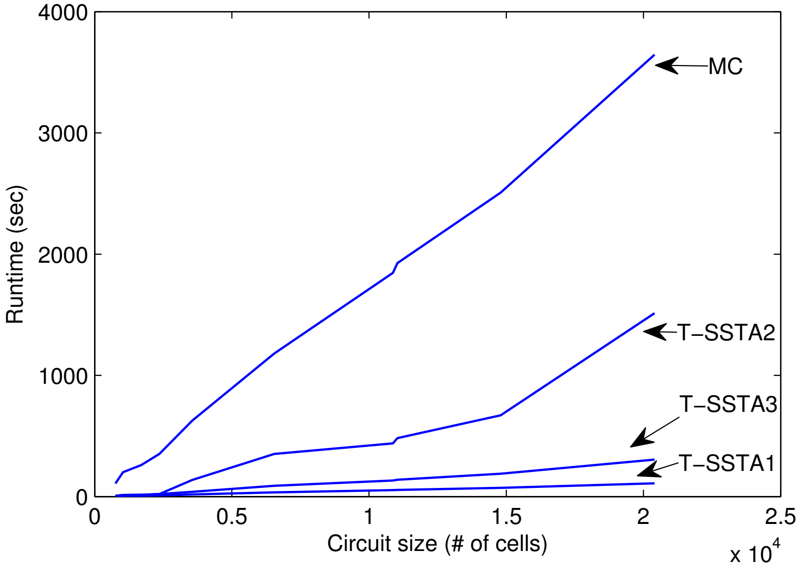

Figure 5.7: Runtime vs. circuit size of different methods.

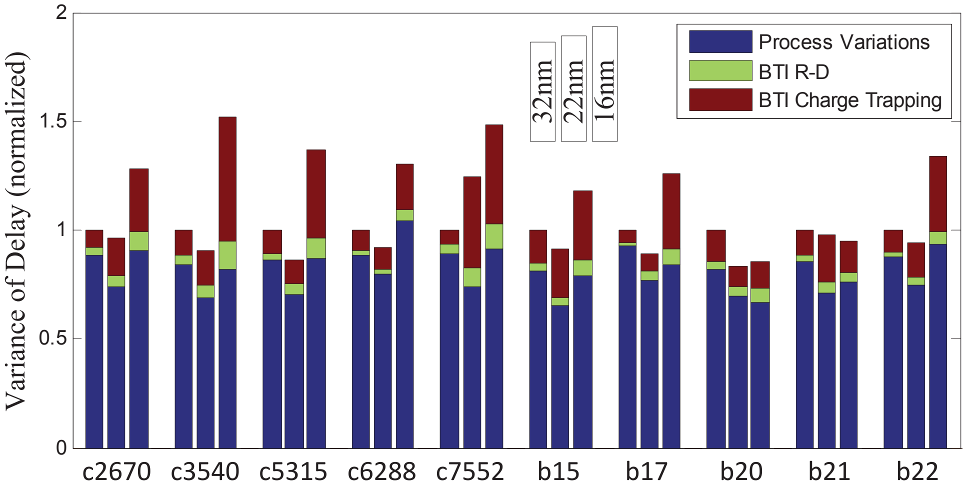

Figure 5.8: Relative contribution of BTI charge trapping, BTI R-D, and process variation to circuit delay variation.

component of BTI variations, and makes a significant contribution to circuit delay variation. In contrast, the BTI variations under the R-D model only introduce a relatively small portion of delay variations. Unlike process variations which have nearly constant influence on delay variation, the impact of BTI variations grows with scaling, becoming increasingly severe in future.

Further, according to the results in Table 5.2 and Fig. 5.8, the circuit level delay variation that can be attributed to BTI variations is not as significant as the single-device ΔV_{th} variation due to BTI effect of a small transistor shown in Fig. 5.1 (b). This is mainly due to the facts that (a) transistors on the critical paths usually have larger than minimum sizes to help with timing, and (b) the average out of randomness of the transistor ΔV_{th} on the critical paths due the sum of delay.

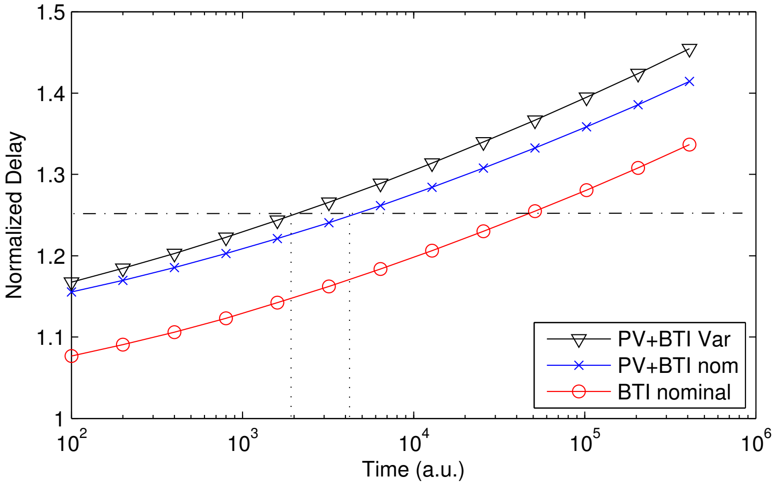

Figure 5.9: Delay degradation vs. time of c5315

Fig. 5.9 presents the circuit delay degradation vs. time curves of benchmark c5315 at 16nm. Three curves are shown for the normalized delay of (1) nominal BTI degradation, without any variation model; (2) $\mu + 3\sigma$ of process variation (PV) and nominal BTI degradation; and (3) $\mu + 3\sigma$ of PV and BTI with variabilities (under both R-D and charge trapping models). The results indicate that BTI degradation and variability, which grow with time, make up the dominant part of total delay degradation, especially at the later point of circuit lifetime. Furthermore, BTI variations has a significant impact on circuit reliability. In this case, the circuit lifetime will be overestimated by over 2× if BTI variations is not considered (lifetime defined as 25% increase of delay from time zero).

5.5 Conclusion

This chapter incorporates both the R-D and charge trapping models of BTI variations into a T-SSTA framework, capturing process variations and path correlations. Experimental results show that the proposed analysis method is fast and accurate. Our results indicate that the charge trapping mechanism, which has been neglected by the EDA field so far, is the dominant source of BTI variations, with significant and growing contributions to circuit timing variations.

