Cross-Layer Resiliency Modeling and Optimization: A Device to Circuit Approach

Zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften

an der Fakultät für Informatik
des Karlsruher Instituts für Technologie (KIT)

genehmigte

Dissertation

von

Saman Kiamehr

Tag der mündlichen Prüfung:	11.05.2015
Referent:	Prof. Dr. Mehdi Baradaran Tahoori KIT, Karlsruhe, Germany
Korreferent:	Dr. Sani Nassif Radyalis, Austin, TX, USA

Saman Kiamehr
Dragonerstr. 9
76185 Karlsruhe

Hiermit erkläre ich an Eides statt, dass ich die von mir vorgelegte Arbeit selbstständig verfasst habe, dass ich die verwendeten Quellen, Internet-Quellen und Hilfsmittel vollständig angegeben haben und dass ich die Stellen der Arbeit - einschlielich Tabellen, Karten und Abbildungen - die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Karlsruhe, Mai 2015

_____

Saman Kiamehr

ACKNOWLEDGEMENT

I am using this opportunity to express my gratitude to everyone who supported me throughout my Ph.D. study.

Foremost, I would like to express my sincere gratitude to my advisor Prof. Mehdi Tahoori for his continuous support, patience, motivation and enthusiasm. His guidance and support helped me overcome many difficulties throughout my research and finish this dissertation. He has been a tremendous mentor for me and I could not have imagined having a better advisor and mentor for my Ph.D. study.

I am also grateful to my co-advisor Dr. Sani Nassif. He provided me with direction, technical support and became more of a friend, than a co-advisor. Dr. Nassif is the one who truly made a difference in my life.

A special thanks to my family. Words cannot express how grateful I am to my mother and father for all of their exceptional sacrifices through my entire life. I would like to express appreciation to my beloved wife Hajar who has been always my support.

I would also like to thank all of the colleagues from the Chair of Dependable Nano Computing (CDNC) for their discussions, feedback, invaluably constructive criticism and friendly advices. In particular, I want to thank Farshad Firouzi, Fabian Oboril and Mojtaba Ebrahimi.

CONTENTS

GLOSSARY

Symbols | B | C | D | E | F | H | L | M | N | P | R | S | T | V

Symbols

V_{DD} Supply voltage.

V_{DS} Drain-source voltage of transistor.

V_{GS} Gate-source voltage of transistor.

V_{SB} Source-body voltage of transistor.

τ_c capture time.

τ_c capture time.

B

BTI Bias Temperature Instability.

C

Capture time Time needed to charge a gate oxide defect during the stress phase.

CHC Channel Hot Carrier.

CMOS Complementary metal-oxide-semiconductor.

CNTFET Carbon Nanotube Field Effect Transistors.

D

Duty cycle Ratio of the time in which transistor is under stress to the total time.

E

EM Electromigration.

Emission time Time needed for the defect to re-emit its charge during the recovery phase.

F

FinFET Fin Filed Effect Transistor.

FIT Failure In Time.

H

HCI Hot Carrier Injection.

L

LER Line-edge Roughness.

M

MBU Multiple Bit Upset.

MOSFET Metal-oxide-semiconductor field effect transistor.

N

NBTI Negative Bias Temperature Instability.

NMOS n-type transistor.

P

PBTI Positive Bias Temperature Instability.

PMOS p-type transistor.

R

RD Reaction-Diffusion.

RDF Random Dopant Fluctuation.

S

SCE Short Channel Effect.

SER Soft Error Rate.

SEU Single-Event Upset.

SOI Silicon On Insulator.

SP Signal Probability.

T

TD Trapping-Detrapping.

TDDB Time Dependent Dielectric Breakdown.

V

VLSI Very Large Scale Integration.

LIST OF FIGURES

LIST OF TABLES

LIST OF OWN PUBLICATIONS INCLUDED IN THESIS

Journal Papers :

[1] S. Kiamehr, M. Ebrahimi, F. Firouzi, and M. Tahoori, "Extending Standard Cell Library for Aging Mitigation," in *IET Computers & Digital Techniques* , 2015.

Conference Papers :

[2] S. Kiamehr, F. Firouzi, and M. Tahoori, "Stacking-based Input Reordering for NBTI Aging Reduction," in *Proceedings of Zuverlässigkeit und Entwurf (ZuE)* , 2011, Germany.

[3] S. Kiamehr, F. Firouzi, and M. Tahoori, "Input and Transistor Reordering for NBTI and HCI Reduction in Complex CMOS Gates," in *Proceedings of Great Lake Symposium on VLSI (GLSVLSI)* , 2012, USA.

[4] S. Kiamehr, F. Firouzi, and M. Tahoori, "Aging-aware Timing Analysis Considering Combined Effects of NBTI and PBTI," in *Proceedings of International Symposium on Quality Electronic Desing (ISQED)* , 2013, USA.

[5] S. Kiamehr, M.Ebrahimi, F. Firouzi, and M. Tahoori, "Chip-level Modeling and Analysis of Electrical Masking of Soft Errors," in *Proceedings of VLSI Test Symposium (VTS)*, 2013, USA.

[6] S. Kiamehr, F. Firouzi, M. Ebrahimi, and M. Tahoori, "Aging-aware Standard Cell Library Design," in *Proceedings of Design, Automation & Test in Europe (DATE)* , 2014, Germany.

[7] S. Kiamehr and M. Tahoori, "A Cross-Layer Approach for Soft Error Analysis of SRAMs in SOI FinFET Technology," in *Workshop on Silicon Errors in Logic-System Effects (SELSE)* , 2014, USA.

[8] S. Kiamehr, T. Osiecki, M.B. Tahoori, and Sani Nassif, "Radiation-Induced Soft Error Analysis of SRAMs in SOI FinFET Technology: A Device to Circuit Approach," in *Proceedings of Design Automation Conference (DAC)* , 2014, USA.

[9] S. Kiamehr, P. Weckx, M.B. Tahoori, B. Kackzer, H. Kukner, P. Raghavan, G. Groeseneken, and F. Catthoor, "The Impact of Process Variation and Stochastic Aging in Nanoscale VLSI," *Submitted to International Test Conference (ITC)* , 2015, USA.

ABSTRACT

Very-large-scale integration (VLSI) circuits have changed our life. We use them in a wide range of applications such as personal computers, smart TVs, and cars. The VLSI circuits are also being used in critical domains, such as air-plains and medical sectors. Therefore, their *reliability*, defined as performing consistently according to their specifications, is very important.

The never ending demand for higher performance and lower power consumption pushes the VLSI industry to further scale the technology down. However, further downscaling of technology at nano-scale leads to major challenges. Reduced reliability is one of them, arising from multiple sources e.g. runtime variations, process variation, and transient errors.

System failures caused by unreliability sources can lead to a wide range of consequences: from financial loses, when the faulty system is used for example for banking application, to even loss of human life when the system is used in transportation or medical sector. To avoid/minimize these consequences, the reliability needs to be modeled, predicted, and mitigated. This mandates the circuit designers to consider reliability as another constraint on top of the conventional constraints such as power and performance.

Designers tried to model different sources of unreliability in order to predict the lifetime of the system. Reliability can be modeled and predicted at different levels of abstractions: from device level, which leads to an accurate but high runtime analysis, up to application level which leads to a fast but inaccurate analysis. However, unreliability sources at nanoscale are more sophisticated and their modeling is more challenging for two main reasons. Firstly, some of the unreliability sources are interdependent and they interact which each other in the way they affect system behavior. If they are modeled and considered separately, it may lead to a wrong reliability prediction. Therefore, there is a need to consider all these sources and their interdependencies to accurately predict the lifetime of the system. Secondly, by technology scaling into deep nano-meter era some of the unreliability sources, e.g. transistor aging, have some intrinsic variabilities. This makes the device level models more complex. Due to this complexity, the runtime of detailed device level analysis becomes even larger which makes it infeasible to be used for large circuits. Therefore, it is important to abstract the stochastic device-level reliability models to be used at higher levels of abstraction.

In order to mitigate the reliability challenges such as transistor aging and process variation, one common approach is to add margins (guard-banding) to the design specifications (e.g. clock cycle) to guarantee the correct performance of the circuit. However, the required margin is increasing by technology scaling which erodes the benefit obtained from down-scaling. Therefore the reliability not only needs to be considered at the end of design flow (where the reliability is predicted and a suitable guard-banding is considered), but also it needs to be addressed in the entire design flow (a reliability-aware circuit design flow) in order to mitigate the unreliability and hence to reduce the amount of the required margin.

The objective of this thesis is to tackle unreliability with a cross layer approach from device up to circuit level. The contribution of this thesis is twofold: i) cross-layer modeling and prediction of reliability and ii) reliability-aware cell and circuit design.

In the first part of the thesis a cross layer modeling approach is proposed to cover a wide

range of reliability challenges (combined effect of process and runtime variations, together with transient errors) affecting the system behavior during lifetime operation. In our cross-layer approach, the information from the device level is abstracted to analyze the reliability at circuit level in order to make a trade-off between accuracy and runtime. Our hierarchical approach enables us to perform reliability analysis with a reasonable runtime while maintaining high accuracy. We try to address the interdependent unreliability sources together in order to reduce the inaccuracy of reliability analysis caused by separate consideration of the interdependent unreliability sources. Moreover, the intrinsic variability of unreliability sources, e.g. transistor aging, is considered in our approach and the stochastic information is abstracted to circuit level information which can be used for the analysis at higher levels of abstraction (e.g. architecture level).

In the second part of the thesis, two novel reliability-aware cell and circuit design techniques are proposed to mitigate the issue of accelerated transistor aging, which can significantly impact the circuit lifetime and the system reliability. In both techniques, the device and the gate level aging information is analyzed, and according to this information, the circuit is modified in order to mitigate the effect of transistor aging on the performance of the circuit. With the help of our proposed aging mitigation techniques, the lifetime of the circuit can be improved with very low amount of area and power overheads.

The results of this study show that our cross-layer modeling approach can accurately capture the combined effect of interdependent unreliability sources and their intrinsic variations on the reliability of the circuit. It is shown that considering the interdependent sources of unreliability separately (like state-of-the-art approaches) might lead to a large inaccuracy in the reliability estimation of the circuit which can eventually lead to an over-design (large overhead) or under-design (low reliability). Moreover, the results of the proposed circuit-level reliability analysis can be used at higher abstraction levels such as architecture level. In addition, it is shown that the proposed reliability-aware cell and circuit design techniques can effectively mitigate the aging effect and hence lead to a lower guard-band with negligible overheads.

ZUSAMMENFASSUNG

Hochintegrierte digitale Schaltkreise haben unser tägliches Leben stark verändert und kommen in einem sehr weiten Anwendungsspektrum zum Einsatz, z.B. in PCs, Fernsehern oder Autos. Darüber hinaus werden sie auch in kritischen Domänen verwendet, etwa in Flugzeugen oder im medizinischen Bereich, weshalb ihre Zuverlässigkeit von besonderer Bedeutung ist.

Die ständig steigenden Anforderungen an die Leistungsfähigkeit und Energieeffizienz zwingt die Schaltkreis-Industrie dazu die Transistorabmessungen immer weiter zu verkleinern. Allerdings bringt dies auch zahlreiche Herausforderungen mit sich. Eine dieser Herausforderungen ist die sinkende Zuverlässigkeit der Schaltkreise verursacht durch eine Vielzahl von Effekten wie z.B. Parameterschwankungen zur Laufzeit, Prozessschwankungen während der Herstellung und transiente Fehler.

Systemfehler die durch solche Zuverlässigkeitsprobleme hervorgerufen werden, können dabei weitreichende Konsequenzen nach sich ziehen, angefangen von finanziellen Verlusten (z.B. im Bankensektor) bis hinzum Verlust von menschlichem Leben falls Systeme im Transport- oder Medizin-Sektor betroffen sind. Um derartige Konsequenzen zu vermeiden ist es daher notwendig, die Zuverlässigkeit zu modellieren, etwaige Probleme vorherzusagen und zu verhindern. Deshalb müssen Schaltkreisentwickler die Zuverlässigkeit als einen weiteren Design-Aspekt, neben den traditionellen Parametern wie Leistungsfähigkeit oder Energiebedarf, miteinbeziehen.

Die Zuverlässigkeit kann dabei auf unterschiedlichen Abstraktionsebenen modelliert werden, angefangen bei der Transistorebene, die zwar sehr genau aber dafür sehr zeitaufwändige Analysen erlaubt, bis hinauf zur Anwendungsebene, die eine schnelle dafür aber auch weniger exakte Untersuchung ermöglicht. Erschwerend kommt hinzu, dass die Zuverlässigkeitsprobleme oftmals eng miteinander verbunden sind, so dass eine unabhängige Modellierung mehrerer Faktoren zu einem falschen Ergebnis führen kann. Deshalb ist es notwendig alle Problemquellen und ihre gegenseitigen Einflüsse mit in die Zuverlässigkeitsbetrachtung einzubeziehen. Eine weitere Schwierigkeit in diesem Zusammenhang sind die intrinsischen Schwankungen einiger Zuverlässigkeitsphänomene (z.B. Transistoralterung) bei Verwendung von extrem kleinen Transistoren. Diese führen dazu, dass die Modelle noch komplexer werden, da stochastische Aspekte mitberücksichtigt werden müssen, was wiederum zu steigenden Analysezeiten führt. Daher ist es sehr wichtig die stochastischen Modelle auf Transistorebene zu abstrahieren um auch auf höheren Abstraktionsebene schnelle aber dennoch akkurate Analysen durchführen zu können.

Um Zuverlässigkeitsprobleme verursacht etwa durch die beschleunigte Transistoralterung oder Prozessvariationen zu vermeiden, führen Designer üblicherweise zusätzliche Sicherheitsmargen ein. Allerdings werden diese Margen mit zunehmender Verkleinerung der Transistorabmessungen immer größer, so dass die Vorteile von kleineren Strukturbreiten immer geringer ausfallen. Aus diesem Grund muss die Zuverlässigkeit der Schaltung bereits während des Designprozesses als zusätzlicher Aspekt miteinbezogen werden, damit die Sicherheitsmargen verkleinert werden können und somit eine höhere Leistungsfähigkeit der Schaltung erreicht werden kann.

Das Ziel dieser Doktorarbeit ist es die Zuverlässigkeitsprobleme mit einem Cross-Layer Ansatz zu adressieren, d.h. unter Berücksichtigung der Abstraktionsebenen zwischen Transistor-

und Schaltkreisebene. Die wesentlichen Beiträge dieser Arbeit sind dabei: i) Cross-Layer Modellierung und Vorhersage von Zuverlässigkeitsproblemen, und ii) Design-Methoden für Gatter und Schaltkreise unter Berücksichtigung der Zuverlässigkeit.

Im ersten Teil der Arbeit wird dazu ein Cross-Layer Modellierungsansatz vorgestellt, der zahlreiche Zuverlässigkeitsprobleme umfasst (Prozess- in Kombination mit Laufzeitschwankungen und transienten Fehlern). Für diesen Ansatz werden Informationen auf Transistorebene abstrahiert um eine Zuverlässigkeitsuntersuchung auf Schaltkreisebene zu ermöglichen, die sowohl schnell als auch hinreichend genau ist. Wir berücksichtigen dabei die Zusammenhänge der einzelnen Zuverlässigkeitsprobleme, um die Ungenauigkeiten einer unabhängigen Betrachtungsweise zu reduzieren. Darüber hinaus werden intrinsische Schwankungen von Zuverlässigkeitsproblemen wie etwa der Transistoralterung mit in die Betrachtung einbezogen, und stochastische Modelle auf Schaltkreisebene entwickelt.

Im zweiten Teil der Arbeit werden zwei neue Designansätze für den Gatter- und Schaltkreis-Entwurf vorgestellt, die die Zuverlässigkeit als Designparameter miteinbeziehen. Hierfür wird insbesondere die beschleunigte Transistoralterung untersucht, die die Lebensdauer und Zuverlässigkeit der Schaltung signifikant beeinträchtigen kann. Basierend auf den gewonnenen Alterungserkenntnissen werden die Gatter und Schaltkreise derart modifiziert, dass die Alterungseinflüsse abgeschwächt werden, und gleichzeitig nur geringe Energie- und Flächenkosten entstehen.

Die Ergebnisse dieser Arbeit zeigen, dass der Cross-Layer Modellierungsansatz den kombinierten Einfluss von sich beeinflussenden Zuverlässigkeitsproblemen sehr genau erfassen kann. In diesem Zusammenhang zeigen die Ergebnisse auch, dass eine unabhängige Untersuchung dieser Zuverlässigkeitsprobleme (wie in der Literatur üblich) zu großen Ungenauigkeit führen kann, was schlussendlich entweder die Zuverlässigkeit beeinträchtigt oder zu unnötig großen Sicherheitsmargen führt. Darüber hinaus wird gezeigt, dass die Gatter- und Schaltkreis-Entwurfsmethoden sehr effektiv sind, und damit deutlich kleineren Sicherheitsmargen verwendet werden können.

CHAPTER

ONE

INTRODUCTION

There are more and more electronics in our nowadays life. Most prominent examples are obviously the devices that connect us, such as smart phones and iPads. However, there is the other class of embedded devices where many people are not aware of how much electronics is involved. Cars wouldn't drive, planes wouldn't fly, factories wouldn't manufacture without the electronic devices. Electronics needs to be reliable, otherwise it would result in minor discomforts (Facebook not working) or catastrophes (planes falling from sky).

Very Large Scale Integration (VLSI) technology is the enabling technology for the wide range of electronic devices we use in our daily life which have changed the way we live nowadays [2]. The VLSI devices now is being used in the personal entertainment systems, automotive industry, medical electronic systems and financial sector. The range of applications is continuously growing which pushes the design and manufacturing to scale the VLSI technology (transistor dimensions) down to obtain more complex systems with smaller size (portable devices), lower power consumption, higher performance and reduced cost [2]. For this purpose, the microelectronic industry has been trying to follow the Moore's law [3], to shrink the device feature size (transistor dimensions) in a way that the number of transistors on a die doubles approximately

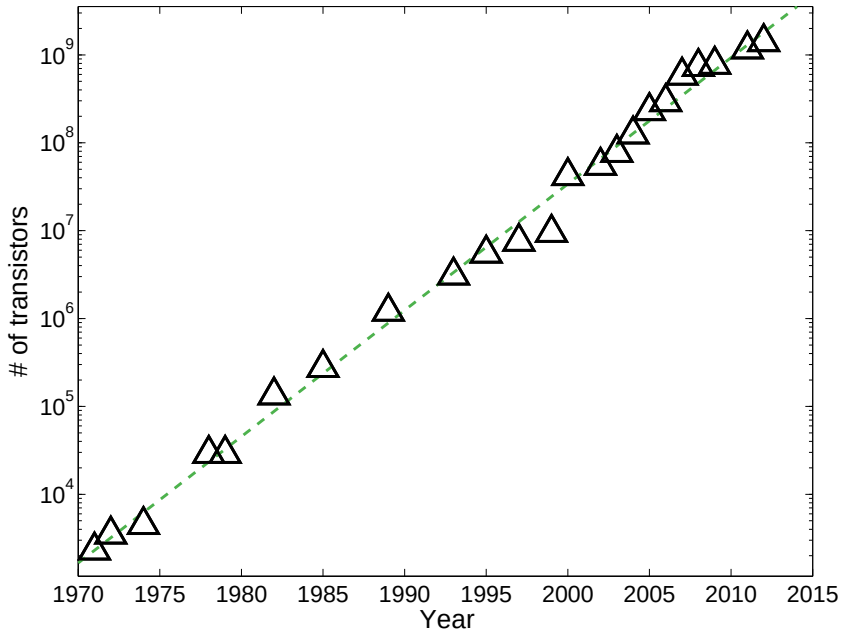


Figure 1.1.: The number of transistors of Intel processors [1]

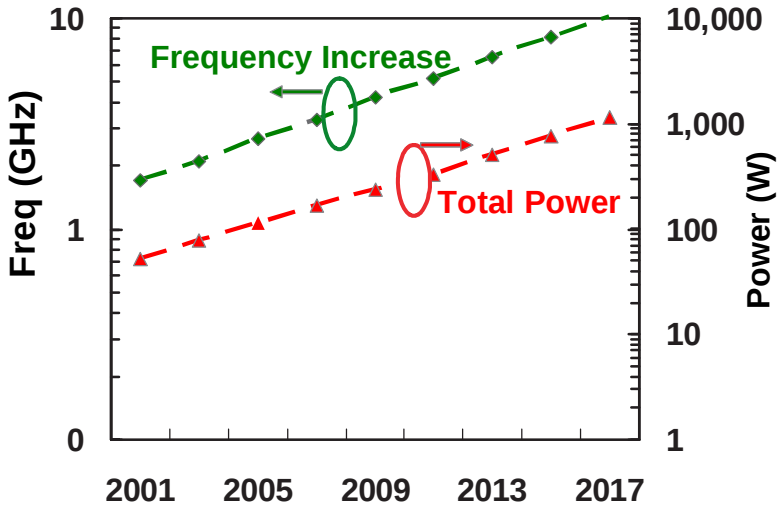


Figure 1.2.: Frequency and power consumption trend for different technology nodes over time [10]

every two years. Figure 1.1 illustrates the number of transistors of Intel processors since 1970 which perfectly shows the consistency with Moore's law.

Although technology scaling provides many opportunities which enables building extremely complex gates and achieving better computing performance, it faces many challenges [4]. Power issue [5] (see Figure 1.2), short channel effect [6], and in particular yield and *reliability* issues [4, 7, 8] are the most important challenges that the VLSI technology is faced due to the technology scaling. Due to the wide range of the VLSI chips applications, the unreliability of these chips can lead to a broad range of consequences from computer crashes and loss of data to financial losses and even loss of human life [9].

Reliability, as an important nano-scale technology issue, can be quantified by the probability that a system operates correctly without a failure until time t [11]:

$$R(t) = P(T > t) = \int_t^{\infty} f(x)dx \tag{1.1}$$

where T is the duration of normal system operation without a failure and $f(x)$ is the failure probability density function.

A failure occurs when for example there is an error in the functionality of the system or a wrong bit value is read or written in the memory cell [12]. The failures happen due to faults which can be categorized into two categories [11, 13]: i) permanent faults: is caused due to physical defects in the components of a circuit (such as wearout) which can permanently alter the circuit function [11–14]. ii) temporal faults: which is caused by transient or intermittent disturbances which affect the circuit for a short period of time (not permanently) [11, 13, 14]. In general these faults can affect the reliability of the VLSI chip as shown in the bathtub curve depicted Figure 1.3.

Technology scaling deep into nanometer era, makes the reliability issues more pronounced. Figure 1.4 shows the predicted unreliability acceleration (increase of failure rate) for different technology nodes. According to the figure, the failure rate increases exponentially as the technology scales down. This bring us to the domain in which the traditional design flows do not work [4] and the designers should rethink how to change the design flow in order to not only optimize the power and performance but also meet the reliability and yield constraints. However, designing a reliable system out of unreliable component is very challenging. *For this purpose, the reliability issues need to be well understood, modeled and mitigated.*

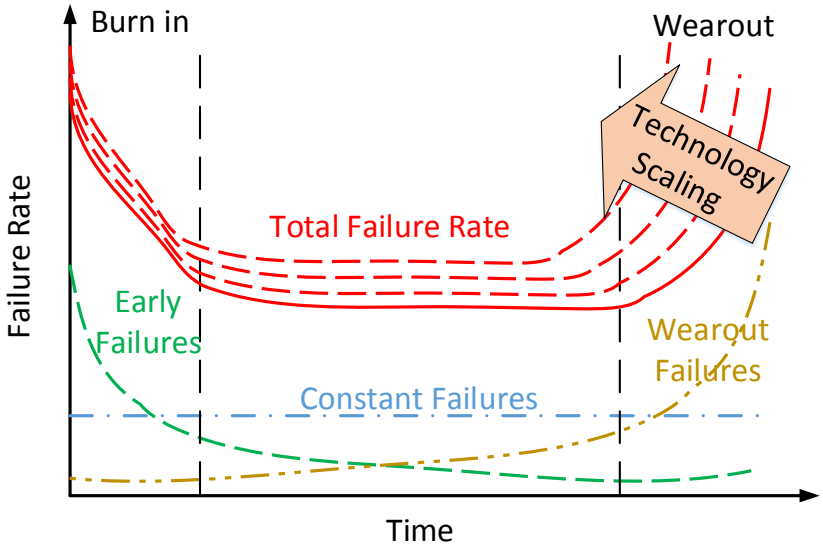


Figure 1.3.: Failure rate Bathtub curve

1.1. Sources of unreliability

Among all the reliability issues in nano-meter technology, *process variation, time dependent variation* and *soft error* are some of the most important issues [7–9].

Soft error is a temporal failure caused by cosmic ray radiation and alpha particle generated from packaging materials. It can cause a flip in the status of memory (see Figure 1.5) or sequential elements of the circuit (e.g. latch and flip-flops). Therefore, several methods have been proposed to protect memory [16, 17] and sequential elements [18] against this type of error. Soft errors can also affect the combinational logic part of the circuit by generating a transient pulse which may be propagated to sequential elements. Although the combinational logic soft error was not a big issue in the previous technology nodes, its contribution is sharply increasing by scaling in deep nano-meter regime [9, 15].

Process variation and time dependent variation lead to a variation in the properties of the

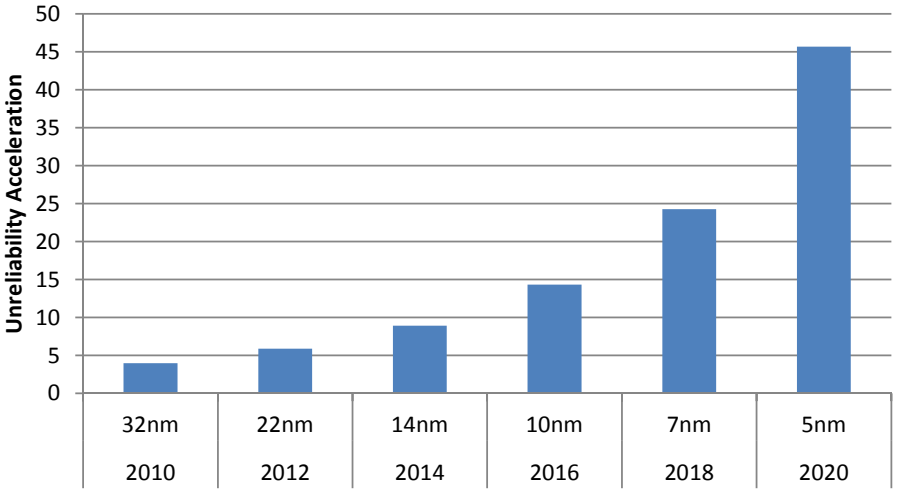


Figure 1.4.: Unreliability acceleration due to scaling for different technology nodes normalized to the failure rate at 32 nm technology node [15]

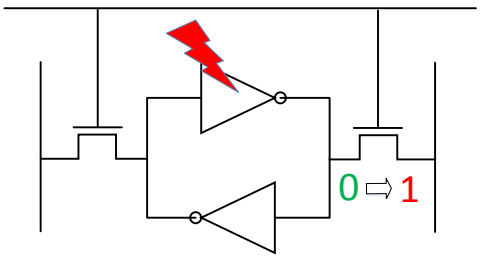


Figure 1.5.: SRAM cell memory affected by a soft error

circuit (e.g. delay and power) compared to the indented specification during the design time. If the timing and/or power constraints are not met due to these variations, a failure might happen. Figure 1.6 shows a simple example in which the timing constraint is not met and a wrong value is stored in the memory or sequential elements leading to a failure.

Process variation is a natural device (transistors) parameter variation (e.g. threshold voltage) among different devices [19]. There are many physical issues leading to process variation, however, the statistical fluctuation of channel dopant is a major cause of variation among devices in deep nano-meter technology [19]. Due to process variation, the performance of the fabricated circuit (e.g. circuit delay) becomes a statistical value as shown in Figure 1.7.

The other set of variability is the time dependent variation. Due to this variation, even two completely identical devices (transistors) with the same initial characteristics may have different characteristics over time according to their different working conditions such as temperature and workload [7]. Transistor aging as one of the most important sources of time dependent variability [7–9] is due to different degradation mechanisms: *Negative Bias Temperature Instability (NBTI)*, *Positive Bias Temperature Instability (PBTI)* and *Hot Carrier Injection (HCI)*.

In order to address different types of variability, a safety timing margin (guard-band) is added to the design to guarantee the reliable operation of the designed system. However, an optimistic or pessimistic guard-banding may lead to a large failure rate or an over design (less performance), respectively (see Figure 1.7).

In summary, all the aforementioned reliability issues are threatening the reliable operation of the system and the situation becomes worse by technology scaling. Figure 1.8 shows the contribution of different reliability issues on the total failure rate of the VLSI circuit for different technologies. According to the figure, the failure rate is increasing. Therefore it is very important to model and mitigate these sources of reliability issues.

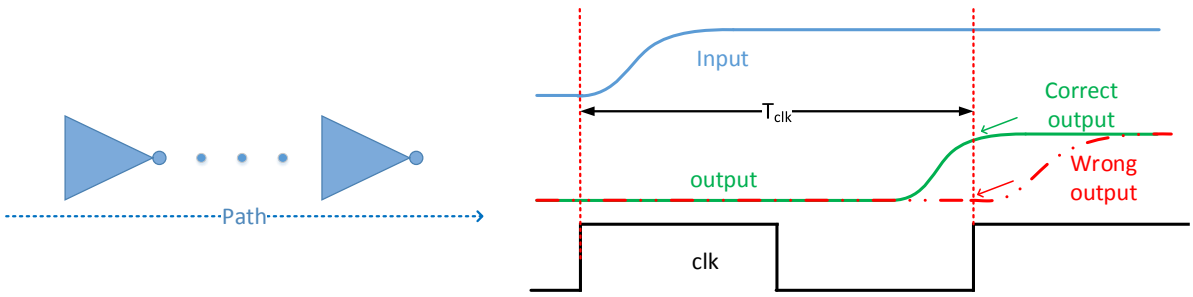


Figure 1.6.: Timing failure due to variation in the path delay

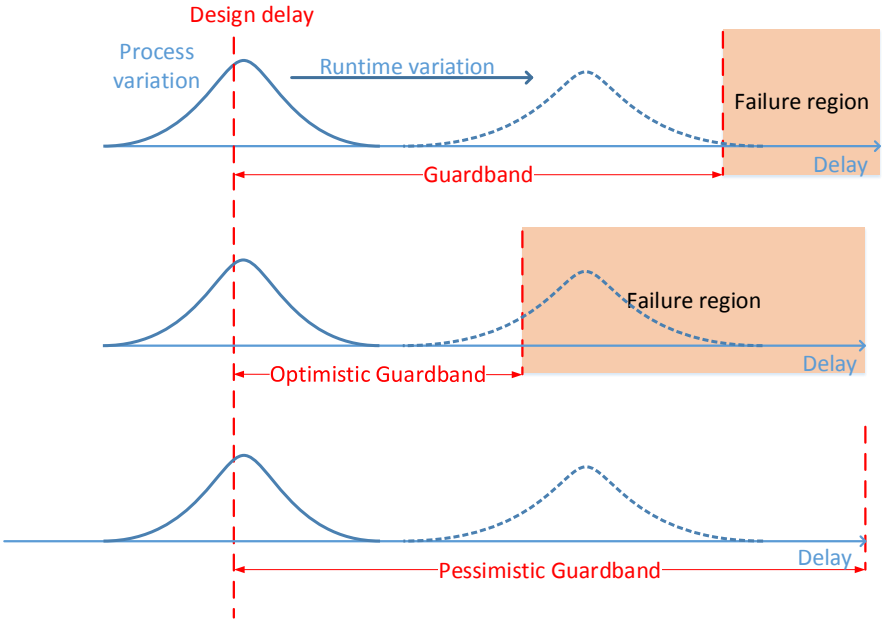


Figure 1.7.: Guard-banding to avoid failures due to process and runtime variation

1.2. Challenges for modeling and mitigation of unreliability

The reliability can be addressed at different levels of abstraction from device level up to application level (see Figure 1.9). At the device level, the reliability is modeled with complex and detailed models which are very hard (almost infeasible) to be used at high levels of abstraction, e.g. architecture level, due to the very large runtime of analysis for large circuits. The situation becomes worse by technology scaling since some of the reliability issues, e.g. transistor aging, have some intrinsic variability [20]. This means that stochastic models need to be used for these reliability issues which makes the models even more complex.

Moreover, some of the unreliability sources have interdependencies. In other words, they affect each other which means that the impact of one issue may be aggravated or alleviated in the presence of other issues. For example, the impact of transistor aging is affected in

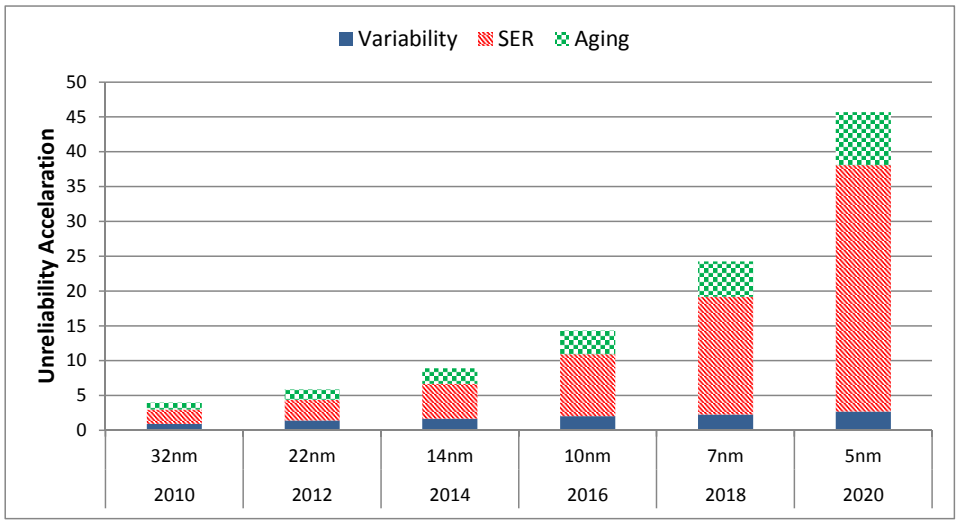


Figure 1.8.: Contribution of different reliability issues on the total unreliability of VLSI circuits [15]

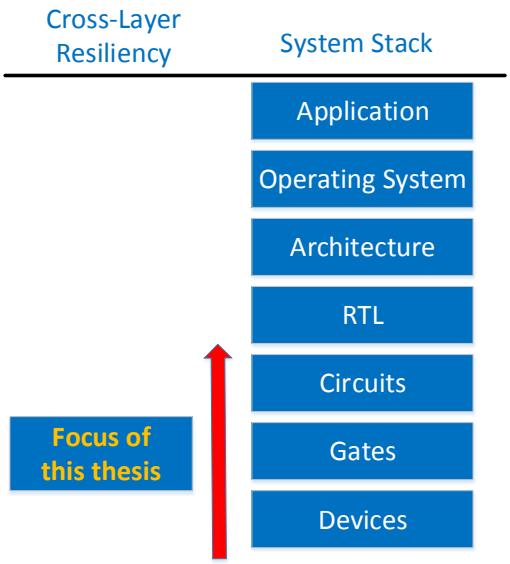


Figure 1.9.: System stack and cross layer resiliency

the presence of process variation. Therefore, it is important to consider the interdependent sources of unreliability together since considering these sources separately may lead to an overestimation or underestimation of the reliability impacts.

Due to the aforementioned complexity of device level modeling and interdependency of different sources of unreliability, it is very challenging to come up with some models which can be used at higher levels of abstraction. Using the complex and time-consuming models considering these interdependencies makes it infeasible to be used for large circuit (e.g. an entire processor) due to issues such as simulation runtime. Therefore, the models need to be simplified, however, the simplification is obtained at the cost of inaccuracy. Putting all together, it is important to come up with scalable models which give a good trade-off between runtime and accuracy.

If the reliability is accurately modeled and predicted, suitable countermeasures can be applied to guarantee the correct performance of the chips. One common practice in the current technologies is to add timing margin (guard-banding) to avoid circuit failures due to reliability challenges such as runtime and process variabilities. However, by technology scaling, the additional required timing margin increases (i.e. lower performance) which erodes the benefit obtained from down-scaling. Therefore, the reliability needs to be considered in the early stages of the design as a constraint in addition to conventional constraints (such as power and delay).

The objective of this thesis is to model and optimize the circuit reliability in the presence of nanoscale unreliability effects with a cross layer approach from device level up to circuit level. The interdependent reliability challenges are addressed together in order to reduce the inaccuracy of circuit reliability analysis caused by separate consideration of the interdependent unreliability sources. In the proposed cross layer approach, the knowledge at different levels of abstraction from device up to circuit level is combined. The hierarchical approach enables us to perform reliability analysis with a reasonable runtime while maintaining high accuracy.

Moreover, several novel techniques to mitigate the issue of accelerated transistor aging, which can significantly impact the circuit lifetime resiliency and the system reliability, are proposed. For this purpose, the standard cells and the circuit are redesigned in a way that the aging effect is reduced. With the help of our proposed aging-aware cell and circuit design, the lifetime of the circuit can be improved with a very low amount of area/power overheads.

1.3. Contribution of this thesis

As mentioned before, the contribution of this thesis is two-fold: i) cross-layer modeling and prediction of reliability. In the proposed modeling, the interdependency of reliability challenges is considered. Moreover, the intrinsic variation of them is also considered. Using our cross-layer approach, the information from device level is abstracted at circuit level which provides suitable information to be used at higher abstraction levels such as architecture level. ii) reliability-aware cell and circuit design in which we try to consider transistor aging, as an important reliability issue, in early stages of cell and circuit design to make them more resilient against aging effect. Using proposed techniques at gate/circuit level, we can either improve the lifetime of the circuit or decrease the amount aging-induced timing guard-band with negligible area and power overheads. In particular, the new contributions of this thesis are as follows:

Cross-Layer Modeling and Prediction of Reliability Issues

- **Chip-Level Modeling and Analysis of Electrical Masking of Soft Errors**: With continuous downscaling of VLSI technologies, logic cells are becoming more susceptible to radiation-induced soft error. Moreover, increasing complexity of VLSI chips at nanoscale results in voltage droop, which is an important source of runtime variabilities, across the chip. The soft error rate of the chip is dependent on the value of supply voltage which is seen by its gate. Therefore, voltage droop may affect the soft error rate of the chip. To consider this dependency, we present a chip-level soft error analysis which accurately considers the impact of voltage droop across the chip.

- **Radiation-Induced Soft Error Analysis of SRAMs in SOI-FinFET technology: A Device to Circuit Approach**: A comprehensive analysis of radiation-induced soft errors of SRAMs designed in SOI FinFET technology is presented. For this purpose, we propose a cross layer approach starting from a 3D simulation of particle interactions in FinFET structures up to circuit level analysis by considering the layout of the memory array. This approach enables us to consider the effect of different factors such as supply voltage and process variation on soft error rate of FinFET SRAM memory arrays.

- **The Impact of Process Variation and Stochastic Aging in Nanoscale VLSI**: With the down-scaling of CMOS technology into deep nano-scale era, the aging effect becomes stochastic due to its widely distributed defect parameters leading to more non-determinism in the functionality of the deeply-scaled circuits. A framework is presented to comprehensively investigate the combined effect of stochastic aging effect and process variation on the performance of the VLSI design at circuit level, by abstracting atomistic aging models (for the stochastic behavior) to the circuit timing analysis flow.

Reliability-aware Cell and Circuit Design

- **Reliability-aware Standard Cell Library Design**: In current VLSI design flow, for each technology node, the building blocks are pre-designed and optimized (at both netlist and layout) and placed in a library, so called *standard cell library*. Then, the circuits are designed and synthesized using these building blocks. Typically, the standard cells are designed considering the area, power and delay without considering the reliability issues, however, the first step to design a reliable circuit is to have reliability-aware standard cell library. In this study, we propose a technique to consider the aging effect in the standard cell library design. Then, we propose a method to adjust the mapping flow according to the new reliability-aware standard cell library.

- **Input and Transistor Reordering for Aging Reduction in Complex CMOS Gates**: In this study, first we show that the order of cell input connections has a considerable impact on transistor aging. Based on this, we redesign the cell connections in order to reduce the aging effect with no effect on the functionality of the circuit and minimal area and power overheads.

1.4. Outline

The rest of this thesis is organized as follows. Chapter 2 as the background chapter is followed by twofold contributions of this thesis: i) cross-layer modeling and prediction of reliability Issues (Chapters 3-5) and ii) reliability-aware cell and circuit design (Chapter 6, 7). Finally the thesis is wrapped out with conclusion and summary in Chapter 8.

After the introduction, in Chapter 2, the necessary background of this thesis is provided. In this chapter, a basic information of transistor structure and functionality is discussed followed by some basic information about the structure of the logic gates and their important properties. Finally, at the end of the section, the important reliability issues and their corresponding models are provided.

After the background section, the thesis continues with its twofold contribution. The first part of contributions (Chapters 3-5) contains a cross-layer modeling and prediction of reliability issues.

In Chapter 3, a new method is proposed to estimate the soft error rate considering voltage droop as an important runtime variability issue which affects the estimation of soft error rate. In this chapter, first the motivation of the work is provided. Afterwards, the related work is discussed. Next, the proposed methodology of soft error rate estimation considering voltage droop is introduced. Then, the experimental results are provided showing the importance of considering voltage droop in soft error estimation. Finally we conclude the chapter.

Chapter 4 provides a cross-layer approach to estimate the soft error rate in SOI-FinFET technology which considers other important issues such as process variation and the effect of supply voltage. In this chapter, first the motivation and contributions of the study are provided. Then, the related work is discussed. Afterwards, the proposed cross-layer approach is provided in three consecutive sections describing the device level, cell level and circuit level analysis, respectively. Then simulation results are provided and finally we conclude the chapter.

At the end of the first part of thesis contribution, Chapter 5 introduces a framework to investigate the combined effect of process variation and stochastic aging effect in FinFET technology at circuit level. In this chapter, after a short introductory about the motivation and contributions of the study, the related work is discussed. Next, the proposed circuit level simulation flow is introduced to obtain the combined effect of stochastic aging and process variation on the circuit delay. Afterwards, the results are provided with a comprehensive discussion about the observations. Finally the achievements are summarized and the chapter is concluded.

The second part of this thesis contributions (Chapter 6, 7) provides new reliability-aware cell and circuit design techniques in order to alleviate the aging effect as an important reliability issue in nano-scale technology nodes.

This part starts with Chapter 6 in which a reliability-aware standard cell library design technique is provided to mitigate the aging effect. In this chapter, first the related work is discussed. Then, the proposed reliability-aware standard cell library is provided. The next section provides a methodology to remap the circuits using the new reliability-aware standard cell library in order to make the circuit more resilient against aging effect. The experimental results are provided afterwards which is followed by conclusion section to summarize the

achievements.

Chapter 7 provides a new input and transistor reordering technique in order to reduce the effect of aging. The first part of the section is an introduction of study including the motivation and contributions. Next, the related work is briefly discussed providing what is missing in state-of-the art. Then, the effect of input and transistor orders on the aging effect of cells are investigated and we show that the order of inputs and transistors has a huge impact on the amount of cell degradation. According to this cell level investigation, the next section provides a circuit level technique to redesign the cell connections and transistors in order to make the circuit more resilient against aging effect. Next, the simulation results are provided and at the end of the chapter, the achievements are summarized and the chapter is concluded.

