Circuits and Systems
Mekelweg 4,
2628 CD Delft
The Netherlands
http://ens.ewi.tudelft.nl/

CAS-2017-4469585

M.Sc. Thesis

A System-Level Aging and Mitigation Assessment Simulation Framework

Evelyn Rashmi Jeyachandra

Abstract

As technology scaling enters the nanometer regime, device aging effects cause quality and reliability issues in CMOS Integrated Circuits (ICs), which in turn shorten its lifetime. Evaluating system aging through circuit simulations is very complex and time consuming. In this thesis, a framework is proposed, which allows for the evaluation of long-term aging effects of ICs and the corresponding measures to counteract premature failure. The focus of this work lies in the abstraction of low-level aging models to system-level models, in order to facilitate swift high-level simulation, without any knowledge of underlying circuit dynamics. Two major aging mechanisms, namely Negative Bias Temperature Instability (NBTI) and Channel Hot Carrier (CHC) degradation are considered for analysis. System-level aging management is performed with the prototype of a System-on-Chip (SoC) including a Management Unit (MU), which counteracts aging by employing Dynamic Voltage Scaling (DVS), Dynamic Frequency Scaling (DFS), and Adaptive Body Biasing (ABB). The simulation platform prototype is based on System-C AMS and a 65-nm technology library. This SoC simulation computes path delay using characterized models, which represent the aged behaviour of individual circuit elements. Results show that the obtained values are within 2% of circuit-level simulation values at the cost of a simulation time which is 15 × lesser than conventional circuit simulators (e.g. Cadence NCSim).

A System-Level Aging and Mitigation Assessment Simulation Framework

Thesis

submitted in partial fulfillment of the
requirements for the degree of

Master of Science

in

Embedded Systems

by

Evelyn Rashmi Jeyachandra
born in Madurai, India

This work was performed in:

Circuits and Systems Group
Department of Microelectronics & Embedded Systems
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology

Delft University of Technology

Delft University of Technology
Department of
Microelectronics & Embedded Systems

The undersigned hereby certify that they have read and recommend to the Faculty of Electrical Engineering, Mathematics and Computer Science for acceptance a thesis entitled **"A System-Level Aging and Mitigation Assessment Simulation Framework"** by **Evelyn Rashmi Jeyachandra** in partial fulfillment of the requirements for the degree of **Master of Science**.

Dated: 12/07/2017

Chairman: _____
prof.dr.ir. A.J. van der Veen

Advisor: _____
dr.ir. T.G.R.M. van Leuken

Committee Members: _____
dr.ir. Nick van der Mijs

_____
prof.dr.ir. Said Hamdioui

_____
dr.ir. Sumeet Kumar

Abstract

As technology scaling enters the nanometer regime, device aging effects cause quality and reliability issues in CMOS Integrated Circuits (ICs), which in turn shorten its lifetime. Evaluating system aging through circuit simulations is very complex and time consuming. In this thesis, a framework is proposed, which allows for the evaluation of long-term aging effects of ICs and the corresponding measures to counteract premature failure. The focus of this work lies in the abstraction of low-level aging models to system-level models, in order to facilitate swift high-level simulation, without any knowledge of underlying circuit dynamics. Two major aging mechanisms, namely Negative Bias Temperature Instability (NBTI) and Channel Hot Carrier (CHC) degradation are considered for analysis. System-level aging management is performed with the prototype of a System-on-Chip (SoC) including a Management Unit (MU), which counteracts aging by employing Dynamic Voltage Scaling (DVS), Dynamic Frequency Scaling (DFS), and Adaptive Body Biasing (ABB). The simulation platform prototype is based on System-C AMS and a 65-nm technology library. This SoC simulation computes path delay using characterized models, which represent the aged behaviour of individual circuit elements. Results show that the obtained values are within 2% of circuit-level simulation values at the cost of a simulation time which is 15 $\times$ lesser than conventional circuit simulators (e.g. Cadence NCSim).

Acknowledgments

Firstly, I would like to thank Prof. Dr. Rene for giving me the opportunity to work under his supervision. Right from patiently helping me get started to the final steps, your counsel was very important in completing this work. Shaping my own course of action in the past nine months gave me many possibilities to pursue, which was quite overwhelming initially. Your expertise, knowledge, and guidance invariably gave me the confidence to shape my work. Thank you, Dr. Rene.