Chapter 6

Conclusion

As the result of continuous technology scaling, the reliability issues in digital VLSI are arising to become major barriers of performance and lifetime. This thesis has focused on the reliability issues of TDDB, BTI, and HC, and has proposed accurate, scalable and efficient approaches to analyze circuit-level performance degradations under these effects, incorporating the effects of process variations. The proposed approaches take advantages of new physics-based device-level models, perform accurate cell-level characterization using SPICE-based methods, and analyze the circuit-level degradations using probabilistic and statistical techniques based on cell library and circuit operational information. This three-level hierarchy helps to achieve best accuracy at each level while effective reducing the computational complexity at the circuit level to be linear to the circuit size.

For TDDB analysis, considering catastrophic failures, we have developed a systematic approach for analyzing the failure probability of large circuits, resulting in some surprising conclusions. Specifically, we have discovered the circuit lifetime was greatly underestimated by traditional methods, while our approach predicts 4–6× longer lifetime by considering the inherent resilience of digital circuits to breakdown. We have also demonstrated that the oxide lifetime of circuits can be improved by using larger sized devices.

For parametric failures due to aging, our analysis of the HC effects has utilized new energy-driven models from the device community, and based on quasistatic characterization of cell libraries to determine the age gain per transition, we have performed efficient circuit-level analysis. This analysis shows that the conventional method has an error that is too large be practical, and that the interaction between process variations and HC effects must be incorporated due to have nonnegligible impact on circuit degradations. This is combined with BTI analysis to efficiently determine the temporal aging properties of a circuit.

In our work on BTI variability analysis, we have incorporated new models for BTI that have very recently gained currency in the device community, and have caused alarm due to the high impact of process variation on the CT effect. Our analysis shows that this impact is strongly attenuated for logic circuits, which reduce variability in aging both due to the tendency to use large devices on critical paths, and the cancellation effects over multiple stages of logic. However, even incorporating such effects, we have discovered that the BTI variability effect is a growing issue with technology scaling and the circuit-level impact requires serious consideration in the future. Although these new device models have not been incorporated together with the new HC models above during the period of this thesis, this is ripe ground for future work and we believe this thesis clearly lays the groundwork for such activity.

Finally, as a byproduct of our work, we have also derived very accurate Gaussian approximations for several functions of Gaussian random variables using the moment matching technique, which is helpful for extending the SSTA framework to include the impact of variability of these reliability effects, as well as their interactions with process variations.

The primary contribution of this thesis has been to raise the level of abstraction for reliability effects from the device level to the level of logic blocks. The presentation in this thesis has been focused on the analysis of combinational logic blocks and does not explicitly consider sequential blocks; however, it is well known that for analyzing timing, the analysis of combinational blocks is a core problem that must be solved, and sequential circuits can then be addressed by considering one block at a time [96]. It is important for such an approach must be supplemented by an analysis of degradation in the clock network and in all synchronizers (flip-flops and latches). While the work in this thesis does not explicitly address these issues, the solutions to these problems are a relatively straightforward extension from the approaches presented in this thesis and the known state of the art in circuit timing, and prior works that address this topic [97, 98] may be extended using approach described in this thesis.

The techniques and models in this thesis are likely to be appropriate for addressing aging problems in several other contexts in the future. For example:

- **Aging issues at other levels of design abstraction**: This thesis makes a first start by addressing such issues at the logic block level, and lays the framework for similar analyses at higher levels of abstraction, such as the core level or the architecture level. At these levels, simplified forms of these device models must be used. To a first order, it is still true that BTI depends on the signal probability, HC effects on the number of transitions, and TDDB on the stress time and area, but the proportionality parameters that characterize aging may change significantly when the mechanisms in this thesis are considered. Moreover, the sensitivity to process and environmental conditions may be very significant, but today's techniques typically do not address these issues at higher levels of design. Such models will affect higher-level design decisions such as architectures [99,100], DVFS schedules [34,35], sleep modes [32,33,37], and design margins [36,41], as well as issues in designing and deploying aging monitors [28, 29] that enable adaptive performance recovery under aging. These are all fertile areas for future work.

- **Aging issues in other types of design blocks** The thesis is focused on digital logic circuits, and does not explicitly address several other types of on-chip structures, such as memory elements (memory cells, sense amplifiers, etc.), interconnects and analog/mixed-signal circuitry [26, 101]. Again, the ideas in this thesis could be used to identify and address aging issues in these elements. Several open problems remain in these domains. For memory cells, for instance, individual transistors are close to minimum-sized and are likely to experience substantial variability under the charge trapping model. For analog blocks, aging must be captured as a function of the bias stress, which, by definition, takes on a continuous range of values rather than the discrete 0/1 values in digital circuits. However, as in the case of the state of the art in digital circuits prior to this thesis, prior analyses of both types of structures have largely relied on older and often obsolete models.

As the device community gains a further understanding of the mechanisms of aging, it is possible that newer models may be available to capture these effects. The fundamental approach in this thesis may be a suitable way to determine the impact of these newer models at higher levels of abstraction.

References

[1] T. Nigam, B. Parameshwaran, and G. Krause. Accurate product lifetime predictions based on device-level measurements. In *Proceedings of the IEEE International Reliability Physics Symposium*, pages 634–639, April 2009.

[2] R. Degraeve, B. Kaczer, A. De Keersgieter, and G. Groeseneken. Relation between breakdown mode and location in short-channel nMOSFETs and its impact on reliability specifications. *IEEE Transactions on Devices and Materials Reliability*, 1(3):163–169, September 2001.

[3] B. Kaczer, S. Mahato, V. V. de Almeida Camargo, M. Toledano-Luque, P. J. Roussel, T. Grasser, F. Catthoor, P. Dobrovolny, P. Zuber, G. Wirth, and G. Groeseneken. Atomistic approach to variability of bias-temperature instability in circuit simulations. In *Proceedings of the IEEE International Reliability Physics Symposium*, pages XT.3.1–XT.3.5, April 2011.

[4] J. H. Stathis. Physical and predictive models of ultrathin oxide reliability in CMOS devices and circuits. *IEEE Transactions on Devices and Materials Reliability*, 1(1):43–59, March 2001.

[5] E. Y. Wu, E. J. Nowak, A. Vayshenker, W. L. Lai, and D. L. Harmon. CMOS scaling beyond the 100-nm node with silicon-dioxide-based gate dielectrics. *IBM Journal of Research and Development*, 46(2/3):287–298, March/May 2002.

[6] K. Chopra, C. Zhuo, D. Blaauw, and D. Sylvester. A statistical approach for full-chip gate-oxide reliability analysis. In *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pages 698–705, November 2008.

[7] C. Zhuo, D. Blaauw, and D. Sylvester. Post-fabrication measurement-driven oxide breakdown reliability prediction and management. In *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pages 441–448, November 2009.

[8] E. Takeda, C. Y. Yang, and A. Miura-Hamada. *Hot-Carrier Effects in MOS Devices*. Academic Press, New York, NY, 1995.