At the end of this thesis, Chapter 8 is provided which summarizes the achievements and concludes the thesis.

Part I.

Background

CHAPTER
TWO

BACKGROUND

2.1. Basic terminology of CMOS technology

Complementary metal-oxide-semiconductor (CMOS) is the main technology being used for constructing integrated circuits. In this section a brief fundamental background of CMOS technology is provided. Further details can be found in [21].

2.1.1. MOSFET transistors

Metal-oxide-semiconductor field effect transistor (MOSFET) is a type of transistor mainly used in CMOS technology. Figure 2.1 shows the structure of a MOSFET transistor. As shown in this figure, a MOSFET consists of Gate (G), Drain (D), Source (S) and Substrate (SB) terminals.

There are two types of MOSFET transistors: n-type MOSFET (referred here as NMOS) and p-type MOSFET (referred here as PMOS). NMOS (PMOS) transistors are made with a p-type (n-type) substrate and their channel contains electrons (holes) as the carriers. The substrate of NMOS (PMOS) transistors are normally connected to ground (V_{DD} node).

MOSFETs are used as switches as shown in Figure 2.2. When the magnitude of the gate-source voltage of the transistor ($|V_{GS}|$) is less than the magnitude of the transistor threshold voltage ($|V_{th}|$), the MOSFET is OFF and there is no connection between source and drain. Otherwise, if $|V_{GS}|$ is bigger than $|V_{th}|$, the transistor is ON and there is a connection between the source and the drain. However, the transistors are not ideal switches. When the transistor

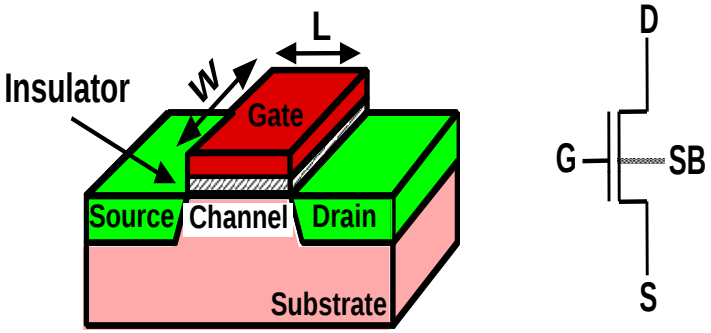


Figure 2.1.: MOSFET structure

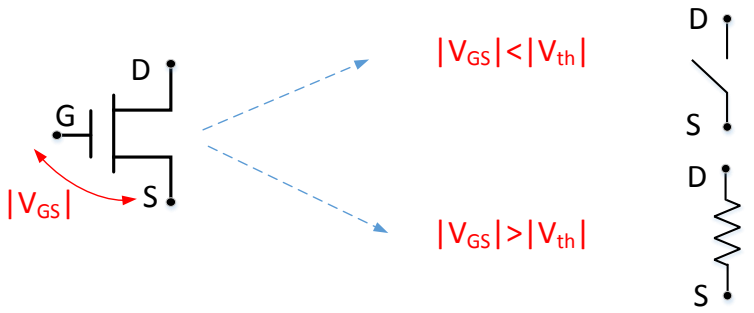


Figure 2.2.: Transistors as switches

is OFF, still a small current is leaked between the drain and the source (so called sub-threshold or leakage current) which is a function of different factors:

$$I_{DS}^{subthreshold} = I_{DS0} \cdot e^{\frac{V_{GS}-V_{th}+\eta V_{DS}-k_\gamma V_{SB}}{n\nu_T}} \left(1 - e^{\frac{-V_{DS}}{\nu_T}}\right) \tag{2.1}$$

where n is a process-dependent term, I_{DS0} is an empirical parameter, ν_T is the thermal voltage, η is a coefficient reflecting drain-induced barrier lowering (DIBL) effect and k_γ is a coefficient reflecting body effect. Moreover, V_{GS}, V_{DS}, and V_{SB} are gate-source, drain-source and source-body voltages, respectively.

On the other hand, when the transistor is ON, there is a resistance between the source and the drain. In this case, transistor could have two different states according to its drain-source voltage (see Figure 2.3):

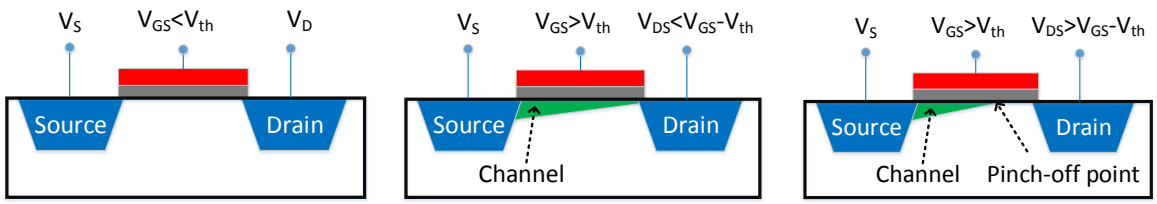
- Linear model ($V_{DS} < V_{GS} - V_{th}$): In this mode the drain-source current is a function of both V_{GS} and V_{DS}.

$$I_{DS}^{ON-lin} = \mu C_{ox} \frac{W}{L} \left(V_{GS} - V_{th} - V_{DS}/2\right) V_{DS} \tag{2.2}$$

- Saturation model ($V_{DS} > V_{GS} - V_{th}$): In this mode the drain-source current is almost independent of V_{DS}:

$$I_{DS}^{ON-sat} = \frac{1}{2}\mu C_{ox} \frac{W}{L} (V_{GS} - V_{th})^2 \tag{2.3}$$

where μ is the mobility of the carriers (electrons in NMOS and holes in PMOS) and C_{ox} is the capacitance per unit area of the gate oxide. W and L are the transistors width and length, respectively.



(a) (b) (c)

Figure 2.3.: Transistor a) OFF state b) ON state linear mode and c) ON state saturation mode

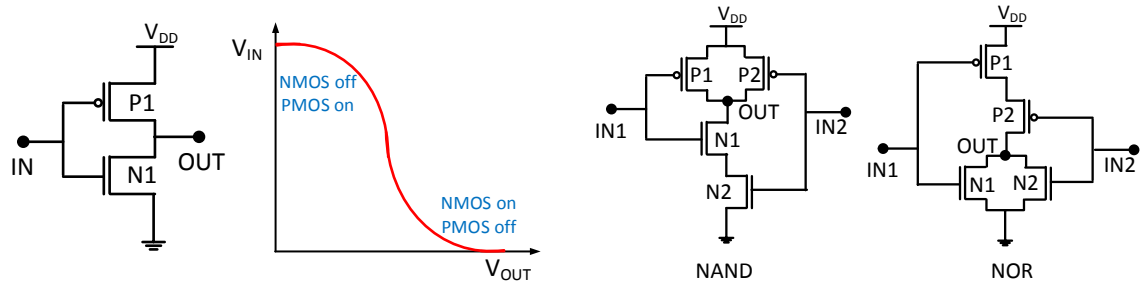


Figure 2.4.: a) CMOS inverter and its DC plot b) two input NAND and NOR CMOS gates

2.1.2. Basic CMOS gates

NMOS can only transfer logic 0 ($V = 0$) ideally and cannot ideally transfer logic 1 ($V = V_{DD}$). For PMOS transistor, the case is the other way around meaning that it can only transfers logic 1 ($V = V_{DD}$) ideally. For this reason, a combination of NMOS and PMOS transistors are used to implement complementary MOS (CMOS) gates. Figure 2.4(a) shows the implementation of a CMOS inverter. As shown in this figure, a PMOS transistor is used as pull-up network to make the output equal to logic 1, when the input is equal to logic 0. On the other hand, an NMOS transistor is used as the pull-down network to convey a logic 0 at the output when the input is equal to logic 1.

Similarly, other types of CMOS gates such as NAND and NOR can be implemented using PMOS and NMOS transistors (see Figure 2.4(b)).

Gate delay

The circuit needs to be designed in way that its delay meets the timing constraint. Therefore, the circuit delay is an important parameter in the design. The circuit delay in turn is a function of its internal gate delays. Therefore, it is important to define the gate delay accurately. For this purpose, the gate propagation delay and the signal transition time are defined as follows (see Figure 2.5):

- **Gate propagation delay**: Gate delay is defined as the time required for the output to reach 50% of its final output level (50% of supply voltage value) when the input changes to 50% of its final input level (50% of the supply voltage value).
- **Signal transition time**: This term is a representative for the slope of the signal transition. There are two types of transition times:

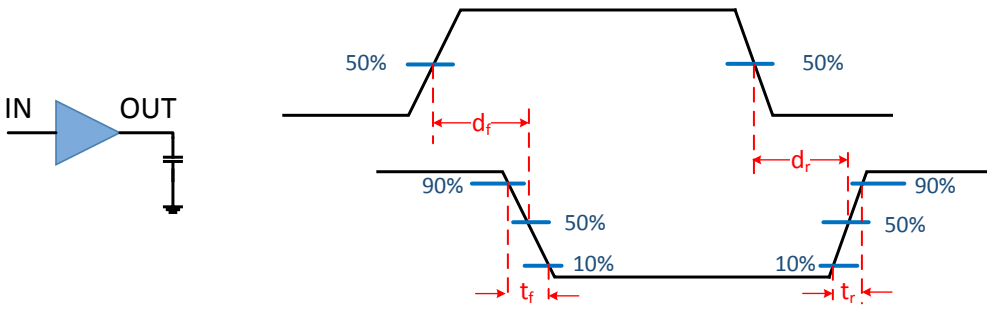
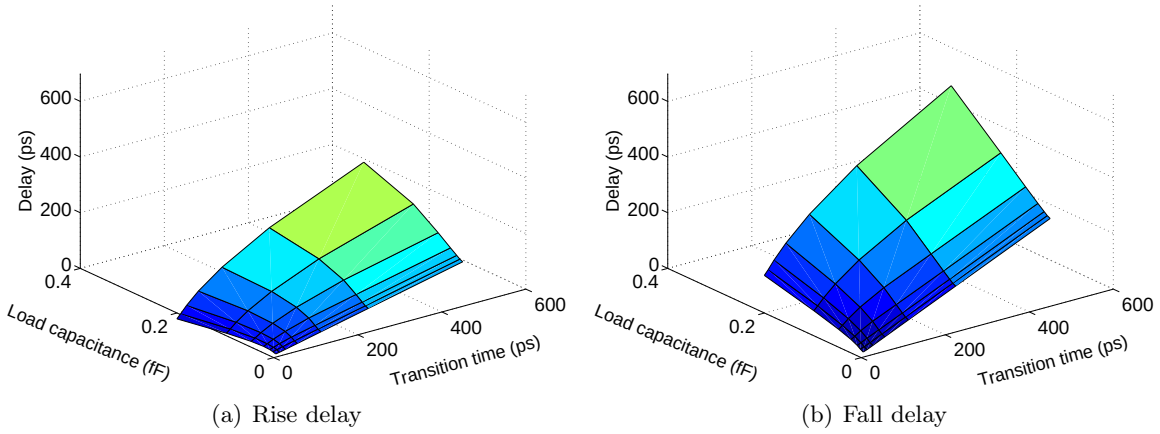


Figure 2.5.: Definition of the gate delay and signal transition time



(a) Rise delay (b) Fall delay

Figure 2.6.: The dependency of a) Rise and b) fall delay of an inverter (using Nangate 45 nm technology library [22]) to output load capacitance and inputs transition time.

1. Rise transition time (t_r): it is defined as the time required for a signal to reach from 10% (30% for more advanced technology nodes) to 90% (70% for more advanced technology nodes) of its final value when the signal has a rise transition.
2. Fall transition time (t_f): it is the time required for a signal to reach from 90% (70% for more advanced technology nodes) to 10% (30% for more advanced technology nodes) of its final value when the signal has a fall transition.

The delay of a gate is function of many parameters. On one side, it is a function of gate internal transistors properties, i.e. transistor width, length and threshold voltage. Therefore, the delay of different gates with different internal transistors properties will be different. Moreover, the gate delay is a function of the gates input transition time and the output load capacitance (see Figure 2.6). This means that the delay of similar gates would be different if they have different load capacitances and input signal transition time. Therefore, in a standard library cell, the delay of each standard cell (gate) is provided as a 2-dimensional *look-up table* (LUT) where one dimension is for output load capacitance and the other one is for input signal transition time.

On top of that, the gate delay is also a function of environmental parameters, e.g. temperature and supply voltage. Temperature affects both carriers mobility and transistor threshold voltage and hence it affects the delay of the gate. The supply voltage impacts the current drawn from transistor (see Equations 2.2 and 2.3) and hence the gate delay.

2.1.3. FinFET transistors

According to the Moore's Law, the number of transistors per chip doubles every two years [3]. In order to keep up with the Moor's law, the semiconductor industry have been scaling the dimension (gate length) of MOSFET transistors for more than 40 years. At the end of 1990's the semiconductor companies started manufacturing a new type of transistors called *Silicon On Insulator (SOI)* [23]. In this technology, a layered silicon-insulator-silicon substrate is used in place of conventional silicon substrates and it has two advantages of reduced parasitic capacitances and enhanced current drive compared to conventional MOSFET [23].

The technology scaling is slowed down recently due to different issues such as *Short Channel Effect (SCE)* and excessively large variations in device properties [25] which makes it infeasible to follow the Moore's law with conventional MOSFET and SOI devices. Therefore, VLSI industry is researching new device structures in order to be able to continue scaling. Among

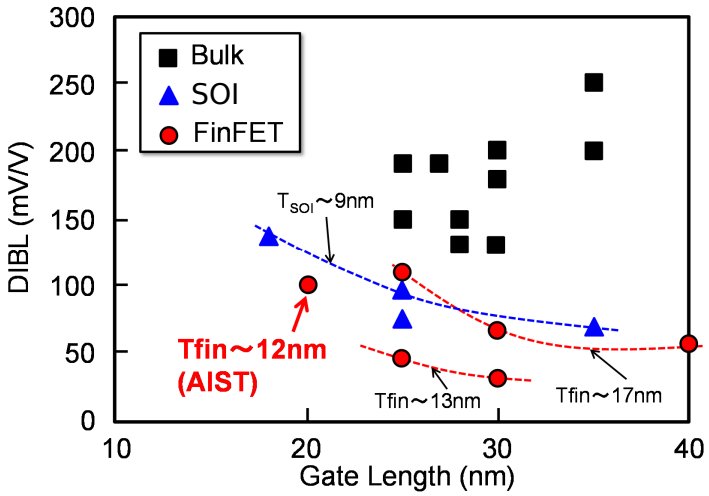


Figure 2.7.: Drain-induced barrier lowering (DIBL) effect in FinFET, SOI and conventional planar bulk transistor structures [24]

different candidates (such as *Carbon Nanotube Field Effect Transistors (CNTFET)*, multi-gate transistors) *Fin Filed Effect Transistor (FinFET)* is one of the most promising structures which is already fabricated by Intel [26], GlobalFoundries [27, 28] and TSMC [29]. This is due to the fact that FinFET exhibits superior immunity to short channel effects. Figure 2.7 shows the drain-induced barrier lowering (DIBL) effect, as an important SCE issue, in three different transistor structures [24]. As shown in this figure, FinFET shows the smallest DIBL (highest SCE immunity). Moreover, the effect of process variation on FinFET device performance is less compared with conventional bulk devices [30, 31].

FinFETs can be fabricated as a bulk device or on SOI. Figure 2.8 shows the structure of a SOI FinFET. As shown in this figure, the gate is wrapped around the channel which provides a better control over the channel and as a result the SCE is less in this type of transistors compared to the conventional planar MOSFET and SOI structures.

2.2. Reliability issues

With down-scaling of CMOS technology into deep nanometer, reliability has become a major issue [32]. In this section, the general sources of reliability issues in current technology nodes are briefly explained.

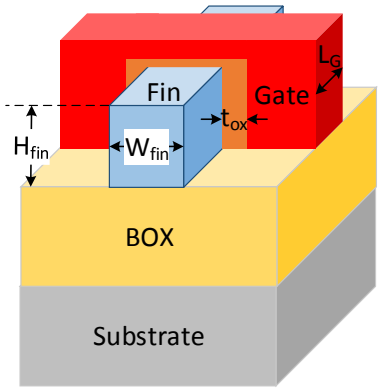


Figure 2.8.: Structure of a SOI FinFET

The sources of unreliability in current technology nodes can be categorized into three different categories: i) variability issues, ii) transient faults and soft errors iii) permanent faults.

Due to variability, the devices/gates/circuits characteristics are different from the intended designed ones. The variability could be due to *"time-zero" variation (process variation)* or *runtime variation*. Process variation is a natural device parameter variation which makes the properties of fabricated devices different from that of designed ones. In other words, due to process variation different similarly designed transistors/gates will perform (operate) differently after fabrication. Due to runtime variation, the transistors/gates properties will change (degrade) during the chip operational lifetime.

Runtime variations are routed in different sources such as voltage variation, temperature variation and transistor aging. The voltage and temperature variations are temporal or spatial according to the place of the transistor/gate and also the workload. Therefore, they cause variation on the properties of different transistors/gates at different location of the circuit and at different time points during the chip operational lifetime.

Transistor aging is the other source of runtime variations caused by different wearout effects such as Bias Temperature Instability (BTI), Hot Carrier Injection (HCI) and soft *Time Dependent Dielectric Breakdown* (soft TDDB). All these effects cause the threshold voltage of the transistor to increase and hence the switching delay of the transistor increases which can eventually lead to a timing failure if the delay of the circuit does not meet the timing constraint.

In order to deal with these sources of variation, guard-banding is a common approach. In this approach, a timing margin is added to the designed clock cycle in order to guarantee the correct operation of the circuit during the operational lifetime. A pessimistic guard-banding leads to a performance loss and an optimistic guard-banding results in a low reliability of the chip. Therefore, the required timing margin needs to be accurately predicted. Figure 2.9 shows the components of the required timing margin for IBM Power7+ processor [33]. As shown in this figure, the main components of the timing margin are uncertainty (e.g. process variation), weraout (transistor aging) and voltage and thermal variations.

The other category of reliability issues is the transient soft errors caused by alpha particles from packaging materials and neutrons in cosmic particles. Transient soft errors do not cause a permanent degradation or fault and it leads to a transient computational error [7]. However, since its nature is random, the detection and correction of this type of errors is very challenging [7]. The soft error can affect memory cell, sequential elements of the circuit and also combinational part of the circuit. Traditionally, only single errors caused by *single event upsets* was considered as the target of detection and correction methods [9]. However, by continuous scaling of transistor dimensions, the probability that multiple nodes of the circuit are affected simultaneously by a strike (*multi-bit upsets* (MBU)) becomes larger which makes the detection and correction even more challenging.

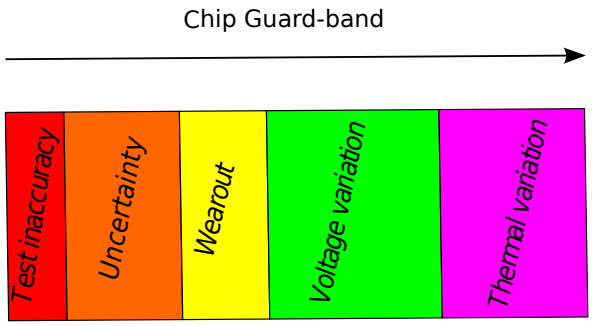


Figure 2.9.: Components of chip guard-band for the IBM Power7+ [33]

Permanent faults is another important category of reliability issues which has been a concern since early days of electronic industry [7]. *Electromigration (EM)* is one of the most important types of permanent faults which can cause an interconnect disconnection due to the transport of material. EM is caused by the movement of ions in an interconnect due to the transfer of the momentum between the carriers and the atoms of the interconnect [34].

Time Dependent Dielectric Breakdown (TDDB) is also a major reliability issue which can lead to permanent fault [35]. The material of transistor gate oxide is degraded when a sufficiently high electric field is applied across the gate oxide which leads to an increase of its conductance. In case of a long term application of electric field a conductive path may be formed in gate oxide leading to an abrupt increase of gate leakage current. This issue is called hard TDDB and it becomes more severe as the gate oxide thickness becomes thinner due to the technology scaling.

In the following sections, some of the reliability issues which are targeted in this thesis will be explained in more details.

2.3. Process variation

The performance of a circuit is a function of its device characteristic and any variation in the characteristic of devices will lead to a deviation of the circuit performance from its intended designed value. This variation is called process variation and it can cause the circuit to fail if the performance of the circuit does not meet the constraint. Process variation can be categorized into two categories: i) *Front-end variability* which is the variations caused by manufacturing process of the device (e.g. transistor length variation) and ii) *Back-end variability* which is the variations caused by manufacturing process of the interconnect [36]. The contribution of these two types of variability is different for various type of reliability concerns (e.g. timing variability or parametric yield) [36]. However, in terms of timing variability, front-end variability is dominant and its contribution in the total path delay is around 90% [37].

2.3.1. Sources of front-end variability

There are different sources of front-end variability, but we will explain the most important issues in the following.

- **Line-edge Roughness (LER)**: LER is the variation of the edge of the gate along its width which is caused due to the lithography variation [36] (see Figure 2.10). LER impacts different device characteristics such as the threshold voltage and the subthreshold current [36, 38].
- **Dielectric thickness variation**: The thickness of the dielectric between the gate and the channel has a large impact on the device characteristic such as the threshold voltage, the drive current and the leakage current [36]. This impact has significantly increased by the continued technology scaling and any variation in the thickness of dielectric will cause a huge variation in the device characteristic [39].
- **Random Dopant Fluctuation (RDF)**: The dopant atoms are placed via ion implantation into the channel. The implantation is such that the number and the location of dopant atoms in the channel is random. This phenomenon is called Random Dopant Fluctuation (RDF) which causes a significant variation in the threshold voltage of the transistor. The effect of RDF on the threshold voltage increases by the technology scaling since the number of dopant atoms in the channel decreases with the scaled dimensions [40]. RDF is the major source of mismatch for identical adjacent devices [38, 41].

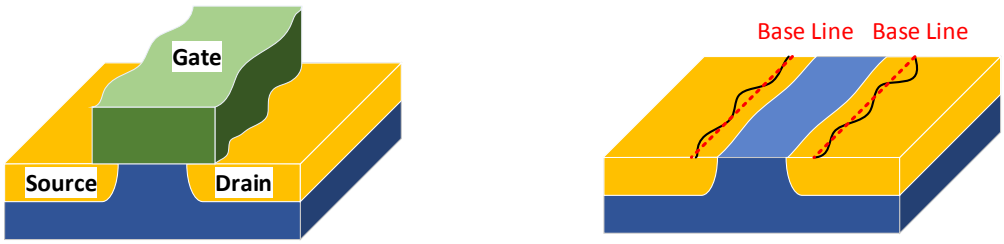


Figure 2.10.: Line-edge Roughness (LER) definition in the transistor

2.3.2. Process variation in emerging technologies

The aim of new emerging devices such as FinFET technology is to decrease short channel effect as mentioned previously in this section. However, they still suffer from some sources of variability. RDF is a major source of threshold voltage variation also for FinFET technology. This is due to the fact the threshold voltage of FinFET devices has a stronger linear dependence on the doping density compared to the conventional MOSFET devices [36]. The other source of threshold voltage variation in this technology is the thickness variation of the silicon fin [36, 42].

2.3.3. Process variation modeling

Since RDF is the major source of variation in advanced MOSFET and FinFET technologies [38, 41] and it significantly impacts the threshold voltage of the transistor, process variation in this thesis is considered as the variation of the threshold voltage. A Gaussian (Normal) distribution is considered for the threshold voltage shift of transistors which has a mean value equal to zero and the standard deviation is obtained using Pelgrom model [36, 38, 43]:

$$\mu_{\Delta Vth}^{PV} = 0 \tag{2.4}$$

$$\sigma_{\Delta Vth}^{PV} = \frac{A_{\Delta Vth}}{\sqrt{WL}} \tag{2.5}$$

where $A_{\Delta Vth}$ is a technology dependent parameter, W is the width and L is the length of the transistor.

2.4. Transistor aging

Transistor aging is one of the major sources of reliability issues in current technologies. The transistor switching delay is degraded over time due to the transistor aging which can eventually cause the circuit to fail if the timing constraint is not met. In this thesis, the focus is on the two major sources of transistor aging which are *Bias Temperature Instability* (BTI) and *Hot Carrier Injection* (HCI). The physical mechanism and modeling of these two effects will be described in more details in the following sections.

2.4.1. Bias Temperature Instability (BTI)

BTI is a wearout phenomenon which gradually degrades the switching delay of the transistor and as a result the circuit delay is degraded over time due to this phenomenon. BTI consists of two similar phenomena: i) *Negative BTI* (NBTI) affecting PMOS transistors and ii) *Positive BTI* (PBTI) affecting NMOS transistors. NBTI was considered as an important reliability issue

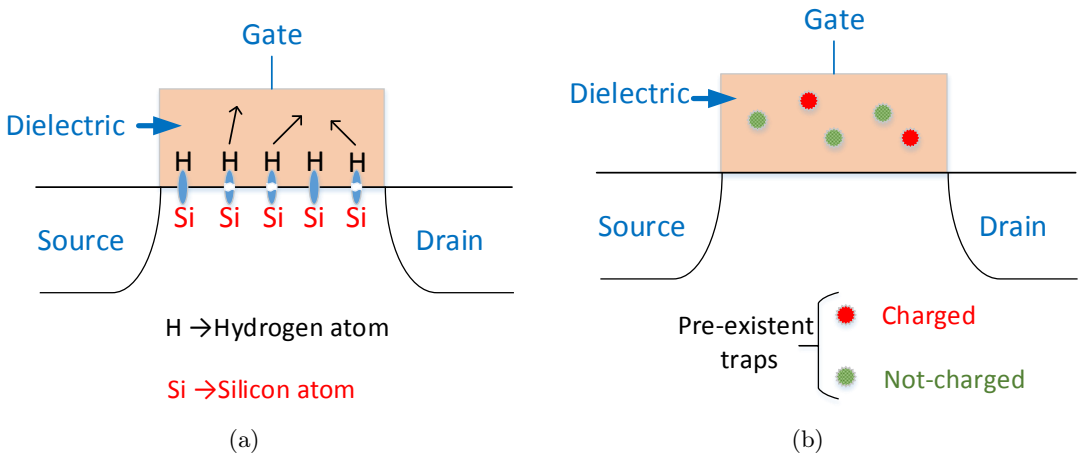


Figure 2.11.: BTI mechanisms: a) Reaction-Diffusion (R-D) mechanism b) Trapping-Detrapping (TD) mechanism

for a long time and PBTI was neglected due its small effect on NMOS transistors, however, by the introduction of high-κ metal-gate technologies, PBTI becomes comparable to NBTI [44, 45].

In general, there are two main models describing this phenomenon: i) *Reaction-Diffusion (RD)* model [46–48] and ii) *Trapping-Detrapping (TD)* model [49, 50]. According to both models, BTI consists of two phases:

- **Stress phase**: the transistor is under NBTI (PBTI) stress if the gate-source of the PMOS (NMOS) transistor is negatively (positively) biased. In other words, the transistor is under stress if it is ON. According to the RD model, in this phase, some of the Si-H bonds at the interface of the channel and the gate oxide are broken leading to the generation of interface traps (Reaction). This reaction is triggered by the carriers in the channel (electrons in NMOS and holes in PMOS). The reaction-generated species (hydrogen atoms or molecules) diffuses inside the gate oxide (diffusion) leading to the generation of traps inside the gate oxide. The generation of these traps at the interface of the channel/gate oxide and inside the gate oxide leads to an increase in the threshold voltage of the transistor. The RD mechanism is shown in Figure 2.11(a). On the other hand, based on the TD model, during the stress phase some pre-existent traps inside the gate oxide captures the charge which leads to an increase in the threshold voltage of the transistor (see Figure 2.11(b)).

- **Recovery phase**: the transistor is in recovery phase if the gate-source bias is removed, i.e. when the transistor is OFF. In this phase, according to the RD model, some of the generated traps are removed since some of the generated hydrogen atoms and molecules diffuses back. According to the TD model, during this phase, some of the traps which captured the charge, re-emit their charge. In general, the threshold voltage of the transistor decreases during the recovery phase, however, it cannot completely compensate the threshold voltage shift due to the stress phase.

There is still a debate about the model which explains the BTI effect better (TD or RD). According to the literature, although the RD model is suitable to accurately predict the stress phase, it fails to cover the recovery phase [49]. It is observed that even after long time stress (1000 sec), threshold voltage drops significantly after 1 second recovery (a very fast recovery) [51]. This fast recovery cannot be explained well by the RD model and it is well explained by TD model [52], however, the RD model is suitable to predict the long term effect of BTI [49].

In previous technology nodes, the BTI effect on transistors was fairly deterministic for a particular workload condition (e.g. temperature and stress) [53]. However, by further down-scaling of the transistor dimensions into deca-nanometer range, the number of defects per device decreases leading to a drastic increase in the time dependent variability of BTI [54]. Thus, it is important to model the stochastic behaviour of BTI in advanced technology nodes. In the following we will explain two BTI models which are used in this thesis in more details. One is a deterministic RD model and the other one is a stochastic atomistic trap-based model.

Deterministic RD model

For the deterministic RD model we exploit the model proposed in [46, 47]. The model is proposed for NBTI effect, but since the mechanism of NBTI and PBTI are the same, we have used similar model to address the PBTI effect.

NBTI can be modeled for two different cases: i) Static NBTI: in which the transistor is under constant stress, and, ii) Dynamic NBTI: in which the transistor alternatively goes to stress (ON) and recovery (OFF) phases. The static NBTI is more severe compared to the dynamic one since the transistor has no time for recovery in the static NBTI (see Figure 2.12(a)). The threshold voltage shift (ΔV_{th}) due to the static NBTI effect can be expressed by:

$$\Delta V_{th}^{static} = A\left((1+\delta)t_{ox} + \sqrt{C(t-t_0)}\right)^{2n} \tag{2.6}$$

$$A = \left(\frac{qt_{ox}}{\epsilon_{ox}}\right)\sqrt[3]{K^2 C_{ox}(V_{GS}-V_{th})\left(exp(\frac{E_{ox}}{E_0})\right)^2} \tag{2.7}$$

where q is the electron charge, E_{ox} is the electric field of the gate oxide, C_{ox} is the oxide capacitance per area and n is a technology dependent factor which is either equal to $1/4$ or $1/6$. The other constants and coefficients are summarized in Table 2.1.

For dynamic NBTI, the ΔV_{th} shift of each stress and recovery phases can be separately expressed by the following equations:

$$Stress: \ \Delta V_{th} = \left(K_v(t-t_0)^1/2 + \sqrt[2n]{\Delta V_{th0}}\right)^{2n} \tag{2.8}$$

$$Recovery: \ \Delta V_{th} = \Delta V_{th0}\left(1 - \frac{2\xi_1 t_e + \sqrt{\xi_2 C(t-t_0)}}{2t_{ox}+\sqrt{Ct}}\right) \tag{2.9}$$

where the constants and coefficients are described in Table 2.1. Equations 2.8 and 2.9 can be exploit to obtain the long term dynamic NBTI-induced V_{th} shift when transistor undergoes alternate stress and recovery phases:

$$\Delta V_{th}^{dynamic} = \left(\frac{\sqrt{K_v^2 \alpha T_{clk}}}{1-\beta_t^{1/2n}}\right)^{2n} \tag{2.10}$$

$$\beta_t = 1 - \frac{2\xi_1 t_e + \sqrt{\xi_2 C(1-\alpha)T_{clk}}}{2t_{ox}+\sqrt{Ct}} \tag{2.11}$$

where T_{clk} is the clock cycle. α in this equation is the *duty cycle* and defined as the ratio of the time in which transistor is under stress to the total time. NBTI-induced ΔV_{th} is a strong function of the duty cycle as shown in Figure 2.12(b).

All the equations and related coefficients and constants are summarized in Table 2.1.

Table 2.1.: RD model of NBTI-induced ΔV_{th}

NBTI-induced ΔV_{th}		
Static	$A\left((1+\delta)t_{ox}+\sqrt{C(t-t_0)}\right)^{2n}$	
Dynamic	Stress	$\left(K_v(t-t_0)^1/2+\sqrt[2n]{\Delta V_{th0}}\right)^{2n}$
	Recovery	$\Delta V_{th0}\left(1-\frac{2\xi_1 t_e+\sqrt{\xi_2 C(t-t_0)}}{2t_{ox}+\sqrt{Ct}}\right)$
	Long-term	$\left(\frac{\sqrt{K_v^2\alpha T_{clk}}}{1-\beta_t^{1/2n}}\right)^{2n}$

Constants and coefficients		
A	$\left(\frac{qt_{ox}}{\epsilon_{ox}}\right)\sqrt[3]{K^2C_{ox}(V_{GS}-V_{th})\left(exp(\frac{E_{ox}}{E_0})\right)^2}$	
K_v	$\left(\frac{qt_{ox}}{\epsilon_{ox}}\right)^3 K^2C_{ox}(V_{GS}-V_{th})\sqrt{C}exp\left(\frac{2E_{ox}}{E_0}\right)$	
E_{ox}	$\frac{V_{GS}-V_{th}}{t_{ox}}$	
C	$T_o^{-1}\cdot exp(-E_a/kT)$	
t_e	if $t-t_0>t_1$	$t_{ox}\sqrt{\frac{t-t_0}{t_1}}-\sqrt{\frac{\xi_2 C(t-t_0)}{2\xi_1}}$
	otherwise	t_{ox}
$E_a(eV)$	0.49	
$E_0(V/nm)$	0.335	
δ	0.5	
$K(s^{-0.25}\cdot C^{-0.5}\cdot nm^{-2})$	8×10^4	
ξ_1	0.9	
ξ_2	0.5	
T_o	10^{-8}	

Stochastic atomistic trap-based model

It is shown that a large portion of the BTI degradation and relaxation during the stress and the recovery phases is due to the charging and discharging of pre-existent gate oxide defects

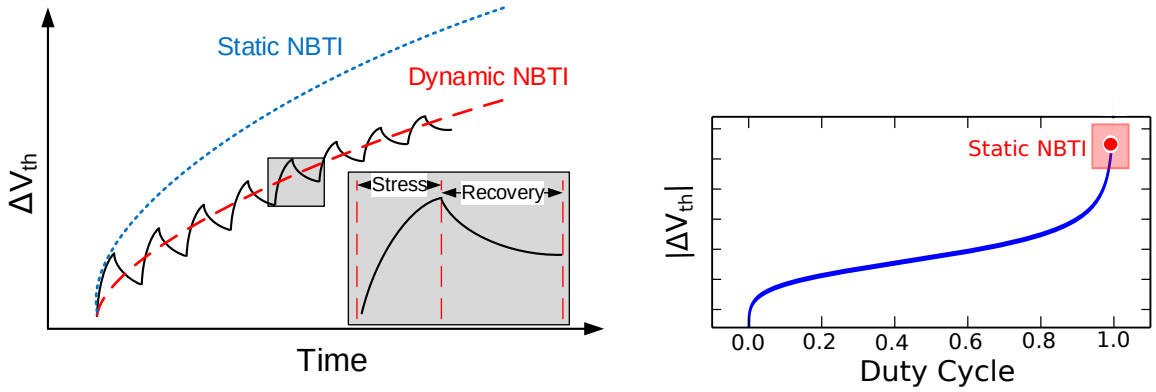


Figure 2.12.: a) Static vs. dynamic NBTI b) the dependency of dynamic NBTI to duty cycle

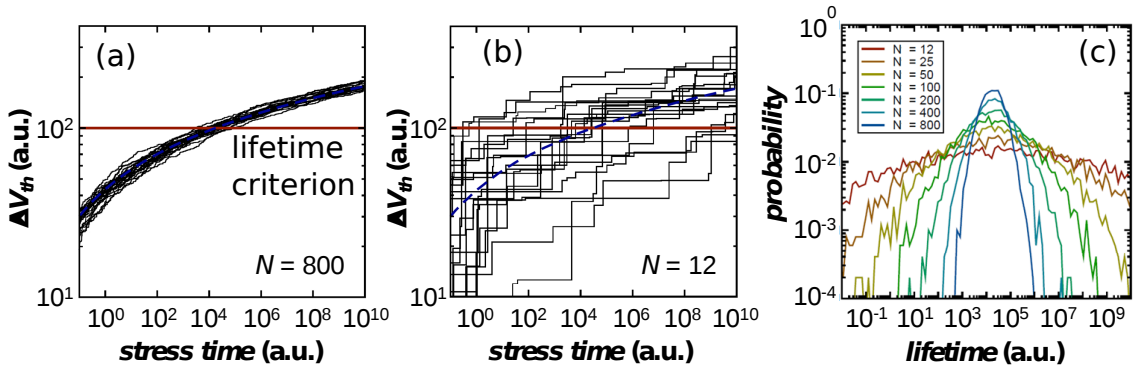


Figure 2.13.: a) BTI effect in large devices b) stochastic behaviour of BTI in deeply-scaled devices and c) lifetime of devices for different technology nodes [20]

[55]. In previous technology nodes, the BTI effect on transistors was fairly deterministic for a particular workload condition (e.g. temperature and stress) due to the large number of defects in the device (see Figure 2.13(a)) . However, by further down-scaling of the transistor dimensions into deca-nanometer range, the number of defects per device decreases leading to a drastic increase in the time dependent variability of BTI [54] (see Figure 2.13(b)). As a result, the lifetime of the device becomes also a *stochastic* value. Figure 2.13(c) shows the lifetime of the device for different technology nodes. As shown in this figure, the lifetime spread of smaller devices with lower number of defects is larger.

Therefore it is important to model the intrinsic variation of BTI. In this thesis, we use the model proposed in [20, 50] for stochastic behaviour of BTI. In this model, each device is characterized by three different factors [50] (see Figure 2.14):

- Number of defects (n)
- Defects capture time (τ_c): it is defined as the time needed to charge a gate oxide defect during the stress phase.
- Defects emission time (τ_c): it is defined as the time needed for the defect to re-emit its charge during the recovery phase.

By knowing these parameters for each device, the total BTI-induced ΔV_{th} of each transistor can be calculated according to Figure 2.14(b). In this model the total number of defects is

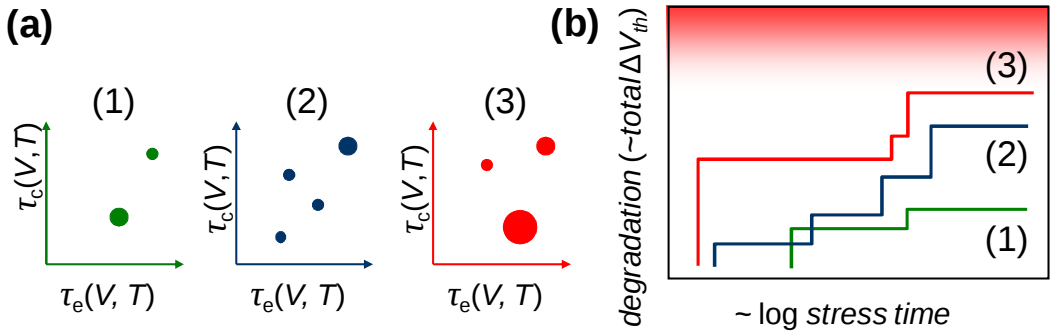


Figure 2.14.: a) Parameters affecting BTI for three different devices b) BTI-induced ΔV_{th} for the three devices [50]

obtained from a Poisson distribution:

$$n = Poiss(N_T) \tag{2.12}$$

$$N_T \propto (L \cdot W) \tag{2.13}$$

where N_T is the mean number of charged (occupied) defects (traps). L and W are the length and width of the transistor. The effect of each occupied trap is obtained from an exponential distribution:

$$\Delta Vth_i = Exp(\eta) \tag{2.14}$$

$$\eta \propto 1/(L \cdot W) \tag{2.15}$$

where η is the average impact of individual defect on threshold voltage ($\propto$1/device area). An analytical description has been derived [50] for the total BTI ΔVth cumulative distribution function as:

$$H_{\eta,N_T}(\Delta Vth) = \sum_{n=0}^{\infty} \frac{e^{-N_T} N_T^n}{n!} [1 - \frac{n}{n!} \Gamma(n, \frac{\Delta Vth}{\eta})] \tag{2.16}$$

This formulation allows for an elegant parametrization of the distribution using the average number of defect N_T and the average impact per defect η which further describes the mean and the variance:

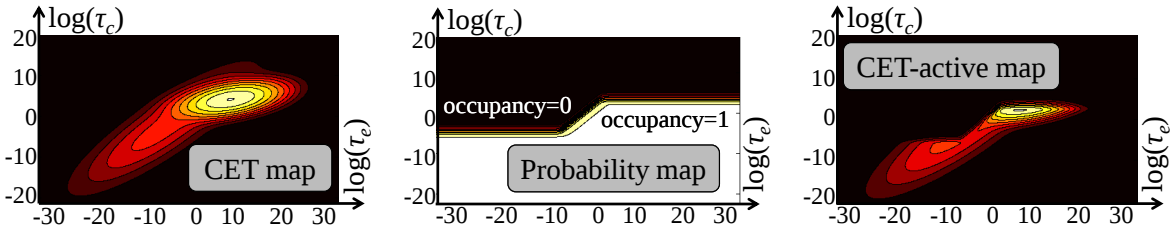
$$\mu_{\Delta Vth} = \langle \Delta Vth \rangle = N_T \eta \tag{2.17}$$

$$\sigma^2_{\Delta Vth} = 2 N_T \eta^2 \tag{2.18}$$

The average impact per defect η can be extracted from experiments [56]. The average number of defect N_T can be calculated using capture/emission time (CET) maps. CET map describes the probability density function of a broadly distributed defect capture and emission times and it is obtained from experimental data [57, 58] (see Figure 2.15(a)). To build the complete CET-map, an analytical 2-component bivariate log-normal mixture distribution is used with a probability density of $f_{CET}(\tau_c, \tau_e)$. By integrating the CET map over the entire time domain the total defect density (n_T) and the mean number of available traps in each device (N_T^{avv}) can be calculated as follows:

$$n_T = \int \int f_{CET}(\tau_c, \tau_e) d\tau_c d\tau_e \tag{2.19}$$

$$N_T^{avv} = W \cdot L \cdot n_T \tag{2.20}$$



(a) (b) (c)

Figure 2.15.: a) CET map b) occupancy probability map c) CET-active map [54]

All of these available traps do not contribute on the total BTI-induced V_{th} shift but those which are charged (occupied). The occupancy probability of each trap (P_{occ}) depends on the applied stress waveform (see Figure 2.15(b)) and can be extracted by the following equation:

$$P_{occ} = \frac{1 - e^{-\frac{\alpha}{f\tau_c}}}{1 - e^{-\frac{1}{f}\left(\frac{\alpha}{\tau_c}+\frac{1-\alpha}{\tau_e}\right)}} \left(1 - e^{-t_{stress}\left(\frac{\alpha}{\tau_c}+\frac{1-\alpha}{\tau_e}\right)}\right) \tag{2.21}$$

where α is the duty cycle (the ratio between the stress time to the total time), f is the frequency, and t_{stress} is the total time. Using this occupancy probability (P_{occ}), the CET-active map can be obtained which shows the distribution of active traps (charged defects) according to the corresponding stress waveform (see Figure 2.15(c)). By integrating the CET-active map over the entire time domain, the average number of defects (N_T) can be obtained by the following equations:

$$\rho = \frac{\int\int f_{CET}(\tau_c, \tau_e) P_{occ}(\tau_c, \tau_e, \alpha, t_{stress}, f) d\tau_c d\tau_e}{\int\int f_{CET}(\tau_c, \tau_e) d\tau_c d\tau_e} \tag{2.22}$$

$$N_T = \rho \cdot N_T^{avv} \tag{2.23}$$

where N_T is the average number of defects as a result of the applied stress waveform. This parameter is used in Equation 2.16 to obtain the CDF of BTI-induced ΔV_{th}.