I would also like to thank Dr. Amir for instilling in me the passion of going an extra step to learn new things. Every one of our discussions provided me with a clearer picture. Thank you, Dr. Amir. Alexander, thank you for listening to my ideas and offering your insights. Sumeet, thank you for always sitting through my presentations with a smile. Your questions helped me look into several new aspects throughout my work.

Antoon, thank you for helping me fix system issues in a heartbeat. Minaksie, thank you for creating a cheerful aura and helping me with all the administrative tasks.

Appa and amma, everything I am, I owe it to you. Thank you for believing in me. Though we are miles apart, your love gives me strength even during the toughest of days. Benin anna and Linda anni, thank you for always being there for me and never letting me doubt myself. I am lucky to be your little sister. Jeba, you have always made my troubles seem lighter than they appear to be. Thank you for all your love and encouragement. Rizwan, I have no words to express how much I am grateful for your support. You are the best person anyone can ever have for a friend.

Evelyn Rashmi Jeyachandra
Delft, The Netherlands
12/07/2017

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

Aggressive scaling down of devices has led to the advent of more powerful and compact inventions. Moore's law states that the number of transistors per chip doubles every two years. As per the prediction, the current average mobile device has close to a billion transistors. However, the industry expects no compromise on the performance. This leads to a situation where shrinking dimensions are not to affect the device life expectancy.

Integrated Circuits (ICs) consist of CMOS devices predominantly, owing to their performance improvement and power reduction as compared to other technologies. Unfortunately, as technology scaling reaches the nanometer regime, CMOS devices are exposed to wide range of quality and reliability issues [1]. Specifically, nanoscale ICs are more susceptible to failure due to increased power density, increased electric fields across the gate dielectric and environmental fluctuations. Ensuring the reliability of ICs at design-time and run-time is of utmost importance. This must be guaranteed for several safety-critical applications such as avionic, automotive, medical, etc.

Scaling and Reliability

Figure 1.1[1] shows the half-pitch and gate length trends in the future years according to the International Technology Roadmap for Semiconductors.

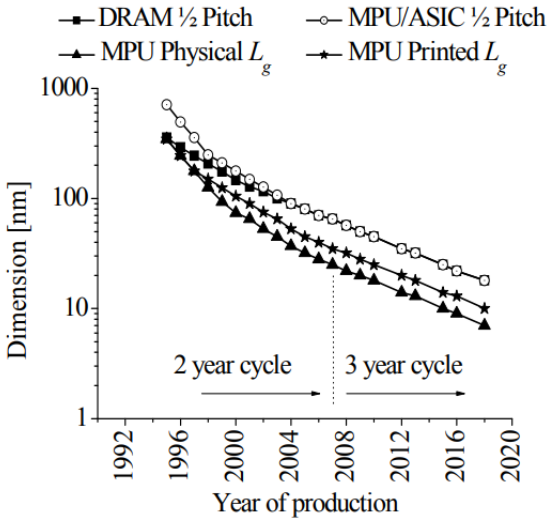


Figure 1.1: Technology half-pitch and gate length trends[1]

As per the trendline, technology generations have shifted from introducing a new technology node once every two years to once every three years. Dimensions lesser than 10 nm are in the imminent future. Device scaling has followed the ideal-scaling principle till the sub-100 nm space [2]. According to this principle, electric fields in the transistor are kept constant by scaling down the device dimensions (transistor length L, transistor width W, oxide thickness t_{ox}) and the supply voltage (V_{DD}) by the same factor. However, for dimensions smaller than 100 nm, supply voltage cannot be scaled on par with the device dimensions, as threshold voltage cannot be lowered further. As the physical thickness of the gate dielectric (t_{ox}) drops below 1.2 nm as in 65 nm process technology nodes, quantum tunneling phenomenon becomes very significant leading to large leakage currents [3]. This problem can be circumnavigated by using high-κ gate dielectric for 45 nm nodes and below. As scaling continues and reaches the predicted limit of 7nm, the adverse effects on reliability become more pronounced.

Owing to non-ideal scaling below 100 nm, the electric field across the gate dielectric is very large, resulting in failure mechanisms like Time Dependent Dielectric Breakdown (TDDB), Hot Carrier Injection (HCI), etc. As more and more devices are packed together, the power density increases. The circuit temperature rises subsequently, and this is worsened by high leakage currents in the devices. High temperature accelerates other failure mechanisms, such as Negative Bias Temperature Instability (NBTI), Positive Bias Temperature Instability (PBTI), etc. These effects substantially reduce the circuit lifetime. Countermeasures need to be taken at both design-time and run-time to overcome premature aging and device failure.