[9] S. Tam, P.-K. Ko, and C. Hu. Lucky-electron model of channel electron injection in MOSFET's. *IEEE Transactions on Electron Devices*, D-31(9):1116–1125, September 1984.

[10] H. Kufluoglu. *MOSFET Degradation due to Negative Bias Temperature Instability (NBTI) and Hot Carrier Injection (HCI) and its Implications for Reliability-Aware VLSI Design*. PhD thesis, Purdue University, West Lafayette, IN, 2007.

[11] S. E. Rauch and G. La Rosa. The energy-driven paradigm of NMOSFET hot-carrier effects. *IEEE Transactions on Devices and Materials Reliability*, 5(4):701–705, December 2005.

[12] C. Guerin, V. Huard, and A. Bravaix. The energy-driven hot-carrier degradation modes of nMOSFETs. *IEEE Transactions on Devices and Materials Reliability*, 7(2):225–235, June 2007.

[13] A. Bravaix, C. Guerin, V. Huard, D. Roy, J. M. Roux, and E. Vincent. Hot-carrier acceleration factors for low power management in DC-AC stressed 40nm NMOS node at high temperature. In *Proceedings of the IEEE International Reliability Physics Symposium*, pages 531–548, April 2009.

[14] V. Huard, C. R. Parthasarathy, A. Bravaix, C. Guerin, and E. Pion. CMOS device design-in reliability approach in advanced nodes. In *Proceedings of the IEEE International Reliability Physics Symposium*, pages 624–633, April 2009.

[15] Z. Liu, B. W. McGaughy, and J. Z. Ma. Design tools for reliability analysis. In *Proceedings of the ACM/IEEE Design Automation Conference*, pages 182–187, 2006.

[16] D. Lorenz, G. Georgakos, and U. Schlichtmann. Aging analysis of circuit timing considering NBTI and HCI. In *Proceedings of the IEEE International On-Line Testing Symposium*, pages 3–8, June 2009.

[17] H. Chang and S. S. Sapatnekar. Statistical timing analysis considering spatial correlations using a single PERT-like traversal. In *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pages 621–625, November 2003.

[18] S. Han, J. Choung, B.-S. Kim, B. H. Lee, H. Choi, and J. Kim. Statistical aging analysis with process variation consideration. In *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pages 412–419, November 2011.

[19] S. E. Tyaginov, I. A. Starkov, O. Triebl, M. Karner, Ch. Kernstock, C. Jungemann, H. Enichlmair, J. M. Park, and T. Grasser. Impact of gate oxide thickness variations on hot-carrier degradation. In *Proceedings of the IEEE International Symposium on the Physical and Failure Analysis of Integrated Circuits*, pages 1–5, July 2012.

[20] K. Ramakrishnan, R. Rajaraman, S. Suresh, N. Vijaykrishnan, Y. Xie, and M.J. Irwin. Variation impact on SER of combinational circuits. In *Proceedings of the IEEE International Symposium on Quality Electronic Design*, pages 911–916, March 2007.

[21] M. A. Alam. A critical examination of the mechanics of dynamic NBTI for PMOSFETs. In *Proceedings of the IEEE International Electronic Devices Meeting*, pages 14.4.1–14.4.4, December 2003.

[22] S. Bhardwaj, W. Wang, R. Vattikonda, Y. Cao, and S. Vrudhula. Predictive modeling of the NBTI effect for reliable design. In *Proceedings of the IEEE Custom Integrated Circuits Conference*, pages 189–192, September 2006.

[23] S. V. Kumar, C. H. Kim, and S. S. Sapatnekar. A finite-oxide thickness-based analytical model for negative bias temperature instability. *IEEE Transactions on Devices and Materials Reliability*, 9(4):537–556, December 2009.

[24] S. Mahapatra, A. E. Islam, S. Deora, V. D. Maheta, K. Joshi, A. Jain, and M. A. Alam. A critical re-evaluation of the usefulness of R-D framework in predicting NBTI stress and recovery. In *Proceedings of the IEEE International Reliability Physics Symposium*, pages 6A.3.1–6A.3.10, April 2011.

[25] R. Zheng, J. Velamala, V. Reddy, V. Balakrishnan, E. Mintarno, S. Mitra, S. Krishnan, and Y. Cao. Circuit aging prediction for low-power operation. In *Proceedings of the IEEE Custom Integrated Circuits Conference*, pages 427–430, September 2009.

[26] S. Park, K. Kang, and K. Roy. Reliability implications of NBTI in digital integrated circuits. *IEEE Design Test of Computers*, PP(99):1, 2009.

[27] K.-C. Wu and D. Marculescu. Aging-aware timing analysis and optimization considering path sensitization. In *Proceedings of the Design, Automation & Test in Europe*, pages 1–6, March 2011.

[28] J. Keane, D. Persaud, and C. H. Kim. An all-in-one silicon odometer for separately monitoring HCI, BTI, and TDDB. In *Proceedings of the IEEE International Symposium on VLSI Circuits*, pages 108–109, June 2009.

[29] P. Singh, E. Karl, D. Sylvester, and D. Blaauw. Dynamic NBTI management using a 45nm multi-degradation sensor. In *Proceedings of the IEEE Custom Integrated Circuits Conference*, pages 1–4, September 2010.

[30] Y. Wang, X. Chen, W. Wang, Y. Cao, Y. Xie, and H. Yang. Gate replacement techniques for simultaneous leakage and aging optimization. In *Proceedings of the Design, Automation & Test in Europe*, pages 328–333, April 2009.

[31] K.-C. Wu and D. Marculescu. Joint logic restructuring and pin reordering against NBTI-induced performance degradation. In *Proceedings of the Design, Automation & Test in Europe*, pages 75–80, April 2009.

[32] A. Calimera, E. Macii, and M. Poncino. NBTI-aware clustered power gating. *ACM Transactions on Design Automation of Electronic Systems*, 16(1):3:1–3:25, November 2010.

[33] K.-C. Wu, D. Marculescu, M.-C. Lee, and S.-C. Chang. Analysis and mitigation of NBTI-induced performance degradation for power-gated circuits. In *Proceedings of the ACM International Symposium on Low Power Electronics and Design*, pages 139–144, August 2011.

[34] E. Mintarno, J. Skaf, R. Zheng, J. B. Velamala, Y. Cao, S. Boyd, R. W. Dutton, and S. Mitra. Self-tuning for maximized lifetime energy-efficiency in the presence of circuit aging. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 30(5):760–773, May 2011.

[35] S. Gupta and S. S. Sapatnekar. GNOMO: Greater-than-NOMinal Vdd operation for BTI mitigation. In *Proceedings of the Asia-South Pacific Design Automation Conference*, pages 271–276, January 2012.