Process variation and stochastic BTI: are they correlated?

Since both process variation and stochastic BTI can affect the threshold voltage of a transistor, it is important to consider the correlation of these two effects for the calculation of the total threshold voltage shift of the transistor considering both effects. According to [20, 59], there is no correlation between BTI-induced threshold voltage shift and process variation. However, there is a strong correlation between the standard deviation quantities of threshold voltage shift of these two variation sources since identical sources are responsible for process variation and stochastic BTI variability [60]. From measurements, independently of the technology [60], the correlation has been found to follow the empirical relation:

$$\sigma^2_{\Delta Vth}(t) = \frac{\mu_{\Delta Vth}}{B} \sigma^2_{Vth_{PV}} \tag{2.24}$$

$$B = 100mV \tag{2.25}$$

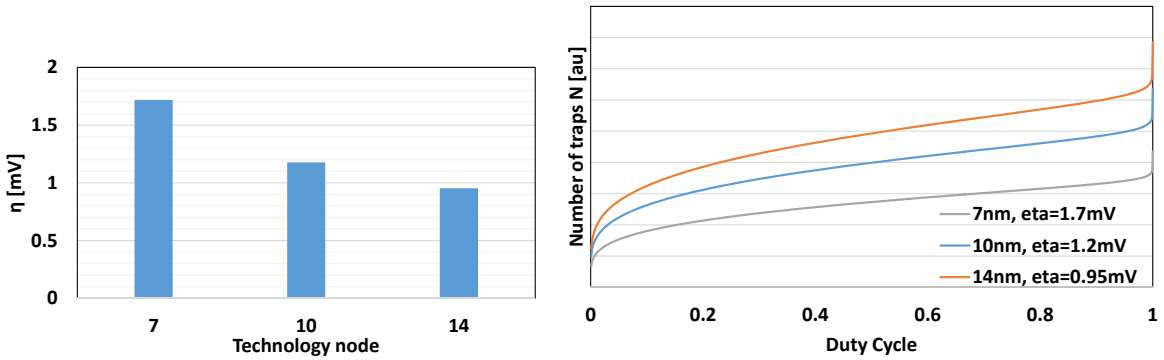
where B is a technology specific parameter. It is important to note that the variances are correlated here, the ΔVth and initial Vth are assumed not to be [20, 59].

Assessing the impact of degradation induced time-dependent variability of the Vth is difficult for future technologies because of the uncertainty on the BTI critical parameters η and N_T .The correlation between process variation and stochastic BTI however gives a powerful predictive method for evaluating existing and future technologies. Combining (2.17) and (2.18) with (2.24), η can be directly derived from the initial process variation:

$$\eta = \frac{1}{2B} \sigma^2_{Vth_{PV}} \tag{2.26}$$

or combining with (2.5)

$$\eta = \frac{A_{\Delta Vth}}{2B\sqrt{WL}} \tag{2.27}$$



(a) (b)

Figure 2.16.: (a) η calculated using (2.27) for different FinFet technologies. (b) Average number of occupied traps as function of DF for different FinFET technologies calculated using (2.21), (2.22) and (2.23)

Thus, for simulating future technologies, η is derived directly from the expected Pelgrom's mismatch parameter $A_{\Delta Vth}$ [61–63] and N_T will be calculated using (2.21), (2.22) and (2.23) with a CET map measured on poly silicon oxynitride (SiON) process technology. The scaling of oxide thickness T_{OX} and stress voltage is incorporated by using a power-law extrapolation for the overdrive electric field E_{OX} (calculated as V_{OV}/T_{INV}.) [64]. Here the Vth degradation is proportional to $(E_{OX})^{\gamma}$, where γ is the voltage acceleration which has a typical value of 3 [65]. Assuming there are no changes in the oxide or oxide quality the extrapolation towards more scaled nodes is done using the following relationship

$$\frac{\langle \Delta Vth_{ref} \rangle}{(E_{OX,ref})^{\gamma}} = \frac{\langle \Delta Vth_{sim} \rangle}{(E_{OX,sim})^{\gamma}} \tag{2.28}$$

As shown in Figure 2.16, values for η and N_T can be readily obtained when using the methodology described above.

2.4.2. Hot Carrier Injection (HCI)

"Hot" carriers are referred to carriers which have a temperature much higher than the lattice temperature. When the transistor is in saturation mode, some of the carriers become "hot" due to the high lateral field and they gain enough energy to overcome channel/gate oxide potential barrier (channel hot carriers) [66]. These channel hot carriers may collide with the silicon atoms in the pinch-off region and generate electron-hole pairs due to the impact ionization. Some of the generated carriers may become "hot" and overcome channel/gate oxide potential barriers [66]. The second type of hot carriers are called avalanche hot carriers.

Both channel and avalanche hot carriers may be injected in the gate oxide and damage it generating traps inside the gate oxide. The gate oxide damage degrades the device characteristic such as the drain current and specially the threshold voltage of the transistor. This phenomenon is called Hot Carrier Injection (HCI) or *Channel Hot Carrier (CHC)* which is an important transistor aging issue in nanometer-technology nodes. The physical mechanism of HCI effect is depicted in Figure 2.17.

HCI issue is observed as a critical issue in Eighties [66] due to the high lateral electric filed in the technologies used in these period of time. However, from the mid-Nineties, the supply voltage started to decrease by the technology scaling to decrease the power consumption issue [67]. As a result the lateral electric field decreased and hence, the HCI effect became less by

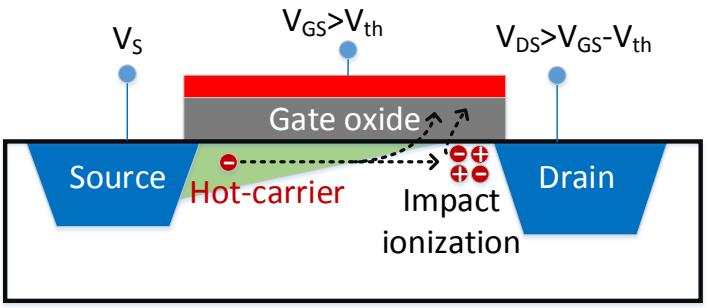


Figure 2.17.: Hot Carrier Injection (HCI) physical mechanism

technology scaling. This trend has stopped in recent technology nodes, due to the fact that the supply voltage scaling is slowing down or stopping due to various reasons such as non-scalability of the threshold voltage and the subthreshold slope, signal-to-noise margin issue and process variation. Therefore, the lateral electric field started to increase and hence HCI again has become an important transistor aging issue [67].

HCI mainly affects NMOS transistors and its effect is negligible in PMOS transistors [68] since in the PMOS transistors fewer hot-carriers are generated. The reason of this is twofold: i) shorter mean free path of the holes and ii) higher oxide barriers for holes.

HCI model

In this section, we explain the HCI model which is used in this thesis. As mentioned previously, the device characteristic such as threshold voltage and subthreshold slope is degraded due to the HCI effect. Here, the model of transistor V_{th} shift as the main effect of HCI is explained (see Figure 2.18). Hot-carriers are generated during logic transition and hence the HCI induced V_{th} degradation is a function of switching frequency [68, 69]:

$$\Delta V_{th} = A_{HCI} \times SW \times f \times e^{\frac{E_{ox}}{E_1}} \times t^{0.5} \tag{2.29}$$

$$E_{ox} = \frac{V_{GS} - V_{th}}{t_{ox}} \tag{2.30}$$

where A_{HCI} is a technology dependent constant, SW is the switching activity factor, and f is the clock frequency. V_{th} and V_{GS} are the threshold voltage and the gate-source voltage of the transistor, respectively. t_{ox} is the oxide thickness, E_1 is a constant equal to $0.8V/nm$ [70] and t is the total time.

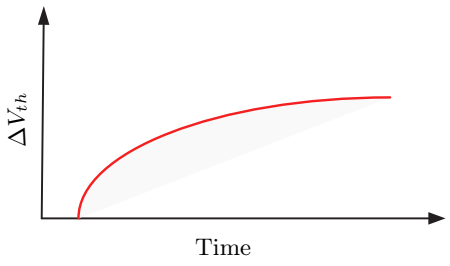


Figure 2.18.: HCI-induced ΔV_{th} over time

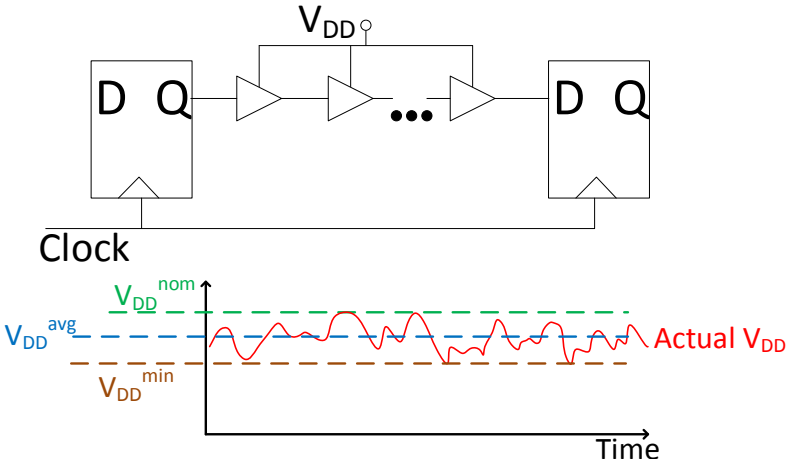


Figure 2.19.: The supply voltage seen by the gates inside the circuit

Moreover, it is shown that HCI effect depends on the temperature [67, 71]. Therefore, the HCI model of Equation 2.29 is modified as follows:

$$\Delta V_{th} = A_{HCI} \times SW \times f \times e^{\frac{-E_a}{kT}} \times e^{\frac{E_{ox}}{E_1}} \times t^{0.5} \tag{2.31}$$

where k is the Boltzmann constant and E_a the activation energy for the charge injection into the gate oxide.

2.5. Voltage droop

During workload execution several nodes switch between 0 and 1 and therefore they draw current from power grids. Due to the current drawn from the power grids, the actual supply voltage seen by individual gates inside the circuit decreases and it could vary from time to time and from gate to gate. This phenomenon is called *voltage droop* which is a strong function of the executed workload (see Figure 2.19). The voltage droop causes the delay and power dissipation to change and in an extreme case, it may even lead to a functional failure.

The effect of voltage droop on the delay of a simple inverter in 45 nm technology node is shown in Figure 2.20. As shown in this figure, a 10% voltage droop can cause the gate delay to increase by more than 20%.

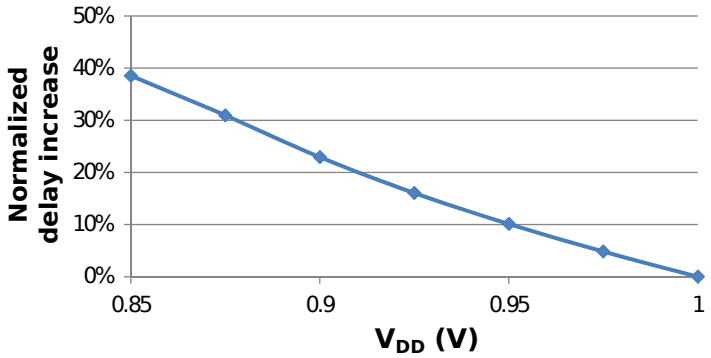


Figure 2.20.: The delay deviation of a simple inverter versus different supply voltage values in a 45 nm technology node

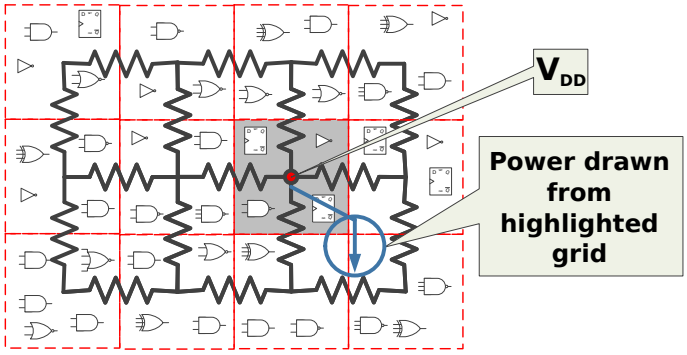


Figure 2.21.: Equivalent R network model of power grid [73]

The voltage droop increases by technology scaling since the frequency as well as power densities is increasing [72]. Moreover, by the technology scaling, the sensitivity of the circuit performance to the voltage droop increases and noise margin decreases since the threshold voltage of transistors does not scale down as fast as supply voltage. As a result, the circuit tolerance to the voltage droop is decreasing as the technology geometry scales down [72].

2.5.1. Voltage droop metrics and important parameters

There are two important metrics for voltage droop as shown in Figure 2.19:

- **Average voltage droop**: which corresponds to the average supply voltage seen by the gates (V_{DD}^{avg}). This value correlates with the voltage droop-induced delay degradation of the circuit. It is shown that the effect of voltage droop on timing of a digital path is equal to applying V_{DD}^{avg} to the gates of the same path [72]. As a result, this metric needs to be considered to set the timing margin of the design.
- **Maximum voltage droop**: which corresponds to the minimum supply voltage seen by the gates (V_{DD}^{min}). This may cause a failure in the behaviour of the gates or memory cells.

Moreover, there are two components contributing in the total amount of voltage droop:

1. ***IR drop***: which is proportional to the level of the current. R represents the resistances of power mesh network, power pads and device package [72].
2. ***Ldi/dt***: which is proportional to the change rate of the current. L represents the inductances of the power mesh network, power pads and device packages [72].

It is shown that the contribution of Ldi/dt is less than that of IR drop [72]. Moreover, Ldi/dt only affects the *maximum voltage droop* and its effect on the *average voltage droop* is negligible [72]. As a result, the contribution of Ldi/dt on the voltage droop-induced timing degradation of the circuit is small and this parameter can be neglected in the modeling.

2.5.2. Voltage droop model

In this thesis, the power grid is modeled as an R network distributed over the die as shown in Figure 2.21. Moreover, for DC analysis, the package model is reduced to a per-connection parasitic resistance [73]. The relationship between voltage (V) and current (I) drawn from each node in the power grid can be written as [74]:

$$V = G^{-1}I \tag{2.32}$$

where G represents the conductance matrix of the power grid. As shown in Figure 2.21, current (I) of a grid is calculated by adding the leakage and dynamic power of the gates. The leakage current of each gate is obtained from leakage LUT according to its load capacitance. The switching activity of each node together with parasitic capacitance information obtained from floor-plan are used to estimate the dynamic power of each gate inside the netlist according to the following equation:

$$power_{dyn} \sim SW \cdot c_l \cdot f \cdot V_{DD}^2 \tag{2.33}$$

where SW, c_l, f, and V_{DD} are the switching activity, load capacitance of the output node, frequency, and the gate supply voltage, respectively.

There is a negative feedback between voltage droop and power consumption of the circuit such that higher power consumption results in a higher voltage droop which in turn reduces the power consumption. Neglecting this effect might result in a considerable inaccuracy in the estimated voltage droop value. To capture this interdependence between the power consumption and the voltage droop, we have used an iterative approach in a way that the extracted voltage droop is used to update the leakage and dynamic power of each cell and in turn the power profile. This power profile again is used to adjust the voltage droop calculation. This loop continues until a convergence is reached.

2.6. Soft error

Soft error is a result of the interaction of particles, such as neutron, alpha radiation-induced particles or proton, with device material leading to a perturbation in the device operation. The perturbation can manifest itself as a bit flip in memory cells or a transient fault in combinational parts of the circuit. This type of error is called "soft" since the device is not permanently damaged and if a new data comes, the device operates correctly again.

Soft error may affect the memory cells and the sequential elements of the circuit by a bit flip. If only one cell is affected by a single particle, the event is called *Single-Event Upset (SEU)*. However, a single energetic particle strike can result in upsets in multiple circuit nodes which is called *Multiple Bit Upset (MBU)*. By the technology scaling, the dimensions become smaller and devices become closer and hence the MBU rate is increasing [9].

Soft error may also affect the combinational logics. In this case, it manifests itself as a transient pulse (a temporal change in the voltage value of the signal). If the wrong value is stored in the sequential elements, the transient pulse leads to an error. However, the wrong value does not necessarily reach to the sequential element and it may be masked due to different masking phenomena. According to [75], more than 90% of the faults are masked and they do not cause an error. In the following, the most important types of masking are explained in more details:

- **Electrical masking**: If the transient pulse propagates through successive gates, it may be attenuated such that it cannot propagate more and it is not latched by the sequential element. This phenomenon is called electrical masking and it is a strong function of gates delay. The electrical masking is shown in Figure 2.22.

- **Logical masking**: The transient pulse induced by radiation is logically masked if it does not affect the output of the logic due to the functionality of the logical gate. For example if a transient pulse occurs at one input of an AND gate and the other input is logically "0" the output will remains unchanged to value of "0". This phenomenon is depicted in the Figure 2.23(a).

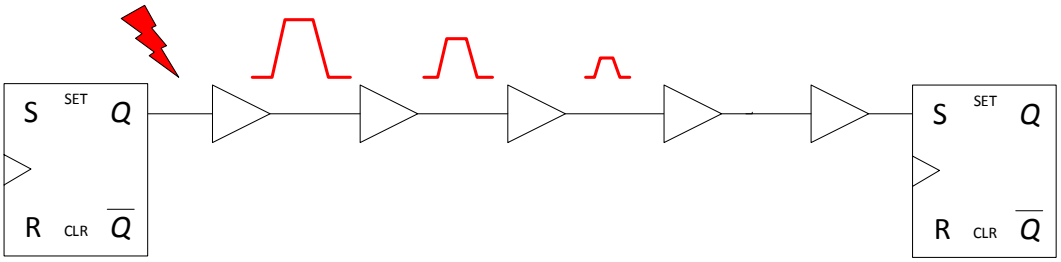


Figure 2.22.: Electrical masking in logical gates

- **Latching window masking**: This type of masking occurs when the transient pulse reaches to the sequential element outside of its latching window. Figure 2.23(b) demonstrates the latching window masking.

2.6.1. Sources of radiation

Radiation at ground level causing soft errors comes from different sources. In general, there are two major types of radiation sources:

1) **Atmospheric radiations**: When a primary cosmic ray (e.g. protons, electrons, photons) enters the atmosphere, it interacts with the molecules of the air leading to the generation of high energy secondary particles (e.g. neutron, hadrons). The neutron is one of the most important ground level radiation sources affecting circuits, leading to the generation of soft errors. Neutrons are not charged and as a result their interactions with materials do not directly create electron-hole pairs. However, their interactions with material lead to the creation of secondary ionizing particles via "indirect ionization" mechanism. The interaction of generated secondary particles and material in turn leads to the generation of electron-hole pairs. Direct ionization from low energy protons is another source of soft errors which has become important for technologies beyond 65 nm [77–79]. Proton has a positive electric charge and its mass is slightly lower than neutron.

Muons are also another important part of the atmospheric radiation at ground level. It is a particle similar to electron with a negative charge but with a much greater mass [80]. The probability of muons interaction with material is very small and the interaction happens only for low energy muons. Like the other charged particles, muons also cause a direct ionization in material. Pions are the other source of atmospheric radiations at ground level. Although they strongly interact with material, their flux density at ground level is very low. Figure 2.24(a) shows the spectrum of atmospheric radiations at ground level [81].

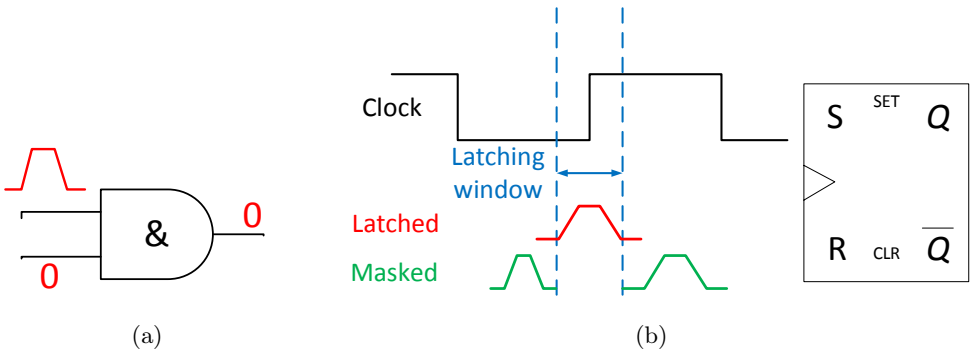


Figure 2.23.: a) Logical masking and b) Latching window masking [76]

2) **Terrestrial radiations**: Alpha-particles are the only terrestrial radiations which cause soft errors in current technologies. An alpha-particle consists of two protons and two neutrons (identical to the helium nuclei). ^{238}U, ^{235}U, and ^{232}Th are the main sources which emit alpha-particles with an energy range of less than 10 MeV (see Figure 2.24(b)). In this thesis, it is assumed that the overall alpha particle emission rate is equal to 0.001 $\alpha/h\ cm^2$ [84].

2.6.2. Basic physical mechanism of soft error

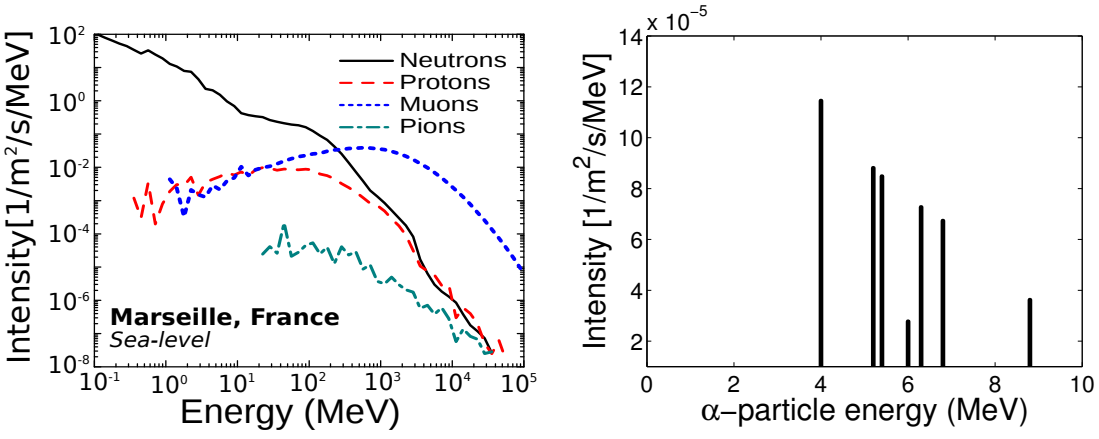
In this section, we briefly describe the physical mechanism of soft error. As mentioned before, there are different types of particles affecting the VLSI devices. These particles can directly lead to ionization of materials if the particle is charged (such as protons and alpha particle). In case of neutron, since the particle is neutral, it cannot directly deposit charge in the material. However, the interaction of neutron and material can lead to the generation of charged particles (secondary particles) which in turn leads to the ionization of the material. In the following, the physical mechanism of soft error caused by charged particles is described.

In general, the passage of a charged particle through the device material can be explained in three different steps [81] which are shown in Figure 2.25:

- **Charge deposition**: When a charged particle strikes the device, it transmits a large amount of energy to the materials due to mainly inelastic interaction [81]. The deposited energy leads to a generation of electron-hole pairs along the particle path (see Figure 2.25(a)). The energy deposited in the material depends on the particle energy and the material. Moreover, the energy needed to generate the electron-hole pairs depends on the material band-gap. For example, the energy required for generation each electron-hole pair in Silicon material is about 3.6 eV. Putting all together, the number of generated electron-hole pairs strongly depends on the particle energy and the struck material.

- **Charge transport**: when the electron-hole pairs are generated due the interaction of particle and material, the generated carriers are transported due to two main mechanisms (see Figures 2.25(b) and 2.25(c)):

  1. Drift: If the generated carriers are in the regions with an electric filed (e.g. channel), the carriers are transported due to the drift mechanism.



(a) (b)

Figure 2.24.: a) Neutron, proton, muons, and pions spectrum at sea level [81, 82], b) alpha-particle spectrum assuming that the emission rate is equal to **0.001** $\boldsymbol{\alpha/h\ cm^2}$ [83]

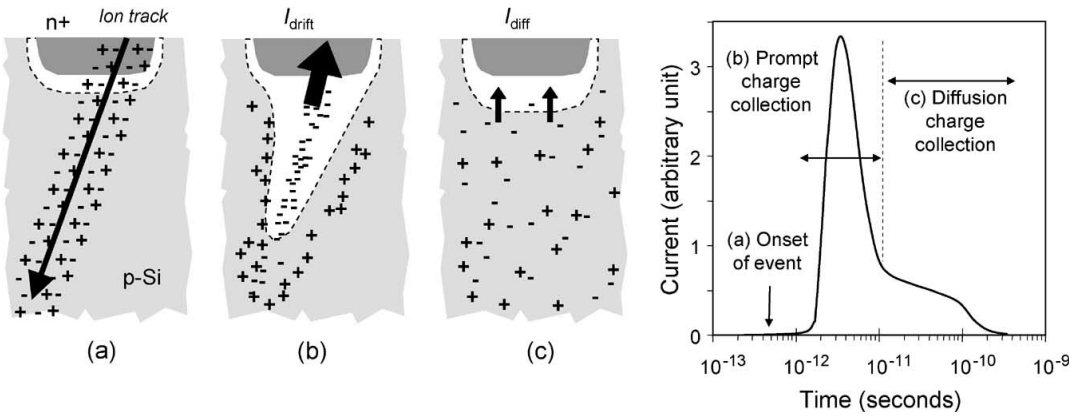


Figure 2.25.: Physical mechanism of soft error caused by the passage of a charged particle [84]

2. Diffusion: If the generated carriers are in the neutral regions, the carriers are transported due to diffusion mechanism from the places with high density of carriers towards the region with less density.

- **Charge collection**: The deposited charge can be collected by the sensitive regions (reversely biased p-n junction due to their strong electric field) and generates a transient current. The generated transient current can lead to a bit-flip in case an SRAM cell or a latch or flip-flop is affected by radiation. In the case of logical gates, the generated transient fault can be propagated through the gates and if it is latched with sequential element, it leads to an error.

2.7. Summary

In summary, in this chapter, a basic necessary information of CMOS as the main technology being used for constructing integrated circuit was provided. For this purpose, first, the transistor structure and its functionality was discussed followed by some basic information about the structure of CMOS logic gates and their properties such as delay. Moreover, we discussed the most important reliability issues in current technology nodes and their corresponding models.

Part II.

Cross-Layer Modeling and Prediction of Reliability Issues

CHAPTER
THREE

CHIP-LEVEL MODELING AND ANALYSIS OF ELECTRICAL MASKING OF SOFT ERRORS

3.1. Overview

With continuous downscaling of VLSI technologies, logic cells are becoming more susceptible to radiation-induced soft error. When radiations affect the logic cells, a transient pulse is generated which may eventually lead to an error if it is propagated and a wrong value is stored in the sequential element. However, the generated transient pulse does not necessarily lead to an error and may be masked due to different masking phenomena such as electrical masking, logical masking and latching window masking. To accurately model soft errors at chip-level, the impact of electrical masking should be accurately considered. Moreover, increasing complexity of VLSI chips at nanoscale results in voltage droop across the chip which impacts the electrical masking. In this chapter, a chip-level electrical masking analysis is presented which accurately considers the impact of voltage droop across the chip. Moreover, a technique based on backward pulse propagation to reduce the runtime of this analysis is proposed.

The rest of the chapter is organized as follows: First, a short introduction of the problem and motivation is presented and the contribution of this work is explained in Section 3.2. Afterwards, previous work is reviewed in Section 3.3. The overall flow of voltage droop-aware SER estimation at chip-level is presented in Section 3.4. The proposed pulse propagation model with consideration of voltage droop is explained in Section 3.5. The backward propagation technique as well as SER estimation algorithm are presented in Section 3.6. Experimental results are presented in Section 3.7. Finally, Section 3.8 concludes the chapter and summarizes the achievements.

3.2. Introduction, motivations and contributions

Aggressive technology scaling of nanoscale VLSI circuits results in a reduction of capacitance per transistor, and as a consequence particles with lower energy, which are far more plentiful, can generate sufficient charge to cause soft errors [76, 85]. Recent experiments reveal that the *SER* of combinational logic in sub-50nm technologies is comparable with that of sequential elements (i.e. latches and flip-flops) and will be dominant factor in the future technologies [86, 87]. Electrical masking is an important phenomenon affecting SER in combinational logic.

The electrical masking strongly depends on the electrical properties of the logic gates and it might be significantly affected by voltage droop. Since voltage droop in recent technologies grows due to increasing current density, its effect and its correlation among gates have to be accurately considered in the electrical masking analysis.

In this chapter, a fast and accurate electrical masking model is proposed for soft error rate estimation at the chip-level with considering the impact of voltage droop. The main contributions can be summarized as following:

- For the first time the effect of correlated voltage droop on the propagation of erroneous transient pulse is considered. For this purpose, a *lookup table (LUT)* based model is proposed. The main advantage of the proposed model is its high accuracy in comparison with the transistor-level SPICE simulation while achieving significant runtime speed-up, providing the required scalability for chip-level analysis of large designs.

- Additionally, a backward propagation technique is proposed that can expedite the SER estimation process for combined analysis of various masking effects. Using this technique, minimum effective pulse width for each node of the circuit is computed and unnecessary propagations of smaller pulses are skipped.

3.3. Related work on electrical masking

Electrical masking is one of the three important masking factors in the combinational circuits. Inaccurate modeling of electrical masking might lead to a remarkable error in the SER estimation results. In this regard, various techniques [88–94] have been presented to model the effect of electrical masking in combinational logic.

The technique presented in [88] is based on the linear RC model of logic gates and the output pulse is computed according to close-formed equations. The equivalent resistances for pull-down and pull-up networks are calculated based on width/length ratio and series/parallel connectivity of transistors. This technique results in a considerable inaccuracy as it cannot handle the non-linear characteristics of the gates for transient pulses.

The pulse propagation equation presented in [89] discretizes the time into small time steps and for each time step the output magnitude is computed based on the previous step output magnitude and pull-up and pull-down currents of the gate. Although this model is very accurate, it is not scalable due to the large number of required time steps.

A parametric waveform model based on the Weibull function is presented in [90]. This technique can accurately model the output pulse when pulse magnitude is less than V_{DD}, however, in case of trapezoidal pulses with full magnitude, it is inaccurate.

In the trapezoidal approximation models [91–94], a transient pulse is modeled by its important parameters such as width, magnitude and slope. These techniques use either empirical equations [91–93] or LUTs generated based on accurate cell characterizations [94]. The method in [91] uses an equation based on linear approximation for propagation of the pulse width and ignores all the other important parameters including magnitude and slope. Another equation-based technique is presented in [92] where propagated pulse magnitude and width is estimated based on the input pulse magnitude and width using fitting parameters. [93] also introduces an equation-based method in which the magnitude and the width of the output pulse is accurately modeled when the input pulse magnitude is equal to V_{DD}. However, it fails to accurately model attenuated pulses where the pulse magnitude is less than V_{DD}.

In this chapter, an LUT-based trapezoidal approximation model is proposed which propagates all three important parameters (i.e. width, magnitude and slope) with a high accuracy. Moreover, the proposed model considers the effect of voltage droop which is completely neglected in all previous study.

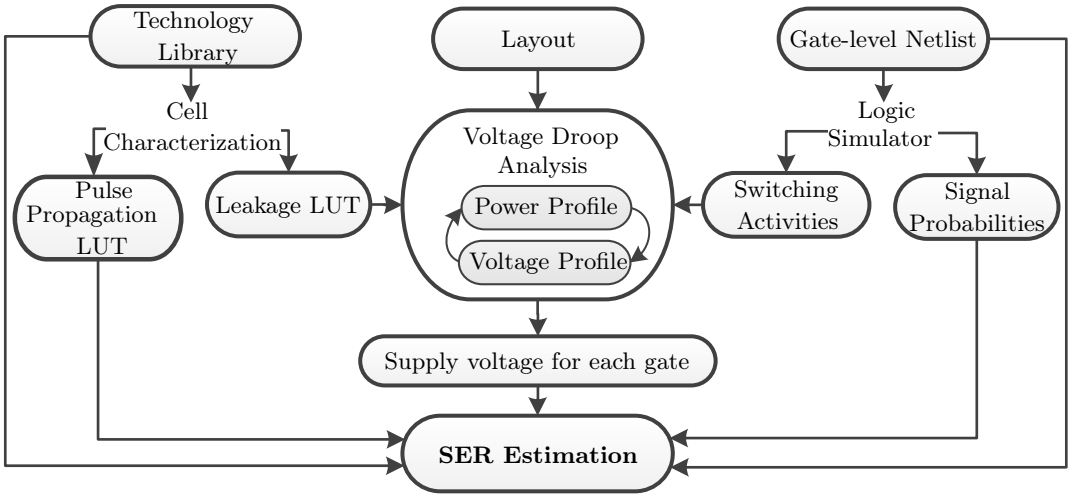


Figure 3.1.: The overall flow of the proposed SER estimation considering voltage droop

3.4. Overall flow

The overall flow of the proposed method for voltage droop-aware SER estimation is presented in this section and then different parts of the flow will be explained in more details in the next sections. Figure 3.1 shows the overall flow of the proposed SER estimation method. First, the standard cells in the library are characterized using accurate SPICE simulations to find the leakage current of each standard cell and *leakage LUT* is constructed. The gate-level netlist of the circuit is given to a logic simulator to obtain the switching activity and signal probability of the internal nodes. The leakage LUT and switching activity in addition to the layout information are given to the voltage droop analysis part in which the layout is divided into some rectangular grids and voltage and power profile of each grid are estimated using an iterative process and after convergence, supply voltage of each grid is extracted. This process is explained in more details in Section 2.5.2.

In order to consider the effect of voltage droop in electrical masking, a pulse propagation model is necessary. To address this issue, in Section 3.5, an LUT-based model is proposed to consider the effect of the voltage droop during pulse propagation through the gates. Finally, the signal probability of the internal nodes and the supply voltage of each gate are given to the SER estimation tool which is described in Section 3.6. In this tool, SER is calculated by considering the effect of the voltage droop on electrical masking while logical masking and latching-window masking factors are also taken into account.

3.5. Pulse propagation modeling

In this section, an accurate LUT-based model is proposed for the transient pulse propagation considering the impact of voltage droop. In the proposed method, the transient pulse is modeled using a trapezoidal waveform as shown in Figure 3.2.

When a transient pulse propagates through a gate, the characteristic of the output pulse is a function of the input pulse width (pw) and the gate rise/fall delay [91–94]. The rise/fall delay of the gate is in turn a function of the gate load capacitance (c_l) and the input pulse rise/fall transition times (t_r/t_f) (see Section 2.1.2 for definitions and more details). Here, instead of rise and fall transition time (t_r/t_f), we use the terms of rise and fall transition slope (s_r/s_f), respectively (see Figure 3.2).

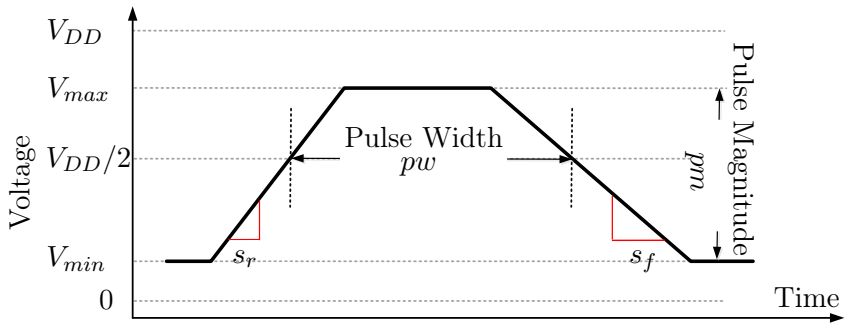


Figure 3.2.: Important parameters of trapezoidal model for transient pulse

In addition to s_r, s_f, pw, and c_l, there are other parameters which also affect the characteristic of the output pulse. These parameters and their effects are explained with the help of Figure 3.3 for a case in which a low-high-low input pulse propagates through an inverting gate (e.g. inverter), however, the similar effects can be observed for other cases.

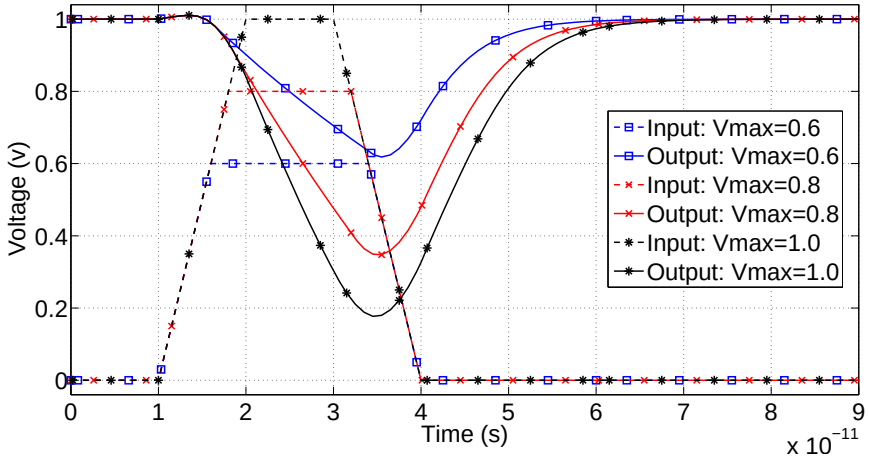
- **Input pulse maximum voltage level (V_{max})**: Since V_{max} of the input pulse affects the drive strength of the gate pull-down network, the gate fall delay and consequently the output pulse characteristic is a function of input pulse V_{max}. Figure 3.3(a) shows the propagation of different input pulses with the same s_r, s_f, and pw but different V_{max} through a simple inverter. As shown in this figure, as V_{max} of the input pulse decreases the fall transition time/delay of the gate increases which eventually results in different output pulses.

- **Input pulse minimum voltage level (V_{min})**: As shown in Figure 3.3(b), different input pulses with the same s_r, s_f and pw but different V_{min} lead to different output pulses. This is due to the fact that different V_{min} result in different drive strengths in the pull-up network during the rise transition which eventually leads to different output pulses.

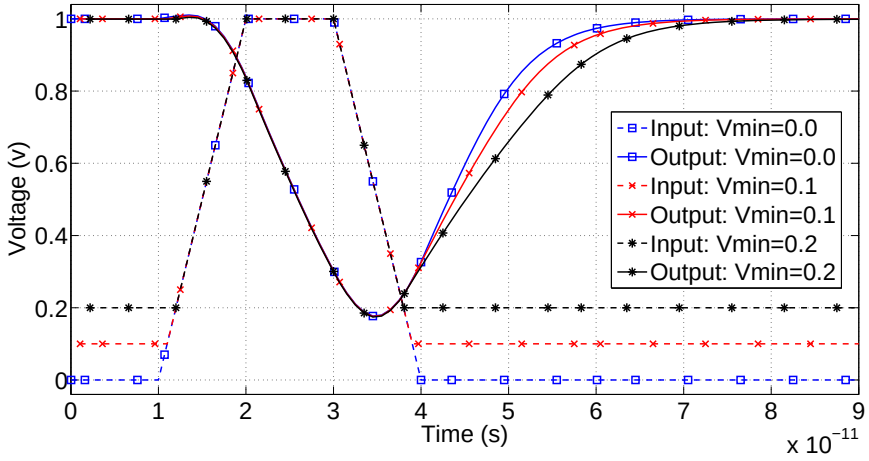
- **Gate supply voltage (V_{DD}^g)**: Supply voltage of the gate not only affects the gate delay, but also affects V_{max} of the output pulse (see Figure 3.3(c)). Since the supply voltage seen by each gate is different due to voltage droop, the effect of V_{DD}^g also has to be considered during transient pulse propagation.

Based on these observations, an accurate LUT-based method is proposed which considers all important factors. In this method, for each standard cell in the library, the cell is characterized and an LUT is built to store the output pulse parameters for different input pulses. In order to characterize a cell, accurate SPICE simulations are performed and all important parameters (pw, V_{min}, V_{max}, s_r, s_f, c_l, and V_{DD}^g) are swept to find the important parameters of the output pulse (pw, V_{min}, V_{max}, s_r, s_f) for each case. The rise and fall delays of the gates (d_r and d_f) are also obtained and stored for each case which can be used in the backward propagation technique explained in Section 3.6.1. If the number of samples for each parameter is large enough, this method becomes highly accurate. To find the output pulse parameters of a special input pulse propagated through a gate, a linear interpolation method is used among the closest existing LUT entries.

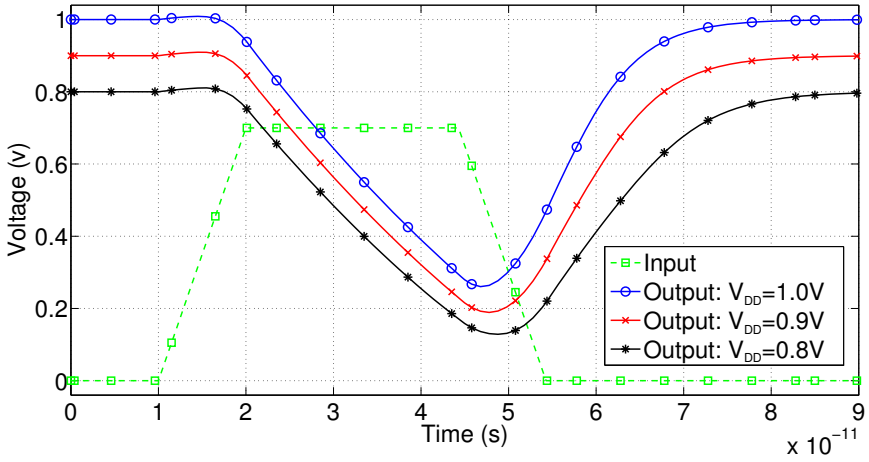
The main advantage of our proposed method is that its accuracy is comparable to the SPICE simulation while it has much less runtime. The characterization process can be accelerated by sampling the parameters more in the critical ranges and less in non-critical ranges. It should be



(a)



(b)



(c)

Figure 3.3.: SPICE simulation results for an inverter; effect of a) input V_{max}, b) input V_{min}, and c) inverter V_{DD}^{g} on output pulse (Input pulses: low-high-low & output pulses: high-low-high)

noted that since the characteristic of the output pulse is a complex function of many parameters, until now, there is no way to represent these LUTs by simple equations. The other advantage of our proposed method over the prior methods is that LUTs could be augmented with other parameters to capture other aspects which can affect the transient pulse propagation such as process variation.