Embedded systems used in critical applications are expected to have a lifetime ranging from 10 years for automotive to 25 years for avionic devices [4]. Elevated aging in new technology nodes affect system life expectancy. Moreover, depending on the environment in which the system operates, its run-time variability may increase, complicating the process of ensuring reliability. Hard real-time systems have to meet strict deadline requirements and deadline misses will result in functional failures. Aging may cause an increase in execution times leading to hard real-time system malfunction. It is, therefore, imperative to model and mitigate reliability flaws from transistor-level up to system-level [4].

1.2 Thesis Goals

A lot of research has been conducted in the analysis and modelling of aging effects at the circuit-level. This thesis aims to attain the following:

- Abstract device-level mathematical aging models to system-level models, expediting the system-level aging assessment process without any knowledge of underlying circuit dynamics. While circuit-level aging analysis is widely prevalent, the same cannot be applied at the system-level. A typical SoC comprises of several units (processor, logic unit, memory, etc.). In such SoCs, using circuit-level simulations to perform system aging assessment can be very complicated and time consum-

ing. Enabling aging assessment at the system-level can overcome this complexity, and minimize the time and resource requirements between the system design and development cycle.

- Include low-level aging effects, such as NBTI and CHC, in the system-level aging assessment process. Design an accurate and fast system aging assessment prototype platform.

- Determine the expected best-case and worst-case corners of system operation, assuming that its critical path is identified initially. Charting the system's working bounds provides an estimate of its reliability factor.

- Maintain system reliability through a Management Unit, which mitigates aging by increasing the supply voltage (V_{DD}), decreasing the operating frequency (f), or adjusting the bulk-to-source voltage (V_{BS}).

1.3 Contributions

The main contributions of the thesis are as follows:

- A prototype of the first system-level aging assessment platform is developed, which does not require any extensive circuit-level simulations.

- Following a bottom-up approach, the models that represent aged behaviour in transistors, combinational cells (gates) and sequential cells (flip-flops) are identified. Various gates exhibit different delay characteristics. They are classified based on their type, and characterized to form a function library. Thus, these circuit-level models are used to model system-level aging.

- An appropriate aging sensor is determined to complete the aging-aware prototype platform. Three prominent sensor architectures are investigated, and the appropriate one is chosen to be characterized and included in the classified library.

- An SoC is simulated by using the library elements to form the critical/near-critical path. The simulation platform is based on System-C AMS and a 65-nm technology. The calculated delay figures are accurate in representing the actual values with a maximum error of 2%. The System-C AMS implementation has a shorter execution time with an approximate speedup of 15 times over conventional circuit simulators (e.g. Cadence NCSim).

- A Management Unit (MU) is designed and integrated with the simulated system. The MU monitors the system for probable delay faults, and adjusts the operating parameters (V_{DD}, f, and V_{BS}) accordingly, thereby allowing the investigation of the mitigation procedures.

1.4 Outline

The remainder of this thesis is organised as follows:

Chapter 2 describes the fundamental causes of aging and the related phenomena. A detailed account on the effects of variability in the nanoscale devices is given. Related work in the analysis of aging mechanisms is presented. The advantages and shortcomings of aging are also discussed.

Chapter 3 details the models used to represent aging in CMOS transistors. Among the various aging mechanisms, the following are chosen - Negative Bias Temperature Instability (NBTI) and Channel Hot Carrier (CHC) degradation. Plots that show the nature of threshold voltage variations at the transistor-level are presented. The dependence of the threshold voltage change on various factors, such as V_{DD}, T and the type of V_{th} device, are illustrated. It also elaborates on the manifestations of aging at the next level of abstraction. The change in transistor V_{th} translates to a corresponding increase in the gate delay (for combinatorial elements), setup time and flip-flop delay (for sequential elements). The dependence of rise and fall propagation gate delay on the input slew rate with aging is evaluated. Additionally, the influence of other factors such as fan-out, fan-in are studied. Various aging sensor architectures are discussed, and the sensitivity of the chosen aging sensor to delay faults due to aging is presented.