[36] S. Gupta and S. S. Sapatnekar. BTI-aware design using variable latency units. In *Proceedings of the Asia-South Pacific Design Automation Conference*, pages 775–780, January 2012.

[37] A. Calimera, E. Macii, and M. Poncino. Design techniques for NBTI-tolerant power-gating architectures. *IEEE Transactions on Circuits and Systems II*, 59(4):249–253, April 2012.

[38] T. Grasser, B. Kaczer, W. Goes, H. Reisinger, T. Aichinger, P. Hehenberger, P.-J. Wagner, F. Schanovsky, J. Franco, P. Roussel, and M. Nelhiebel. Recent advances in understanding the bias temperature instability. In *Proceedings of the IEEE International Electronic Devices Meeting*, pages 4.4.1–4.4.4, December 2010.

[39] R. da Silva and G. I. Wirth. Logarithmic behavior of the degradation dynamics of metal-oxide-semiconductor devices. *Journal of Statistical Mechanics: Theory and Experiment*, 2010(04):P04025, April 2010.

[40] G. I. Wirth, R. da Silva, and B. Kaczer. Statistical model for MOSFET bias temperature instability component due to charge trapping. *IEEE Transactions on Electron Devices*, 58(8):2743–2751, August 2011.

[41] J. B. Velamala, K. Sutaria, T. Sato, and Y. Cao. Physics matters: Statistical aging prediction under trapping/detrapping. In *Proceedings of the ACM/IEEE Design Automation Conference*, pages 139–144, June 2012.

[42] S. E. Rauch. The statistics of NBTI-induced V_T and β mismatch shifts in pMOSFETs. *IEEE Transactions on Devices and Materials Reliability*, 2(4):89–93, December 2002.

[43] H. Reisinger, T. Grasser, W. Gustin, and C. Schlünder. The statistical analysis of individual defects constituting NBTI and its implications for modeling DC- and AC-stress. In *Proceedings of the IEEE International Reliability Physics Symposium*, pages 7–15, May 2010.

[44] B. Kaczer, T. Grasser, P. J. Roussel, J. Franco, R. Degraeve, L. Ragnarsson, E. Simoen, G. Groeseneken, and H. Reisinger. Origin of NBTI variability in deeply scaled pFETs. In *Proceedings of the IEEE International Reliability Physics Symposium*, pages 26–32, May 2010.

[45] M. Toledano-Luque, B. Kaczer, P. J. Roussel, T. Grasser, G. I. Wirth, J. Franco, C. Vrancken, N. Horiguchi, and G. Groeseneken. Response of a single trap to AC negative bias temperature stress. In *Proceedings of the IEEE International Reliability Physics Symposium*, pages 4A.2.1–4A.2.8, April 2011.

[46] R. Degraeve, G. Groeseneken, R. Bellens, J. L. Ogier, M. Depas, P. J. Roussel, and H. E. Maes. New insights in the relation between electron trap generation and the statistical properties of oxide breakdown. *IEEE Transactions on Electron Devices*, 45(4):904–911, April 1998.

[47] J. H. Stathis. Percolation models for gate oxide breakdown. *Journal of Applied Physics*, 86(10):5757–5766, November 1999.

[48] F. Crupi, B. Kaczer, R. Degraeve, A. De Keersgieter, and G. Groeseneken. A comparative study of the oxide breakdown in short-channel nMOSFETs and pMOSFETs stressed in inversion and in accumulation regimes. *IEEE Transactions on Devices and Materials Reliability*, 3(1):8–13, March 2003.

[49] S. Cheffah, V. Huard, R. Chevallier, and A. Bravaix. Soft oxide breakdown impact on the functionality of a 40 nm SRAM memory. In *Proceedings of the IEEE International Reliability Physics Symposium*, pages 704–705, April 2011.

[50] R. Fernández, J. Martin-Martinez, R. Rodriguez, M. Nafria, and X. H. Aymerich. Gate oxide wear-out and breakdown effects on the performance of analog and digital circuits. *IEEE Transactions on Electron Devices*, 55(4):997–1004, April 2008.

[51] J. Suñé, G. Mura, and E. Miranda. Are soft breakdown and hard breakdown of ultrathin gate oxides actually different failure mechanisms? *IEEE Electron Device Letters*, 21(4):167–169, April 2000.

[52] S. Tsujikawa, M. Kanno, and N. Nagashima. Reliable assessment of progressive breakdown in ultrathin mos gate oxides toward accurate TDDB evaluation. *IEEE Transactions on Electron Devices*, 58(5):1468–1475, May 2011.

[53] Y. H. Lee, N. Mielke, M. Agostinelli, S. Gupta, R. Lu, and W. McMahon. Prediction of logic product failure due to thin-gate oxide breakdown. In *Proceedings of the IEEE International Reliability Physics Symposium*, pages 18–28, March 2006.

[54] B. Kaczer, R. Degraeve, M. Rasras, K. Van de Mieroop, P. J. Roussel, and G. Groeseneken. Impact of MOSFET gate oxide breakdown on digital circuit operation and reliability. *IEEE Transactions on Electron Devices*, 49(3):500–506, March 2002.

[55] F. N. Najm. A survey of power estimation techniques in VLSI circuits. *IEEE Transactions on VLSI Systems*, 2(4):446–455, December 1994.

[56] R. Burch, F. N. Najm, P. Yang, and T. N. Trick. A Monte Carlo approach for power estimation. *IEEE Transactions on VLSI Systems*, 1(1):63–71, March 1993.

[57] R. Rodríguez, J. H. Stathis, and B. P. Linder. A model for gate-oxide breakdown in CMOS inverters. *IEEE Electron Device Letters*, 24(2):114–116, February 2003.

[58] Hua Wang, M. Miranda, F. Catthoor, and Dehaene Wim. Impact of random soft oxide breakdown on SRAM energy/delay drift. *IEEE Transactions on Devices and Materials Reliability*, 7(4):581–591, December 2007.

[59] B. Kaczer, R. Degraeve, A. De Keersgieter, K. Van de Mieroop, V. Simons, and G. Groeseneken. Consistent model for short-channel nMOSFET after hard gate oxide breakdown. *IEEE Transactions on Electron Devices*, 49(3):507–513, March 2002.

[60] J. Segura, C. Benito, A. Rubio, and C. F. Hawkins. A detailed analysis and electrical modeling of gate oxide shorts in MOS transistors. *Journal of Electronic Testing*, 8(3):229–239, June 1996.