3.6. Pulse propagation and SER estimation

In this section, a method based on backward propagation is presented to find the minimum effective pulse for each node with respect to various masking effects. This method can reduce electrical masking analysis complexity for the overall chip-level SER estimation. We first introduce the proposed backward propagation technique and then explain the overall SER estimation algorithm.

3.6.1. Backward transient pulse propagation

A remarkable fraction of transient pulses are masked in combinational gates and do not propagate to the flip-flops inputs. The key idea behind the backward propagation technique is to find the *Minimum Effective Pulse width* (MEP) for each node of the circuit and avoid unnecessary propagation of transient pulses weaker than MEP. MEP of a node is defined as the transient pulse width which all pulses with lower width are definitely masked and cannot be propagated and latched at the sequential elements. This means, after computation of the MEP using backward analysis, it can be predicted whether a specific transient pulse (generated due to particle strike at this node or propagated from another struck node) is going to be propagated to flip-flops or primary outputs. This can significantly reduce the runtime of SER estimation by skipping the propagation of narrow pulses.

In order to calculate the MEPs, the netlist is first levelized from primary inputs and flip-flop outputs. Then, starting from nodes with highest level (i.e. closest to flip flops and latches), the MEP for each node is calculated according to the following rules:

- If the node only drives a flip-flop, transient pulse should have enough strength to be latched in the flip-flop. A transient pulse is latched in a flip-flop if its width is greater than $t_{setup} + t_{hold}$ [76] where t_{setup} and t_{hold} are the setup time and hold time of the flip-flop, respectively.

- If the node only drives a combinational gate, its MEP is calculated according to MEP of the gate output node using reverse propagation equations. For a Low-High-Low (LHL) pulse propagating from an inverting gate type (e.g. NAND, NOR, and INV), the output pulse width is obtained by the following equation [93]:

$$pw_{out} = pw_{in} - d_r + d_f \tag{3.1}$$

  The equation above can be rewritten as:

$$pw_{in} = pw_{out} - d_f + d_r \tag{3.2}$$

  since the MEP of a node is defined as the transient pulse width which all the pulses with lower width are definitely masked, the equation can be rewritten as:

$$MEP_{in}^{LHL} = MEP_{out} - d_f^{max} + d_r^{min} \tag{3.3}$$

  where d_f^{max} (d_r^{min}) is the largest (smallest) fall (rise) delay of the gate for a special load capacitance and different input rise (fall) delay. d_r^{max} and d_f^{min} are obtained from LUT

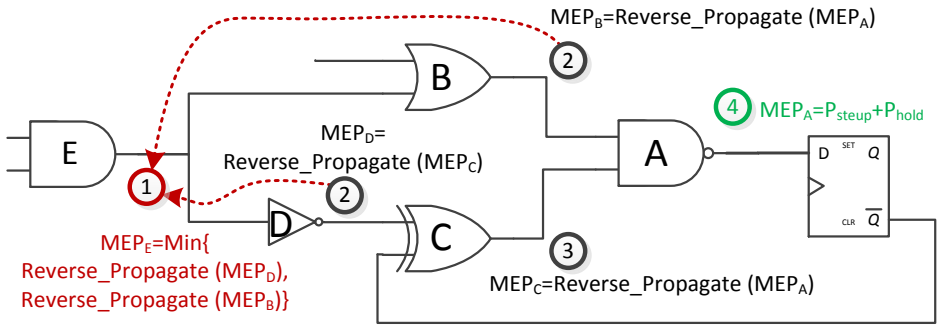


Figure 3.4.: An Example of Backward Propagation of MEPs

which is discussed in Section 3.5. MEP for High-Low-High (HLH) pulse can be obtained similarly and the overall MEP of the node is the minimum of MEP_{in}^{LHL} and MEP_{in}^{HLH}. A similar process is done to obtain the MEP of the nodes which drive the non-inverting gates (e.g. AND, OR).

- For the node that drives multiple gates and/or flip-flops, MEP for each gate and flip-flop is calculated and minimum of these MEPs is considered as the node MEP.

Figure 3.4 shows an example to show how MEPs are calculated. The output of each gate is marked with its level. The proposed backward propagation is a general technique and can be applied to all other electrical masking techniques in the literature to reduce their runtime.

3.6.2. SER estimation algorithm

The overall algorithm of the proposed SER estimation considering all three masking factors (logical, electrical, and latching-window) is demonstrated in Algorithm 1. The logical and latching-window masking models are adopted from [95] and [76], respectively. For the electrical masking, the proposed LUT-based technique for propagation of trapezoidal pulse parameters is employed.

In the SER estimation algorithm, the netlist is first levelized from primary inputs and flip-flop outputs (line 2). Next, by performing the backward_propagation procedure (line 3), the MEPs are extracted (according to the details discussed in Section 3.6.1). Then, for each gate, the GATE_SER procedure is performed to calculate the gate SER with respect to a specific pulse (lines 4-6).

The GATE_SER procedure, shown in Algorithm 1 (lines 8-22), gets a target gate and a target pulse as input and propagates the target pulse from the target gate output towards sequential elements inputs. In this procedure, a priority queue (PQ) based on the level of the gates is used to keep the waiting list of the gates in the forward cone that should be processed. In the beginning, the priority queue is filled with the successor gates of the target gate (line 10). Next, in each step, the gate with the lowest level is extracted from the priority queue (line 12) and after the calculation of the logical masking probability (line 13), based on the input pulse(s) the parameters of the output pulse are calculated (line 14). If the output pulse has a zero error probability or the output pulse is weaker than the MEP of the output node, the error is masked and has no chance to propagate to the sequential elements inputs from this path (line 15-16). Otherwise, we continue to propagate the effect of this pulse through this path by adding all successor gates of the gate to priority queue. The operation of processing gates with highest priority (lowest level) and adding successor gates continues until no gate remains in the priority queue. When priority queue is empty, all the gates in the forward cone are processed

Algorithm 1 SER Estimation Algorithm

```
1: procedure SER (netlist, pulse)
2:     levelize(netlist)
3:     minp ← backward_propagation()
4:     for each gate in netlist do
5:         GATE_SER(gate, pulse)
6:     end for
7: end procedure
8: procedure GATE_SER (target_gate, target_pulse)
9:     pulse[target_gate] ← target_pulse
10:    PQ.insert(all successors of target_gate)
11:    while PQ is not empty do
12:        gate ← PQ.pop()
13:        err[gate_out] ← calculate_logical_masking(gate)
14:        pulse[gate_out] ← propagate_transient_pulse(gate)
15:        If err[gate_out] > 0 and pulse[gate_out] > minp[gate_out]
16:            PQ.insert(all successors of gate)
17:    end while
18:    for each ff in netlist do
19:        latching_probability[ff] ← (pw[ff_in]−t_setup −t_hold)/T_clk
20:    end for
21:    failure_probability[ff] ← latching_probability[ff] × err[ff_in]
22:    return failure_probability
23: end procedure
```

and pulse in propagated to the sequential elements inputs. Finally, latching probability for each sequential element is computed (line 19).

3.7. Experimental results

In order to show the accuracy and scalability of the proposed LUT-based method, SER of OpenRISC 1200 (OR1200) and Leon2 processor as well as four benchmark circuits selected form ITC99 are evaluated using this method. These benchmarks have up to hundred thousands of gates which is indicative of scalability of the proposed method.

Each benchmark circuit is synthesized with Synopsys Design Compiler using 45 nm Nangate library and its layout is extracted using Cadence SOC Encounter. Accurate SPICE simulations are performed in order to characterize the standard cells to build suitable LUTs for both leakage current and pulse propagation. Supply voltage of each gate is extracted using voltage droop model presented in Section 2.5.2 and then SER of each circuit is computed according to Algorithm 1.

3.7.1. Accuracy of LUT-based method

The sampling points for the targeted 45 nm technology are summarized in Table 3.1. Using non-uniform sample intervals and putting more samples on sensitive range of each parameter leads more accurate results. Based on the observation from the experiments, the sensitive range for each parameter (output parameters variations in this range are high) are obtained. Since the inaccuracy/runtime of LUT-based method strongly depends on the number of samples and

Table 3.1.: Sampling Points of LUT-based Method for 45 nm Standard Cell Library (Nominal V_{DD} is 1.0)

Parameter	Unit	Sampling Points
c_l	fF	$0, 2, 4, 8, 16, 32, 64$
t_r and t_f	ps	$0, 2, 4, 8, 16, 32, 64$
pw	ps	$5, 10, 15, 25, 50, 70, 100, 200$
V_{min}	V	$0.0, 0.2, 0.4$
V_{max}	V	$0.6, 0.8, 1.0$
V_{DD}^g	V	$0.8, 0.9, 1.0$

the sampling method, using non-uniform sample intervals for each parameter results in a good trade-off between accuracy and runtime.

To validate the effectiveness of the LUT-based method, it is compared with the SPICE simulation in terms of runtime and accuracy. Since SPICE simulation is infeasible for large benchmarks, a set of critical paths (1000 paths) are extracted from all benchmark circuits. For each path, different input pulses with different parameter (pw, V_{min}, V_{max}, t_r, t_f) are given to the primary input of the path and the pulse width at the primary output is obtained using both LUT-based method and SPICE simulation. To better compare the accuracy of pulse propagation, only the effect of electrical masking is considered, i.e. the effect of logical masking was ignored. To do so, the inputs of each gate, which are not in the path between primary input to primary output, are set to a logical constant guaranteeing that the input pulse is propagated through the gate. Table 3.2 shows the accuracy/runtime comparison of our LUT-based method and SPICE simulation for 10 selected paths. Considering all 1000 paths, our LUT-based model provides more than 99% accuracy comparing with SPICE simulation while it is more than 100x faster in average. It is worth to mention that, for larger circuits our LUT-based method scales much better, since the runtime of our LUT-based increases linearly with the circuit size.

Table 3.2.: Accuracy/runtime of LUT-based model in comparison with SPICE simulation for transient pulse propagation

Selected Paths	Path Length (# of gates)	Accuracy of LUT-based	LUT-based Runtime Speed-up
PATH1	27	99.5%	49x
PATH2	23	99.3%	94x
PATH3	14	99.0%	84x
PATH4	17	99.1%	81x
PATH5	25	98.9%	142x
PATH6	26	99.4%	120x
PATH7	15	99.2%	117x
PATH8	12	99.7%	125x
PATH9	18	99.3%	142x
PATH10	13	99.5%	240x

Table 3.3.: Error due to Neglecting Voltage Droop on SER

Pulse Width	b17	b18	b19	b22	Leon2 Core	OR1200 Core	Average
20ps	9.3%	96.9%	43.2%	152.3%	25.5%	15.0%	57.0%
30ps	1.6%	7.1%	13.6%	7.6%	18.9%	0.9%	8.3%
50ps	0.3%	1.8%	5.1%	3.1%	4.7%	0.1%	2.5%
90ps	0.0%	0.1%	1.3%	1.1%	1.4%	0.0%	0.6%

3.7.2. Effect of voltage droop on overall SER

Based on our observation, average fluctuation of supply voltage in these benchmarks is between 1.7% and 8.5% of nominal V_{DD}. Table 3.3 shows the relative error in SER estimation if the effect of voltage droop is neglected. The error is obtained based on the following equation:

$$error = |SER_{WVD} - SER_{WOVD}|/SER_{WVD} \tag{3.4}$$

Where SER_{WVD} and SER_{WOVD} are SER estimation results with and without consideration of voltage droop, respectively.

According to the table, neglecting the effect of voltage droop results in a larger error for the smaller pulse widths. This is due to the fact that pulse propagation is more sensitive to the gate delay when the width of input pulse is relatively small. On the other hand, if the width of input pulse is large enough, it can be propagated completely without being affected by the gate which means the sensitivity of the propagation to the voltage droop is negligible for pulses with large enough pulse widths.

3.7.3. Runtime

The runtime of the proposed LUT-based SER estimation technique for evaluating 20,000 gates in each benchmark is reported in Table 3.4 for two cases. The first case is without considering the backward propagation technique and the second one is with backward propagation technique. As it can be seen, backward propagation reduces the runtime of the proposed method by 14.1%. It should be mentioned that in LUT-based method by decreasing the number of samples the simulation runtime exponentially decreases meaning that the runtime of this method can be reduced at the expense of higher inaccuracy.

Table 3.4.: Runtime of Proposed Electrical Masking Method
*W/O-B: Without Backward Propagation Technique
**W-B: With Backward Propagation Technique

Benchmark Circuit	# of Gates	Runtime of SER Estimation [s]		Runtime Improvement
		W/O-B*	W-B**	
b17	27k	638	554	13.2%
b18	88k	801	728	9.1%
b19	165k	864	688	20.3%
b22	40k	1340	1250	6.7%
OpenRISC	85k	339	301	11.2%
Leon2	60k	2114	1600	24.3%
Average				14.1%

3.8. Summary

In this chapter, the effect of voltage droop on the SER of combinational logic is investigated. Based on the experiments, neglecting the effect of voltage droop leads up to 152% inaccuracy in the overall SER estimation. Also an LUT-based model for transient pulse propagation is proposed considering voltage droop. The results show that the model is highly accurate while it is scalable to handle large circuits, as the experiments are performed on the benchmarks containing up to hundred thousand gates. Moreover, a backward propagation technique is proposed to further improve the runtime of overall SER estimation. According to the results the runtime is improved by 14.1% in average using this technique.

CHAPTER
FOUR

RADIATION-INDUCED SOFT ERROR ANALYSIS OF SRAMS IN SIO-FINFET TECHNOLOGY

4.1. Overview

While the focus of previous chapter was to model the soft errors in logic gates of combinational circuits, in this chapter we try to address this issue in the memory.

For this purpose, in this chapter, a comprehensive analysis of radiation-induced soft errors is presented for SRAMs designed in SOI FinFET technology. For this purpose, we propose a cross layer approach starting from a 3D simulation of particle interactions in FinFET structures up to circuit level analysis by considering the layout of the memory array. This approach enables us to consider the effect of different factors such as supply voltage and process variation on SER of FinFET SRAM memory arrays. Our analysis shows that proton-induced soft errors are becoming important and comparable to the SER induced by alpha-particles especially for low supply voltages (low power applications). Moreover, we observe that the ratio of MBU to SEU for alpha-particle radiation is much higher than that of proton. The chapter is organized as follows: First, the motivation of this work is presented and the contribution of this work is briefly explained in Section 4.2. Next, the related work is discussed in Section 4.3. In Section 4.4 the overall flow of this work is presented. The interactions of different particles with device material are described in Section 4.5. The effect of transient current (generated from interaction of particles and the material) on SRAM cells is described in Section 4.6. Next, in Section 4.7, the methodology to obtain the SER is explained considering the placement of the transistors in the circuit layout. Simulation results are presented in Section 4.8. Finally, Section 4.9 provides the summary of the achievements and concludes the chapter.

4.2. Introduction, motivations and contributions

Further scaling of the planar bulk CMOS technology beyond 22nm is expected to be difficult due to short channel effects [32]. To overcome the scaling limits of this conventional CMOS technology, FinFET is one of the most promising candidates [23, 96]. This is due to the fact that FinFET exhibits superior immunity to short channel effects. Moreover, the effect of process variation on FinFET device performance is less compared with conventional bulk devices [30, 31]. FinFETs can be fabricated as a bulk device or on SOI. However, FinFET has been mainly fabricated on SOI [97, 98] due to lower junction capacitance, higher mobility, and voltage gain with reduced mismatch compared to bulk devices [99].

One of the most important issues which has to be studied carefully in these structures is the effect of radiation-induced soft errors.

In this chapter, a comprehensive analysis of radiation-induced soft errors of SOI FinFET technology is presented in order to obtain SEU as well as MBU rates in SRAM-based memory arrays. In order to estimate failure rates, a cross layer approach, which combines simulations at three different levels, is used: 1) 3-D analysis of particle passage through Fin structure, 2) SRAM cell characterization, and 3) 3-D memory array layout analysis. First, the number of electrons generated by the passage of particles through the matter is obtained. Afterwards, using the number of generated electrons, the SER of SRAM cells is estimated using circuit level simulations considering the location of different transistors in the layout of the memory. We use a hierarchical approach to be able to perform this three-level analysis with a reasonable runtime. The effect of process variation and supply voltage are considered in the SER analysis. Moreover, the contributions of MBU and SEU to the total SER are obtained. Since this work was performed in a collaboration with IBM, the focus of this work is on SOI FinFET technology.

4.3. Related work

Although the radiation effect is well studied for conventional bulk CMOS technologies, there are few studies on FinFET technology. The previous studies can be categorized into two groups:

1. *Device level studies* [100–104]: In [100], the behavior of three different multi-gate transistor structures (Double gate, FinFET and Gate-All-Around MOSFET) under heavy ion irradiation is analyzed using 3-D device simulations. A 3-D TCAD device model is used in [101] to describe the I-V characteristic of SOI FinFET transistors and their transient response to radiation. In [102], the critical charge and single event upset sensitivities of SRAM cells in bulk and SOI FinFET technologies are obtained and compared using TCAD simulations. The Neutron-induced charge collection is estimated and compared for bulk FinFET and conventional CMOS using a mixed-mode 3-D TCAD in [103]. Authors in [104] present a model to estimate the transient charge collection induced by energetic particles for the SOI FinFET technology.

2. *Circuit level studies* [105–107]: The critical charge of a SRAM cell, a simple inverter and a logic chain is obtained for bulk CMOS and FinFET technologies in [105]. The authors used a double exponential current source model [108] to estimate the Soft Error Rate (SER) at a node. In [107], different FinFET technologies are studied using a 3-D TCAD tool to obtain the minimum radiation dose required to flip SRAM cells.

All the aforementioned studies either focus on device or circuit level. The studies at device level are more accurate. However, due to the high runtime of these techniques, it is intractable to apply them in order to obtain the results at circuit level. On the other hand, the circuit level studies, suffer from a lack of information at the device level. In [106], sea-level SER is investigated for three different structures of FinFET (Si, III-V and III-V tunnel FET) technologies considering device and circuit level. In that paper, first, the transient current profile is evaluated using device simulations. Then, according to the critical charge extraction, the electrical and latching window masking effects are studied. The focus of that work was only on the neutron-induced soft error of circuits designed in bulk FinFET technology. They reported SER of a single SRAM cell and there is no information about the contribution of Single Event Upset (SEU) and Multiple Bit Upset (MBU) rates of SRAM arrays. Moreover, the effect of process variation is not considered in their investigation.

Another approach, orthogonal to the simulation-based studies mentioned above, is to perform radiation experiments to obtain SER for FinFET technology. The authors in [109] reported on measured radiation-induced SER of memory and logic devices in a 22nm bulk Tri-gate

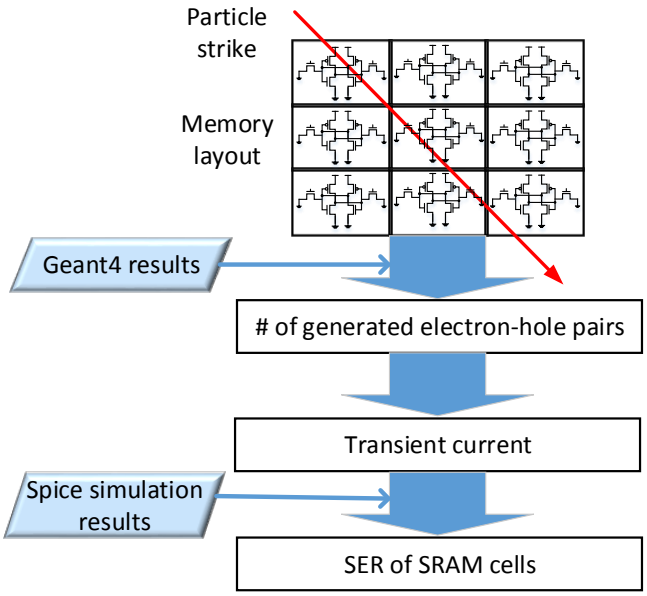


Figure 4.1.: The overall flow of SER estimation

technology. In [110], the charge collection is investigated for bulk PMOS FinFETs using wafer Two-Photon Absorption (TPA) experiments.

4.4. Overall flow

In this section, the flow of SER estimation is described (see Figure 4.1). When a particle strikes the memory array, it affects some of the transistors inside the memory layout leading to the generation of electron-hole pairs inside those transistors. The generated electron-hole pairs inside the affected transistors lead to parasitic transient current pulses which can eventually flip the state of the SRAM cells. In order to obtain the transient current pulses, two steps have to be performed:

1. A Monte Carlo (MC) simulation of the interaction of the particle and the 3-D material structure needs to be performed to obtain the number of generated electron-hole pairs for different particles energies and the results are stored in look-up tables (LUTs).

2. The number of generated electron-hole pairs has to be converted to a transient current pulse.

All the aforementioned device level steps will be described in Section 4.5.

The next step is to convert the transient current pulses to the Probability Of Failure (POF) of individual SRAM cells considering process variation. For this purpose, SPICE simulations are performed for different transient current pulse magnitudes to obtain POF for each case and the corresponding data is stored in POF LUTs. The details of these steps are explained in Section 4.6. An MC 3-D simulation is performed using the transient current and POF LUTs to obtain SEU/MBU rates of SRAM-based memory array. This step is detailed in Section 4.7.

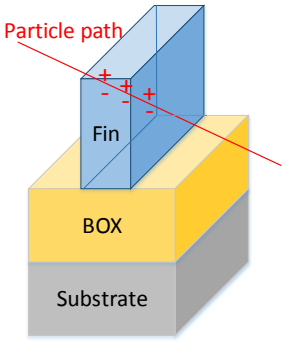


Figure 4.2.: 3-D structure of SOI FinFET: the gate, drain and source are not shown in the figure

4.5. 3-D Analysis of particle strike with fin structure

4.5.1. Interaction of particles and material

In this work, the Geant4 toolkit [111, 112] is used to simulate the interaction of particles and devices. Geant4 is a platform for Monte-Carlo simulation of the passage of particles through the matter. In this work, the target material is the 3-D structure of a single Fin in a transistor (with dimensions provided in [113]). For this target material, the number of generated electron-hole pairs due to particle strike for different energy ranges is obtained (see Figure 4.2) and stored in LUTs. For this purpose, 10 million MC simulations are performed with different particle directions and positions for each particular energy. The runtime of Geant4 simulations is relatively high (a few hours for 10 Million simulations for the FinFET under investigation), however, it should be noted that the simulations have to be performed only once to build up LUTs. Figure 4.3 shows the average number of electrons generated in a single Fin of transistor due to its interaction with alpha-particle and proton for different energy ranges.

4.5.2. Radiation-induced parasitic transient current pulse

When a particle hits a transistor, some electron-hole pairs are generated. These charges can be transported in the device leading to a parasitic current which in turn can affect the device. In general, the devices containing a reversed p-n junction are sensitive to the particle strikes. This

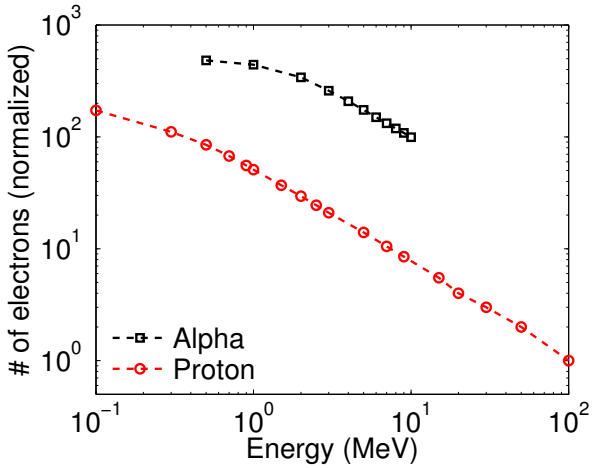


Figure 4.3.: The normalized number of electrons generated by the interaction of alpha-particle and proton with a Fin

is due to the fact that these types of devices have a strong electric field in the depletion layer of the p-n junction which can lead to a collection of deposited charge as a transient current pulse. The shape of the generated parasitic current is strongly dependent of the technology and can be obtained by 3-D simulations of the device. However, in this work, we consider a simplistic model for the parasitic current which is explained in the following.

Normally the parasitic current is generated due to two different mechanisms:

1. *Diffusion*: If the particle hits the substrate of the transistor, the generated charges are collected by sensitive nodes due to the diffusion of the carriers. However, in SOI FinFET technology, the diffusion current can be neglected due to the Buried Oxide (BOX) between the substrate and Fin (see Figure 4.2).

2. *Drift*: If the particle passes through the sensitive area between source and drain of the transistor (Fin), the generated electron-hole pairs are collected due to the electric field between source and drain (drift mechanism) and create a parasitic current. In order to model this type of current in SOI FinFET technology, different factors have to be considered:

   - **Particle passage time (τ_p)**: It is defined as the time which takes for the particle to pass through the Fin.
   $$\tau_p = \frac{w_{Fin}}{v_p} \tag{4.1}$$
   where v_p is the speed of the particle and w_{Fin} is the width of Fin. τ_p is less than 1 fs (femto second) for alpha particle. For proton, τ_p is approximately 10 times smaller than that of alpha-particle.

   - **Transit time (τ)**: It is defined as the average time required for an electron to travel between source and drain.
   $$\tau = \frac{L_{Fin}^2}{\mu_e \cdot V_{DS}} \tag{4.2}$$
   where L_{Fin} is the length of Fin, μ_e is the electron mobility and V_{DS} is the voltage between drain and source (which is equal to V_{DD} for sensitive transistors). The transit time (τ) for the transistor shown in Figure 4.2 with a supply voltage equal to 1V is more than 10 fs. Since τ is much larger than τ_p, we can assume that when the particle passes through the Fin, all electron-hole pairs are generated at the same time and start being collected due to the drift mechanism.

   - **Recombination time (τ_r)**: It determines the rate of the electron-hole recombination. τ_r ranges from 1ns to 1ms in Si, which is much larger than τ. Therefore, we can assume that the recombination of the electron-hole pairs is negligible at these dimensions and the electron-hole pairs are collected before a major recombination happens.

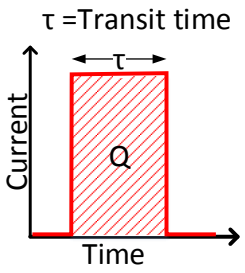


Figure 4.4.: Parasitic current model

Considering all aforementioned parameters and their relations, we model the parasitic current as a current pulse (see Figure 4.4) with a width equal to τ and an amplitude equal to:

$$I = \frac{Q}{\tau} = \frac{n_e \cdot e}{\tau} \tag{4.3}$$

where n_e is the number of electron-hole pairs generated in the Fin and e is the charge of a single electron.

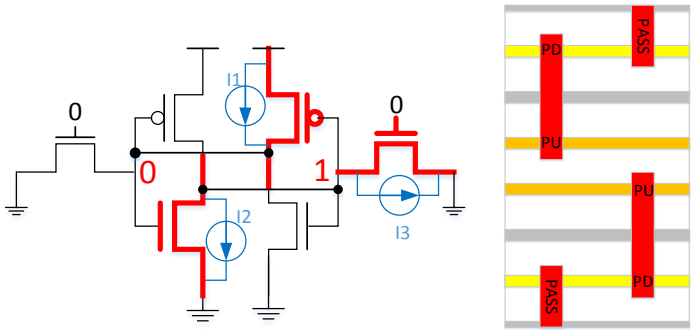
4.6. SRAM cell soft error characterization

In this section, we briefly describe how to obtain the radiation-induced Probability Of Failure (POF) of a single SRAM cell considering process variation. For this purpose, we consider a 6T SRAM cell designed with SOI FinFET technology as shown in Figure 4.5.

As mentioned in Section 4.5.2, transistors containing a reversed p-n junction are sensitive to particle strikes. In other words, the sensitive transistors to radiation in an SRAM cell are the ones which are in OFF state with $V_{DS} = V_{DD}$ (transistors shown with red-bold lines in Figure 4.5(a)). If one or multiple of these sensitive transistors are struck with a particle, parasitic current pulses are generated in these transistors which may eventually lead to a change in the state of the SRAM cell.

POF of an SRAM cell is a function of the current pulse magnitude, the supply voltage and the number of struck transistors. If process variation is neglected, the POF for a given strike scenario becomes a deterministic binary value in which '0' means the parasitic current pulse does not lead to a flip, and '1' means that the parasitic transient current causes the SRAM cell to flip. For the case in which process variation is considered, POF of the struck SRAM cell for a particular current pulse magnitude becomes a probability value between 0 and 1 ([0.0 1.0]). In this case, to obtain POF, we consider the threshold voltage variation by performing 1000 MC simulations based on accurate SPICE simulations using the current model described in Section 4.5.2. These POFs are obtained for different supply voltages, current pulse magnitudes, and all possible combinations of current pulses (for $I1$, $I2$, $I3$ or any combination of these three currents in Figure 4.5(a)) and then stored in LUTs.

We performed some experiments to find out the effect of transient current pulse shape on POF obtained from SRAM cell characterization. The SPICE simulation results show that POFs have no sensitivity to the current pulse width. In other words, two current pulses with different pulse widths but similar charge (area under the I-t curve) cause the same POF of the



(a) (b)

Figure 4.5.: a) 6T SRAM cell: sensitive transistors to soft error are shown with red-bold color b) Layout of 6T SRAM cell

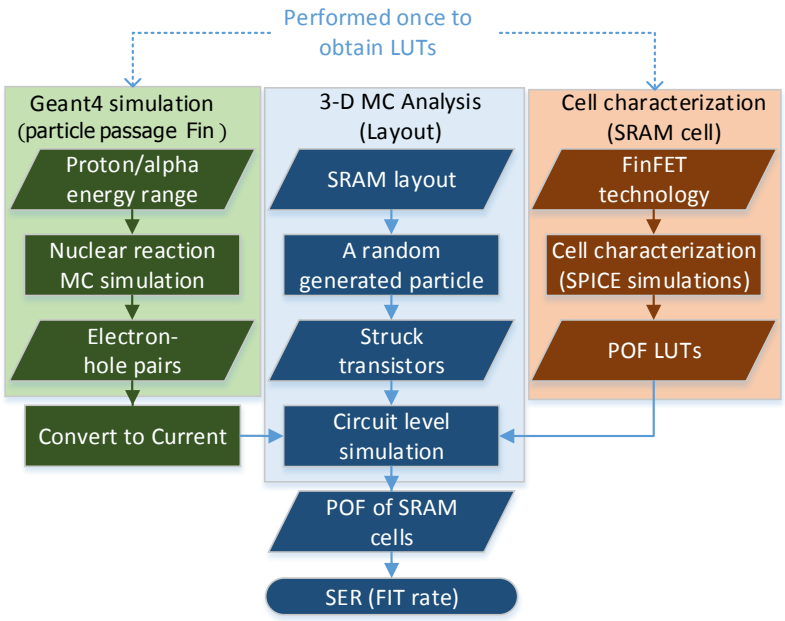


Figure 4.6.: The overall flow to obtain the SER of SRAM-based memory array for SOI FinFET technology

SRAM cell. Moreover, we obtained the POF results applying a triangular current shape with different current pulse widths. The results show that although the effect of particular current pulse shape (triangular vs rectangular) is more than the current pulse width, it is still negligible. In other words, the most important parameter for SPICE simulation is the generated charge (the area under the current pulse curve).

4.7. 3-D memory array analysis

In this section, the methodology to estimate the SER for the entire SRAM array is explained. Since a single particle strike can hit multiple Fins (in different cells), we obtain SEU as well as MBU rates by taking the SRAM array layout into consideration. Figure 4.6 shows the overall flow to obtain the SER of SRAM-based memory array for SOI FinFET technology. The flow to obtain electron-hole pairs/current LUTs is explained in previous sections. In this section we mainly focus on the 3-D MC simulation of the layout (the middle part of the flow shown in Figure 4.6).

4.7.1. Probability Of Failure (POF) due to a particle strike with a particular energy range

In order to estimate the overall SER due to particle strike, we first need to perform an MC simulation to obtain POF of different cells due to the strike of a particle with a particular energy. The steps of obtaining the POFs of different cells for a simple 2×2 SRAM array which is shown in Figure 4.7 are explained as follows:

1. A random particle with a random direction and position is generated. Based on the angle and the direction of the generated particle, the struck Fins (transistors) can be found by a simple 3-D analysis considering the 3-D layout of SRAM array and the position of Fins/transistors inside the layout ($\{m1, m2, m3\}$ transistors in cell1 and $\{m4, m5\}$ in cell2 in this example).

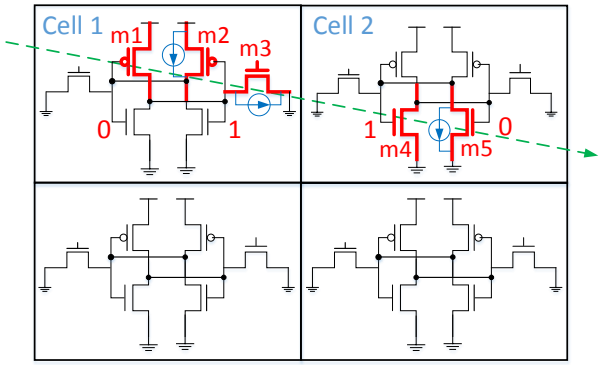


Figure 4.7.: The flow of obtaining POF of different SRAM cells inside a SRAM array for a particle with a particular energy range

2. In the next step, the number of generated electron-hole pairs for the affected transistors is obtained from LUTs (using Geant4 as explained in Section 4.5.1) according to the particle energy.

3. If the affected transistors are the sensitive ones in the SRAM cell ($\{m2, m3\}$ in cell1 and $\{m5\}$ in cell2), the number of electron-hole pairs are converted to the parasitic current pulse as explained in Section 4.5.2 (see Figure 4.7).

4. POFs of SRAM cells are obtained according to their parasitic current pulses using SPICE LUTs (obtained using the approach explained in Section 4.6).

5. The total POF as well as POF of SEU and MBU are computed for SRAM array as a function of cell POFs, as shown below:

$$POF_{tot} = 1 - \prod_{i} (1 - POF(celli)) \tag{4.4}$$

$$POF_{SEU} = \sum_{i} [POF(celli) \cdot \prod_{j \neq i} (1 - POF(cellj))] \tag{4.5}$$

$$POF_{MBU} = POF_{tot} - POF_{SEU} \tag{4.6}$$

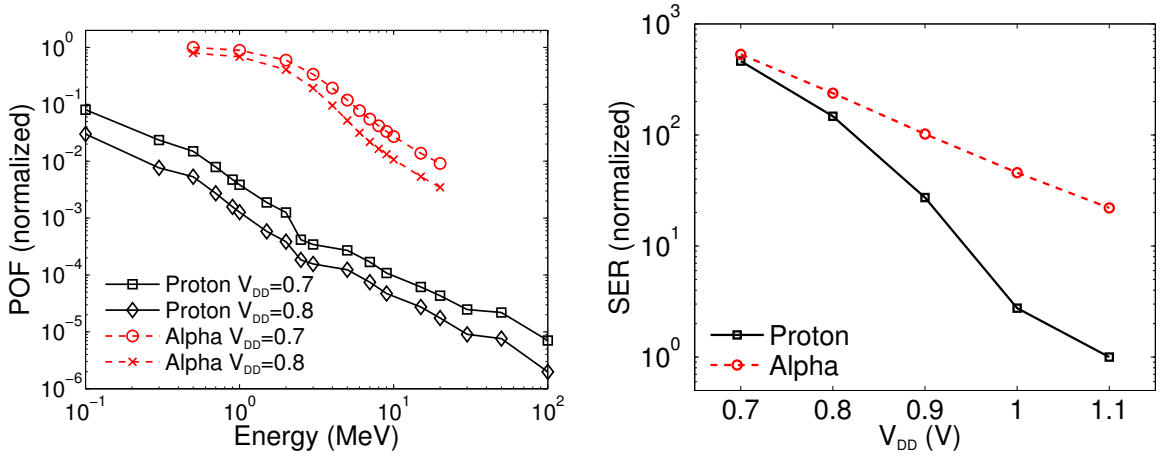
6. Steps 1-5 are performed iteratively for particles with the same energy range and different random directions and positions. For a particle with a particular energy range, the overall POFs (POF_{tot}, POF_{SEU}, and POF_{MBU}) of SRAM array are obtained by the averaging over all iterations.

4.7.2. Failure In Time (FIT) rate calculation

After obtaining the POF of different particles with different energy ranges, the next step is to obtain the *Failure In Time (FIT)* rate of the memory array. In order to obtain FIT rate, the following equation is used:

$$SER(FIT) = \int POF(E) \cdot Flux(E) \cdot Lx \cdot Ly \cdot dE \tag{4.7}$$

In this equation, POF is the probability of failure of the particle at a particular energy, E, which is obtained as explained in Section 4.7.1. $Flux$ is the flux of the particle which can be obtained according to Figure 2.24. Lx and Ly are the dimensions of the memory array. Since it



(a) (b)

Figure 4.8.: a) Normalized POF of SRAM-based memory array with a V_{DD}=0.7V due to strike of proton and alpha-particle b) Normalized FIT rate due to proton and alpha-particle radiation

is not possible to obtain POF over all energy ranges, we need to discretize the energy spectrum of the particle to different ranges. Therefore, Equation 4.7 can be rewritten as follows:

$$SER(FIT) = \sum POF(E) \cdot IntFlux(E) \cdot Lx \cdot Ly \tag{4.8}$$

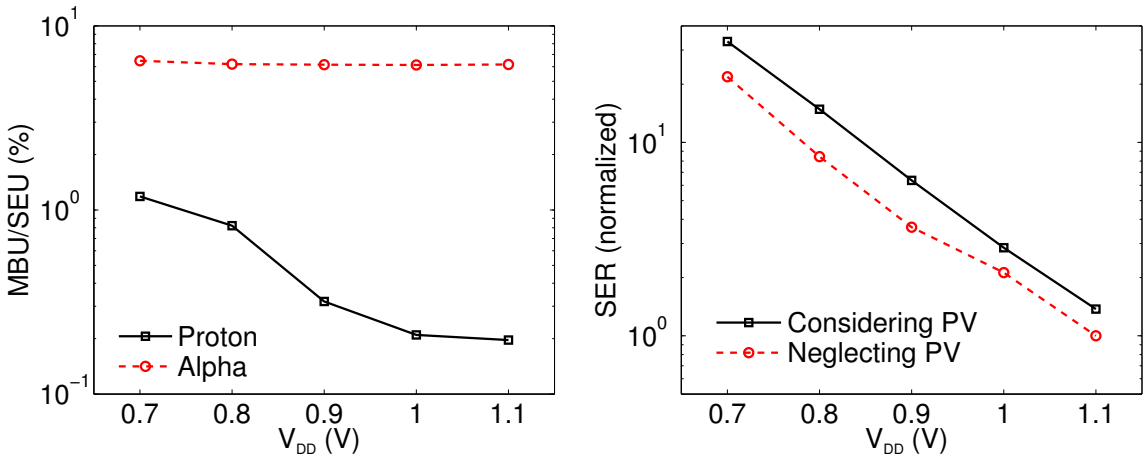
where E is the representative energy of each range and IntFlux(E) is the integral flux of the particle at that range.

4.8. Simulation results

In this section, the simulation results are presented. The Geant4 toolkit is used in this work to build electron-hole pairs LUTs. A 14 nm SOI FinFET technology library [70] is used for SPICE simulations. After obtaining the number of electron-hole pairs and POF LUTs, a 3-D MC simulation (10 Million iterations) is performed as described in Section 4.7 to estimate SEU and MBU rates for a 9×9 SRAM array with a layout which is shown in Figure 4.5(b). The device parameters are obtained from [113]. It should be noted that the runtime of the entire process for a 9×9 memory array for 10 Million MC simulations is around 2 hours. Such an array size is large enough to obtain a realistic ratio for MBU vs. SEU and there is no need to explicitly consider larger arrays. The simulation results are presented in the following. All presented results in this section are normalized.

POF of different particles at different energy ranges Figure 4.8(a) shows the total POF of the SRAM-based memory array with $V_{DD} = 0.7V$ and $V_{DD} = 0.8V$ due to alpha-particle and proton strike. For this experiment, we assume that the particle definitely hits the layout of the memory array under investigation.

As shown in this figure, the POF due to alpha-particle is much larger than that of proton, as more electron-hole pairs are generated by alpha-particles (see Figure 4.3). Moreover, POF decreases for both particles for higher particle energies, since, according to Figure 4.3, for higher particle energies less electron-hole pairs are generated. In addition, the POF increases with decreasing V_{DD} for both particle types. This is due to the fact that SRAM cells are more sensitive to the soft error at lower supply voltages. In other words, some particles which have



(a) (b)

Figure 4.9.: a) MBU vs. SEU b) Effect of process variation on alpha-particle induced SER estimation (SER results are normalized)

no effect on SRAM cells with higher supply voltages, may flip the SRAM cell value at lower supply voltages.

Overall FIT rate of memory array Figure 4.8(b) shows the total SER due to proton and alpha-particle strike for various supply voltage values. As shown in this figure, the overall SER increases by decreasing the supply voltage. Moreover, the results show that the SER due to proton radiation is comparable to that of alpha-particle for $V_{DD} = 0.7$. Although the POF due to proton strike is much smaller than that for alpha particle, the overall SERs of proton and alpha particles are comparable at lower supply voltages ($V_{DD} = 0.7$) due to the higher flux of protons at ground level. The proton-induced SER decreases with an extremely higher rate for larger supply voltages compared to alpha-particle-induced SER. This implies that proton-induced soft error is important especially for low power applications (lower V_{DD}).

SEU vs. MBU Figure 4.9(a) shows the MBU to SEU ratio for alpha and proton particles. As shown in this figure, MBU/SEU ratio of protons (less than 2%) is much smaller than that of alpha-particles (6-7%). This is due to the fact that SEU rate is proportional to POF^1 while MBU is proportional to POF^n ($n > 1$). As a result, since the proton-induced POF of SRAM cells is much smaller than that induced by alpha-particles, the difference between MBU rate of alpha particle and protons is much larger than the difference between SEU rates of these particles. Moreover, alpha-induced MBU/SEU ratio has almost the same value for different supply voltages while this ratio decreases with increasing V_{DD} for protons. This is rooted in the fact that POF is less sensitive to V_{DD} for alpha particles (according to Figure 4.8(a)) especially in the range of energies which is important at sea level (less than 10Mev). However, the sensitivity to V_{DD} is much higher for protons. The other reason is that the mass of alpha particles is four times bigger than that of protons, therefore, at the same speed the kinetic energy of alpha particles is four times bigger than that of protons.

Effect of process variation Figure 4.9(b) shows the total SER due to alpha-particle strike for two cases:

1. Neglecting process variation: For this case, SPICE simulations are run for the nominal case. The outputs of the simulations for different current pulses (i.e. each simulation

case) are binary values: '1' (for the BIT flip) and '0' (no flip). The overall POFs are obtained by averaging over all these binary values for 10 million iterations.

2. Considering process variation: In this case, as explained in the flow, the process variation is considered in SPICE simulations and probabilistic POF values (between 0.0 and 1.0) are reported for different parasitic currents (i.e. each simulation case).