Chapter 4 details the implementation specifics. The proposed system-level aging assessment and mitigation platform is described. The System-C AMS implementations of gates, flip-flops and aging sensor are elaborated. The possible mitigation measures are discussed, and the management unit implementation that executes these measures is explained.

Chapter 5 discusses the simulation results for three individual circuits. The system's behaviour with best-case and worst-case conditions, and its response to the mitigation measures are plotted.

Chapter 6 summarizes the thesis outcomes. Recommendations for future endeavours, based on the results, are proposed.

2 Background

2.1 CMOS Device Failure

Potential failures in CMOS Intergrated Circuits (ICs) can be classified as intrinsic failures, extrinsic failures and electrical stress failures [5]. Intrinsic failures arise from manufacturing discrepancies at the silicon or die level [6]. Extrinsic failures usually occur at the interconnection or packaging level. Improper electrical connections between the IC pads and external power sources lead to potential extrinsic faults. Electrical Over Stress (EOS), and Electrostatic Discharge (ESD) due to improper handling may lead to electrical stress failures. The course of an IC's lifetime is usually modelled by the bathtub curve, as shown in Figure 2.1.

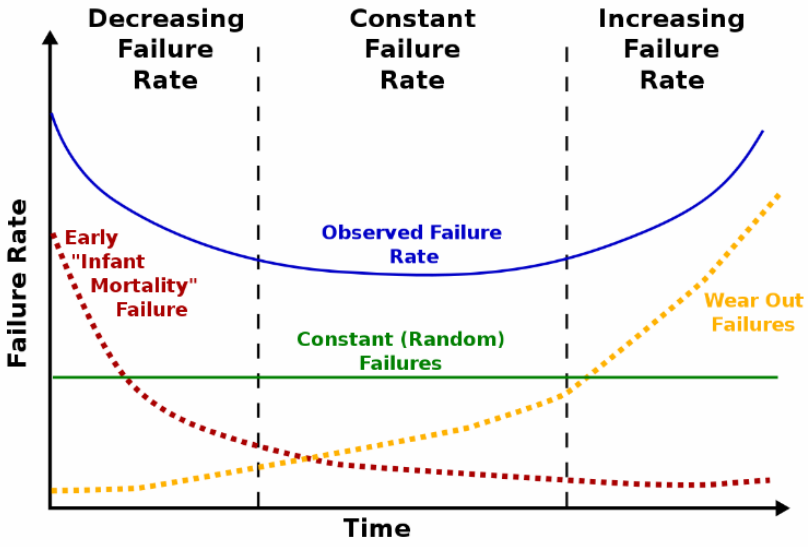


Figure 2.1: Bathtub curve showing failure rate vs. time

As seen from Figure 2.1, there are different kinds of failures in three distinct periods. The first stage represents "infant mortality", wherein, the failure rate is high due to manufacturing defects. Subjecting the devices to a burn-in process (stressing at high temperature and high supply voltage) eliminates devices prone to infant mortality before they are shipped. During its lifetime, the device enters the "normal operation state" of the curve, which is almost constant. Towards the end of the lifetime, the failure rate increases as the device experiences "wearout". As the technology nodes evolve, the bathtub curve moves higher on the graph, and the normal operation state has a nearly increasing failure rate trend. This is mainly attributed to time-dependent intrinsic failure mechanisms due to aging.

2.2 Variability

Device aging stems from transistors that have been in operation for a long period of time, resulting in their wearout. This is accelerated by other mechanisms that are becoming increasingly prominent with shrinking device dimensions. Currently, manufacturers set clock speeds to a safe limit, so that the wearout is less obvious. Some of the recent trends that contribute to faster aging are scaled down device dimensions, high clock speeds, high operating frequencies and heat generation, new fabrication processes, very small difference between the threshold voltage and supply voltage, etc.

Variability is becoming one of the leading causes for chip failures and delayed schedules in nanometer technologies [2]. The continued miniaturization of circuits leads to increased variations in transistor parameters. This makes analog and digital circuits susceptible to premature failure. Variability reduces the circuit's ability to deliver the correct functionality within the specified time frame. Variability has several root causes [7], namely,

- Static variability - Variations in process parameters such as oxide thickness (t_{ox}), gate length (L) and gate width (W).
- Temporal variability - Variations in operating voltage and temperature, aging, radiation effects, crosstalk, IR-drop, etc.