[61] X. Lu, Z. Li, W. Qiu, D. M. H. Walker, and W. Shi. A circuit level fault model for resistive shorts of MOS gate oxide. In *Proceedings of the IEEE International Workshop on Microprocessor Test and Verification*, pages 97–102, September 2004.

[62] K. Shubhakar, K. L. Pey, S. S. Kushvaha, M. Bosman, S. J. O'Shea, N. Raghavan, M. Kouda, K. Kakushima, Z. R. Wang, H. Y. Yu, and H. Iwai. Nanoscale electrical and physical study of polycrystalline high-k dielectrics and proposed reliability enhancement techniques. In *Proceedings of the IEEE International Reliability Physics Symposium*, pages 786–791, April 2011.

[63] Predictive Technology Model. Available: `http://www.eas.asu.edu/~ptm/`.

[64] J. Xiong, V. Zolotov, and L. He. Robust extraction of spatial correlation. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 26(4):619–631, April 2007.

[65] R. Degraeve, G. Groeseneken, R. Bellens, M. Depas, and H. E. Maes. A consistent model for the thickness dependence of intrinsic breakdown in ultra-thin oxides. In *Proceedings of the IEEE International Electronic Devices Meeting*, pages 863–866, December 1995.

[66] C. Clark. The greatest of a finite set of random variables. *Operations Research*, 9:85–91, 1961.

[67] A. A. Abu-Dayya and N. C. Beaulieu. Comparison of methods of computing correlated lognormal sum distributions and outages for digital wireless applications. In *Proc. VTC*, volume 1, pages 175–179, June 1994.

[68] A. Srivastava, S. Shah, K. Agarwal, D. Sylvester, D. Blaauw, and S. Director. Accurate and efficient gate-level parametric yield estimation considering correlated variations in leakage power and performance. In *Proceedings of the ACM/IEEE Design Automation Conference*, pages 535–540, 2005.

[69] Berkeley Logic Synthesis and Verification Group. Abc: A system for sequential synthesis and verification, release 70930. Available: `http://www.eecs.berkeley.edu/~alanmi/abc/`.

[70] Nangate 45nm Open Cell Library. Available: `http://www.nangate.com/`.

[71] International Technology Roadmap for Semiconductors, 2008 update. Chapter of process integration, devices and structures. Available: `http://www.itrs.net/`.

[72] J. P. Fishburn and A. E. Dunlop. TILOS: A posynomial programming approach to transistor sizing. In *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pages 326–328, November 1985.

[73] S. S. Sapatnekar, V. B. Rao, P. M. Vaidya, and S. M. Kang. An exact solution to the transistor sizing problem for CMOS circuits using convex optimization. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 12(11):1621–1634, November 1993.

[74] J. G. Ecker. Geometric programming: Methods, computations and applications. *SIAM Review*, 22(3):338–362, July 1980.

[75] The MOSEK Optimization Software. `http://www.mosek.com/`.

[76] K. Hess, L. F. Register, W. McMahon, B. Tuttle, O. Aktas, U. Ravaioli, J. W. Lyding, and I. C. Kizilyalli. Theory of channel hot-carrier degradation in MOSFETs. *Physica B*, 272:527–531, 1999.

[77] S. E. Rauch, F. Guarin, and G. La Rosa. High-V_{gs} PFET DC hot-carrier mechanism and its relation to AC degradation. *IEEE Transactions on Devices and Materials Reliability*, 10(1):40–46, March 2010.

[78] B. C. Paul, K. Kang, H. Kufluoglu, M. A. Alam, and K. Roy. Impact of NBTI on the temporal performance degradation of digital circuits. *IEEE Electron Device Letters*, 26(8):560–562, August 2005.

[79] S. V. Kumar, C. H. Kim, and S. S. Sapatnekar. An analytical model for negative bias temperature instability. In *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pages 493–496, November 2006.

[80] S. Nassif. Delay variability: Sources, impact and trends. In *Proc. ISSCC*, pages 368–369, February 2000.

[81] L. Zhang, W. Chen, Y. Hu, and C. C. Chen. Statistical timing analysis with extended pseudo-canonical timing model. In *Proceedings of the Design, Automation & Test in Europe*, pages 952–957, March 2005.

[82] Y. Taur and T. H. Ning. *Fundamentals of modern VLSI devices.* Cambridge University Press, New York, NY, 1998.

[83] A. Agarwal, D. Blaauw, and V. Zolotov. Statistical timing analysis for intra-die process variations with spatial correlations. In *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pages 900–907, November 2003.

[84] C. Visweswariah, K. Ravindran, K. Kalafala, S. G. Walker, S. Narayan, D. K. Beece, J. Piaget, N. Venkateswaran, and J. G. Hemmett. First-order incremental block-based statistical timing analysis. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 25(10):2170–2180, October 2006.

[85] W. Wang, V. Balakrishnan, B. Yang, and Y. Cao. Statistical prediction of NBTI-induced circuit aging. In *Proceedings of the International Conference on Solid State and Integrated Circuits Technology*, pages 416–419, October 2008.

[86] Y. Lu, L. Shang, H. Zhou, H. Zhu, F. Yang, and X. Zeng. Statistical reliability analysis under process variation and aging effects. In *Proceedings of the ACM/IEEE Design Automation Conference*, pages 514–519, July 2009.

[87] S. Han and J. Kim. NBTI-aware statistical timing analysis framework. In *Proceedings of the IEEE International SoC Conference*, pages 158–163, September 2010.

[88] K. Kang, S. P. Park, K. Roy, and M. A. Alam. Estimation of statistical variation in temporal NBTI degradation and its impact on lifetime circuit performance. In *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pages 730–734, November 2007.

[89] B. Vaidyanathan, A.S. Oates, and Y. Xie. Intrinsic NBTI-variability aware statistical pipeline performance assessment and tuning. In *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pages 164–171, November 2009.

[90] M. J. M. Pelgrom, A. C. J. Duinmaijer, and A. P. G. Welbers. Matching properties of MOS transistors. *IEEE Journal of Solid-State Circuits*, 24(5):1433–1439, October 1989.

[91] A. J. Bhavnagarwala, Xinghai Tang, and J. D. Meindl. The impact of intrinsic device fluctuations on CMOS SRAM cell stability. *IEEE Journal of Solid-State Circuits*, 36(4):658–665, April 2001.

[92] W. Wang, V. Reddy, A. T. Krishnan, R. Vattikonda, S. Krishnan, and Y. Cao. Compact modeling and simulation of circuit reliability for 65-nm CMOS technology. *IEEE Transactions on Devices and Materials Reliability*, 7(4):509–517, December 2007.

[93] J. H. Pollard. *A Handbook of Numerical and Statistical Techniques: With Examples Mainly from the Life Sciences.* Cambridge University Press, New York, NY, 1977.