As shown in this figure, neglecting the effect of process variation leads to an underestimation of SER (up to 45%). The simulation results also show the same trend for proton-induced SER.

4.9. Summary

In this chapter, the SER of SRAM-based memory array in SOI FinFET technology is investigated using a cross layer approach. In our approach we used information from device level (interaction of particles and materials), circuit-level cell characterization, and array-level 3D simulations. This study can be summarized as:

1. SER is higher for lower supply voltages.
2. SER due to proton strike is comparable to that for alpha-particle strike at very low supply voltages (low power applications).
3. MBU/SEU ratio is relatively higher for alpha radiation compared to that for protons.
4. Neglecting the effect of process variation leads to an underestimation of SER.

CHAPTER
FIVE

THE IMPACT OF PROCESS VARIATION AND STOCHASTIC AGING IN NANOSCALE VLSI

5.1. Overview

In the previous parts of this dissertation, we have addressed soft errors as an important source of unreliability. In this chapter, we focus on accelerated transistor aging, as another important reliability challenge. In particular we address NBTI, which is the most severe runtime variability challenge, in combination with process variation, to estimate the required timing guard-band for a reliable design.

With the down-scaling of CMOS technology into deep nano-scale era the NBTI effect becomes stochastic due to its widely distributed defect parameters leading to more non-determinism in the functionality of the deeply-scaled circuits. This chapter presents a framework to comprehensively investigate the combined effect of stochastic NBTI and process variation on the performance of the VLSI design at circuit level, by abstracting atomistic NBTI models (for the stochastic behavior) to the circuit timing analysis flow. Simulation results, performed in a 7 nm FinFET technology, show that the stochastic behavior of NBTI can result in a significant increase of the guard-band. Moreover, our analysis reveals that stochastic NBTI and process variation should be considered together, otherwise it can lead to a major overestimation of the mean value of the delay degradation. The rest of the chapter is organized as follows. First, the motivation and contribution of this work is briefly discussed in Section 5.2. Next, the related work is discussed in Section 5.3. Afterwards, Section 5.4 introduces the flow of the proposed stochastic NBTI/process variation-aware timing analysis framework. Simulation results are presented in Section 5.5 and finally, Section 5.6 concludes the chapter and summarizes the achievements.

5.2. Introduction, motivations and contributions

In previous technology nodes , the NBTI effect on transistors was fairly deterministic for a particular workload condition (e.g. temperature and stress) and it was effectively mitigated by adding timing guard-band to the design by "predicting" worst case workload and stress conditions in the field [53]. However, by further down-scaling of transistor dimensions into deca-nanometer range, the number of defects per device decreases leading to a drastic increase in the time dependent variability of NBTI [54]. As a result, the delay degradation due to this *non-deterministic* NBTI effect becomes also *stochastic* and the timing guard-band needs to be obtained according to the far tail of the distribution (e.g. $\mu+3\sigma$) to guarantee reliable operation in the field. Accurate stochastic timing analysis of the circuit becomes very important in this

case since over and under margining can lead to significant performance or yield loss (timing failure), respectively. The matter is aggravated when it is combined with process variation, adding another degree of non-determinism. Thus, it is important to analyze the stochastic behavior of NBTI combined with process variation affecting the VLSI circuits performance.

In this work, we propose a framework for analyzing the combined effect of stochastic NBTI together with process variation during circuit timing analysis. Using the framework, a comprehensive analysis is performed to obtain the contribution of process variation, stochastic NBTI and their combined effect on the total variation of the circuit delay during its operational lifetime.

The proposed timing analysis is performed on a set of ISCAS85 benchmark circuits [114] with a 7nm FinFET technology library [70]. The simulation results show:

- Stochastic NBTI effect leads to an asymmetric (non-normal) distribution of circuit delay degradation.
- The effect of process variation on the mean of the delay degradation is less than that of stochastic NBTI while it leads to more variability on the distribution.
- The stochastic behavior of NBTI can result in a significant increase of the guard-band compared to a deterministic case.
- Considering these sources of variation separately, leads to a considerable overestimation of the delay degradation mean value.

In this work NBTI is only considered since for deeply scaled FinFET devices NBTI is expected to dominate over PBTI. For recent FinFET technologies metal gate work function tuning and using a fully depleted body has successfully reduced the electric field in the high-k layer and has been shown to be an effective mitigation approach for PBTI. Nonetheless, similar models and statistical flow can be leveraged to obtain the effect of stochastic PBTI as well.

5.3. Related work

There have been some studies to analyze the stochastic behavior of NBTI effect. In [115], the authors proposed a statistical analysis framework to obtain the combined effect of process variation and NBTI on the delay distribution. For this purpose, the mean and standard deviation of the gate delays are propagated to obtain the mean and standard deviation of the circuit delay as a normal distribution. However, we will later show that the distribution of circuit delay considering the combined effect of NBTI and process variation is not a normal distribution. In their framework, NBTI is considered as a random process because the NBTI effect is a function of the transistor initial threshold voltage as well as oxide thickness [47] and these two process parameters have variation due to process variation. In order to model this, they used a deterministic NBTI model proposed in [47] and by considering stochastic behavior of the oxide thickness and the threshold voltage, the NBTI effect can also be expressed as a stochastic parameter. However, it is shown that the NBTI effect has an intrinsic variation [54]. This means that even two identical transistors with the same oxide thickness, initial threshold voltage, and stress condition may have different NBTI effect. This intrinsic variability of NBTI which is not considered in [115] can be modeled by statistical atomistic models.

Since the existing atomistic models are complex, they can be used at device level [48, 116, 117] up to gate/cell level [118] including memory cells such as SRAM [54]. However, implementing the stochastic behavior of NBTI into circuit level timing analysis poses severe restrictions on circuit size and/or simulation time [54]. Moreover, as shown later in the results section, normal distribution models for stochastic NBTI (as an attempt to abstract out atomistic models)

are fairly inaccurate. There have been few studies of stochastic NBTI at circuit level and due to the aforementioned challenges at this level, they either consider a simple chain of inverters [20] or only a single critical path [119]. However, the analysis of single path does not properly reflect the statistical NBTI-induced distribution of circuit delay since the interaction of several critical paths in well-balanced circuits are neglected. Authors in [120] studied NBTI-induced performance degradation of 32-bit adders. However, the analysis is only performed at the $+3\sigma$ corner. Moreover, none of the existing work considers the combined effect of process variation and stochastic NBTI, as our analysis shows they cannot be considered separately.

5.4. Circuit level simulation flow

Figure 5.1 shows the flow of the proposed stochastic NBTI and process variation aware timing analysis. As shown in this figure, the first step is to characterize the standard cells for different values of their internal transistor threshold voltage shift. Then, the corresponding delay/transition time *look-up tables* (LUTs) are stored in an extended library. The extended library is used later by timing analysis tool to obtain the timing information of the gates.

Besides, a logic simulator is used to calculate the internal signal probabilities (the probability that the signal is equal to binary value of "one") and the NBTI parameters of transistors accordingly.

Finally, a *Monte-Carlo* (MC) simulation is performed to obtain the distribution of NBTI and process variation-induced delay degradation. For this purpose, for each MC iteration, ΔVth samples are calculated for transistors according to their NBTI-model parameters. Then, the gates timing information is updated in the library and a *Static Timing Analysis* (STA) is performed to calculate the circuit delay.

5.4.1. Library cell characterization

In the standard cell library, the timing information of a cell (e.g. rise/fall delays and transition times) is stored for different input transition times and load capacitances as 2-dimensional LUTs for the nominal threshold voltage value (Vth). These LUTs are used in STA tools to obtain the circuit delay. However, the Vth of the internal transistors of the cell may change

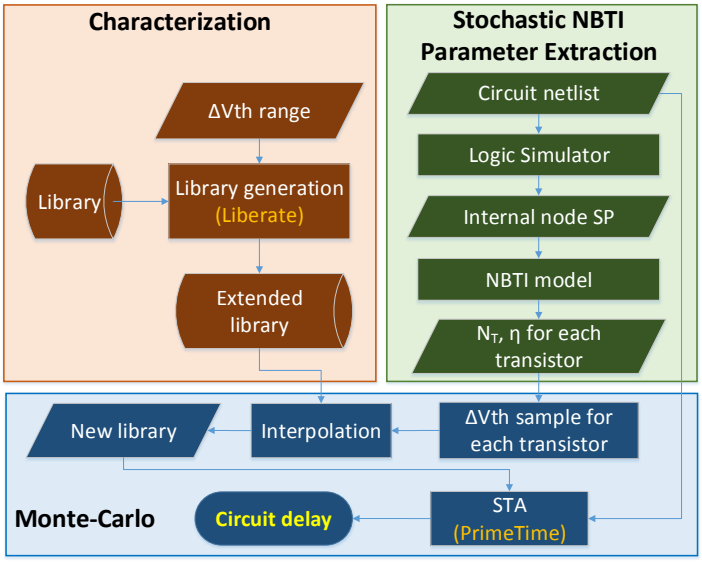


Figure 5.1.: Flow of proposed stochastic NBTI and process variation aware timing analysis framework

due to process/runtime variations. In order to be capable of performing the STA considering process and runtime variations, we need to characterize the cells for different combination of ΔVth of their transistors and store the timing information LUTs in an extended library.

For example, instead of having only one LUT in the library for INVX1 cell, we will have different LUTs for different $\text{INVX1}_{\Delta Vthn\Delta Vthp}$ instances where $\Delta Vthn$ and $\Delta Vthp$ are the corresponding threshold voltage shift values of NMOS and PMOS transistors, respectively. However, it is not possible to characterize the cells for all combinations of ΔVths (the library size goes to infinity). Therefore, ΔVths range is discretized and the library is extended for different combination of ΔVths according to the discretization. In this work, the minimum, maximum and discretization step values of ΔVth are -60mv, 120mv and 10mv, respectively.

5.4.2. Stochastic NBTI parameter extraction

For a given circuit netlist, a logic simulator is used to obtain the internal node signal probabilities. Next, the duty cycle of all internal transistors are calculated according to the internal node signal probabilities. Then, for each transistor, based on its duty cycle and feature size, the parameters of the atomistic NBTI model (η and n) are obtained using the model introduced in Section 2.4.1. These parameters are used in the MC simulation to obtain the samples of NBTI-induced ΔVth for each iteration.

5.4.3. Monte-carlo simulation

ΔVth sampling

The first step of the MC simulation is to obtain samples for ΔVth of all internal transistors due to the stochastic NBTI effect and process variation. For sampling the NBTI-induced ΔVth, for each iteration (i) and for each transistor Tj, we get a sample ($n_{i,Tj}$) from a Poisson distribution with a mean of N_{Tj} , where N_{Tj} is the mean number of transistor defects obtained from Equations 2.22 and 2.23. This sample ($n_{i,Tj}$) represents the number of defects in the transistor T_j for iteration i of MC. Then, the total NBTI-induced ΔVth is obtained by:

$$\Delta Vth_{i,Tj}^{NBTI} = \sum_{k=1}^{n_{i,Tj}} \Delta Vth_k \tag{5.1}$$

where ΔVth_k is the threshold voltage shift for each defect obtained by sampling from an exponential distribution with a mean of η (the mean threshold voltage shift due to each defect) obtained from Equations 2.26 and 2.27. Afterwards, the process variation-induced threshold voltage shift sample ($\Delta Vth_{i,Tj}^{PV}$) is obtained from a normal distribution (according to Section 2.3.3) with a mean equal zero and standard deviation obtained from Equation 2.5 and the total threshold voltage shift of each iteration for each transistor is calculated with:

$$\Delta Vth_{i,Tj} = \Delta Vth_{i,Tj}^{NBTI} + \Delta Vth_{i,Tj}^{PV} \tag{5.2}$$

Here a simple superposition is performed to obtain the total threshold voltage shift. This is due to the fact that according to [20, 59], there is no correlation between NBTI-induced threshold voltage shift and process variation.

Stochastic NBTI and process variation aware STA

After obtaining the ΔVths of all transistors for each iteration, the delay and transition time LUTs of each gate in the netlist is calculated by interpolating the LUTs of the extended library according to ΔVth values and a new library is built according to these LUTs. Finally, a

STA tool is used to obtain the circuit delay using the updated timing LUTs of the gates. By performing such a STA during each iteration of MC, the delay degradation distribution of the circuit is calculated.

5.5. Results and discussion

5.5.1. Simulation setup, terms and definitions

In this section the results of the proposed stochastic NBTI and process variation aware STA are presented for ISCAS85 benchmark circuits [114] as well as three inverter chains with 1 (chain1), 9 (chain9) and 39 (chain39) inverters, respectively. A 7 nm FinFET technology node library [70] is used. Cadence Altos Liberate [121] is exploited for characterization to obtain the extended library and Synopsys PrimeTime [122] is used for STA. All the results are presented as normalized ΔD in percentage calculated from:

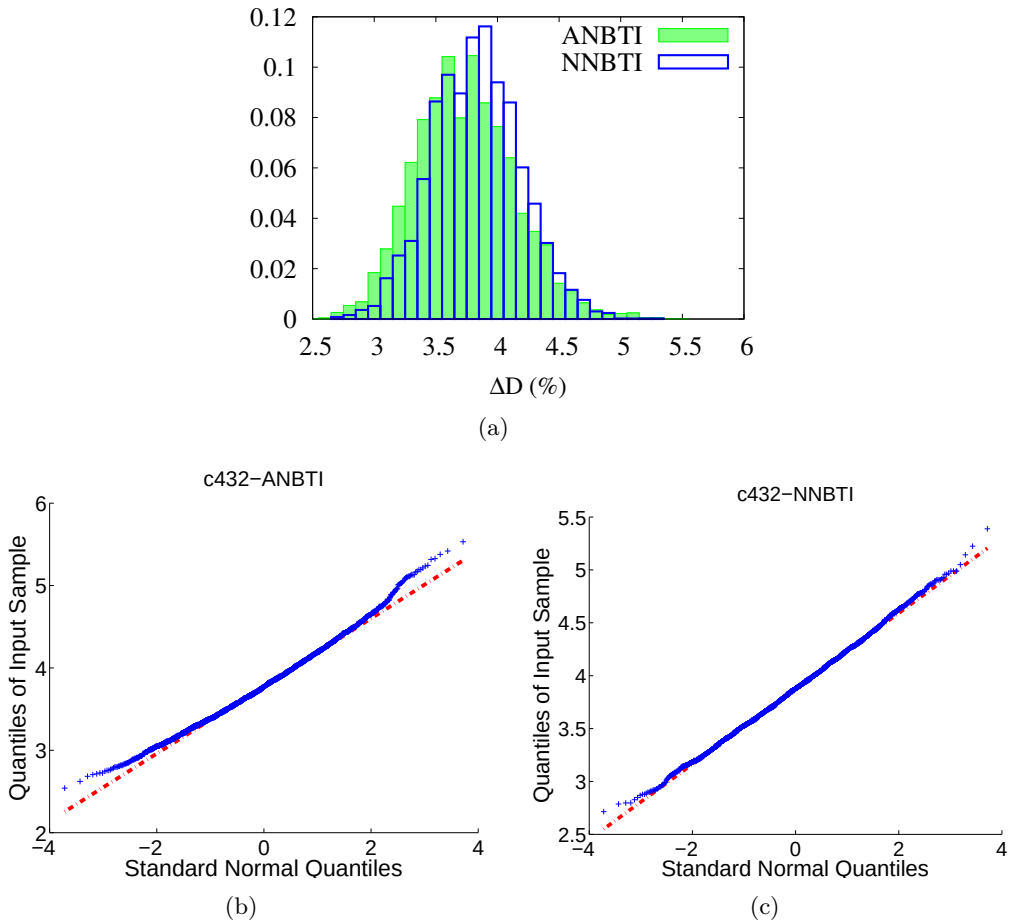
$$normalized\ \Delta D = \frac{D - D_0}{D_0} \tag{5.3}$$

where D is the delay of the circuit considering process/aging variations and D_0 is the nominal delay at time zero. There are some terms used later in the results discussion which are defined as follows:

- **Number of critical paths**: it shows the number of critical paths contributing in the post-aging delay distribution of the circuit.
- **Skewness**: it is a measure of the asymmetry of the probability distribution of a random variable. It is equal to zero for a random variable with a normal distribution.
- **Kurtosis**: it is a measure of the tails weight of a distribution. Kurtosis value of a variable with normal distribution is equal to 3. Higher kurtosis means that the weight of infrequent extreme deviations on the variance is more compared to that of frequent modestly sized deviations.

5.5.2. Atomistic NBTI model vs equivalent normal NBTI model

Figure 5.2(a) shows the *Probability Density Function* (PDF) of NBTI-induced ΔD distribution of the c432 benchmark circuit for two different stochastic NBTI models: i) atomistic NBTI (ANBTI) model and ii) equivalent normal NBTI (NNBTI) model using mean and standard deviation according to Equations 2.17 and 2.18, respectively. As shown in this figure, the shapes of the distributions are different specially in the tails. This is also shown in QQ-plots of the ΔD distribution (see Figure 5.2(b) and 5.2(c)). According to these figures, the NBTI-induced ΔD distribution using the atomistic model is not a normal distribution specially in the tails while a normal model for NBTI leads to a distribution of delay degradation which is very similar to a normal distribution. Table 5.1 also summarizes the information of ΔD distribution for these two models of stochastic NBTI behavior. According to the table, the mean and standard deviation of the ΔD distributions are almost similar for these two models, however, the difference between the skewness metric is significant, for instance as large as 3X for c880 benchmark circuit. Moreover, according to the table, the required timing margin for guard-banding is 30% larger on average for the case of stochastic NBTI ($\mu + 3\sigma$ of delay degradation) compared to that of a deterministic analysis (μ of delay degradation) which shows the importance of the stochastic NBTI analysis at circuit level.



(a) (b) (c)

Figure 5.2.: a) Probability density function of NBTI-induced ΔD for atomistic NBTI (ANBTI) and normal NBTI (NNBTI) b) QQ-plot of atomistic NBTI (ANBTI) c) QQ-plot of normal NBTI (NNBTI)

5.5.3. Effect of process variation vs stochastic NBTI

Figure 5.3 shows the PDF of the ΔD distribution for three cases considering i) only NBTI ii) only process variation and iii) combined effect, for c880 and c1355 benchmark circuits. As shown in this figure, the mean of ΔD is larger for NBTI effect while the distribution of process variation is wider (more variability due to process variation). Moreover, the mean and standard deviation of combined effect is larger than those of both cases when they are considered separately.

Since the threshold voltage shift due to NBTI and process variation for transistors are uncorrelated [20, 59], the mean and sigma of the overall effect on the threshold voltage shift of the transistor can be estimated by the following equations:

$$\mu_{ANBTI+PV} = \mu_{ANBTI} + \mu_{PV}$$
$$\sigma^2_{ANBTI+PV} = \sigma^2_{ANBTI} + \sigma^2_{PV}$$

If process variation and stochastic NBTI are considered separately, the same set of equations shall be used to obtain the mean and sigma of the ΔD distribution at circuit level. However, according to the results of the ISCAS85 benchmark circuits (see Table 5.1), the following

Table 5.1.: Information of the normalized ΔD distribution for four different cases i) normal NBTI (NNBTI) ii) atomistic NBTI (ANBTI) iii) process variation (PV) and iv) combined effect of process variation and NBTI (ANBTI+PV)

Benchmark	# of critical paths				Mean				Standard deviation				Skewness				Kurtosis			
	NNBTI	ANBTI	PV	ANBTI+PV	NNBTI	ANBTI	PV	ANBTI+PV	NNBTI	ANBTI	PV	ANBTI+PV	NNBTI	ANBTI	PV	ANBTI+PV	NNBTI	ANBTI	PV	ANBTI+PV
chain1	1	1	1	1	4.32	4.23	-0.21	3.56	1.7	1.75	3.86	4.23	0.04	0.51	0.14	-0.09	2.72	3.41	2.44	2.8
chain9	1	1	1	1	4.67	4.6	0.27	4.15	0.87	0.89	1.51	1.7	0	0.27	0.09	-0.03	2.96	3.04	3.09	2.88
chain39	1	1	1	1	4.7	4.63	0.28	4.15	0.43	0.44	0.73	0.82	-0.03	0.14	0.01	-0.02	3.07	2.98	2.89	2.93
c17	1	1	3	3	2.82	2.71	1.73	3.88	1.27	1.33	2.6	2.82	0.24	0.64	0.21	0.12	2.69	3.42	2.83	2.86
c432	15	17	25	30	3.88	3.79	1.26	4.48	0.36	0.41	0.79	0.86	0.11	0.33	0.2	0.13	3.02	3.25	3.1	2.94
c499	35	35	94	103	3.57	3.49	1.32	4.09	0.34	0.36	0.66	0.69	0.2	0.25	0.36	0.32	2.91	3.06	3.29	3
c880	2	3	5	5	4.16	4.07	0.72	4.13	0.36	0.37	0.74	0.8	0.07	0.21	0.07	0.07	2.97	3.24	3.08	2.98
c1355	54	84	635	610	3.8	3.71	1.69	4.5	0.3	0.32	0.58	0.62	0.25	0.26	0.27	0.14	3.02	3.04	3.16	3.08
c1908 (unbalance)	4	4	7	7	3.38	3.28	0.67	3.35	0.3	0.32	0.68	0.73	0.11	0.2	0.05	0.02	2.94	2.99	2.98	3.05
c1908 (balanced)	185	234	1065	1106	3.61	3.52	2.01	4.59	0.28	0.31	0.56	0.57	0.36	0.43	0.23	0.22	3.25	3.38	3.13	3.05
c2670 (unbalanced)	6	6	9	9	4.26	4.15	1.04	4.45	0.39	0.42	0.94	1.04	0.11	0.26	0.11	0.09	2.9	3.09	2.96	3.09
c2670 (balanced)	24	31	74	84	3.38	3.31	1.52	4.21	0.37	0.4	0.72	0.74	0.3	0.38	0.28	0.3	3.15	3.34	3.13	3.14

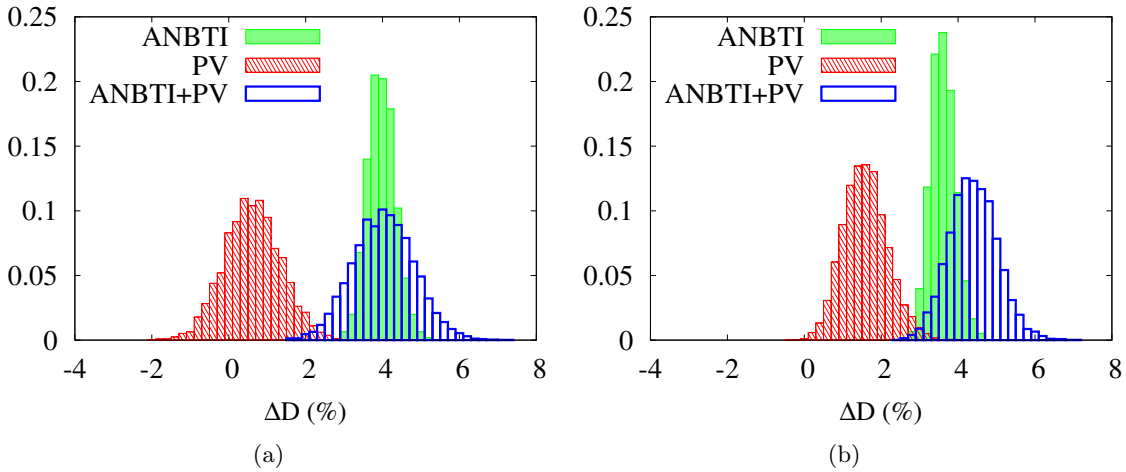


Figure 5.3.: ΔD distribution of a) c880 and b) c1355 considering atomistic NBTI (ANBTI), PV and combined effects (ANBTI+PV)

inequalities are valid at the circuit level when the combined effect is considered:

$$\begin{aligned} \mu_{PV} < & \mu_{ANBTI} < \mu_{ANBTI+PV} < \mu_{ANBTI} + \mu_{PV} \\ & \sigma^2_{ANBTI} < \sigma^2_{PV} < \sigma^2_{ANBTI+PV} \\ & \sigma^2_{ANBTI+PV} \neq \sigma^2_{ANBTI} + \sigma^2_{PV} \end{aligned}$$

which means that considering these two sources of variation separately, always leads to an overestimation of the mean value of the ΔD distribution (17% on average) and an error (for some circuits it leads to an underestimation and for others an overestimation) in the estimation of standard deviation (4% error in average).

Figure 5.4 shows the error in the timing margin estimation if these sources of variation are

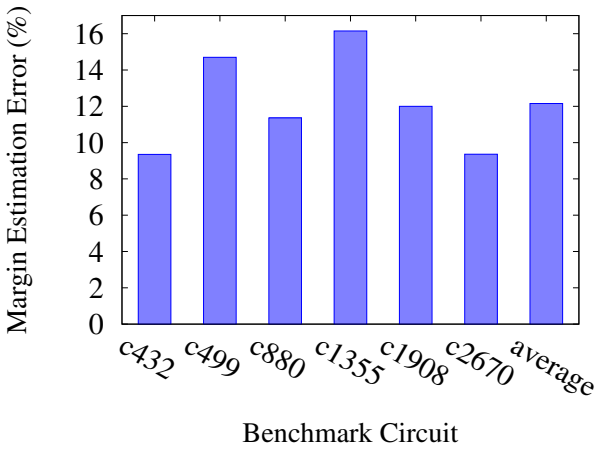


Figure 5.4.: Timing margin error due to separately consideration of process variation and stochastic NBTI compared to combined effect

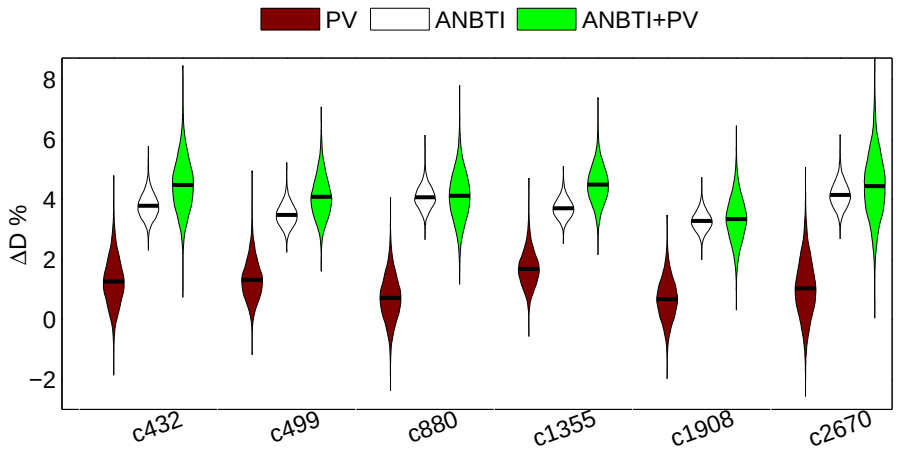


Figure 5.5.: Violin plot of NBTI and process variation-induced ΔD

considered separately and the error is calculated according to the following equations:

$$margin_{seperated} = \mu_{ANBTI} + \mu_{PV} + 3\sqrt{(\sigma^2_{ANBTI} + \sigma^2_{PV})}$$
$$margin_{combined} = \mu_{ANBTI+PV} + 3\sigma_{ANBTI+PV}$$
$$Error = \frac{margin_{seperated} - margin_{combined}}{margin_{combined}}$$

According to the Figure 5.4, considering these two sources of variation separately lead to an overestimation of the timing margin for guard-banding by 13% on average.

Figure 5.5 shows the violin-plot of the ΔD distributions for different benchmark circuits. According to this figure, NBTI and process variation lead to different amount of variation on the circuit delay for different benchmark circuits. The combined effect of NBTI and process variation can lead to a worst-case delay degradation ranging between 6-8%.

5.5.4. Effect of balanced paths in complex circuits

According to Table 5.1, when the number of cells in the inverter chain increases, the ΔD distribution becomes closer to a normal distribution (see skewness and Kurtosis of chain1, chain9 and chain39 in the table). In other words, for the longer paths, ΔD distribution becomes closer to a normal distribution which is consistent with *central limit theorem* (see Figure 5.6 (a)-(c)).

However, more than one path contribute to the ΔD distribution of the circuit (see # of CP in the table) in a typical circuit. Therefore, reporting only the ΔD distribution of one path [119] may reduce the accuracy of the timing analysis. This is because the delay of the circuit is obtained by a *maximum* operation across the delay of all the critical paths. Since the *maximum* operation is not a linear function, the maximum of two normal distributions is not necessarily a normal distribution. According to the table, for more balanced circuits (the ones with more critical paths) the skewness of NBTI-PV induced ΔD distribution is larger (see Figure 5.6(d)). To show the effect of balanced paths more clearly, we synthesized c1908 and c2670 circuits in two different ways i) unbalanced, by putting a loose timing constraint for synthesis and ii) balanced, by putting a very tight timing constraint for synthesis. According to the table, the skewness of the ΔD distribution in balanced circuit is 10X and 3X larger than that of unbalanced one for c1908 and c2670, respectively.

5.5.5. Effect of workload

Figure 5.7 illustrates the effect of workload on the amount of ANBTI-induced delay degradation. For this purpose, the signal probabilities of all internal inputs are swept from 0.0 to 1.0 with a step of 0.25. Figure 5.7(a) shows the results for a simple inverter. As shown in this figure, by increasing the signal probability (decreasing equivalent duty cycle), the mean and standard deviation of delay distribution decrease. This is consistent with Figure 2.16 in which the number of traps increases as the duty cycle increases. However, as shown in Figure 5.7(b), the workload has negligible effect on the distribution of ANBTI-induced delay degradation of a chain of inverter with 10 inverters in a row. This is due to the structure of inverter chain circuit in which the average duty cycle of internal transistor is approximately equal to 0.5 for all values of input signal probability and as a result, the input signal probability has negligible effect on the distribution delay degradation. Figure 5.7(c) and 5.7(d) shows the results for two different benchmark circuits of c880 (with a balanced structure) and c499 (with an unbalanced structure). According to these figures, the effect of input signal probability on the delay degradation distribution is a strong function of circuit structure.

To further investigate the effect of the circuit structure, the effect of the workload on the amount of ANBTI-induced delay degradation for two versions (balanced and unbalanced) of a similar benchmark circuit (c2670) is obtained and the results are illustrated in Figure 5.8. According to the figure, the sensitivity of the NBTI-induced delay degradation to the input

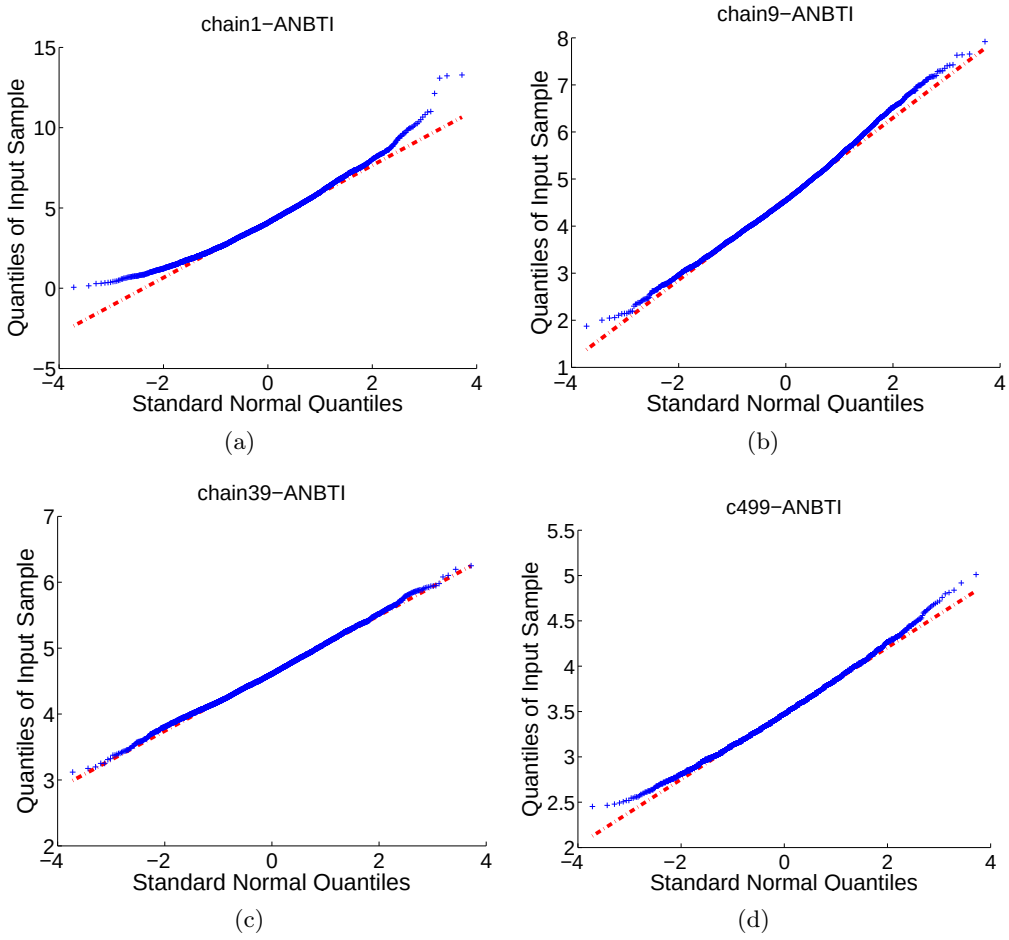
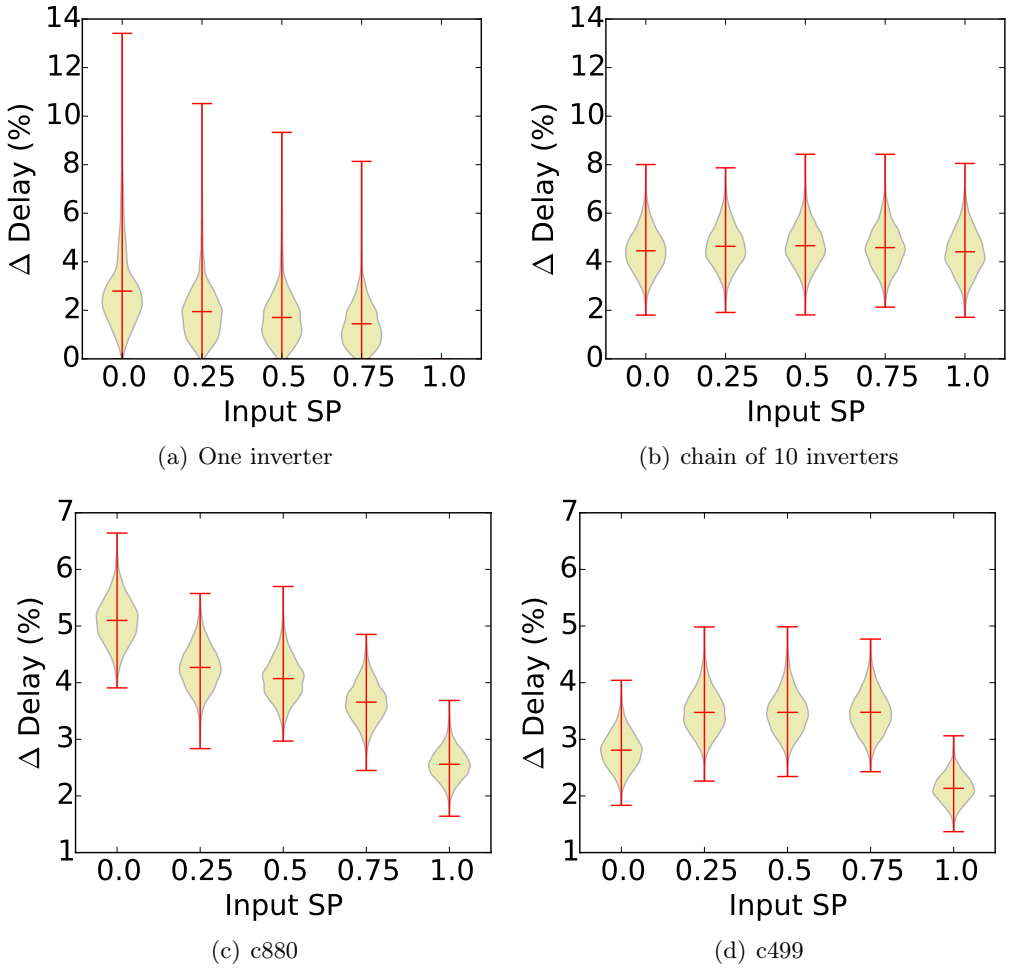


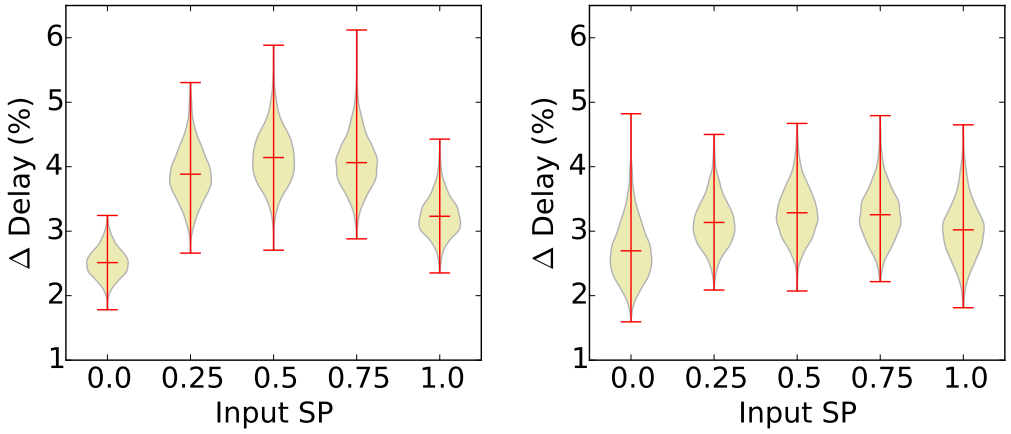
Figure 5.6.: QQ-plot of different circuits with different number of levels and critical paths



(a) One inverter (b) chain of 10 inverters

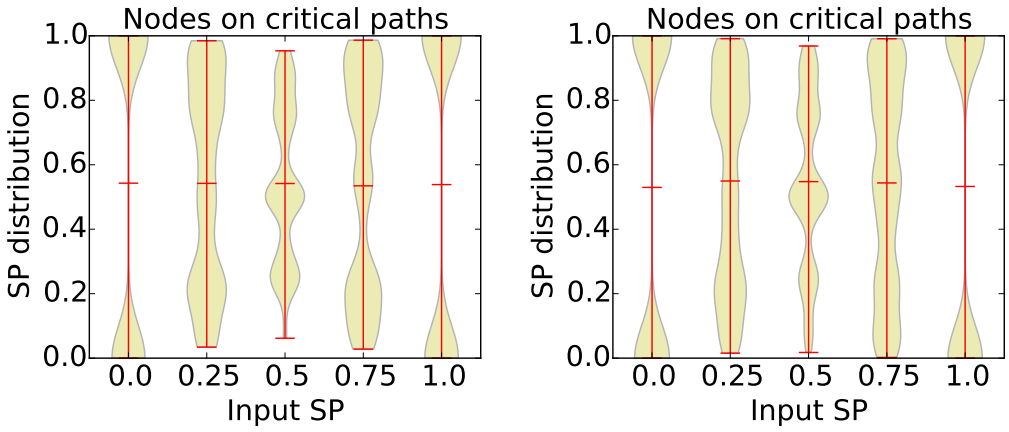
(c) c880 (d) c499

Figure 5.7.: Effect of workload on ANBTI-induced ΔD distribution of the circuit



(a) unbalanced c2670 (b) balanced c2670

Figure 5.8.: Effect of workload on ANBTI-induced ΔD distribution of c2670 circuit for a) unbalanced and b) balanced versions



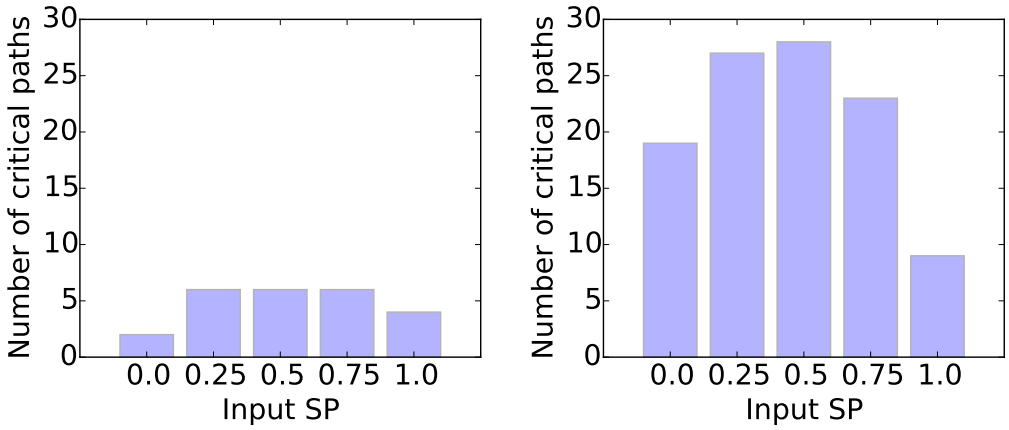
(a) unbalanced c2670 (b) balanced c2670

Figure 5.9.: Effect of workload on signal probability distribution of the nodes on critical paths for c2670 circuit for a) unbalanced and b) balanced versions

signal probability for the unbalanced version of the circuit is larger compared to that of balanced one.

In order to explain this observation, first, we obtained the signal probability distribution of the internal nodes, placed on the most critical paths, for these two versions of the circuit. The results are illustrated in Figure 5.9. According to the figure, the mean and the shape of the signal probability distributions are very similar for both versions of the circuit and hence, the difference between the sensitivity of the NBTI-induced delay degradation to the input signal probability is not related to this factor.

We also obtained the number of "aging" critical paths, contributing in the post-aging delay distribution of the circuit. As depicted in Figure 5.10, the number of aging critical paths is different for various workloads (input signal probabilities). Moreover, the number of aging critical paths is much larger in balanced version of the circuit compared to that of unbalanced one. Since the delay of the circuit is obtained by a maximum operation across the delay of all the aging critical paths, the circuit delay distribution is affected by the number of aging critical paths. Figure 5.11 show the dependency of the mean value of the maximum of n



(a) unbalanced c2670 (b) balanced c2670

Figure 5.10.: Number of critical paths for c2670 circuit for a) unbalanced and b) balanced versions

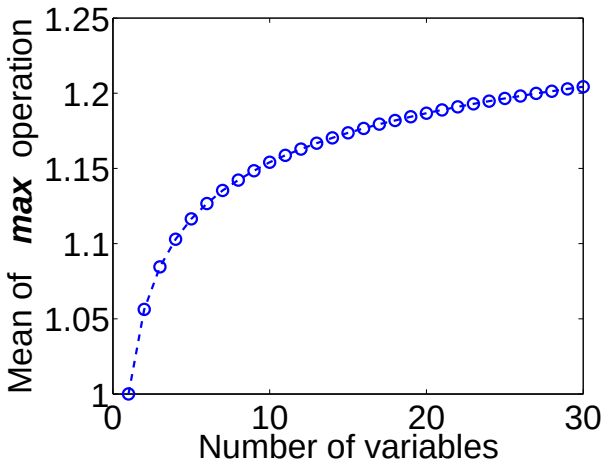


Figure 5.11.: The mean value of the maximum of n random variables ($E[\max_{i=1}^{n} V_i]$). The random variables have a mean value of 1 and standard deviation of 0.1. The values of standard deviation and mean are set according to the maximum NBTI induced delay degradation of a single path which is less than 10%.

random variables ($E[\max_{i=1}^{n} V_i]$) for different values of n. Based on the figure, by increasing n the mean value of the maximum increases, however, the mean value saturates for large amount of variables ($n > 10$). Therefore, for balanced version of the circuit, the sensitivity of the post-aging delay distribution to the number of aging critical paths, and hence to input signal probability, is much less compared to that of unbalanced one.