Device aging is associated with operation-dependent variability, specifically, PVT-variations. Aging effects can be measured in transistors one-by-one using microscopic electrodes. However, this method is very expensive, and given the large number of transistors in a chip, it is virtually impossible to probe every transistor during its lifetime in order to evaluate the remaining lifetime of the chip. Thus, it is imperative to model aging effects using predictive approaches so that methodologies that mitigate these effects can be framed. Normally, manufacturers handle aging degradation by building high design margins, using wide guard-bands, or both. Employing such methods is very conservative and does not exploit the full potential of improved performance that nanoscale devices have to offer. By incorporating aging models, the circuit behavior can be predicted at design time. Accordingly, counter-measures can be charted to extend the chip lifetime during run-time.

2.3 Reliability Framework

The reliability-aware aging framework is presented in Figure 2.2. It is built on a firm understanding of the physical and thermodynamic processes that lead to failure mechanisms, such as Negative and Postive Bias Temperature Instability (NBTI/PBTI), Channel Hot Carrier (CHC) degradation, Time Dependent Dielectric Breakdown (TDDB), Electromigration (EM), Thermal Cycling (TC), etc. With the aid of the knowledge of these wearout processes, models that predict device-level and circuit-level degradation can be used in high-level reliability assessment.

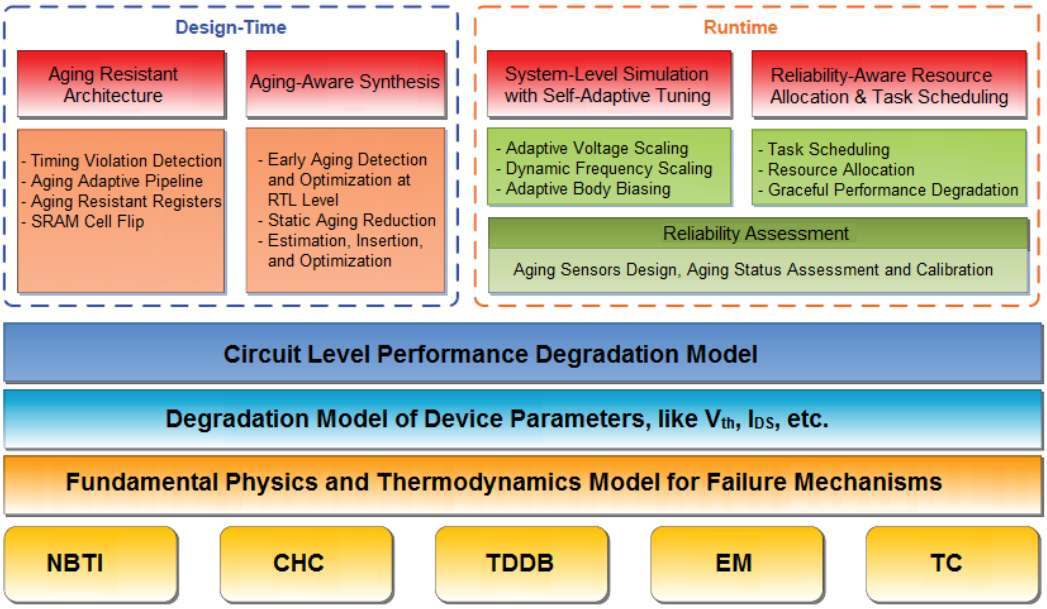


Figure 2.2: Framework for reliability [5]

The remainder of this chapter elaborates the various aging failure mechanisms at the transistor level, the physics behind such phenomena, and the damages to the circuit.

2.4 CMOS Aging Mechanisms

2.4.1 Negative Bias Temperature Instability

Negative Bias Temperature Instability is a primary reliability concern for pMOS transistors. This increases the threshold voltage of the transistor, thereby slowing it down. A pMOS transistor has n-type body with p-type source and drain regions. Thus, the majority carriers are holes. The pMOS transistor turns on when a negative gate-to-source voltage is applied, which creates an inversion layer near the gate that acts as the conducting channel. Extensive research has been made into the NBTI phenomenon. Two main theories that explain both PBTI and NBTI mechanisms are

- Reaction-Diffusion (R-D model)
- Trapping-Detrapping (T-D model)

A few differences that exist between the two models are as follows - the trap-based model can capture fast trapping behaviour and, thus, can differentiate between different input sequences, which the R-D model lacks [8]. Simulations based on the T-D model can assess degradation for shorter intervals effectively. The detailed nature of the model causes a major bottleneck for long-term simulations [8]. On the other hand, the R-D model simplifies BTI interpretation [8], hastening long-term simulation. The following explanation of both the models is based on the NBTI effect in pMOS transistors. However, the same models can be used to describe the PBTI mechanism

in nMOS transistors.