[94] T. Grasser, B. Kaczer, W. Goes, T. Aichinger, P. Hehenberger, and M. Nelhiebel. A two-stage model for negative bias temperature instability. In *Proceedings of the IEEE International Reliability Physics Symposium*, pages 33–44, April 2009.

[95] G. Math, C. Benard, J. L. Ogier, and D. Goguenheim. Geometry effects on the NBTI degradation of PMOS transistors. In *IEEE Integrated Reliability Workshop (IRW) Final Report*, pages 1–15, October 2008.

[96] S. Sapatnekar. *Timing*. Kluwer Academic Publishers, Boston, MA, 2004.

[97] A. Chakraborty, G. Ganesan, A. Rajaram, and D. Z. Pan. Analysis and optimization of NBTI induced clock skew in gated clock trees. In *Proceedings of the Design, Automation & Test in Europe*, pages 296–299, April 2009.

[98] W. Liu, S. Miryala, V. Tenace, A. Calimera, E. Macii, and M. Poncino. NBTI effects on tree-like clock distribution networks. In *Proceedings of the Great Lakes Symposium on VLSI*, pages 279–282, May 2012.

[99] U. R. Karpuzcu, B. Greskamp, and J. Torrellas. The BubbleWrap many-core: Popping cores for sequential acceleration. In *Proceedings of the IEEE/ACM International Symposium on Microarchitecture*, pages 447–458, December 2009.

[100] M. Loghi, H. Mahmood, A. Calimera, M. Poncino, and E. Macii. Energy-optimal caches with guaranteed lifetime. In *Proceedings of the ACM International Symposium on Low Power Electronics and Design*, pages 141–146, July 2012.

[101] X. Wang, P. Jain, D. Jiao, and C. H. Kim. Impact of interconnect length on BTI and HCI induced frequency degradation. In *Proceedings of the IEEE International Reliability Physics Symposium*, pages 2F.5.1–2F.5.6, April 2012.

Appendix A

Proof of Theorem 1

Since failures of different logic cells are independent, the circuit-level FP at time t, $\mathrm{Pr}_{\mathrm{fail}}^{(\mathrm{ckt})}(t)$, can be calculated as:

$$\begin{aligned}
\mathrm{Pr}_{\mathrm{fail}}^{(\mathrm{ckt})}(t) &= 1 - \prod_{i \in \mathrm{NMOS}} \left(1 - \mathrm{Pr}_{\mathrm{fail}}^{(i)}(t)\right) \\
&= 1 - \prod_{i \in \mathrm{NMOS}} \left(1 - \mathrm{Pr}_{(\mathrm{fail}|\mathrm{BD})}^{(i)} \mathrm{Pr}_{\mathrm{BD}}^{(i)}(t)\right)
\end{aligned}$$

Here, $\mathrm{Pr}_{\mathrm{fail}}^{(i)}(t)$ represents the probability that NMOS transistor i in the circuit fails at time t, which implies two facts: first, transistor i breaks down at t, an event that has probability $\mathrm{Pr}_{\mathrm{BD}}^{(i)}(t)$, and second, the breakdown causes a logic failure, which is captured with the cell-level FP $\mathrm{Pr}_{(\mathrm{fail}|\mathrm{BD})}^{(i)}$ from Section 2.3.2. Substituting (2.3) above:

$$\mathrm{Pr}_{\mathrm{fail}}^{(\mathrm{ckt})}(t) = 1 - \prod_{i \in \mathrm{NMOS}} \left(1 - \mathrm{Pr}_{(\mathrm{fail}|\mathrm{BD})}^{(i)} \left(1 - \exp\left(-\left(\frac{\gamma_i t}{\alpha}\right)^{\beta} a_i\right)\right)\right). \qquad \text{(A.1)}$$

This equation gives the circuit FP, incorporating considerations related to the effective stress time and to whether a breakdown event in a transistor causes a cell-level failure. It can further be simplified. For simplicity, we will use the following abbreviated notations: denote $\mathrm{Pr}_{\mathrm{fail}}^{(\mathrm{ckt})}(t)$ by P_f, $\mathrm{Pr}_{(\mathrm{fail}|\mathrm{BD})}^{(i)}$ by p_i, and $(\frac{\gamma_i t}{\alpha})^{\beta} a_i$ by μ_i. Then, taking the logarithm of (A.1):

$$\ln(1 - P_f) = \sum_{i \in \mathrm{NMOS}} \ln\left(1 - p_i\left(1 - \exp\left(-\mu_i\right)\right)\right). \qquad \text{(A.2)}$$

Using first-order Taylor expansions, first using $\exp(-x) = 1 - x$ for $x = \mu_i$, and then $\ln(1 - x) = -x$ for $x = p_i\mu_i$, the approximation is further simplified as

$$\ln(1 - P_f) \approx \sum_{i \in \text{NMOS}} \ln(1 - p_i\mu_i) \approx -\sum_{i \in \text{NMOS}} p_i\mu_i. \tag{A.3}$$

In other words, resubstituting the full forms of P_f, p_i, and μ_i, we get the simplified closed-form formula of the FP as:

$$\text{Pr}_{\text{fail}}^{(\text{ckt})}(t) = 1 - \exp\left(-\left(\frac{t}{\alpha}\right)^{\beta} \sum_{i \in \text{NMOS}} \text{Pr}_{(\text{fail}|\text{BD})}^{(i)} \gamma_i^{\beta} a_i\right). \tag{A.4}$$

For this problem, $0 \leq p_i \leq 1$ and $0 < \mu_i \ll 1^1$. Thus the conditions $|x| \leq 1, x \neq 1$ for the Taylor expansion of $\ln(1 - x)$ are satisfied, and the approximations with first-order Taylor expansions are quite accurate since the high order terms $\text{O}(x^2)$ are much smaller.

We can convert (A.4) to the following form:

$$\begin{aligned} W &= \ln\left(-\ln\left(1 - \text{Pr}_{\text{fail}}^{(\text{ckt})}(t)\right)\right) & \text{(A.5)} \\ &= \beta \ln\left(\frac{t}{\alpha}\right) + \ln \sum_{i \in \text{NMOS}} \text{Pr}_{(\text{fail}|\text{BD})}^{(i)} \gamma_i^{\beta} a_i. & \text{(A.6)} \end{aligned}$$

[1] The region of interest for circuit failure is usually at the lower end, e.g. $P_f < 0.1$. Due to the weakest-link property, the breakdown probability of each individual cell p_i in a large circuit must be very small, which implies that μ_i is very small and must be far less than 1 (considering $\mu_i = 1$ implies that $p_i = 0.632$ for a unit-size device). These approximations are validated by experimental results in Section 2.6.