It should be noted that, the number of *near critical paths* (the paths with slack less than 5% of the circuit delay) at design time of the balanced version of c2670 is more than 800, but since the NBTI effect asymmetrically impacts the circuit, the number of post-aging critical paths is reduced to less than 30 paths for this version of the circuit. For the unbalanced version, these numbers are 56 and 6, respectively. This means that even for very large circuits, if the design of the circuit is not well-balanced, the number of post-aging critical path might become very small, leading to a high sensitivity of NBTI-induced delay degradation to the executed workload.

5.5.6. Runtime of the proposed variation-aware timing analysis

Table 5.2 shows the runtime of the proposed variation-aware timing analysis for 10000 MC samples. As shown in this table, the runtime of the technique increases linearly with the size of the circuit. It should be noted that this framework is not optimized for runtime but it enables us to perform a comprehensive analysis of stochastic NBTI and process variation at circuit

Table 5.2.: Runtime of proposed variation-aware timing analysis

Benchmark	# of gates	runtime (s)
c17	6	3,450
c432	220	15,508
c499	539	32,600
c880	566	29,240
c1355	585	38,202
c1908	610	34,952
c2670	1,101	57,192

level with a high accuracy by abstracting detailed device-level atomistic models to logic-level circuit timing analysis. The results of this analysis can be used as a baseline for faster (but less accurate) timing analysis at higher level of abstraction or better scalability for larger circuits.

Moreover, there are multiple ways to improve the existing flow. One way is to decrease the number of MC iterations by leveraging smart approaches such as importance sampling or quasi MC simulation [123]. The other opportunity is to decrease the runtime of each MC simulation (iteration). For this purpose, the actual gate delay can be updated for each MC iteration according to the values of ΔVth samples (e.g. using Standard Delay File (SDF)) instead of updating the entire gates delay LUTs and performing STA for each case.

5.6. Summary

By further down-scaling, the NBTI effect becomes stochastic which has a significant impact on guard-banding for aging. This chapter presented a framework to analyze the effect process variation and stochastic NBTI on the timing of logic-level circuits. Simulation results show that the stochastic behavior of NBTI can result in a significant increase in the guard-band by up to 30% compared to the deterministic case. Moreover, results reveal that the effect of process variation on the mean of the delay degradation is less than that of stochastic NBTI while it leads to a more variability on the distribution. In general, the analysis reveals that there is a need to consider the stochasticity as a part of the standard timing analysis flow and the stochastic nature of NBTI mandates a more complex guard-banding in contrast to existing static guard-banding solutions.

Part III.

Reliability-aware Cell and Circuit Design

CHAPTER
SIX

RELIABILITY-AWARE STANDARD CELL LIBRARY DESIGN

6.1. Overview

While the focus of the previous chapters was the modeling of reliability, the focus of the following chapters is to design reliable cells and circuits in order to mitigate the aging effect.

Typically, standard cells in the library are optimized according to the design time delay, however, due to the asymmetric effect of BTI, the rise and fall delays might become significantly imbalanced over the lifetime. In this chapter, a method is proposed to mitigate the BTI effect by balancing the rise and fall delays of the standard cells at the excepted lifetime. We find an optimal tradeoff between the increase in the size of the library and the lifetime improvement (timing margin reduction) by non-uniform extension of the library cells for various ranges of the input signal probabilities. The simulation results reveal that our technique can prolong the circuit lifetime by around 150% with a negligible area overhead. Moreover, we investigate the effect of different realistic workloads on the distribution of internal node signal probabilities. This is done to obtain the sensitivity of our static (design time) approach to different workloads during system lifetime. The results show that our proposed approach is still efficient if the workload changes during the runtime.

The rest of the chapter is organized as follows. Related work is discussed in Section 6.2. The proposed BTI-aware cell sizing idea is explained in Section 6.3. Afterwards, the proposed methodology is described in Section 6.4. Simulation results are presented in Section 6.5 and finally, Section 6.6 concludes the chapter.

6.2. Related work

In order to mitigate BTI, there are two main categories of techniques: 1) sense and adapt (at runtime) and 2) model, predict, and margin (at design time). In the former approach, the circuit behavior has to be monitored at runtime and according to the feedback of monitors an adaptive technique is applied to compensate the degradation due to aging. Body biasing [124–126], clock and power gating [127–129], and dynamic voltage and frequency scaling (DVFS) [130, 131] are some of the well-known methods in this category [132]. In the later approach, the proper operation of the circuit at the expected lifetime is guaranteed by a *guard-banding* method [53], i.e. adding additional timing margin. For effective guard-banding, the timing margin has to be accurately predicted.

The guard-banding method can be combined with aging mitigation techniques at design and run-time such as gate sizing [133–137] and input vector control [138–142] in order to improve lifetime or reduce the amount of required timing margin. Input vector control is a design

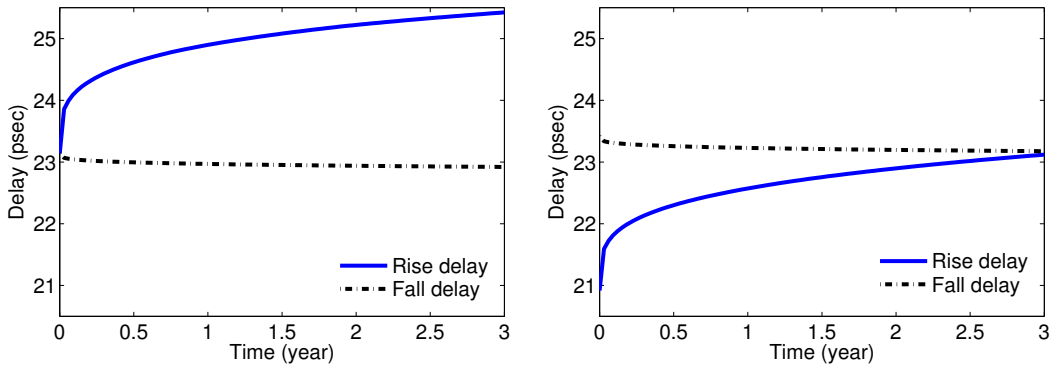
time technique in which a suitable input vector is applied during standby-mode of the circuit to maximize the recovery phase of aging critical gates. Gate sizing is another design time technique to mitigate the aging effect. In this approach, the gates placed in aging critical paths are upsized in order to prevent timing failures due to aging. For this purpose, the larger gates, which are available in the standard cell library, are used. However, since the aging effect is not considered in the standard cell library design, the area and/or power overhead of upsizing is high. Therefore, there is need to redesign/extend the standard cell library in order to reduce the overheads of gate sizing method.

In the standard cell library design, the transistors in each gate are sized in a way that the rise and fall delays become equal for a typical load capacitance and transition time [143]. However, BTI asymmetrically affects the rise and fall delays of the gates according to their transistors duty cycle (which in turn is a function of its input *Signal Probability (SP)* [47]). In other words, the rise and fall delays of a cell become significantly imbalanced during operational time.

To address this issue, one approach is to size the transistors in an efficient way to reduce aging effect. A transistor level sizing approach based on Lagrangian relaxation technique is proposed in [144]. However, it is not based on standard cell library design and it gives different sizing for each gate inside the circuit which makes it infeasible for standard cell-based design. In [145], an NBTI/process variation aware standard cell design method is presented. However, the effect of input SP is neglected and a constant SP of 50% is assumed for NBTI-induced Vth shift calculation. An NBTI-aware basic cell design is also proposed in [146]. However, no proper methodology to consider the effect of input SPs (uneven BTI-induced degradation) is provided. As will be shown in our simulation results, neglecting the effect of SP impacts the efficiency of such method especially in advanced technology nodes where both NBTI and PBTI matter.

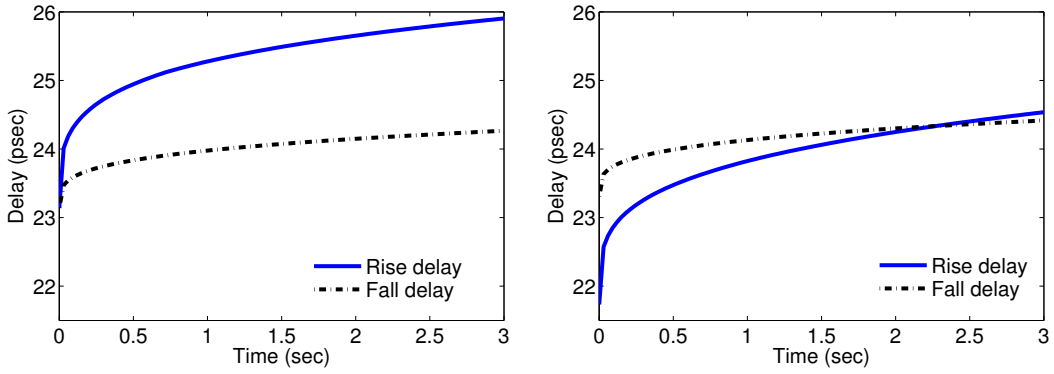
6.3. Aging-aware cell sizing

In the typical library cell design, the optimal ratio of Wp (width of the PMOS transistor) to Wn (width of the NMOS transistor) is adjusted in order to balance the rise and fall delays of the gate [143]. However, due to BTI effect, the threshold voltage of transistors degrades unevenly leading to unequal rise and fall delays of the gates at the end of the expected lifetime. Figure 6.1(a) shows the rise and fall delays of a simple inverter over the time for a duty cycle equal to 0.5 considering only NBTI effect. As shown in this figure, although the rise and fall



(a) Time-zero-balanced (b) Lifetime-balanced

Figure 6.1.: Effect of Wp/Wn optimization on NBTI-induced delay degradation of an inverter with input SP=0.5



(a) Time-zero-balanced (b) Lifetime-balanced

Figure 6.2.: Effect of Wp/Wn optimization on BTI-induced delay degradation of an inverter with input SP=0.1

delays are equal at design time ($time = 0$), they diverge after 3 years. This is due to the fact that NBTI effect leads to an increase in the threshold voltage of PMOS transistor leading to an increase in the rise delay and a decrease in the fall delay of the gate.

Our objective is to design the cell (by changing the Wp/Wn ratio), in a way that its rise and fall delays become equal at the end of the expected lifetime (see Figure 6.1(b)). As shown in this figure, by optimizing Wp/Wn ratio to balance rise and fall delays at expected lifetime, at the expense of upsizing only PMOS transistor in the gate, a better post-aging delay is achieved.

Figure 6.2 shows similar results but considering BTI (NBTI and PBTI) effect for an inverter with input SP of 0.1. Since the *duty cycle* (DC) of the PMOS transistor ($DC_{NBTI} = 1 - SP_{in} = 0.9$) is higher than that of the NMOS transistor ($DC_{PBTI} = SP_{in} = 0.1$), the rise delay degradation is higher than the fall delay degradation. As a result, the Wp/Wn ratio has to increase compared to the typical mode in order to have the same rise and fall delays at the expected lifetime.

Since BTI effect is a function of DC (see Figure 2.12(b), the optimized BTI-aware Wp/Wn ratio for each cell is a function of input duty cycle (and hence SP). Figure 6.3 shows the optimized NBTI-aware Wp/Wn ratio for different input SPs normalized to the case that aging effects are not considered. Since for the NBTI case $DC_{NBTI} = 1 - SP_{in}$, for smaller SPs the NBTI effect is larger and, as a result, the PMOS transistor in the pull-up network has to be designed larger to compensate NBTI effect.

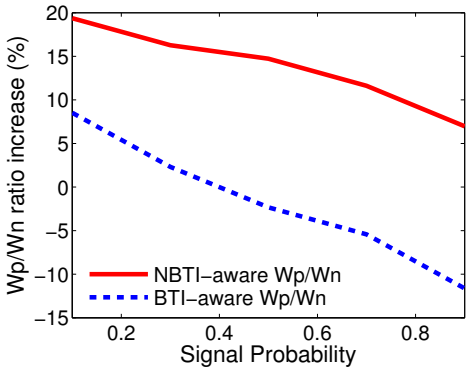


Figure 6.3.: Optimized Wp/Wn ratio increase for different signal probabilities normalized to the case that it is optimized for time=0

Figure 6.3 also shows BTI-aware Wp/Wn ratio but for the case where both NBTI and PBTI effects are considered. As shown in this figure, the optimized Wp/Wn ratio is more sensitive to SP compared to the case in which only NBTI effect is considered. Another observation is that, for the case of BTI effect, if $SP = 0.5$ the Wp/Wn ratio is almost equal to the case that aging effect is neglected. In fact, when $SP = 0.5$, both NMOS and PMOS transistors are almost under the same stress and as a result both transistors degrade at almost the same pace. Therefore, their ratio in this case is almost equal to the typical case. Moreover, for the larger SPs the Wp/Wn ratio is less than the typical case. Since we consider a constant Wn and for larger SPs ($SP > 0.5$) the NMOS transistor degrades more than PMOS transistor, the fall delay becomes larger than the rise delay and hence we can decrease Wp to make the rise and fall delays equal in order to save area and power. It should be noted that, in this case we only gain area/power, however, BTI is not mitigated.

The efficiency of this approach is demonstrated in the following by the example circuit given in Figure 6.4. The circuit is an inverter chain with the primary input SP of 0.1.

Figure 6.4(a) shows the rise and fall delays when the time-zero-balanced (typical) library cells are used. As shown in this figure, although the rise and fall delays of the path are the same at design time, they diverge significantly throughout the lifetime. There are two reasons for this significant imbalance during operational time:

1. The rise delays of inverter 1 and 3 become larger than their fall delay during operational time ($Dr(inv1) > Df(inv1)$ & $Dr(inv3) > Df(inv3)$ at time=3 years) as these inverters suffer more from NBTI effect rather than PBTI.

2. The situation for inverters 2 and 4 is the opposite since their fall delays increase over the lifetime as they are mostly under PBTI stress. ($Df(inv2) > Dr(inv2)$ & $Df(inv4) > Dr(inv4)$ at time=3 years).

Therefore, the total rise and fall delays of the path which can be obtained by Equation (6.1) become significantly imbalanced.

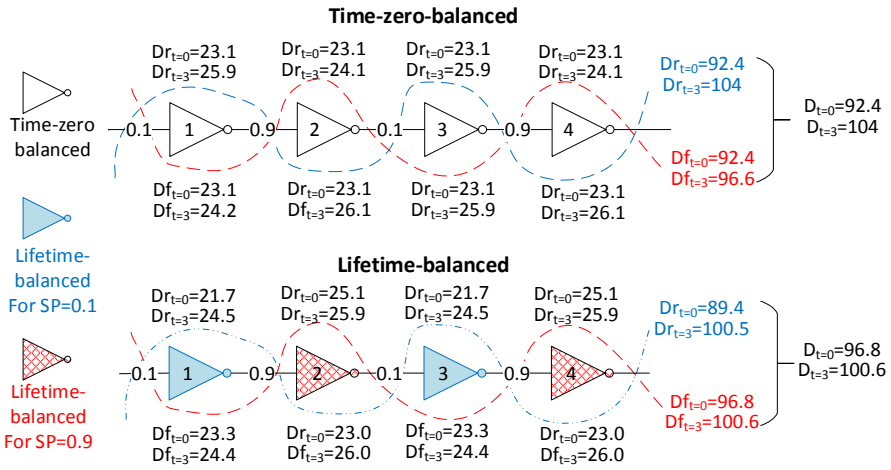
$$\begin{aligned} D_{rise} &= D_f(inv1) + D_r(inv2) + D_f(inv3) + D_r(inv4) \\ D_{fall} &= D_r(inv1) + D_f(inv2) + D_r(inv3) + D_f(inv4) \end{aligned} \tag{6.1}$$

Figure 6.4(a) also shows the rise and fall delays when the lifetime-balanced library cells are used. As shown for this case, the rise and fall delays of the path become similar at the end of the operational lifetime of the circuit and the overall delay (100.6ps) is less than the case in which the time-zero-balanced library cells are used (104ps).

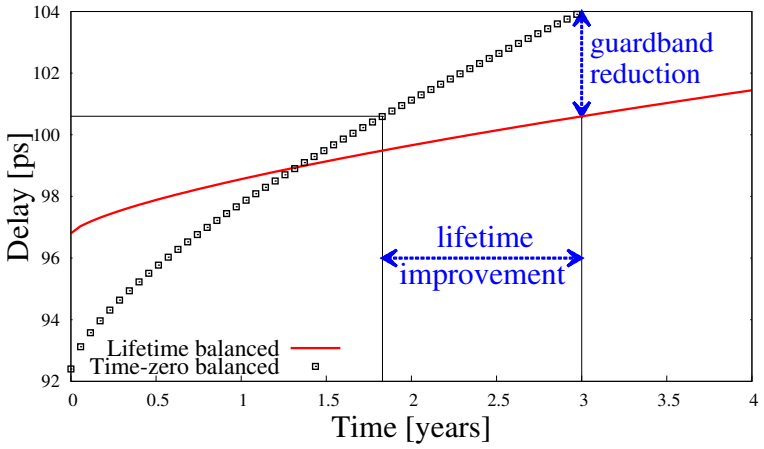
$$D_{lifetime-balanced}(100.6ps) < D_{time-zero-balanced}(104ps) \tag{6.2}$$

This is obtained by the upsizing of PMOS transistors for the inverters under NBTI stress (higher Wp/Wn ratio for inverters 1 and 3 with smaller SPs according to Figure 6.3). On the other side, this upsizing is compensated (in terms of power and area) by downsizing the PMOS transistors of the other inverters (inverters 2 and 4) which are more under PBTI stress (lower Wp/Wn ratio for larger SPs according to Figure 6.3). Due to the downsizing, the sum of transistors width (which is a representative for area and power) for the lifetime-balanced library cells is equal to that of time-zero-balanced scenario while it has better delay after 3 years.

Based on the observation above, in order to reduce aging effect, we can optimize the Wp/Wn ratio of different gates (i.e. library cell redesign) according to their input SPs, in order to make their rise and fall delays equal at their expected lifetime.



(a)



(b)

Figure 6.4.: A simple circuit to show the efficiency of aging-aware standard cell sizing: a) time-zero-balanced vs lifetime-balanced mapping b) delay of lifetime-balanced vs time-zero balanced

In order to address different types of variability (e.g. transistor aging), a safety margin (guard-band) is added to the design to guarantee the reliable operation of the designed circuit. Therefore, the overall clock cycle is obtained by the following equation:

$$Tclk = D_0 + GB \tag{6.3}$$

where D_0 is the time-zero delay, GB is the guard-band and $Tclk$ is the clock period. According to Figure 6.4(b), although the delay of the circuit at time zero may even become larger, it becomes smaller at the expected lifetime:

$$\begin{aligned} D_0^{LTB} &> D_0^{TZB} \\ GB^{LTB} &< GB^{TZB} \\ D_0^{LTB} + GB^{LTB} &< D_0^{TZB} + GB^{TZB} \\ &\Longrightarrow Tclk^{LTB} < Tclk^{TZB} \end{aligned} \tag{6.4}$$

where LTB and TZB are the abbreviations for lifetime-balanced and time-zero balanced, respectively. Therefore, with the reduction in the amount of aging-induced timing margin, the clock period is reduced in overall.

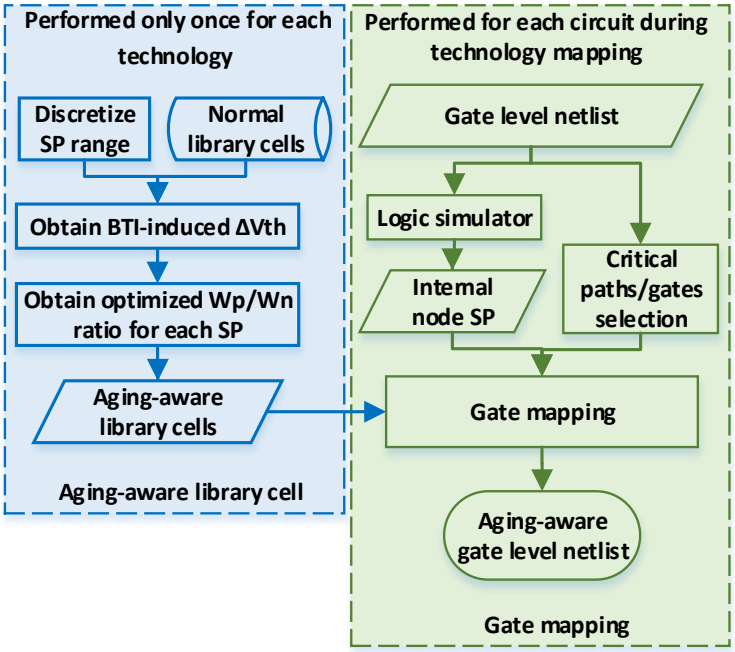


Figure 6.5.: Overall flow of proposed aging-aware standard cell library design

6.4. Cell library redesign and mapping

Figure 6.5 shows the overall flow of the proposed methodology. It consists of two phases: i) aging-aware standard cell library redesign, and ii) circuit library mapping using the new library cells. The first phase is the aging-aware standard cell library redesign in which the library cells are optimized for different SPs considering BTI effect. This step is done only once for each technology in order to build the aging-aware library. The second phase is circuit library mapping using the new library cells. The gate level netlist is given to a logic simulator to obtain the SPs of all internal nodes. According to internal node SPs, the gates are replaced with optimized aging-aware designed library cells. The details of each phase are explained next.

6.4.1. Aging-aware cell library

We propose aging-aware standard cell library redesign, in which the library cells are optimized for different SPs considering the BTI effect. According to Figure 6.3, the optimized Wp/Wn ratio is a function of the SPs of cell inputs. However, it is not possible to extend the library for all combinations of SPs. For this purpose, the SP range ([0.0, 1.0]) is discretized and for each combination of these SP values a new library cell is added and optimized by finding a suitable Wp/Wn ratio for that range using SPICE simulations.

In order to obtain the optimized Wp/Wn ratio, first the BTI-induced ΔVth for all internal transistors of the cell is calculated according to the particular SP value. Then, the Wp/Wn ratio of the cell is swept using a binary search to obtain the best ratio leading to equal rise and fall delays. For example, if we discretize the SP range to {[0.0 0.2), [0.2 0.4), [0.4 0.6), [0.6 0.8), [0.8 1.0]}, then for a simple inverter ($INVX1$) we need to extend the library with five additional cells: {$INVX1\_0.1$, $INVX1\_0.3$, $INVX1\_0.5$, $INVX1\_0.7$, $INVX1\_0.9$} and for each cell the Wp/Wn are obtained to have equal rise and fall delays at the expected lifetime. Here, {0.1, 0.3, 0.5, 0.7, 0.9} are the representative SPs for each range and the library cell for each range is optimized according to its representative SP value. In order to build the

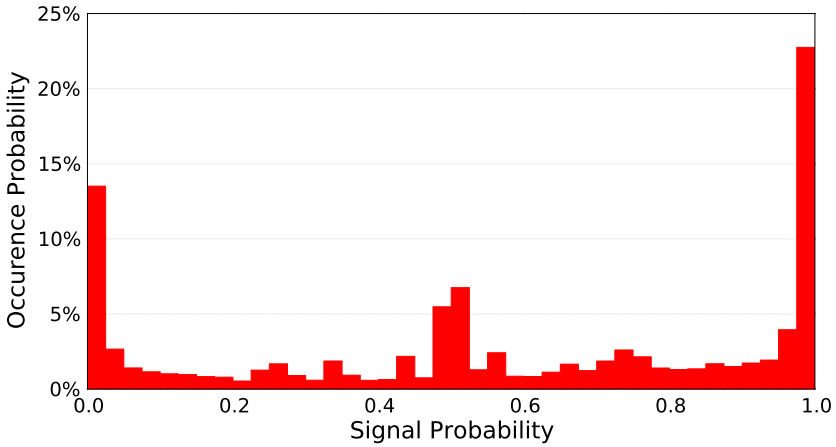


Figure 6.6.: The histogram of internal node SP distribution for ISCAS89 benchmark circuits (over all benchmarks)

library, each cell is characterized to obtain the delay and leakage Look-up tables (LUTs). By increasing the library size, the characterization time/effort increases accordingly, however, it should be noted that the aging-aware library cell design and characterization are done only once for each technology.

Library size increase and non-uniform SP sampling: If the SP range ([0.0 1.0]) is discretized to m intervals, for each cell with n inputs, m^n cells are added to the library. In other words, the number of new library cells increases exponentially with m. More sampling points for the input SP range may increase the efficiency of this approach in terms of delay balancing, however, it leads to a very large size of the library. This makes the approach infeasible for industrial-scale libraries which contain more than 1000 cells. Moreover, a high resolution of SP sampling makes the technique very sensitive to process variations. This implies that a suitable discretization resolution has to be considered for a reasonable trade-off between the efficiency of the method and the library size. For this purpose, two important parameters have to be considered:

i) The sensitivity of BTI-induced Vth shift to SP: As shown in Figure 2.12(b), the BTI-induced Vth shift has different sensitivities to the SP (duty cycle) in different range of SPs. Therefore, more samples (at least one sample) has to be considered for more sensitive ranges (e.g. [0.0 0.1] range in Figure 2.12(b)).

ii) Distribution of the SPs of internal nodes in typical circuits: The SPs of the internal nodes in typical circuits are not uniformly distributed over the entire range ([0.0 1.0]). Figure 6.6 shows the histogram of SP distribution for internal nodes of ISCAS89 benchmark circuits. As shown in this figure, SP values around 0.0, 0.5, and 1.0 are more frequent. Therefore, in a non-uniform sampling, more samples have to be chosen in ranges where the probability of occurrence in the typical circuits is higher.

Considering these two factors, a non-uniform discretization and sampling can be used in order to keep the sampling points as few as possible, while maintaining a high efficiency for this technique in terms of aging mitigation.

6.4.2. Technology mapping using aging-aware standard cell library

Once the aging-aware cell library is constructed, it can be used for the technology mapping phase for different circuits. In order to obtain suitable standard cells for each circuit, we start from a netlist mapped into the original aging-unaware library. Then, the gate level netlist is given to a logic simulator to obtain the internal SPs. According to the input SPs of each

gate, a new cell with the closest set of input SPs from the new library is chosen to replace the initial cell. For example, if we have a two-input NAND gate with the SPs of 0.15 and 0.73 for its inputs, according to the discretization example of previous subsection, this NAND gate will be replaced with the $NAND\_0.1\_0.7$ aging-aware cell. In order to minimize the area/power overhead, this remapping is done only for the critical gates (gates which are in the critical/near-critical paths) since the others have no contribution to the delay of the circuit.

6.5. Simulation results

In this section, we show the efficiency of our proposed method by comparing it to the time-zero-balanced library cell design as well as scenarios in which a representative SP is considered for all gates. We also investigate the tradeoff between library size increase and aging mitigation by considering various sample sizes and strategies. The impact of different workloads is also discussed.

6.5.1. Simulation setup and flow

Figure 6.7 shows the details of our flow to obtain the simulation results. The gate-level netlist, is obtained by synthesizing the ISCAS89 benchmark circuits using Nangate 45 nm library [22] containing 42 cells (INVERTER, BUFFER and two inputs AND, OR, NAND, NOR, , XOR, and XNOR gates). The worst case BTI-induced delay degradation is assumed to be 10% in 3 years and the parameters of the BTI model are set accordingly using the deterministic model explained in Section 2.4.1.

The first step is to conduct the aging-aware library cell design as proposed in Section 6.4.1. Here, we consider four different scenarios for the discretization of the SP range:

1) Uniform sampling with 5 points (U5): In this scenario, the SP range is discretized uniformly to 5 ranges: {[0.0 0.2), [0.2 0.4), [0.4 0.6), [0.6 0.8), [0.8 1.0]} with representative sampling points of 0.1, 0.3, 0.5, 0.7, 0.9. The number of logical standard cells (AND, OR, BUFFER, NAND, NOR, INVERTER, XOR, and XNOR) is increased by around 50X (from 42 to 2010 cells). Such an increase in the library size makes it infeasible to be used for industrial applications.

2) Non-uniform sampling with 3 points (NU3): Another option for SP discretization is to use a non-uniform sampling. For this purpose we put more SP samples in the range that BTI is more sensitive to SP changes (for example the range [0.0 0.1] according to Figure 2.12(b)). Another important factor which has to be considered is the SP distribution for different logic gates of a typical circuit. For this scenario, based on Figure 6.6, we only consider 3 samples for SP ($\{0.1, 0.5, 0.9\}$) for 3 ranges of {[0.0 0.2), [0.2 0.8), [0.8 1.0]} in order to decrease the library size in comparison with the previous scenario. In this case, the library size consists of 522 cells which is almost 4 times smaller than that of the previous scenario.

3) Non-uniform sampling with 2 points (NU2): For this case we only consider 2 samples of SP ($\{0.1, 0.9\}$) for two ranges {[0.0 0.5), [0.5 1.0]} in order to further decrease the library size. In this case, the aging-aware library consists of 192 cells (more than 10 times reduction compared to the first scenario).

4) Non-uniform worst case sampling with 2 points (NU2W): For this case we also consider 2 samples of SP ($\{0.1, 0.5\}$) for two ranges {[0.0 0.5), [0.5 1.0]}. The library size is equal to previous case, however, in this case all the gates with input SP larger than 0.5 are mapped to cell with SP=0.5 in order to upsize the PMOS transistors (according to Figure 6.3) to further reduce BTI-degradation compared to NU2.

In order to obtain the results for the case in which the effect of SP is neglected in library cell design (e.g. the method in [146]), we considered two cases. In the first case, we assumed that

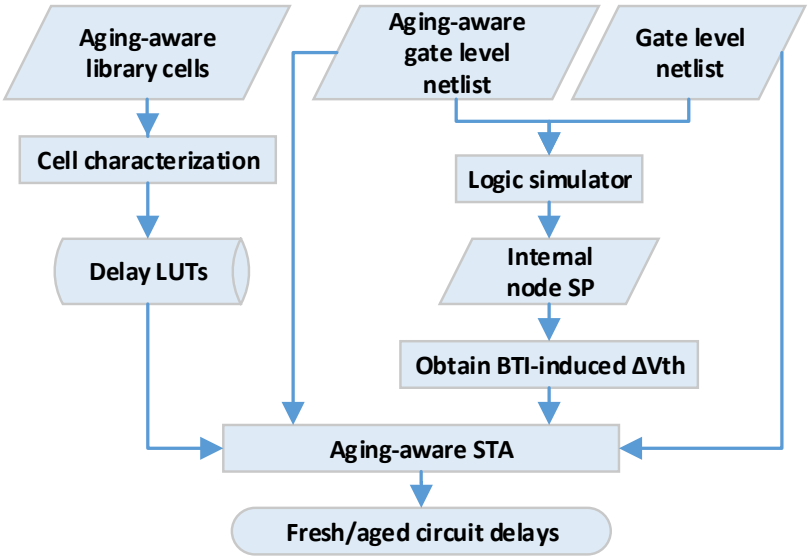


Figure 6.7.: Overall flow to obtain simulation results

all cells are optimized with input SP of 0.5. For the second case, we considered a worst-case approach. For this purpose, according to Figure 6.3, we consider very small SP of 0.1 in order to upsize the PMOS transistors to mitigate BTI.

We used accurate SPICE simulations in order to optimize the standard cells by finding the best Wp/Wn ratio for all scenarios. After obtaining the new aging-aware cells, each cell is characterized to obtain delay LUTs for different load capacitances, transition times, and ΔVth. This means that for each standard cell we have generated $n_T + 2$ dimensional LUTs, where n_T is the number of transistors inside that gate, and the two other dimensions are related to the load capacitance and the transition time. Since the SP range ([0.0 1.0]) is discretized and for each range of SPs a suitable standard cell is added to library, the characterization step has to be done only once in order to obtain the delay LUTs for each standard cell.

Besides, the benchmark circuit is synthesized to obtain the gate-level netlist. In the mapping phase, only the critical gates in the gate-level netlist are replaced with the new cells, as described in Section 6.4.2, to obtain the aging-aware gate-level netlist. Here, the critical/near critical paths which have the delay more than 90% of circuit delay are selected in order to find the critical gates. Next, the netlist mapped into the original library and the one mapped into the aging-aware library are given to a logic simulator to obtain the SPs of internal nodes. Then the BTI-induced ΔVth of all the transistors are obtained according to the model proposed in [47]. Finally, the gate level netlists, ΔVth values, and the delay LUTs are given to an aging-aware Static Timing Analysis (STA) tool, similar to the one proposed in [147], to obtain the fresh and aged circuit delays. For the aging-aware STA, a method similar to the one proposed in [147] is employed.

6.5.2. Aging mitigation

The simulation results are shown in Table 6.1. The area is approximated by the summation of all transistor widths inside the circuit. Therefore, it is not only a representative for the area, but also shows the trend of the power consumption. As shown in this table, uniform sampling with 5 points (U5) and non-uniform sampling with 3 points (NU3) scenarios lead to 26.7% and 25.4% timing margin reduction (167% and 155% lifetime improvement), respectively. This implies that while NU3 needs much smaller library size compared to U5, their efficiencies

Table 6.1.: The efficiency of our technique compared to the normal standard cell library design in terms of lifetime improvement and area overhead. **U5**: Uniform sampling with 5 points ({0.1,0.3,0.5,0.7,0.9}), **NU3**: Non-uniform sampling with 3 points({0.1,0.5,0.9}), **NU2**: Non-uniform sampling with 3 points({0.1,0.9}), **NU2W**: Non-uniform worst case sampling with 2 points({0.1,0.5}),.

Benchmark	# of	Timing margin reduction				Lifetime improvement						Area overhead					
Circuit	gates	U5	NU3	NU2	NU2W	U5	NU3	NU2	NU2W	SP=0.5	SP=0.1	U5	NU3	NU2	NU2W	SP=0.5	SP=0.1
s400	228	35.7%	32.8%	29.7%	27.3%	239.3%	211.4%	183.4%	162.3%	-13.2%	79.6%	0.1%	-0.1%	0.1%	0.4%	-0.1%	0.8%
s444	249	39.2%	38.8%	35.3%	27.2%	275.0%	271.0%	235.1%	162.0%	-13.7%	82.4%	0.2%	-0.1%	-0.3%	0.3%	-0.1%	0.9%
s420	280	18.1%	13.9%	2.9%	10.6%	94.4%	68.3%	12.2%	49.4%	-14.5%	23.7%	0.0%	-0.1%	0.0%	0.5%	-0.1%	0.9%
s526	330	29.2%	34.2%	27.6%	17.2%	178.6%	224.3%	165.3%	88.4%	-8.1%	33.6%	0.2%	0.1%	0.0%	0.3%	-0.1%	0.6%
s510	373	18.7%	13.5%	0.9%	13.2%	98.2%	65.9%	3.7%	64.2%	-11.9%	36.0%	0.0%	0.0%	0.0%	0.2%	-0.1%	0.5%
s832	584	25.4%	24.3%	11.1%	18.0%	147.2%	139.0%	52.6%	94.1%	-12.6%	39.6%	0.0%	0.0%	0.1%	0.2%	0.0%	0.4%
s953	683	37.3%	39.1%	31.7%	20.3%	255.0%	274.3%	200.6%	109.7%	-11.7%	78.2%	0.0%	0.0%	0.0%	0.1%	0.0%	0.3%
s1196	797	23.4%	22.1%	12.0%	19.6%	132.0%	121.9%	57.5%	104.8%	-13.3%	61.2%	0.0%	0.0%	0.0%	0.4%	-0.1%	1.0%
s1423	824	31.8%	33.9%	25.5%	22.4%	201.8%	221.5%	147.9%	124.8%	-15.9%	93.8%	0.0%	-0.1%	-0.1%	0.3%	-0.1%	0.7%
s1238	881	52.3%	36.2%	25.8%	26.5%	438.5%	244.3%	150.1%	155.9%	-13.8%	127.4%	0.1%	0.0%	0.0%	0.1%	0.0%	0.3%
s1488	902	24.0%	21.0%	17.3%	14.8%	136.1%	114.3%	89.4%	73.4%	-2.8%	31.8%	0.0%	0.0%	0.0%	0.2%	0.0%	0.3%
s9234	1725	32.4%	29.5%	22.4%	20.5%	207.5%	181.2%	124.2%	110.8%	-11.3%	71.2%	0.0%	-0.1%	0.0%	0.3%	-0.1%	0.6%
s5378	2926	36.2%	38.3%	28.7%	28.7%	244.1%	265.5%	174.2%	174.4%	-13.4%	135.6%	0.0%	0.0%	0.0%	0.1%	0.0%	0.2%
s13207	4074	9.8%	29.4%	19.9%	19.9%	45.4%	180.7%	106.9%	106.7%	-10.6%	56.9%	0.3%	0.1%	0.0%	0.2%	0.0%	0.5%
s15850	5244	21.9%	23.7%	8.1%	13.6%	120.7%	133.8%	36.5%	66.8%	-8.4%	36.6%	0.0%	0.0%	0.0%	0.3%	-0.1%	0.6%
s38584	18142	26.2%	19.8%	14.4%	18.1%	153.8%	106.0%	71.5%	94.5%	-12.5%	52.6%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
average		26.7%	25.4%	14.2%	18.0%	167.2%	155.3%	85.5%	97.1%	-11.8%	56.8%	0.1%	0.0%	0.0%	0.3%	-0.1%	0.6%

in terms of lifetime improvement and timing margin reduction are comparable. This shows that the lifetime improvement saturates when the number of SP samples exceeds a particular limit. Both scenarios have negligible area overhead. There are two reasons for that: i) the aging-aware technology mapping is only performed for critical gates, and ii) for the gates with large input SPs, the PMOS transistors are down-sized in order to save area/power overhead (according to Figure 6.3) . The results for NU2 scenario shows that with a much smaller library size, 14% timing margin reduction (85% lifetime improvement) is obtained. However, NU2W gives better results compared to NU2 in terms of timing margin (lifetime improvement) with the same library size at the expense of a small area/power overhead (0.3%). This shows the importance of the proper SP sampling for aging-aware cell library design.

Compared to other alternative (in which the SP distribution of internal node is neglected [146]), considering a fixed SP of 0.5 leads to even worse lifetime compared to the original library cell design, although it might be beneficial only when NBTI effect is considered. For the other scenario (fixed SP of 0.1), the lifetime improvement is much less than all the four scenarios above. Moreover, for this worst case scenario, the area/power overhead is higher (0.6%).

To account for wearout mechanisms, the clock frequency has to be set according to delay of the circuit at the expected lifetime (not at t=0), by adding aging-induced timing margins. This means that the circuit performance is determined by the post-aging delay. Although our proposed method may even lead to a higher circuit delay at t=0 (up to 2%), it provides an overall performance improvement by reducing the post-aging delay and its associated timing margin, as shown in Table 6.1. For the case where the performance is fixed, the proposed technique results in an improvement in the circuit lifetime.

6.5.3. Library size

According to the results, using more sampling points results in better lifetime improvement at the expense of library size explosion. However, the non-uniform sampling with fewer points (e.g. NU3) provides comparable improvement with much reduced library size. For further reduction of library size, we found that NU2W is a good tradeoff which provides 97% improvement in the lifetime with around 4X increase of library size. Another alternative solution for SP sampling is a hybrid non-uniform sampling. In this scenario, the cells with more inputs (e.g. more than 2 inputs) will have fewer sampling points (e.g. 2 sampling points), while the cells with fewer inputs (e.g. 1 or 2 inputs) use more sampling points (e.g. 3 non-uniform sampling points) in order to balance library size and lifetime improvement.

Effect of the workload

For the simulations results presented in Table 6.1, we assumed that the primary input SPs are 0.5 for both the gate mapping phase and the SP calculation of internal nodes (with which BTI-induced ΔVth values are obtained accordingly). However, different workloads result in different primary input SPs, and internal SPs accordingly, observed by the circuit during its operational lifetime. To account for this on the efficiency of our methodology, we performed two sets of experiments.

In the first experiment, the gates are mapped (optimized) according to the primary input SP of 0.5 but the internal nodes SPs (and BTI-induced Vth shifts) are calculated for the primary input SP of 0.2. In the second experiment, the primary input SPs used for the mapping phase and the internal SP (and degradation) calculation were chosen as 0.2 and 0.5, respectively, i.e. the reverse situation as in the first experiment. The results show that the efficiency of U5 decreases by around 25% (from 167% lifetime improvement to 127%) when different primary input SPs are used for the mapping phase and the delay degradation calculation. However,

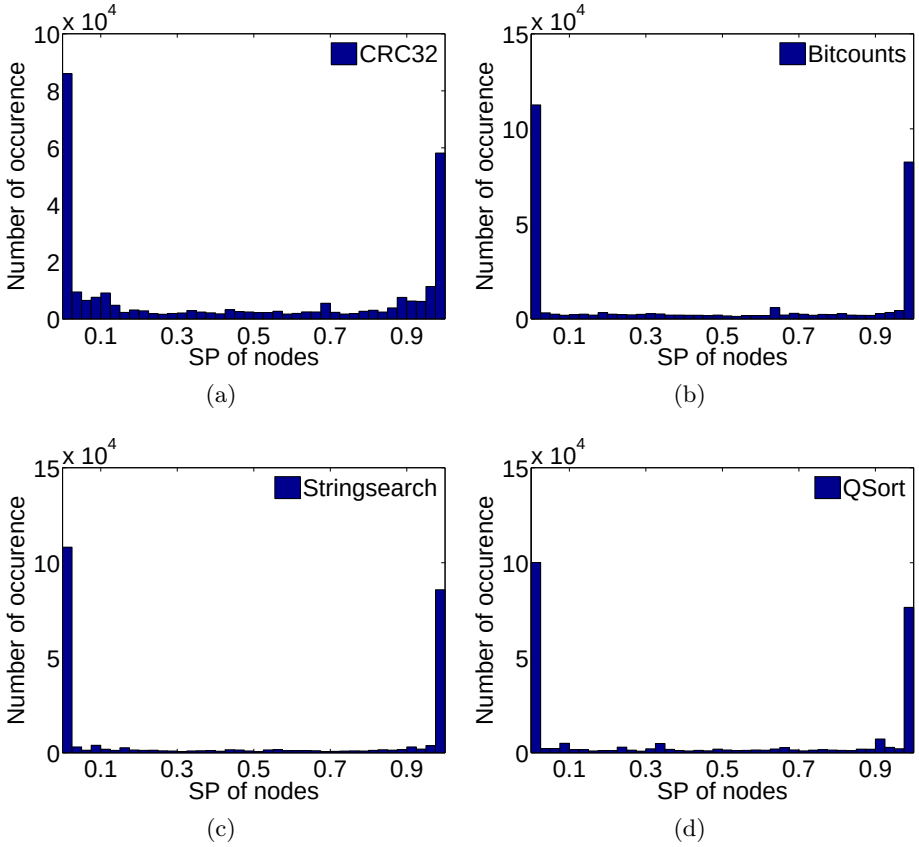


Figure 6.8.: Histogram of the internal node SPs of different applications

this has negligible effect on the efficiency of the methods with the fewer SP samples (less than 1% decrease in the lifetime improvement of NU2W). Therefore, while the optimization is done at "design time" using a particular assumption over the primary input SPs, the method still remains effective as the workload changes during runtime.

Effect of workloads in a real system

As a case study of applying proposed approach to a real system, we extracted the SP distribution for different workloads running on OR1200 processor. The idea is to show whether a design time decision which maps the gates to different cells based on a "typical" SP distribution known at design time, is still valid when real workloads exhibit changes for SPs and the system switches to different workloads during its lifetime. OR1200 is an in-order processor with five stages pipeline which implements Harvard architecture, i.e. has separate instruction and data cache. This processor is synthesized with Design Compiler using Nangate 45 nm standard cell library. The final netlist has 30,986 gates and 2,693 flip-flops. During the post-synthesis simulation, several benchmarks from MiBench benchmark suite are executed on this processor and the activity of all signals is dumped in VCD format. By analyzing the corresponding VCD file of each workload, we extracted the SP of all signals during the execution of that workload. Figure 6.8 shows the histogram of SP of all signals for different workloads running on this processor. As shown in this figure, SP values around [0.0, 0.1] and [0.9, 0.1] are more frequent which needs to be considered in the SP sampling of our proposed approach.

Moreover, as shown in the flow of proposed approach, the gate mapping phase is done according to the SP of the internal nodes. However, the SP of internal nodes might change

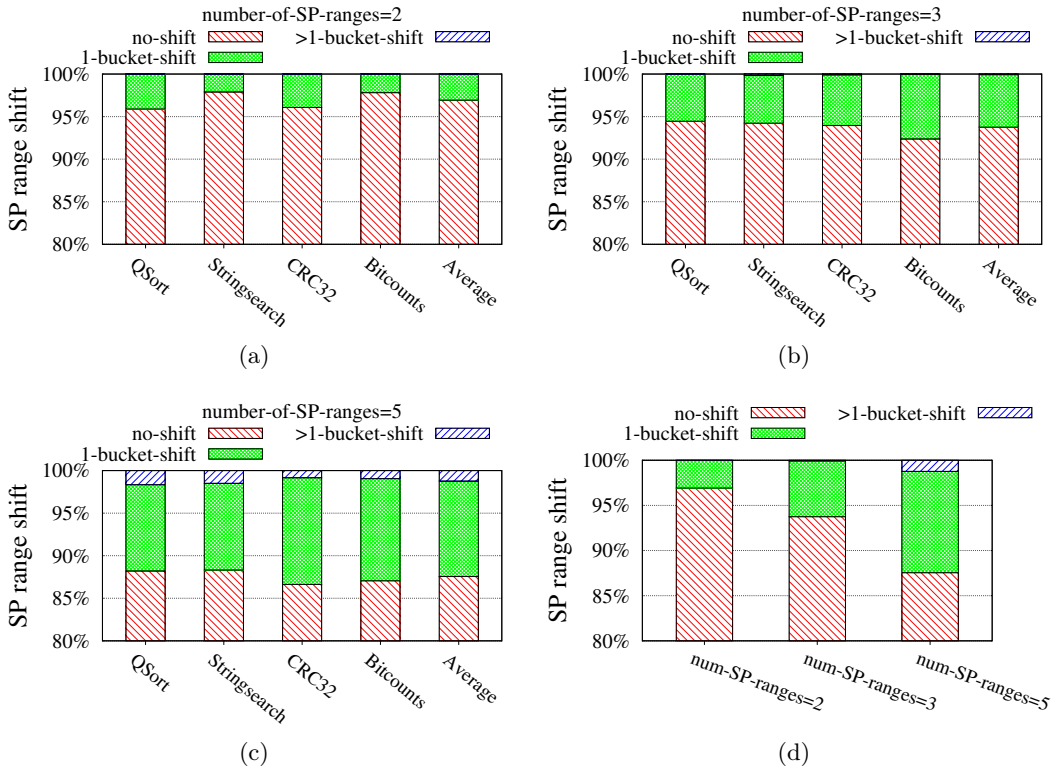


Figure 6.9.: Effect of different workload on the SP range of internal nodes when the number of SP ranges is equal to a) 2 b) 3 c) 5 d) the average for all cases. SP range shift is shown for [80% 100%] range.

from one workload to another. This means that for different workloads we need to have different mapping which is not possible since the gate mapping has to be done only once at design time. Therefore, we propose to perform the gate mapping phase according to the average of internal node SPs over all workloads. However, the efficiency of our approach, for a particular workload, strongly depends on the internal node SP distribution of that workload compared to that of average case. In other words, if there is a huge difference between the SP distribution of one workload compared to average case, the efficiency of our approach would be very limited.

To investigate this issue, we perform a set of simulations to compare the internal node SP distribution of different workloads and the average case. For this purpose, we obtain the SP range of the internal nodes for the average case and each particular workload to check whether the internal node SP remains in the same range or it is shifted to other ranges. This is done for all discretization scenarios introduced in Section 6.5.1.

Figure 6.9.a shows the results for the case in which we have only two SP ranges (NU2 and NU2W). As shown in this figure, for all workloads more than 95% of the internal nodes remain in the same SP bucket compared to the average case. Figure 6.9.b and 6.9.c show similar results for the case in which the entire SP range is divided to 3 and 5 buckets (NU3 and U5), respectively.

Figure 6.9.d compares the results for U5, NU3, NU2 (NU2W) scenarios. As shown in this figure, while in NU2 scenario, more than 95% of the internal nodes remain in the same SP bucket compared to the average case, this percentage is around 85% for U5 scenario. This means that when the SP range is discretized to more buckets, there will be more shifts across different buckets when considering different workloads. However, the efficiency of our approach will be less affected in NU2 and NU2W scenarios for different workloads compared to the U5

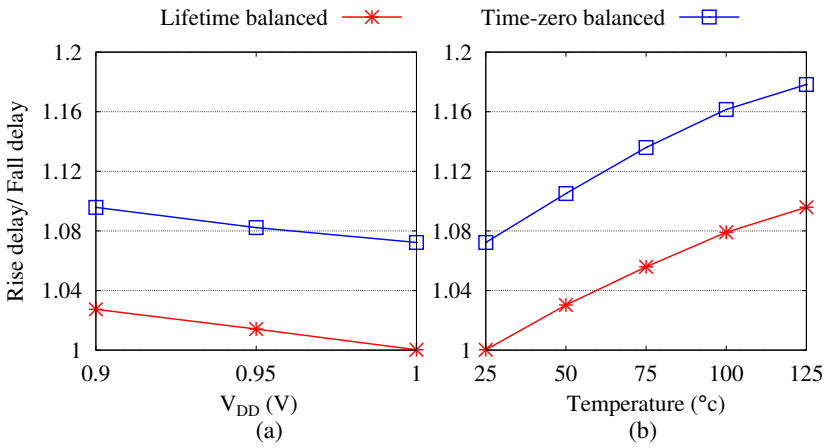


Figure 6.10.: Effect of a) voltage and b) temperature variation on the rise/fall delay ratio of a simple inverter with input SP of 0.1 after 3 years

scenario. As also shown in Section 6.5.3, less discretization of the SP range, using NU2 and NU2W methods, results in a better tradeoff between library size and lifetime improvement, which is also supported by this analysis.

6.5.4. Effect of voltage and temperature variation

As discussed previously, the main idea of this work is to balance the rise and fall delays of standard cells at the end of their expected lifetime rather than the design time according to the input SP of the cell. However, the BTI effect is not only dependent of input SP, but also the supply voltage and the temperature. Since different circuits (and hence cells inside the circuits) may have different working temperatures and supply voltages, the efficiency of our proposed approach may be affected. To investigate this issue, the rise to fall delay ratio of a simple inverter is obtained for different corners of temperature and supply voltage values for two cases of proposed lifetime-balanced cell and time-zero balanced cell. It should be noted that the optimization is done for the nominal corner of supply voltage ($V_{DD} = 1V$) and room temperature (25°c).

As shown in Figure 6.10, although the cell becomes less balanced for higher temperature and lower V_{DD} values, still our approach provides more balanced rise and fall delay values (the ratio is more close to 1). This is due to the fact that the temperature and voltage variations affect both pull-up and pull-down network and their aging rate with almost the same rate. Therefore, making the cell more balances at one corner (nominal corner of supply voltage and room temperature) is helpful for other corners as well. This confirms that our proposed approach is still useful in the different corners of temperature and supply voltage variations.

6.6. Summary

In this chapter, we proposed a BTI-aware library cell design to mitigate the BTI effect. The main idea is to balance the rise and fall delays of a cell by considering the target lifetime delay degradations instead of time-zero delays. We also presented a technology mapping technique in which the critical gates in the circuit are mapped to suitable cells within this aging-aware library based on their input signal probabilities. The simulation results show that our technique can improve the lifetime by approximately 150% with negligible area/power overheads. Our experiments also show that the proposed approach remains effective even when the system workload changes during the operational lifetime.

CHAPTER
SEVEN

INPUT AND TRANSISTOR REORDERING FOR AGING REDUCTION IN COMPLEX CMOS GATES

7.1. Overview

In previous chapter, an aging mitigation technique was proposed in which the cells were re-designed by changing the aspect ratio of the sizes of PMOS and NMOS transistors. In this chapter, we propose a complementary aging mitigation technique by changing the placement of internal transistors of the standard cells and their inputs.

For this purpose, we investigate the stacking effect of transistors on aging and propose a novel input/transistor reordering approach to alleviate the effect of NBTI and HCI during the active mode operation of the circuit. According to the results, the aging effect is postponed by increasing the operational lifetime for selected ISCAS benchmarks by 14.1%, in average, while it has an extremely negligible effect on delay, area, and power compared to the original cell input ordering. In addition, we observed that, neglecting the HCI effect in Input/Transistor reordering method reduces the efficiency of the method by 54%.

The rest of this chapter is organized as follows. In Section 7.2 a short introductory of motivation and contributions of this work is presented. In Section 7.3, the related work is discussed. The stacking effect on NBTI and HCI, and stacking-aware aging model are presented in Section 7.4. The proposed input and transistor reordering technique is described in Section 7.5. The experimental results are discussed in Section 7.6. Finally, Section 7.7 concludes the chapter.

7.2. Introduction, motivations and contributions

NBTI and HCI increase the absolute value of the PMOS and NMOS threshold voltages respectively and as a result, the performance of the circuit decreases over time. Eventually, when the delay of the circuit exceeds the timing constraints, the circuit fails due to these two phenomena. Consequently, the failure rate in the field can significantly increase and hence the operational lifetime of CMOS VLSI chips is reduced.

Connection of multiple transistors in series is referred to Transistor *stacking*. For example NMOS transistors are stacked in NAND and PMOS transistors are stacked in NOR gates in CMOS implementation. The way the inputs are connected to different transistors in the stack has a considerable impact on the NBTI and HCI effects. The NMOS (PMOS) transistor stacking increases (decreases) the source voltage of the upper (lower) transistors in the stack as well as lowers the absolute value of the gate-source voltage of these transistors. Since NBTI and

HCI exponentially depend on the transistor threshold voltage as well as gate-source voltage, stacking has a strong impact on these effects.

In this chapter, a novel approach called *Input and Transistor Reordering* (ITR) is presented to reduce the transistor aging effect during circuit operation (active mode). By exploiting this approach, the gate-source voltages of transistors change and as a result the impact of NBTI/HCI on internal transistors of the gates can be reduced while providing the same output.

The key contributions of this work are: First, We propose a new simplified stacking-aware NBTI and HCI aging model. Based on the models, we present an algorithm to precisely calculate NBTI and HCI considering stacking effect. Next, a novel input/transistor reordering technique for the CMOS complex gates is proposed to reduce the impact of HCI and NBTI. The proposed methodology is general and can be applied to circuits which consist of both simple gates (e.g. NAND, NOR, etc.) and complex gates (i.e. the gates which contain both series and parallel combinations of transistors). Moreover, most of the literature has focused only on reduction of NBTI effect in circuits built with simple gates. However, in this work for the first time, both HCI and NBTI effects are considered which leads to 54% improvement in total lifetime in comparison with considering the NBTI effect only.

7.3. Related work

There is a considerable amount of work for alleviating the effect of transistor aging at various design levels. Most of the literature primarily focuses on NBTI effect, since it was considered as the dominant factor of the transistor aging and overall delay increase [148]. Gate sizing, supply voltage regulation, threshold tuning and guard-banding are some of the well-known aging mitigation methods trying to compensate the timing degradation due to NBTI [140, 149]. In the gate sizing approach, an appropriate size for each transistor is assigned, considering its expected probability of being stressed [137]. However, the major shortcoming of this method is a considerable area and power overhead because of transistor up-sizing. In [125], the threshold voltage increase of PMOS transistors due to NBTI is compensated by using a forward body biasing technique. NBTI induced delay degradation is tracked by a monitoring circuit and based on the measured value, body biasing is applied to mitigate NBTI effect.

Another category of work focuses on alleviating the impact of the NBTI and mitigate the delay degradation. In [129] a power gating method is proposed to reduce the power consumption of a circuit as well as the NBTI effect. The main drawback of this technique is long wake-up latency which makes power gating infeasible when the idle time of a circuit is not long enough or could not be predicted. An NBTI-aware dynamic voltage and frequency scaling (DVFS) is proposed in [131]. The disadvantages of this technique are additional area and power overhead for the NBTI monitoring sensors. *Input Vector Control (IVC)* [150] and internal node control [151] approach have been used to alleviate the impact of NBTI by controlling the state of the internal gates.

A stacking aware gate-level NBTI delay degradation model is presented in [152]. This work was limited to NBTI and does not consider HCI. However, as the channel length is scaling aggressively beyond 65nm, the effect of HCI should be considered as well [153]. In [154], a stacking-aware pin reordering approach is proposed to reduce the effect of NBTI. However, it was limited only to simple stacking cases (i.e. not complex CMOS gates) and also it did not take HCI effect into account.

7.4. Transistor stacking and aging

When multiple transistors are connected in series, the drain-source voltage (V_{DS}) of each transistor is smaller than V_{DD}. As a result, the magnitude of the threshold voltage increases due to *Drain Induced Barrier Lowering* (DIBL). Furthermore, the absolute gate-source voltage (V_{GS}) of upper (lower) transistors is smaller in a series structure in pull-down (pull-up) network. This phenomenon is called stacking effect. Since there is an exponential relation between aging effects and both gate-source and threshold voltages of transistors, the stacking effect has a noticeable impact on both NBTI and HCI. In the following, we investigate the impact of the stacking effect on transistor aging of simple (i.e. NAND, NOR) and complex CMOS gates. *Complex gates* are gates which consist of combination of parallel and series transistors in pull-up/pull-down network. Based on this definition, simple gates (NAND, NOR) are a sub-set of complex gates. We define a Super Transistor (ST) as a virtual transistor that might be either a *Super Series Transistor* (SST), *Super Parallel Transistor* (SPT), NMOS transistor (NT) or PMOS transistor (PT). An SST is referred as a virtual transistor which is an ordered list of several ST in series. An SST is ON when all its internal STs are ON. We also define an SPT as a virtual transistor which consists of several STs in parallel. An SPT is ON when at least one of its internal parallel STs is ON. In conclusion, the above definitions are represented by the following notations:

$$\begin{aligned} ST &= SST \mid SPT \mid NT \mid PT \\ SST_i &= (ST_i^1, ..., ST_i^n) \\ SPT_j &= \{ST_j^1, ..., ST_j^m\} \end{aligned} \tag{7.1}$$

Using aforementioned definitions, all complex gates can be reduced to a simple structure which has only series or parallel STs. We also refer to simple-SST as an SST which contains only transistors in series. On the other hand, complex-SST is that containing transistors and super transistors. Based on these definitions, since the transistors in a simple-SST are symmetrical, reordering the inputs of transistors, does not affect the functionality of simple-SST. In contrast, input reordering (connecting the different inputs to different transistors) changes the functionality of complex-SSTs, because they are not necessarily symmetric.

7.4.1. Stacking effect on NBTI

Since NBTI only affects PMOS transistors, stacking effect on NBTI should be considered only for pull-up network. To investigate the stacking effect, first, we consider a simple 3-input NOR gate as shown in Figure 7.1. In a 3-input NOR gate, 8 different cases may occur. Here we investigate the four most important cases.

1. The three inputs are equal to zero (Figure 7.1(a)), $V_X = V_{DD}$ and $V_Y = V_{DD}$. Therefore, $V_{GS}^{P1} = V_{GS}^{P2} = V_{GS}^{P3} = -V_{DD}$. In this case, the three PMOS transistors are in stress mode.

2. The most upper input is equal to zero. Therefore $V_X = V_{DD}$ (Figure 7.1(b)). Since the gate of $P2$ is connected to V_{DD}, gate-source voltage of $P2$ is equal to zero. In other words, $P1$ is in stress mode and $P2$ is in recovery mode. Moreover, since $V_Y \approx 0$, the gate-source voltage of $P3$ is almost equal to zero. Therefore, although the input of this transistor is equal to zero, it is in recovery mode.

3. The input of transistor $P1$ is equal to one and the other inputs are zero (Figure 7.1(c)). Because of the stacking effect, the resistance of $P1$ is much larger than the resistance of

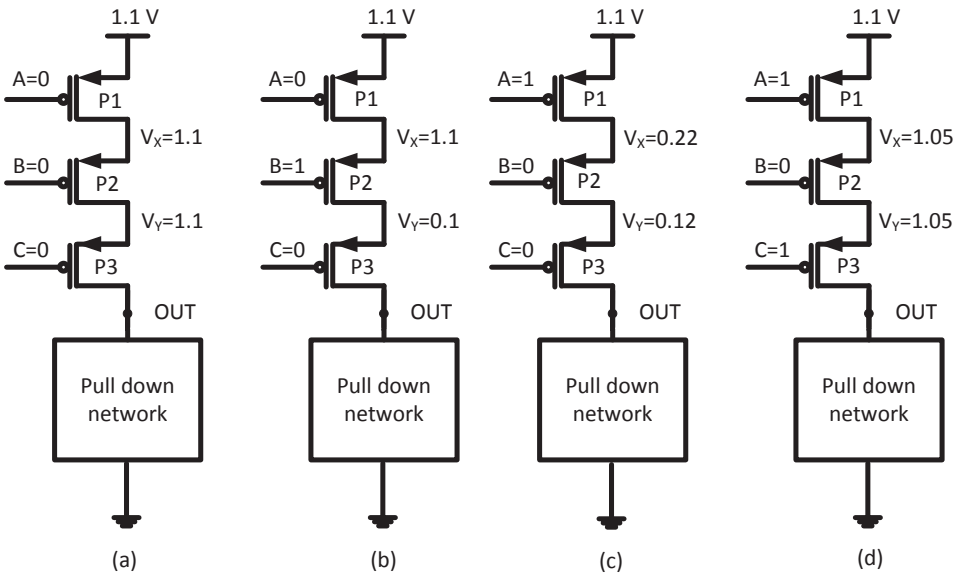


Figure 7.1.: Stacking effect in a 3 input NOR gate (voltages are calculated by HSPICE simulations)

$P2$ and $P3$. As a result, $V_X \approx 0$ and hence the gate-source voltage of the lower transistors ($P2$ and $P3$) are equal to zero. Therefore, both lower transistors are in the recovery mode although $V_G^{P2} = V_G^{P3} = 0$.

4. In this case, $P1$ and $P3$ are OFF and $P2$ is ON (Figure 7.1(d)). Due to the stacking effect, the resistance of $P3$ is much bigger than the resistance of $P1$. Therefore, $V_X = V_Y \approx V_{DD}$. As a result, $V_{GS}^{P2} = -V_{DD}$ and hence $P2$ is under stress. The other transistors are in recovery mode.

It can be concluded that the state of a transistor (stress or recovery) depends not only on its input, but also on the state of its upper and lower transistors. For example transistor $P2$ is in recovery and stress mode in Figure 7.1(c) and Figure 7.1(d) respectively, although the input

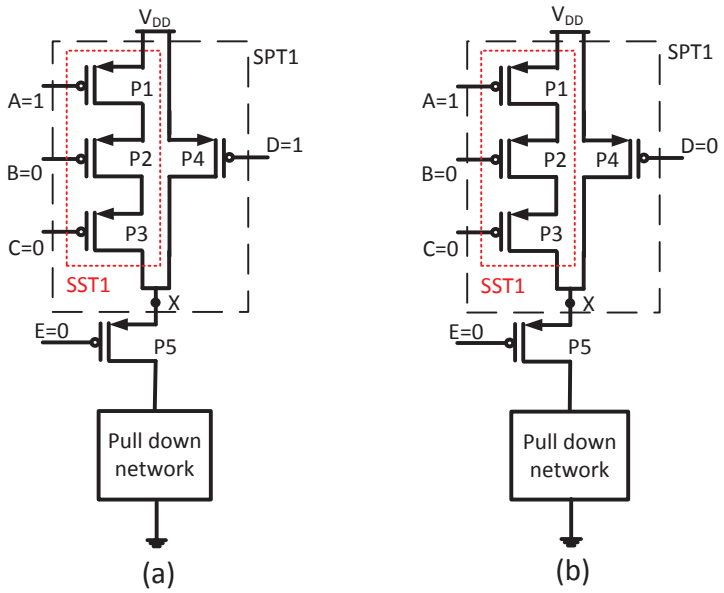


Figure 7.2.: Stacking effect in a complex gate

of $P2$ is the same.

Now to better explain the stacking effect for a more complex gate, consider an example illustrated in Figure 7.2. In this example SST1 consists of transistors $P1$, $P2$ and $P3$ and SPT1 consists of $SST1$ and $P4$.

The situation in SST1 is almost the same as the pull-up network of a 3-input NOR gate. However, SST1 has an interaction with other transistors in the complex gate. To investigate the interaction between $SST1$ and $P4$ consider Figure 7.2. In Figure 7.2(a), $P1$ and $P4$ are OFF and the other transistors are ON. In this situation the parallel transistor of SST1 ($P4$) has no effect on $SST1$ and similar to the 3-input NOR (Figure 7.1(c)), all transistors inside $SST1$ are in recovery mode. On the other hand, if transistor $P4$ is ON, the voltage of node X is equal to V_{DD}. As a result, the voltages of all internal nodes in $SST1$ are equal to V_{DD}. Therefore, in contrast to previous case, $P2$ and $P3$ are in stress mode. Finally we analyze the stacking effect on transistor $P5$. The pull-up network of this complex gate can be considered as a two-series transistors ($P5$ and $SPT1$). As a result, $P5$ is in stress mode if only the corresponding input is zero and $SPT1$ is ON. The Stress-Recovery (SR) flowchart illustrated in Figure 7.3 summarizes the above cases for an N-input complex gate.

7.4.2. Stacking effect on HCI

Since HCI affects NMOS transistors, the stacking effect on HCI should be considered only in a pull-down network of gates, such as NAND gate. Please note that HCI occurs at the transition on the gate input of NMOS transistors [155]. To describe the stacking effect on HCI in more details, consider a simple two input NAND gate shown in Figure 7.4. In this example, a falling transition is considered to explain the effect of stacking on HCI, however, a similar explanation is valid for the case of input rise transition.

For the lower NMOS transistor (N2), because its source voltage is equal to zero, each falling transition of its gate input results in a falling transition in V_{GS}. Consequently, the effective activity factor of N2 used to obtain HCI-induced V_{th} shift with Equation (2.31) is equal to the activity factor of the input connected to it (Figure 7.4(a)). However, for the upper transistor (N1) two different situations can occur depending on the status of N2 (Figures 7.4 (b) and (c)). When the input of N2 is connected to V_{DD}, this transistor is ON and hence $V_{S1} = 0$. Therefore each falling transition of IN2 leads to a falling transition in V_{GS}^{N2}. On the other hand, when $IN2 = 0$, V_{GS}^{N1} remains constant even when there is a falling transition on IN1. During a falling transition of IN1, the initial value of IN1 is equal to 1 and IN2 is equal to zero. In this situation, N1 is ON and has a lower resistance in comparison to N2 which is OFF. By this observation, $V_{S1} = V_{DD}$ and consequently the initial value of V_{GS}^{N1} is zero. When IN1 switches to zero, both transistors become OFF. Since the resistance of N1 is much larger than the resistance of N2 (due to the body effect), V_{S1} becomes zero. This means V_{GS1} remains constant and does not switch. It can be concluded that, when IN2 is zero, the switching of IN1 does not result in a switching of V_{GS}^{N1}. Consequently, the effective switching activity of the transistor N1 is less than the switching activity of its input. In other words, the lower transistor (N2) masks some switchings of the upper transistor which causes N1 to suffer less from HCI. Considering the above cases and by using the definitions of the super-transistors (SST and SPT), the flowchart of Figure 7.3 (red bold one) shows how the stacking affects HCI in an N input complex gate.

7.5. Reordering methodology

Since transistor aging is considerably affected by the order in which the transistors are placed in a complex gate and their relative input values, our key idea is to reorder the inputs of each

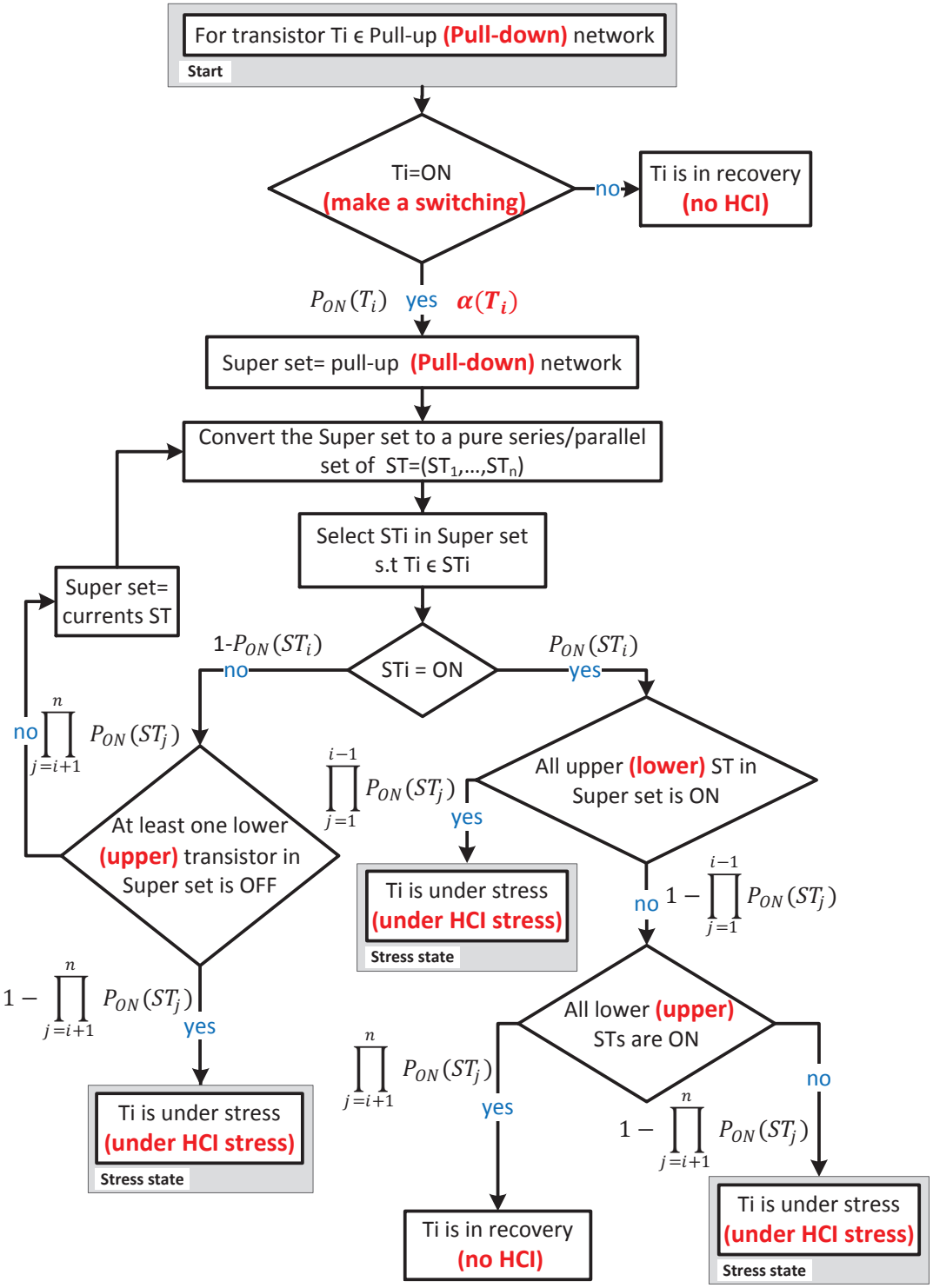


Figure 7.3.: Stress-Recovery (SR) flowchart for NBTI (HCI) in an N input complex gate. For $ST = (ST_1, ..., ST_n)$ if $i < j$, ST_i is closer to power-line

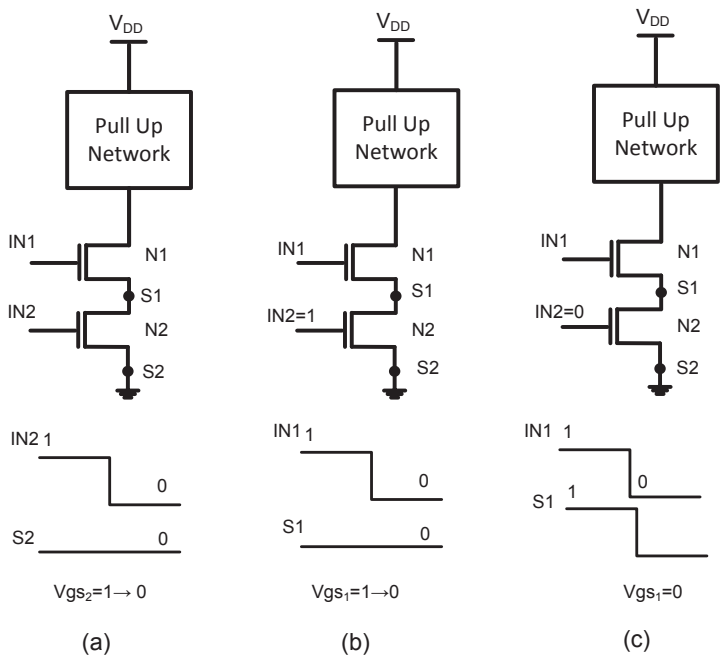


Figure 7.4.: Stacking effect in a NAND gate

gate (which input is connected to which transistor in the gate connection) without changing the functionality of the gate to decrease transistor aging during the active mode. Another technique which can improve the efficiency of the input reordering approach is Super-Transistor (ST) reordering. In this work, this is done for pull-up network for NBTI reduction, as well as pull-down network for HCI reduction. For Input and Transistor reordering we need to calculate the ON/OFF probability of each ST based on the *Signal Probability (SP)* of the inputs (i.e. the probability of being one for a signal value). According to the definitions presented in Equation (7.1) and the definitions of ST, SST, and SPT the following equations are extracted:

$$
\begin{aligned}
&P_{ON}(NT) = SP(NT)\\
&P_{ON}(PT) = 1 - SP(PT)\\
&P_{ON}(SST_i) = \prod_{k=1}^{n}(P_{ON}(ST_i^k)\\
&P_{ON}(SPT_j) = 1 - \prod_{l=1}^{m}(1 - P_{ON}(ST_j^l))
\end{aligned}
\tag{7.2}
$$

7.5.1. NBTI reduction

To describe our NBTI reduction method, consider Figure 7.5. We assume that the input vector is $ABC = 100$. If the inputs are connected in a way depicted in Figure 7.5(a), then based on the rules described in the previous section, all three transistors are in recovery mode (see Figure 7.1(c)). On the other hand, if the inputs are connected in a way illustrated in Figure 7.5(b) only the two lower transistors are in recovery mode (see Figure 7.1(b)). It can be concluded that for the input $ABC = 100$, the first interconnection between the inputs and transistors is more favorable than the second one, in terms of NBTI. Based on this observation, the main idea of our proposed method is minimizing the NBTI effect of a circuit by means of reordering the inputs connections of transistors in a pull-up network of gates. In the NBTI-aware input

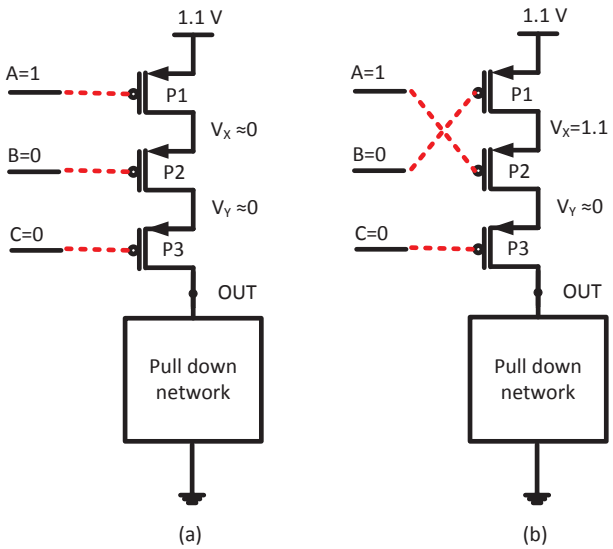


Figure 7.5.: Input reordering in a 3 input NOR gate

reordering approach, all possible combinations of the input connections of series transistors in each simple-SST, are considered and the order resulting in the minimum NBTI is chosen (reordering inputs of transistors $P1$, $P2$, and $P3$ in Figure 7.2). Since, the series transistors in a simple-SST are symmetrical, the order of the inputs does not affect the functionality of the gate. Input reordering can be improved by using a method called ST reordering. In this technique, all possible permutation of ST in each complex-SST are considered and the ordering leading to minimum NBTI effect is selected. Since the input reordering changes the functionality of complex-SST, in ST reordering method, the ST (not their inputs) are reordered. For example in Figure 7.2 reordering $SPT1$ and $P5$ may change the overall NBTI effect on the complex gate.

The following steps, show the general methodology of the NBTI-aware input/ST reordering for an N-input complex gate:

1. The pull-up network is converted to a set of SSTs and SPTs.
2. For each complex-SST, all permutations of ST ordering are considered
3. For each permutation, all possible orders of inputs are considered
4. For each case, the effective duty cycle (the time ratio between stress time to the total time) of each transistor is calculated
5. The case which has the minimum $\sum_i (D_{c-eff}(P_i))^n$ is selected

The effective duty cycle (D_{c-eff}) represents the NBTI effect on a transistor using Equations (2.10). The NBTI status of a transistor (stress/recovery) can be determined according to the ON/OFF status of this transistor and the other transistors in the gate. The following steps show the process of the effective duty cycle calculation of each transistor in a complex gate:

1. Construct a Stress-Probability tree from the stress-recovery (SR) flowchart. The nodes and the edges of this tree are defined as follow:
   - Node: traversed ST in the SR flowchart
   - Edge: branch probability in the SR flowchart

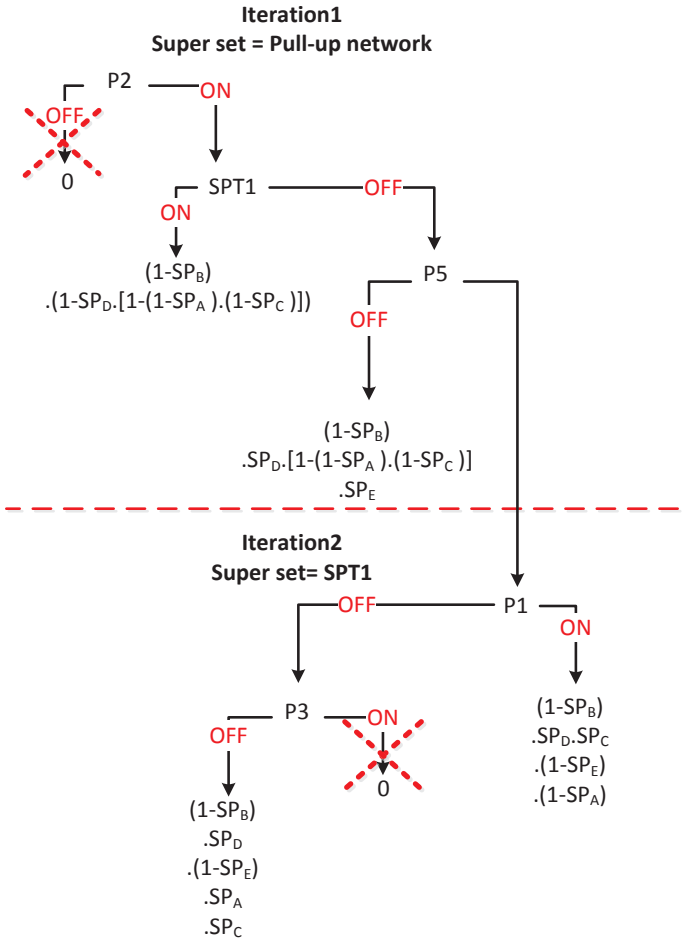


Figure 7.6.: Stress-probability tree: effective duty cycle calculation of $P2$

2. Remove leaves (end edges) leading to recovery

3. $Prob\_leaf$ is equal to product of probability of edges on the path from the root to that leaf

4. $D_{c-eff} = \Sigma Prob\_leaf$

As an example, Figure 7.6 shows the steps of the effective duty-cycle calculation for transistor $P2$ in Figure 7.2. It should be noted that, for only input reordering, the minimum overall effective duty cycle (minimum NBTI) is obtained by connecting the input with a higher SP to the transistor with a higher position in series structure.

Table 7.1.: Effective Activity Factor of a 2-input NAND gate

Transistor	Effective Activity Factor
$N1$	$\alpha_{IN1}.SP_{IN2}$
$N2$	α_{IN2}

7.5.2. HCI reduction

As described in Section 7.4.2, in a 2-input NAND gate when the input of the lower NMOS transistor is zero, the falling transition of the gate-source voltage of the upper transistor is masked. This phenomenon leads to an HCI reduction of the upper transistor. The effective switching activity, α_{eff}, of a 2-input NAND gate (see Figure 7.4) can be calculated based on Table 7.1. The key idea is to reorder the inputs/ST to minimize the total gate delay degradation due to HCI. According to the Equation (2.31), HCI has a linear relation with the switching activity. Consequently, the HCI minimization problem of a gate is equivalent to a minimization of the total effective switching activities. Based on this fact, the input reordering of a two-input NAND gate for HCI minimization can be formulated as shown below.

$$Minimize(\alpha_{eff}(N1) + \alpha_{eff}(N2)) = Minimize(\alpha_A \times SP_B + \alpha_B)$$

where α is the switching activity factor of an input and SP is the signal probability. Therefore, input reordering is done in a way that $\alpha_{eff}(N1) + \alpha_{eff}(N2)$ is minimized.

The general flow of the HCI-aware input/ST reordering for an N-input complex gate is similar to the one for NBTI reduction (explained in Section 7.5.1), except that here, the objective is to find the case which has the minimum $\sum_i \alpha_{eff}(N_i)$. The α_{eff} of each transistor is a representative of its HCI effect based on the Equation (2.31). When a transistor makes a switching, depending on the ON/OFF status of the other transistors in the gate, this transistor might experience HCI effect. The α_{eff} of each transistor is calculated by a similar flow as the effective duty cycle (D_{c-eff}) calculation explained in Section 7.5.1.

7.6. Experimental results

We have evaluated the efficiency of the proposed methods with ISCAS'85 and ISCAS'89 benchmark circuits. The overall proposed methodology for NBTI and HCI reduction by means of stacking-based input reordering is summarized below.

First, a logic synthesis tool is used to map a benchmark circuit to a gate-level netlist. In our experiments, we used Synopsys Design Compiler as the synthesis tool to map the circuits to SAED 90nm standard cell library. The gate delays are extracted from the standard cell library as well. Afterwards, the extracted gate-level description of the circuit is given to a logic simulator and signal probabilities and activity factors of all internal nodes are calculated. To

Table 7.2.: Delay degradation of benchmark circuits

Circuit	No Stacking $\Delta delay(ps)$	Stacking $\Delta delay(ps)$	Error
C432	146.03	135.42	14%
C880	155.57	131.31	23%
C1355	81.11	71.83	19%
C2670	100.89	91.17	18%
C3540	201.94	185.28	14%
C5315	40.49	36.31	16%
C6288	623.99	589.06	12%
Average			16.6%

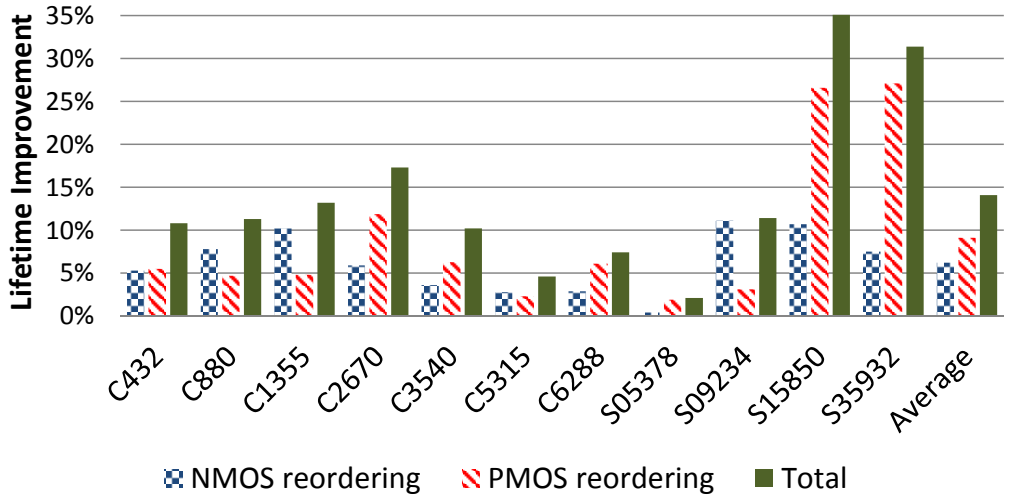


Figure 7.7.: Lifetime improvement using the proposed input and transistor reordering technique

alleviate the effect of NBTI and HCI, the input/component of pull-up and pull-down network are reordered based on the methodology proposed in the Section 7.5 to minimize the total effective duty cycle and activity factor, respectively. Finally, the obtained aging-aware netlist and original netlist are further processed to analyze the effectiveness of the proposed technique on lifetime of the circuit and its overheads as well. For this set of experiments, the NBTI and HCI models, explained in Sections 2.4.1 and 2.4.2 are respectively used.

Table 7.2 illustrates the effect of considering stacking on NBTI and HCI delay degradation models. Based on these results, traditional methods which do not consider the stacking effect overestimate the delay degradation more than 16.6% on average. This overestimation leads to pessimistic analysis in aging mitigation techniques resulting in over-design in terms of unnecessary power consumption, performance and area overhead.