Appendix B

Cell-Level Characterization under Variations

Under process variations, the I-V characteristics of driver cell and load cell can be expressed using first-order Taylor expansion as

$$\begin{aligned} I_{\text{dr}}(V_{\text{dr}}) &= I_{\text{dr}}^0 + \frac{\partial I_{\text{dr}}}{\partial V_{\text{dr}}}\Delta V_{\text{dr}} + \sum_{i\in\text{driver}} \frac{\partial I_{\text{dr}}}{\partial q_i}\Delta q_i \qquad \text{(B.1)} \\ I_{\text{in}}(V_{\text{in}}, x_{\text{BD}}) &= I_{\text{in}}^0 + \frac{\partial I_{\text{in}}}{\partial V_{\text{in}}}\Delta V_{\text{in}} + \sum_{j\in\text{load}} \frac{\partial I_{\text{in}}}{\partial q_j}\Delta q_j \\ &\quad + \frac{\partial I_{\text{in}}}{\partial \epsilon_r}\epsilon_r + \frac{\partial I_{\text{in}}}{\partial x_{\text{BD}}}\Delta x_{\text{BD}} \qquad \text{(B.2)} \\ V_{\text{out}}(V_{\text{in}}, x_{\text{BD}}) &= V_{\text{out}}^0 + \frac{\partial V_{\text{out}}}{\partial V_{\text{in}}}\Delta V_{\text{in}} + \sum_{j\in\text{load}} \frac{\partial V_{\text{out}}}{\partial q_j}\Delta q_j \\ &\quad + \frac{\partial V_{\text{out}}}{\partial \epsilon_r}\epsilon_r + \frac{\partial V_{\text{out}}}{\partial x_{\text{BD}}}\Delta x_{\text{BD}} \qquad \text{(B.3)} \end{aligned}$$

Here X^0 denotes the nominal value of parameter X when not considering variations, and q_i, q_j stand for the process parameters (i.e. W, L, and T_{ox}) of the transistors in the driver cell and load cell, respectively. All the first-order derivatives $\partial x/\partial y$ can be calculated in the precharacterization procedure in Algorithm 1 and stored in LUTs.

From (2.11), we know that $I_{\text{dr}}(V_{\text{dr}}) = I_{\text{in}}(V_{\text{in}}, x_{\text{BD}})$, $I_{\text{dr}}^0 = I_{\text{in}}^0$, $V_{\text{dr}} = V_{\text{in}}$, and

$\Delta V_{\mathrm{dr}} = \Delta V_{\mathrm{in}}$, thus from (B.1) and (B.2) we get

$$\left(\frac{\partial I_{\mathrm{in}}}{\partial V_{\mathrm{in}}} - \frac{\partial I_{\mathrm{dr}}}{\partial V_{\mathrm{dr}}}\right)\Delta V_{\mathrm{dr}} + \frac{\partial I_{\mathrm{in}}}{\partial x_{\mathrm{BD}}}\Delta x_{\mathrm{BD}} + \frac{\partial I_{\mathrm{in}}}{\partial \epsilon_r}\epsilon_r + \sum_{j\in\mathrm{load}} \frac{\partial I_{\mathrm{in}}}{\partial q_j}\Delta q_j - \sum_{i\in\mathrm{driver}} \frac{\partial I_{\mathrm{dr}}}{\partial q_i}\Delta q_i = 0 \tag{B.4}$$

To calculate the impact of variations on driver failure, $x^{\mathrm{dr}}_{\text{fail-s}}$ and $x^{\mathrm{dr}}_{\text{fail-d}}$, we have $V_{\mathrm{dr}} = V^{\mathrm{dr}}_{\mathrm{TH}}$, hence $\Delta V_{\mathrm{dr}} = 0$, therefore Δx_{BD} can be solved from (B.4) as

$$\Delta x_{\mathrm{BD}} = \left(\sum_{i\in\mathrm{driver}} \frac{\partial I_{\mathrm{dr}}}{\partial q_i}\Delta q_i - \sum_{j\in\mathrm{load}} \frac{\partial I_{\mathrm{in}}}{\partial q_j}\Delta q_j - \frac{\partial I_{\mathrm{in}}}{\partial \epsilon_r}\epsilon_r\right) / \frac{\partial I_{\mathrm{in}}}{\partial x_{\mathrm{BD}}}$$

To calculate the impact of variations on load failure, $x^{\mathrm{ld}}_{\text{fail-s}}$ and $x^{\mathrm{ld}}_{\text{fail-d}}$, we have $V_{\mathrm{out}} = V^{\mathrm{out}}_{\mathrm{TH}} = V^{0}_{\mathrm{out}}$, therefore (B.3) can be rewritten as

$$\frac{\partial V_{\mathrm{out}}}{\partial V_{\mathrm{in}}}\Delta V_{\mathrm{dr}} + \frac{\partial V_{\mathrm{out}}}{\partial x_{\mathrm{BD}}}\Delta x_{\mathrm{BD}} + \sum_{j\in\mathrm{load}} \frac{\partial V_{\mathrm{out}}}{\partial q_j}\Delta q_j + \frac{\partial V_{\mathrm{out}}}{\partial \epsilon_r}\epsilon_r = 0 \tag{B.5}$$

Then using (B.4) and (B.5) the unknowns ΔV_{dr} and Δx_{BD} can be solved. The FP components in (2.23) are obtained using solved Δx_{BD}'s and (2.7). This variation-aware cell-level analysis approach can be fully integrated to Algorithm 2.

Appendix C

Logarithm of a Gaussian RV

For $x \sim N(\mu_x, \sigma_x^2)$, given $\mu_x \gg \sigma_x > 0$ so that $x > 0$ is always true, its logarithm $y = \ln x$ can be approximated linearly as $y = c + kx$. In order to get better accuracy, the following moment-matching method is used.

For $y = \ln x$, we want to approximate it as $y' {\sim} N(\mu_y, \sigma_y^2)$. Therefore $x' = \exp(y')$ has a lognormal distribution with first two moments

$$\begin{aligned} u_1 &= \exp(\mu_y + \sigma_y^2/2) \\ u_2 &= \exp(2\mu_y + 2\sigma_y^2) \end{aligned} \tag{C.1}$$

By matching the first two moments of x' and x: $u_1 = \mu_x$, $u_2 = \sigma_x^2 + \mu_x^2$, we can get the distribution of y as

$$\begin{aligned} \mu_y &= 2\ln\mu_x - \frac{1}{2}\ln(\sigma_x^2 + \mu_x^2) \\ \sigma_y^2 &= \ln(\sigma_x^2 + \mu_x^2) - 2\ln\mu_x \end{aligned} \tag{C.2}$$

Therefore the coefficients for the linear form $y = c + kx$ are $k = \sigma_y/\sigma_x$ and $c = \mu_y - \mu_x\sigma_y/\sigma_x$.