The experimental results of our proposed input/component reordering technique are shown

Table 7.3.: Effect of Input Reordering on Lifetime

Circuit	Lifetime Improvement NMOS Reordering	Lifetime Improvement PMOS Reordering	Total lifetime Improvement	Delay Overhead	Power Overhead	Area Overhead
C432	5.3%	5.5%	10.8%	0.85%	0.12%	-0.10%
C880	7.8%	4.7%	11.3%	0.00%	-0.05%	0.00%
C1355	10.2%	4.8%	13.2%	-0.38%	-1.70%	0.00%
C2670	5.9%	11.9%	17.3%	0.12%	-0.01%	0.00%
C3540	3.6%	6.3%	10.2%	-0.06%	-0.19%	0.00%
C5315	2.8%	2.3%	4.6%	-0.43%	-0.01%	0.00%
C6288	2.9%	6.1%	7.4%	0.00	0.00%	0.00%
S05378	0.4%	1.9%	2.1%	0.45%	0.12%	0.00%
S09234	11.1%	3.1%	11.4%	0.37%	-0.45%	0.00%
S15850	10.7%	26.6%	35.1%	-.08%	0.13%	0.00%
S35932	7.5%	27.1%	31.4%	-0.67%	0.23%	0.00%
Average	6.2%	9.1%	14.1%	0.00%	-0.15%	0.00%

in Figure 7.7. The new lifetime after reordering is calculated and the percentage increase is reported in this figure. According to this figure, the choice of suitable order of inputs can significantly impact the delay degradation due to NBTI and HCI. Based on these results, in average, the lifetime can be extended by 14.1%. According to the results reordering the inputs/components for both NBTI and HCI improves the lifetime of the circuit by 54% in average comparing to reordering for only NBTI effect. It should be noted that, this method may change the pre-aging delay of the circuit. It is due to the fact that, the delay optimization tools use reordering based on the arrival time of the signal to minimize the pre-aging delay. However, since we reorder inputs/STs to minimize aging, it may have a side-effect on the pre-aging delay. The delay overhead of the proposed technique is reported in Table 7.3. According to the results, the proposed method does not significantly affect the pre-aging delay of the circuits and the introduced overhead is negligible. In addition, area and power overhead of the proposed methodology of reordering are reported which indicates that our method does not incur significant area and power overhead.

7.7. Summary

In this work, we exploit the stacking effect on transistor aging and propose an input reordering approach to reduce the effect of transistor aging during the active mode. This was done for both PMOS and NMOS stacks as well as for complex CMOS gates, to reduce the effect of NBTI and HCI, respectively. This approach comes at negligible area, delay, or power overhead. Our experiments for ISCAS benchmark circuits show that this method can increase the lifetime by 14.1%, in average.

CHAPTER
EIGHT

SUMMARY AND CONCLUSIONS

With the continuous down scaling of VLSI technology nodes, the reliability has become an important design constraint. The reliability issues can cause the system to fail which in turn can lead to catastrophic consequences such as financial loses and even loss of human life according to the application of VLSI circuit.

There are different sources of unreliability at nano-meter technology nodes such as soft error, transistor aging, voltage droop and process variation. It is shown that by further technology down scaling, the effect of these sources of unreliability becomes larger, leading to either a shorter lifetime or higher failure rate during lifetime operation. Therefore, it is important to model, predict and mitigate different sources of unreliability.

There are various challenges in the modeling as well as mitigation of the sources of unreliability. Two main challenges of the modeling are: i) some of the sources of unreliability are interdependent ii) some of the reliability effects, such as transistor aging, have some intrinsic variability. The main difficulty is to have the right trade-off between runtime and accuracy of reliability modeling abstraction. In terms mitigation, the common practice is to add additional timing guard-band to the design specifications in order to guarantee the correct operation of the circuit. By technology scaling, the amount of additional guard-band increases such that it decreases the associated benefits. Therefore, it is mandatory to address the important sources of unreliability in the early stages of the design by adding reliability constraint on top of the conventional design constraints such as power and performance.

In this thesis, the reliability problem is tackled from a device to circuit level perspective. In the modeling part of thesis, a cross-layer approach is used to model different sources of unreliability, their interdependencies and their intrinsic variabilities. Our cross-layer approach enables us to perform circuit level analysis with an acceptable runtime keeping the device level information (such as interdependency of the sources) leading to a high accuracy of our analysis. Using our cross-layer approach we provide circuit-level results of the combined effect of unreliability sources which can be used at higher levels of abstraction (e.g. RTL and architecture levels). Moreover, we show that separate consideration of these interdependent phenomena can lead to a high amount of inaccuracy. The main observations of the first part (modeling and prediction part) of the thesis can be summarized as follows:

- Voltage droop affects the gates delay and hence the electrical masking of radiation-induced soft error. Therefore, it is important to consider the combined effect of soft error and voltage droop since the soft error rate is strongly dependent on the supply voltage of the gates.

- Soft error due to proton strike is comparable to that for alpha-particle strike at very low supply voltages (low power applications) in SOI-FinFET technology. Moreover, the ratio

of multiple to single bit upsets is relatively higher for alpha radiation compared to that for protons. We also show that, neglecting the effect of process variation leads to an underestimation of soft error rate.

- Stochastic NBTI effect leads to an asymmetric (non-normal) distribution of delay degradation for a circuit designed with FinFET technology. It is also shown that the stochastic behavior of NBTI, as an important transistor aging effect, can result in a significant increase of the guard-band compared to a deterministic case. Finally, we show that considering process variation and stochastic NBTI separately, leads to a considerable overestimation of the delay degradation mean value.

In the mitigation part of the thesis, two methods for reliability-aware cell and circuit design are introduced in order to mitigate the effect of transistor aging on the performance of the circuit. For this purpose, the device and gate level analysis is performed and accordingly the circuit structure/gates is modified in order to efficiently alleviate the effect of aging with a low power and area overhead without affecting the functionality of the circuit. The main achievements of mitigation techniques of the thesis is summarized as follows:

- An aging-aware library cell design approach is proposed to balance the rise and fall delay of the cell at its expected lifetime rather than design time. Using this technique, the lifetime can be improved by 150% with negligible area/power overheads.

- An input and transistor reordering method is provided to alleviate the effect of NBTI and HCI on the circuit delay. It is shown that, using the proposed approach, the lifetime of the circuit can be improved by 14% with a negligible power and area overhead.

The results of this study show that the proposed cross-layer modeling approach can accurately capture the combined effect of interdependent reliability issues and their intrinsic variations on the reliability of the circuit. Moreover, it is shown that the proposed aging-aware cell and circuit design can effectively mitigate the aging effect and hence leads to a lower guard-band with negligible overheads. The trends show that by further downscaling of the technology, the unreliability effects and their intrinsic variabilities become more pronounced meaning the modeling and mitigation techniques proposed in this thesis can be used for future technology nodes. Moreover, the provided circuit level information can be leveraged at higher levels of abstraction (such as architecture level) to model and mitigate the reliability using effective approaches at those levels.

BIBLIOGRAPHY

[1] L. Jia, Z. Liu, Y. Qin, M. Zhao, and L. Diao, *Proceedings of the 2013 International Conference on Electrical and Information Technologies for Rail Transportation (EITRT2013)-Volume I.* Springer Science & Business, 2014, vol. 287.

[2] W. Wolf, *Modern VLSI design: system-on-chip design.* Pearson Education, 2002.

[3] G. E. Moore *et al.*, "Cramming more components onto integrated circuits," 1965.

[4] M. Nicolaidis, L. Anghel, Y. Zorian, T. Karnik, K. Bowman, J. Tschanz, S.-L. Lu, C. Tokunaga, A. Raychowdhury, M. Khellah *et al.*, "Design for test and reliability in ultimate cmos," in *Proceedings of the Conference on Design, Automation and Test in Europe*, pp. 677–682, 2012.

[5] D. J. Frank, "Power-constrained cmos scaling limits," *IBM Journal of Research and Development*, vol. 46, no. 2.3, pp. 235–244, 2002.

[6] B.-g. Park, S. W. Hwang, and Y. J. Park, *Nanoelectronic Devices.* CRC Press, 2012, vol. 1.

[7] M. A. Alam, K. Roy, and C. Augustine, "Reliability-and process-variation aware design of integrated circuits- a broader perspective," in *Reliability Physics Symposium (IRPS), 2011 IEEE International*, pp. 4A–1, 2011.

[8] J. Henkel, L. Bauer, N. Dutt, P. Gupta, S. Nassif, M. Shafique, M. Tahoori, and N. Wehn, "Reliable on-chip systems in the nano-era: Lessons learnt and future trends," in *Proceedings of the 50th Annual Design Automation Conference*, p. 99, 2013.

[9] S. Mitra, K. Brelsford, Y. M. Kim, H.-H. Lee, and Y. Li, "Robust system design to overcome cmos reliability challenges," *Emerging and Selected Topics in Circuits and Systems, IEEE Journal on*, vol. 1, no. 1, pp. 30–41, 2011.

[10] S. Borkar, "Thousand core chips: a technology perspective," in *Proceedings of the 44th annual Design Automation Conference*, pp. 746–749, 2007.

[11] L.-T. Wang, C. E. Stroud, and N. A. Touba, *System-on-chip test architectures: nanometer design for testability.* Morgan Kaufmann, 2010.

[12] T. Brozek, *Micro-and Nanoelectronics: Emerging Device Challenges and Solutions.* CRC Press, 2014.

[13] M. Bushnell and V. D. Agrawal, *Essentials of electronic testing for digital, memory and mixed-signal VLSI circuits.* Springer Science and Business Media, 2000, vol. 17.

[14] A. Avižienis, "Design of fault-tolerant computers," in *Proceedings of the November 14-16, 1967, fall joint computer conference*, pp. 733–743, 1967.

[15] H. Nguyen, "Resiliency challenges in future communications infrastructure," in *Proceedings of the Communications Quality and Reliability Workshop*, May 2014.

[16] D. G. Mavis and P. H. Eaton, "Soft error rate mitigation techniques for modern microcircuits," in *IEEE international reliability physics symposium*, pp. 216–225, 2002.

[17] S. Mitra, N. Seifert, M. Zhang, Q. Shi, and K. S. Kim, "Robust system design with built-in soft-error resilience," *Computer*, no. 2, pp. 43–52, 2005.

[18] N. Seifert, V. Ambrose, B. Gill, Q. Shi, R. Allmon, C. Recchia, S. Mukherjee, N. Nassif, J. Krause, J. Pickholtz *et al.*, "On the radiation-induced soft error performance of hardened sequential elements in advanced bulk cmos technologies," in *Reliability Physics Symposium (IRPS), 2010 IEEE International*, pp. 188–197, 2010.

[19] A. Zjajo, *Stochastic Process Variation in Deep-Submicron CMOS*. Springer, 2014.

[20] 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 *et al.*, "Atomistic approach to variability of bias-temperature instability in circuit simulations," in *Reliability Physics Symposium (IRPS), 2011 IEEE International*, pp. XT–3, 2011.

[21] N. H. Weste and D. M. Harris, *CMOS VLSI design: a circuits and systems perspective*. Pearson Education India, 2005.

[22] "Nangate," http://www.nangate.com/.

[23] J.-P. Colinge, *FinFETs and other multi-gate transistors*. Springer, 2008.

[24] M. Masahara, "Advanced finfet process technology," in *National Institute of Advanced Industrial Science and Technology*.

[25] J. G. Fossum and V. P. Trivedi, *Fundamentals of Ultra-thin-body MOSFETs and FinFETs*. Cambridge University Press, 2013.

[26] http://www.intel.com/content/www/us/en/silicon-innovations/intel-22nm-technology.html, accessed: 2015-01-08.

[27] http://www.globalfoundries.com/technology-solutions/leading-edge-technology/14-lpe-lpp, accessed: 2015-01-08.

[28] http://www.globalfoundries.com, accessed: 2015-01-08.

[29] http://www.tsmc.com/english/dedicatedFoundry/technology/16nm.htm, accessed: 2015-01-08.

[30] C. Ma, B. Li, L. Zhang, J. He, X. Zhang, X. Lin, and M. Chan, "A unified finfet reliability model including high k gate stack dynamic threshold voltage, hot carrier injection, and negative bias temperature instability," in *ISQED*, 2009.

[31] C. Manoj, M. Nagpal, D. Varghese, and V. R. Rao, "Device design and optimization considerations for bulk finfets," *T-ED*, vol. 55, no. 2, pp. 609–615, 2008.

[32] "International technology roadmap of semiconductors (itrs)," http://www.itrs.net.

[33] S. Taylor *et al.*, "Power7+: Ibm's next generation power microprocessor," in *Hot Chips*, vol. 24, 2012.

[34] R. Reis, Y. Cao, and G. Wirth, *Circuit Design for Reliability*. Springer, 2014.

[35] J. B. Bernstein, M. Gurfinkel, X. Li, J. Walters, Y. Shapira, and M. Talmor, "Electronic circuit reliability modeling," *Microelectronics Reliability*, vol. 46, no. 12, pp. 1957–1979, 2006.

[36] M. Orshansky, S. Nassif, and D. Boning, *Design for manufacturability and statistical design: a constructive approach*. Springer, 2007.

[37] M. Orshansky, C. Spanos, and C. Hu, "Circuit performance variability decomposition," in *Statistical Metrology, 1999. IWSM. 1999 4th International Workshop on*, pp. 10–13, 1999.

[38] K. Kuhn, C. Kenyon, A. Kornfeld, M. Liu, A. Maheshwari, W.-k. Shih, S. Sivakumar, G. Taylor, P. VanDerVoorn, and K. Zawadzki, "Managing process variation in intel's 45nm cmos technology." *Intel Technology Journal*, vol. 12, no. 2, 2008.

[39] M. Koh, W. Mizubayashi, K. Iwamoto, H. Murakami, T. Ono, M. Tsuno, T. Mihara, K. Shibahara, S. Miyazaki, and M. Hirose, "Limit of gate oxide thickness scaling in mosfets due to apparent threshold voltage fluctuation induced by tunnel leakage current," *Electron Devices, IEEE Transactions on*, vol. 48, no. 2, pp. 259–264, 2001.

[40] H. Mahmoodi, S. Mukhopadhyay, and K. Roy, "Estimation of delay variations due to random-dopant fluctuations in nanoscale cmos circuits," *Solid-State Circuits, IEEE Journal of*, vol. 40, no. 9, pp. 1787–1796, 2005.

[41] K. J. Kuhn, "Reducing variation in advanced logic technologies: Approaches to process and design for manufacturability of nanoscale cmos," in *Electron Devices Meeting, 2007. IEDM 2007. IEEE International*, pp. 471–474, 2007.

[42] S. Xiong and J. Bokor, "Sensitivity of double-gate and finfetdevices to process variations," *Electron Devices, IEEE Transactions on*, vol. 50, no. 11, pp. 2255–2261, 2003.

[43] M. J. Pelgrom, A. C. Duinmaijer, A. P. Welbers *et al.*, "Matching properties of mos transistors," *IEEE Journal of solid-state circuits*, vol. 24, no. 5, pp. 1433–1439, 1989.

[44] S. Zafar, Y. Kim, V. Narayanan, C. Cabral, V. Paruchuri, B. Doris, J. Stathis, A. Callegari, and M. Chudzik, "A comparative study of nbti and pbti (charge trapping) in sio2/hfo2 stacks with fusi, tin, re gates," in *VLSI Technology, 2006. Digest of Technical Papers.*

[45] J. H. Stathis, M. Wang, and K. Zhao, "Reliability of advanced high-k/metal-gate n-fet devices," *Microelectronics Reliability*, vol. 50, no. 9, pp. 1199–1202, 2010.

[46] 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," *Device and Materials Reliability, IEEE Transactions on*, vol. 7, no. 4, pp. 509–517, 2007.

[47] S. Bhardwaj, W. Wang, R. Vattikonda, Y. Cao, and S. Vrudhula, "Predictive modeling of the nbti effect for reliable design," in *Custom Integrated Circuits Conference, 2006. CICC'06. IEEE*, pp. 189–192, 2006.

[48] T. Naphade, N. Goel, P. Nair, and S. Mahapatra, "Investigation of stochastic implementation of reaction diffusion (rd) models for nbti related interface trap generation," in *Reliability Physics Symposium (IRPS), 2013 IEEE International*, pp. XT–5, 2013.

[49] V. Huard, C. Parthasarathy, C. Guerin, T. Valentin, E. Pion, M. Mammasse, N. Planes, and L. Camus, "Nbti degradation: From transistor to sram arrays," in *Reliability Physics Symposium, 2008. IRPS 2008. IEEE International*, pp. 289–300, 2008.

[50] B. Kaczer, T. Grasser, P. J. Roussel, J. Franco, R. Degraeve, L.-A. Ragnarsson, E. Simoen, G. Groeseneken, and H. Reisinger, "Origin of nbti variability in deeply scaled pfets," in *Reliability Physics Symposium (IRPS), 2010 IEEE International*, pp. 26–32, 2010.

[51] C. Shen, M.-F. Li, C. Foo, T. Yang, D. Huang, A. Yap, G. Samudra, and Y. Yeo, "Characterization and physical origin of fast vth transient in nbti of pmosfets with sion dielectric," in *Electron Devices Meeting, 2006. IEDM'06. International*, pp. 1–4, 2006.

[52] T. Grasser, B. Kaczer, W. Goes, H. Reisinger, T. Aichinger, P. Hehenberger, P.-J. Wagner, F. Schanovsky, J. Franco, M. T. Luque *et al.*, "The paradigm shift in understanding the bias temperature instability: from reaction–diffusion to switching oxide traps," *Electron Devices, IEEE Transactions on*, vol. 58, no. 11, pp. 3652–3666, 2011.

[53] V. Reddy, J. M. Carulli, A. T. Krishnan, W. Bosch, and B. Burgess, "Impact of negative bias temperature instability on product parametric drift." in *International Tset Conference (ITC)*, pp. 148–155, 2004.

[54] P. Weckx, B. Kaczer, M. Toledano-Luque, T. Grasser, P. J. Roussel, H. Kukner, P. Raghavan, F. Catthoor, and G. Groeseneken, "Defect-based methodology for workload-dependent circuit lifetime projections-application to sram," in *Reliability Physics Symposium (IRPS), 2013 IEEE International*, pp. 3A–4, 2013.

[55] T. Grasser, B. Kaczer, W. Goes, H. Reisinger, T. Aichinger, P. Hehenberger, P.-J. Wagner, F. Schanovsky, J. Franco, P. Roussel *et al.*, "Recent advances in understanding the bias temperature instability," in *Electron Devices Meeting (IEDM), 2010 IEEE International*, pp. 4–4, 2010.

[56] J. Franco, B. Kaczer, M. Toledano-Luque, P. J. Roussel, J. Mitard, L.-A. Ragnarsson, L. Witters, T. Chiarella, M. Togo, N. Horiguchi *et al.*, "Impact of single charged gate oxide defects on the performance and scaling of nanoscaled fets," in *Reliability Physics Symposium (IRPS), 2012 IEEE International*, pp. 5A–4, 2012.

[57] H. Reisinger, T. Grasser, W. Gustin, and C. Schlunder, "The statistical analysis of individual defects constituting nbti and its implications for modeling dc-and ac-stress," in *Reliability Physics Symposium (IRPS), 2010 IEEE International*, pp. 7–15, 2010.

[58] T. Grasser, P.-J. Wagner, H. Reisinger, T. Aichinger, G. Pobegen, M. Nelhiebel, and B. Kaczer, "Analytic modeling of the bias temperature instability using capture/emission time maps," in *Electron Devices Meeting (IEDM), 2011 IEEE International*, pp. 27–4, 2011.

[59] D. Angot, V. Huard, L. Rahhal, A. Cros, X. Federspiel, A. Bajolet, Y. Carminati, M. Saliva, E. Pion, F. Cacho *et al.*, "Bti variability fundamental understandings and impact on digital logic by the use of extensive dataset," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, pp. 15–4, 2013.

[60] M. Toledano-Luque, B. Kaczer, J. Franco, P. Roussel, M. Bina, T. Grasser, M. Cho, P. Weckx, and G. Groeseneken, "Degradation of time dependent variability due to interface state generation," in *VLSI Technology (VLSIT), 2013 Symposium on*, pp. T190–T191, June 2013.

[61] T. Matsukawa, Y. Liu, W. Mizubayashi, J. Tsukada, H. Yamauchi, K. Endo, Y. Ishikawa, S. O'uchi, H. Ota, S. Migita, Y. Morita, and M. Masahara, "Suppressing vt and gm variability of finfets using amorphous metal gates for 14 nm and beyond," in *Electron Devices Meeting (IEDM), 2012 IEEE International*, pp. 8.2.1–8.2.4, Dec 2012.

[62] A. Veloso, G. Boccardi, L.-A. Ragnarsson, Y. Higuchi, J. Lee, E. Simoen, P. Roussel, M. Cho, S. Chew, T. Schram, H. Dekkers, A. Van Ammel, T. Witters, S. Brus, A. Dangol, V. Paraschiv, E. Vecchio, X. Shi, F. Sebaai, K. Kellens, N. Heylen, K. Devriendt, O. Richard, H. Bender, T. Chiarella, H. Arimura, A. Thean, and N. Horiguchi, "Highly scalable effective work function engineering approach for multi-vt modulation of planar and finfet-based rmg high-k last devices for (sub-)22nm nodes," in *VLSI Technology (VLSIT), 2013 Symposium on*, pp. T194–T195, June 2013.

[63] X. Yuan, T. Shimizu, U. Mahalingam, J. Brown, K. Habib, D. Tekleab, T.-C. Su, S. Satadru, C. Olsen, H.-W. Lee, L.-H. Pan, T. Hook, J.-P. Han, J.-E. Park, M.-H. Na, and K. Rim, "Transistor mismatch properties in deep-submicrometer cmos technologies," *Electron Devices, IEEE Transactions on*, vol. 58, no. 2, pp. 335–342, Feb 2011.

[64] M. Cho, J.-D. Lee, M. Aoulaiche, B. Kaczer, P. Roussel, T. Kauerauf, R. Degraeve, J. Franco, L. Ragnarsson, and G. Groeseneken, "Insight into n/pbti mechanisms in sub-1-nm-eot devices," *Electron Devices, IEEE Transactions on*, vol. 59, no. 8, pp. 2042–2048, Aug 2012.

[65] J. Franco, B. Kaczer, P. Roussel, J. Mitard, S. Sioncke, L. Witters, H. Mertens, T. Grasser, and G. Groeseneken, "Understanding the suppressed charge trapping in relaxed- and strained-ge/sio2/hfo2 pmosfets and implications for the screening of alternative high-mobility substrate/dielectric cmos gate stacks," in *Electron Devices Meeting (IEDM), 2013 IEEE International*, pp. 15.2.1–15.2.4, Dec 2013.

[66] K.-L. Chen, S. A. Saller, I. A. Groves, and D. B. Scott, "Reliability effects on mos transistors due to hot-carrier injection," *Solid-State Circuits, IEEE Journal of*, vol. 20, no. 1, pp. 306–313, 1985.

[67] 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 *Reliability Physics Symposium, 2009 IEEE International*, pp. 531–548, 2009.

[68] A. Tiwari and J. Torrellas, "Facelift: Hiding and slowing down aging in multicores," in *Microarchitecture, IEEE/ACM Int'l Symposium*, pp. 129–140, 2008.

[69] E. Takeda, C. Y.-W. Yang, and A. Miura-Hamada, *Hot-carrier effects in MOS devices*. Academic Press, 1995.

[70] "Predictive Technology Model," http://ptm.asu.edu/.

[71] W.-K. Yeh, W.-H. Wang, Y.-K. Fang, and F.-L. Yang, "Temperature dependence of hot-carrier-induced degradation in 0.1 μm soi nmosfets with thin oxide," *Electron Device Letters, IEEE*, vol. 23, no. 7, pp. 425–427, 2002.

[72] K. Arabi, R. Saleh, and M. Xiongfei, "Power supply noise in socs: metrics, management, and measurement," *Design & Test of Computers, IEEE*, vol. 24, no. 3, pp. 236–244, 2007.

[73] S. Nassif, "Power grid analysis benchmarks," in *Proceedings of the Asia and South Pacific Design Automation Conference (ASPDAC)*, pp. 376–381, 2008.

[74] K. Haghdad and M. Anis, "Power yield analysis under process and temperature variations," *IEEE Transactions on Very Large Scale Integration Systems (TVLSI)*, no. 99, pp. 1–10, 2011.

[75] S. Mukherjee, *Architecture design for soft errors*. Morgan Kaufmann, 2011.

[76] P. Shivakumar, M. Kistler, S. W. Keckler, D. Burger, and L. Alvisi, "Modeling the effect of technology trends on the soft error rate of combinational logic," in *Proceedings of the International Conference on Dependable Systems and Networks*, pp. 23–26, 2002.

[77] B. D. Sierawski, J. A. Pellish, R. A. Reed, R. D. Schrimpf, K. M. Warren, R. A. Weller, M. H. Mendenhall, J. D. Black, A. D. Tipton, M. A. Xapsos *et al.*, "Impact of low-energy proton induced upsets on test methods and rate predictions," *T-NS*, vol. 56, no. 6, pp. 3085–3092, 2009.

[78] D. F. Heidel, P. W. Marshall, K. A. LaBel, J. R. Schwank, K. P. Rodbell, M. C. Hakey, M. D. Berg, P. E. Dodd, M. R. Friendlich, A. D. Phan *et al.*, "Low energy proton single-event-upset test results on 65 nm soi sram," *T-NS*, vol. 55, no. 6, pp. 3394–3400, 2008.

[79] K. P. Rodbell, D. F. Heidel, H. H. Tang, M. S. Gordon, P. Oldiges, and C. E. Murray, "Low-energy proton-induced single-event-upsets in 65 nm node, silicon-on-insulator, latches and memory cells," *T-NS*, vol. 54, no. 6, pp. 2474–2479, 2007.

[80] http://en.wikipedia.org/wiki/Muon, accessed: 2015-01-21.

[81] J.-L. Autran, S. Semikh, D. Munteanu, S. Serre, G. Gasiot, and P. Roche, "Soft-error rate of advanced sram memories: Modeling and monte carlo simulation," *Numerical Simulation: From Theory to Industry*, 2012.

[82] F. Lei, S. Clucas, C. Dyer, and P. Truscott, "An atmospheric radiation model based on response matrices generated by detailed monte carlo simulations of cosmic ray interactions," *Nuclear Science, IEEE Transactions on*, vol. 51, no. 6, pp. 3442–3451, 2004.

[83] G. A. Sai-Halasz, M. R. Wordeman, and R. H. Dennard, "Alpha-particle-induced soft error rate in vlsi circuits," *T-ED*, vol. 29, no. 4, pp. 725–731, 1982.

[84] R. C. Baumann, "Radiation-induced soft errors in advanced semiconductor technologies," *Device and Materials Reliability, IEEE Transactions on*, vol. 5, no. 3, pp. 305–316, 2005.

[85] V. Chandra and R. Aitken, "Impact of technology and voltage scaling on the soft error susceptibility in nanoscale cmos," in *Proceedings of IEEE International Symposium on Defect and Fault Tolerance of VLSI Systems*, pp. 114–122, 2008.

[86] B. Gill, N. Seifert, and V. Zia, "Comparison of alpha-particle and neutron-induced combinational and sequential logic error rates at the 32nm technology node," in *IEEE International Reliability Physics Symposium*, pp. 199–205, 2009.

[87] N. N. Mahatme, S. Jagannathan, T. D. Loveless, L. W. Massengill, B. L. Bhuva, S.-J. Wen, and R. Wong, "Comparison of combinational and sequential error rates for a deep submicron process," *IEEE Transactions on Nuclear Science*, vol. 58, no. 6, pp. 2719–2725, 2011.

[88] K. Mohanram, "Simulation of transients caused by single-event upsets in combinational logic," in *Proceedings of IEEE International Test Conference (ITC)*, pp. 1–9, 2005.

[89] F. Wang and Y. Xie, "Soft error rate analysis for combinational logic using an accurate electrical masking model," *IEEE Transaction on Dependable and Secure Computing*, vol. 8, no. 1, pp. 137–146, 2011.

[90] R. R. Rao, K. Chopra, D. T. Blaauw, and D. M. Sylvester, "Computing the soft error rate of a combinational logic circuit using parameterized descriptors," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 26, no. 3, pp. 468–479, 2007.

[91] Y. S. Dhillon, A. U. Diril, and A. AChatterjee, "Soft-error tolerance analysis and optimization of nanometer circuits," in *Proceedings of the IEEE/ACM International Conference on Design, Automation and Test in Europe*, pp. 288–293, 2005.

[92] M. Omaña, G. Papasso, D. Rossi, and C. Metra, "A model for transient fault propagation in combinatorial logic," in *Proceedings of International On-Line Testing Symposium (IOLTS)*, pp. 111–115, 2003.

[93] R. Rajaraman, J. S. Kim, N. Vijaykrishnan, Y. Xie, and M. J. Irwin, "Seat-la: A soft error analysis tool for combinational logic," in *Proceedings of the 19th International Conference on VLSI Design*, pp. –, 2006.

[94] B. Zhang, W. Wang, and M. Orshansky, "Faser: Fast analysis of soft error susceptibility for cell-based designs," in *Proceedings of the 7th International Symposium on Quality Electronic Design*, pp. 755–760, 2006.

[95] M. Fazeli, S. G. Miremadi, H. Asadi, and S. N. Ahmadian, "A fast and accurate multi-cycle soft error rate estimation approach to resilient embedded systems design," in *International Conference on Dependable Systems and Networks*, pp. 131–140, 2010.

[96] T. Chiarella, L. Witters, A. Mercha, C. Kerner, M. Rakowski, C. Ortolland, L.-Å. Ragnarsson, B. Parvais, A. De Keersgieter, S. Kubicek *et al.*, "Benchmarking soi and bulk finfet alternatives for planar cmos scaling succession," *Solid-State Electronics*, vol. 54, no. 9, pp. 855–860, 2010.

[97] B. Yu, L. Chang, S. Ahmed, H. Wang, S. Bell, C.-Y. Yang, C. Tabery, C. Ho, Q. Xiang, T.-J. King *et al.*, "Finfet scaling to 10 nm gate length," in *IEDM*, pp. 251–254, 2002.

[98] A. Bansal, S. Mukhopadhyay, and K. Roy, "Device-optimization technique for robust and low-power finfet sram design in nanoscale era," *T-ED*, vol. 54, no. 6, pp. 1409–1419, 2007.

[99] T. Chiarella, L. Witters, A. Mercha, C. Kerner, R. Dittrich, M. Rakowski, C. Ortolland, L.-A. Ragnarsson, B. Parvais, A. De Keersgieter *et al.*, "Migrating from planar to finfet for further cmos scaling: Soi or bulk?" in *ESSCIRC*, pp. 84–87, 2009.

[100] K. Castellani-Couli, D. Munteanu, J. Autran, V. Ferlet-Cavrois, P. Paillet, and J. Baggio, "Analysis of 45-nm multi-gate transistors behavior under heavy ion irradiation by 3-d device simulation," *T-NS*, vol. 53, no. 6, pp. 3265–3270, 2006.

[101] M. Turowski, A. Raman, and W. Xiong, "Accurate modeling of soi multi-gate fets and their transient response to radiation," in *ULIS*, pp. 137–140, 2012.

[102] D. Ball, M. Alles, R. Schrimpf, and S. Cristoloveanu, "Comparing single event upset sensitivity of bulk vs. soi based finfet sram cells using tcad simulations," in *SOI Conference*, pp. 1–2, 2010.

[103] Y.-P. Fang and A. S. Oates, "Neutron-induced charge collection simulation of bulk finfet srams compared with conventional planar srams," *T-DMR*, vol. 11, no. 4, pp. 551–554, 2011.

[104] L. Artola, G. Hubert, and R. Schrimpf, "Modeling of radiation-induced single event transients in soi finfets," in *IRPS, 2013*, pp. SE–1.

[105] F. Wang, Y. Xie, K. Bernstein, and Y. Luo, "Dependability analysis of nano-scale finfet circuits," in *Emerging VLSI Technologies and Architectures*, pp. 6–pp, 2006.

[106] H. Liu, M. Cotter, S. Datta, and V. Narayanan, "Technology assessment of si and iii-v finfets and iii-v tunnel fets from soft error rate perspective," in *IEDM*, pp. 25–5, 2012.

[107] V. Ramakrishnan and R. Srinivasan, "Soft error study in double gated finfet-based sram cells with simultaneous and independent driven gates," *Microelectronics Journal*, vol. 43, no. 11, pp. 888–893, 2012.

[108] R. C. Baumann, "Soft errors in advanced semiconductor devices-part i: the three radiation sources," *T-DMR*, vol. 1, no. 1, pp. 17–22, 2001.

[109] N. Seifert, B. Gill, S. Jahinuzzaman, J. Basile, V. Ambrose, Q. Shi, R. Allmon, and A. Bramnik, "Soft error susceptibilities of 22 nm tri-gate devices," *Nuclear Science, IEEE Transactions on*, vol. 59, no. 6, pp. 2666–2673, 2012.

[110] F. El-Mamouni, E. Zhang, N. Pate, N. Hooten, R. Schrimpf, R. Reed, K. Galloway, D. McMorrow, J. Warner, E. Simoen *et al.*, "Laser-and heavy ion-induced charge collection in bulk finfets," *T-NS*, vol. 58, no. 6, pp. 2563–2569, 2011.

[111] S. Agostinelli, J. Allison, K. e. Amako, J. Apostolakis, H. Araujo, P. Arce, M. Asai, D. Axen, S. Banerjee, G. Barrand *et al.*, "Geant4- a simulation toolkit," *Nuclear instruments and methods in physics research section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 506, no. 3, pp. 250–303, 2003.

[112] J. Allison, K. Amako, J. Apostolakis, H. Araujo, P. A. Dubois, M. Asai, G. Barrand, R. Capra, S. Chauvie, R. Chytracek *et al.*, "Geant4 developments and applications," *T-NS*, vol. 53, no. 1, pp. 270–278, 2006.

[113] X. Wang, B. Cheng, A. Brown, C. Millar, J. Kuang, S. Nassif, and A. Asenov, "Statistical variability and reliability and the impact on corresponding 6t-sram cell design for a 14-nm node soi finfet technology," *IEEE Design and Test*, vol. PP, no. 99, 2013.

[114] http://web.eecs.umich.edu/ jhayes/iscas.restore/.

[115] 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 46th Annual Design Automation Conference*, pp. 514–519, 2009.

[116] J. Bhaskarr Velamala, K. B. Sutaria, H. Shimizu, H. Awano, T. Sato, G. Wirth, and Y. Cao, "Compact modeling of statistical bti under trapping/detrapping," *Electron Devices, IEEE Transactions on*, vol. 60, no. 11, pp. 3645–3654, 2013.

[117] H. Kukner, S. Khan, P. Weckx, P. Raghavan, S. Hamdioui, B. Kaczer, F. Catthoor, L. Van der Perre, R. Lauwereins, and G. Groeseneken, "Comparison of reaction-diffusion and atomistic trap-based bti models for logic gates," *Device and Materials Reliability, IEEE Transactions on*, vol. 14, no. 1, pp. 182–193, 2014.

[118] S. Khan, S. Hamdioui, H. Kukner, P. Raghavan, and F. Catthoor, "Bti impact on logical gates in nano-scale cmos technology," in *International Symposium on Design and Diagnostics of Electronic Circuits & Systems (DDECS)*, 2012.

[119] V. V. Camargo, B. Kaczer, G. Wirth, T. Grasser, and G. Groeseneken, "Use of ssta tools for evaluating bti impact on combinational circuits," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 22, no. 2, pp. 280–285, 2014.

[120] H. Kukner, P. Weckx, S. Morrison, P. Raghavan, B. Kaczer, F. Catthoor, L. V. d. Perre, R. Lauwereins, and G. Groeseneken, "Nbti aging on 32-bit adders in the downscaling planar fet technology nodes," in *Digital System Design (DSD), 2014 17th Euromicro Conference on*, pp. 98–107, 2014.

[121] http://www.altos-da.com/.

[122] http://www.synopsys.com.

[123] V. Veetil, D. Sylvester, and D. Blaauw, "Efficient monte carlo based incremental statistical timing analysis," in *Design Automation Conference, 2008. DAC 2008. 45th ACM/IEEE*, pp. 676–681, 2008.

[124] S. V. Kumar, C. H. Kim, and S. S. Sapatnekar, "Adaptive techniques for overcoming performance degradation due to aging in cmos circuits," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,*, vol. 19, no. 4, pp. 603–614, 2011.

[125] Z. Qi and M. R. Stan, "Nbti resilient circuits using adaptive body biasing," in *Proceedings of the 18th ACM Great Lakes symposium on VLSI*, pp. 285–290, 2008.

[126] H. Mostafa, M. Anis, and M. Elmasry, "Adaptive body bias for reducing the impacts of nbti and process variations on 6t sram cells," *IEEE Transactions on Circuits and Systems I: Regular Papers,*, vol. 58, no. 12, pp. 2859–2871, 2011.

[127] A. Calimera, E. Macii, and M. Poncino, "Nbti-aware clustered power gating," *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, vol. 16, no. 1, p. 3, 2010.

[128] F. Oboril and M. B. Tahoori, "Extratime: Modeling and analysis of wearout due to transistor aging at microarchitecture-level," in *Dependable Systems and Networks (DSN), 2012 42nd Annual IEEE/IFIP International Conference on*, pp. 1–12, 2012.

[129] A. Calimera, E. Macii, and M. Poncino, "Nbti-aware power gating for concurrent leakage and aging optimization," in *Proceedings of the 14th ACM/IEEE international symposium on Low power electronics and design*, pp. 127–132, 2009.

[130] L. Zhang and R. P. Dick, "Scheduled voltage scaling for increasing lifetime in the presence of nbti," in *Asia and South Pacific Design Automation Conference(ASP-DAC)*, pp. 492–497, 2009.

[131] M. Basoglu, M. Orshansky, and M. Erez, "Nbti-aware dvfs: a new approach to saving energy and increasing processor lifetime," in *Proceedings of the 16th ACM/IEEE international symposium on Low power electronics and design*, pp. 253–258, 2010.

[132] J. Fang, S. Gupta, S. V. Kumar, S. K. Marella, V. Mishra, P. Zhou, and S. S. Sapatnekar, "Circuit reliability: from physics to architectures," in *Proceedings of the International Conference on Computer-Aided Design*, pp. 243–246, 2012.

[133] B. Paul, K. Kang, H. Kufluoglu, M. A. Alam, and K. Roy, "Negative bias temperature instability: Estimation and design for improved reliability of nanoscale circuits," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 26, no. 4, pp. 743–751, 2007.

[134] W. Wang, Z. Wei, S. Yang, and Y. Cao, "An efficient method to identify critical gates under circuit aging," in *International Conference on Computer-Aided Design*, pp. 735–740, 2007.

[135] J. Chen, S. Wang, and M. Tehranipoor, "Efficient selection and analysis of critical-reliability paths and gates," in *Proceedings of the ACM Great Lakes symposium on VLSI (GLSVLSI)*, pp. 45–50, 2012.

[136] F. Oboril and M. Tahoori, "MTTF-Balanced pipeline design," in *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 1–6, 2013.

[137] X. Yang and K. Saluja, "Combating nbti degradation via gate sizing," in *International Symposium on Quality Electronic Design (ISQED)*, pp. 47–52, 2007.

[138] Y. Wang, X. Chen, W. Wang, V. Balakrishnan, Y. Cao, Y. Xie, and H. Yang, "On the efficacy of input vector control to mitigate nbti effects and leakage power," in *International Symposium on Quality Electronic Design (ISQED)*, pp. 19–26, 2009.

[139] F. Firouzi, S. Kiamehr, and M. B. Tahoori, "Nbti mitigation by optimized nop assignment and insertion," in *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 218–223, 2012.

[140] J. Abella, X. Vera, and A. Gonzalez, "Penelope: The NBTI-aware processor," in *International Symposium on Microarchitecture*, pp. 85–96, 2007.

[141] D. Bild, R. Dick, and G. Bok, "Static NBTI reduction using internal node control," *ACM Transactions on Design Automation of Electronic Systems*, vol. 17, no. 4, p. 45, 2012.

[142] Y. Wang, H. Luo, K. He, R. Luo, H. Yang, and Y. Xie, "Temperature-aware NBTI modeling and the impact of input vector control on performance degradation," in *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 1–6, 2007.

[143] J. M. Rabaey, A. P. Chandrakasan, and B. Nikolic, *Digital integrated circuits*. Prentice hall Englewood Cliffs, 2002, vol. 2.

[144] K. Kang, H. Kufluoglu, M. Alain, and K. Roy, "Efficient transistor-level sizing technique under temporal performance degradation due to nbti," in *International Conference on Computer Design (ICCD)*, pp. 216–221, 2007.

[145] S. Basu and R. Vemuri, "Process variation and nbti tolerant standard cells to improve parametric yield and lifetime of ics," in *IEEE Computer Society Annual Symposium on VLSI*, pp. 291–298, 2007.

[146] M. B. da Silva, V. V. Camargo, L. Brusamarello, G. I. Wirth, and R. da Silva, "Nbti-aware technique for transistor sizing of high-performance cmos gates," in *IEEE Latin-American Test Workshop (LATW)*, pp. 1–5, 2009.

[147] F. Firouzi, S. Kiamehr, M. Tahoori, and S. Nassif, "Incorporating the impacts of workload-dependent runtime variations into timing analysis," in *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 1022–1025, 2013.

[148] D. Schroder and J. Babcock, "Negative bias temperature instability: Road to cross in deep submicron silicon semiconductor manufacturing," *Journal of Applied Physics*, vol. 94, p. 1, 2003.

[149] R. Vattikonda, W. Wang, and Y. Cao, "Modeling and minimization of PMOS NBTI effect for robust nanometer design," in *Proc. Design Automation Conf.*, pp. 1047–1052, 2006.

[150] F. Firouzi, S. Kiamehr, and M. Tahoori, "A linear programming approach for minimum nbti vector selection," in *Proceedings of the 21st edition of the great lakes symposium on Great lakes symposium on VLSI*, pp. 253–258, 2011.

[151] D. Bild, G. Bok, and R. Dick, "Minimization of NBTI performance degradation using internal node control," in *Proc. Design, Automation and Test in Europe Conf.*, pp. 148–153, 2009.

[152] H. Luo, Y. Wang, K. He, R. Luo, H. Yang, and Y. Xie, "A novel gate-level NBTI delay degradation model with stacking effect," *Integrated Circuit and System Design. Power and Timing Modeling, Optimization and Simulation*, pp. 160–170, 2007.

[153] K. Kim, H. Nan, and K. Choi, "Adaptive HCI-aware power gating structure," in *ISQED Int'l Symposium*, pp. 219–224, 2010.

[154] K. Wu and D. Marculescu, "Joint logic restructuring and pin reordering against nbti-induced performance degradation," in *Proceedings of the Conference on Design, Automation and Test in Europe*, pp. 75–80, 2009.

[155] K. Chen, S. Saller, and R. Shah, "The case of AC stress in the hot-carrier effect," *Electron Devices, IEEE Transactions on*, vol. 33, no. 3, pp. 424–426, 2005.