Appendix D

Square root of a Gaussian RV

For $x \sim N(\mu_x, \sigma_x^2)$, given $\mu_x \gg \sigma_x > 0$ so that $x > 0$ is always true, its square root $y = x^n,\ n = \frac{1}{2}$ can be approximated as another Gaussian distribution $y \sim N(\mu_y, \sigma_y^2)$, with its mean μ_y and variance σ_y^2 calculated as follows.

Since $x = y^2$, we have

$$\begin{aligned} E(x) &= E(y^2) \qquad (D.1) \\ E(x^2) &= E(y^4) \qquad (D.2) \end{aligned}$$

As both x and y have Gaussian distribution,

$$\begin{aligned} \mu_x &= \mu_y^2 + \sigma_y^2 \qquad (D.3) \\ \mu_x^2 + \sigma_x^2 &= \mu_y^4 + 6\mu_y^2\sigma_y^2 + 3\sigma_y^4 \qquad (D.4) \end{aligned}$$

The mean and variance of y is solved to be

$$\begin{aligned} \mu_y &= \left(\mu_x^2 - \frac{1}{2}\sigma_x^2\right)^{\frac{1}{4}} \qquad (D.5) \\ \sigma_y^2 &= \mu_x - \left(\mu_x^2 - \frac{1}{2}\sigma_x^2\right)^{\frac{1}{2}} \qquad (D.6) \end{aligned}$$

Appendix E

Power Function of a Gaussian RV

In the case of n is a real number with arbitrary value[1] , the relation $y = x^n$ can be rewritten as

$$y = \exp\left(n \ln\left(x\right)\right) \tag{E.1}$$

In Appendix C, $q = n \ln(x)$ is approximated as Gaussian with

$$\begin{aligned} \mu_q &= n \left(2 \ln \mu_x - \frac{1}{2} \ln\left(\sigma_x^2 + \mu_x^2\right)\right) && \text{(E.2)} \\ \sigma_q^2 &= n^2 \left(\ln\left(\sigma_x^2 + \mu_x^2\right) - 2 \ln \mu_x\right) && \text{(E.3)} \end{aligned}$$

Therefore $y = \exp(q)$ has a lognormal distribution with

$$\begin{aligned} \mu_y &= \exp\left(\mu_q + \frac{1}{2}\sigma_q^2\right) && \text{(E.4)} \\ \sigma_y^2 &= \left(\exp\left(\sigma_q^2\right) - 1\right) \exp\left(2\mu_q + \sigma_q^2\right) && \text{(E.5)} \end{aligned}$$

This lognormal distribution is approximated as Gaussian, with the same mean and variance.

[1] The HC time exponent is generally modeled to be 1/2, however the exact value may vary for a specific technology.

Appendix F

Linear Approximation of x^n

For $y = x^n$ in which x is a Gaussian RV with a RV space expression, we want to approximate y as a Gaussian RV in the same space in order to simplify the circuit analysis. In this scenario, y is approximated as a linear relation with x,

$$y \doteq k \cdot x + c \tag{F.1}$$

Since x and y are both Gaussian with known or derived mean and variance, the value of k and c are

$$k = \frac{\sigma_y}{\sigma_x} \tag{F.2}$$

$$c = \mu_y - \frac{\mu_x \sigma_y}{\sigma_x} \tag{F.3}$$

Experiments have shown that for $\sigma_x \leq 0.1\mu_x$, Gaussian distribution is a good approximation for $y = x^n$, and the proposed methods calculate its mean and variance with very good accuracy. For $n = 0.3$–0.7, the error of μ_y and σ_y is less than 1% for $\sigma_x = 0.1\mu_x$.

Appendix G

Product and Division of Gaussian RVs

For $y = x_1 x_2$ (or x_1/x_2), where $x_1 \sim N(\mu_{x_1}, \sigma_{x_1}^2)$ and $x_2 \sim N(\mu_{x_2}, \sigma_{x_2}^2)$ are Gaussian random variables with correlation coefficient ρ_{x_1,x_2}, it can be approximated as Gaussian in following steps. First, y is rewritten as

$$\begin{aligned} y &= \exp(q), & \text{(G.1)} \\ \text{where } q &= \ln x_1 \pm \ln x_2 & \text{(G.2)} \end{aligned}$$

According to Appendix C, $\ln x$ is approximated as a linear function of x, we can get the approximation

$$q \doteq k_1 x_1 \pm k_2 x_2 + c_1 \pm c_2, \qquad \text{(G.3)}$$

in which k_1, k_2, c_1, and c_2 is obtained as

$$\begin{aligned} k_i &= \sqrt{\ln\left(\sigma_{x_i}^2 + \mu_{x_i}^2\right) - 2\ln \mu_{x_i}} / \sigma_{x_i} & \text{(G.4)} \\ c_i &= 2\ln \mu_{x_i} - \frac{1}{2}\ln\left(\sigma_{x_i}^2 + \mu_{x_i}^2\right) - \mu_{x_i} k_i. & \text{(G.5)} \end{aligned}$$

So q is Gaussian with mean and variance

$$\begin{aligned} \mu_q &= k_1 \mu_{x_1} \pm k_2 \mu_{x_2} + c_1 \pm c_2 & \text{(G.6)} \\ \sigma_q^2 &= k_1^2 \sigma_{x_1}^2 + k_2^2 \sigma_{x_2}^2 + 2 k_1 k_2 \sigma_{x_1} \sigma_{x_2} \rho_{x_1,x_2} & \text{(G.7)} \end{aligned}$$

Alternatively, the mean and variance of q can be obtained directly by using the random space interpretation of q. The rest is similar to previous section where a lognormal RV is approximated as Gaussian. Since $y = \exp(q)$ is lognormal with mean μ_y and variance σ_y^2 following (E.4,E.5), it is approximated as Gaussian using linear function

$$\begin{aligned} y &\doteq kq + c \\ &= k(k_1 x_1 \pm k_2 x_2 + c_1 \pm c_2) + c, && \text{(G.8)} \\ \text{where } k &= \frac{\sigma_y}{\sigma_q}, c = \mu_y - \frac{\mu_q \sigma_y}{\sigma_q} && \text{(G.9)} \end{aligned}$$

Experiments have verified that for $\sigma_x \leq 0.1\mu_x$, Gaussian distribution is a good approximation for $y = x_1 x_2$ (or x_1/x_2), and the proposed methods calculate its mean and variance with very good accuracy. This method can easily extend to the case of product of three or more Gaussian random variables.