Reaction-Diffusion Model

When pMOS transistors are manufactured, most silicon (Si) atoms bond with oxygen to form the silicon dioxide (SiO_2) insulator between the n-well and the gate. However, every manufacturing environment may not be ideal. Due to certain inaccuracies in the processes, some of the silicon atoms may bond with hydrogen atoms at the interfacial layer (layer between the n-well and the SiO_2 substrate).

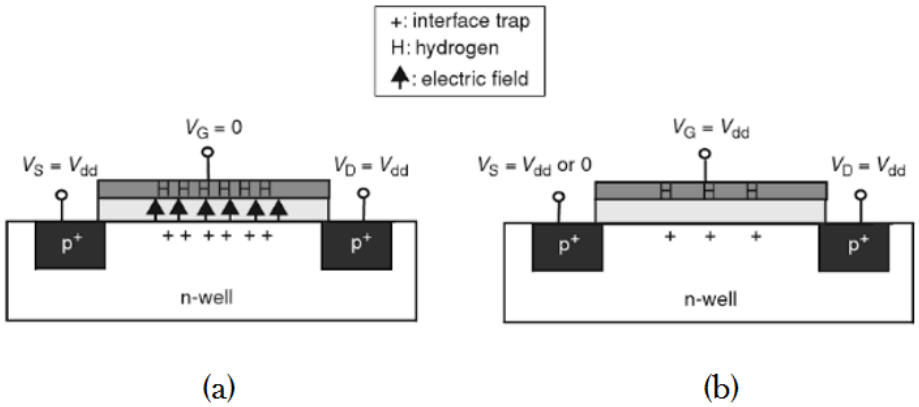


Figure 2.3: R-D model of NBTI in pMOS transistor (a) Stress Phase (b) Recovery Phase [9]

During the transistor operation, a negative gate-to-source voltage is applied so it turns ON ($V_G = 0$, $V_S = V_{DD}$). Negative V_{GS} repels the electrons in the n-well so as to form holes below the SiO_2 substrate, forming a positive conducting channel (holes) between the source and drain, thereby, turing the transistor ON. This, coupled with a high temperature, causes the hydrogen atoms to break free from their bond with silicon atoms. Since the gate is negatively charged, the hydrogen atoms crowd near the gate, leaving holes (due to dangling silicon atoms) at the interfacial layer, as shown in Figure 2.3(a). This is called the **stress phase** of NBTI.

When the negative gate voltage is removed ($V_G = V_{DD}$ and $V_S = V_{DD}$), the hydrogen atoms drift back toward the Si-SiO_2 interface, recombining with the silicon atoms to form Si-H bonds again. Therefore, the NBTI phenomenon is partially reversible. This phase is called the **recovery phase**, as shown in Figure 2.3(b). The recovery phase can be accelerated by applying a positive gate voltage ($V_G = V_{DD}$ and $V_S = 0$). However, as seen from Figure 2.3(b), the recovery is not complete, and a few hydrogen atoms still remain near the gate permanently. This leads to an accumulation of holes in the substrate [9]. The stress-recovery combination is referred to as **dynamic NBTI**. On the other hand, if the transistor is constantly under stress, it is said to undergo **static NBTI**. Dynamic NBTI is more accurate in representing actual transistor operation because a transistor does not remain ON always.

Post the recovery phase, the few holes at the interfacial layer that are not recovered, are filled by electrons from the n-well. Thus, when the transistor is turned ON the next time, higher voltage needs to be applied to repel those electrons in order to form the conducting channel below the substrate. Specifically, this voltage increase essentially translates to an elevated threshold voltage. As the transistor is in continuous operation, the threshold voltage undergoes degradation, which is elevated at high temperatures. Ultimately, this worsens the switching speed of the transistor [9].

Trapping-Detrapping Model

The trapping-detrapping model is proposed as an alternative to explain the Bias Temperature Instability phenomenon. One of the primary advantages of using T-D based models is that, they exhibit a logarithmic dependence on the time evolution of BTI, thereby avoiding an overly pessimistic guard band for reliability [10].

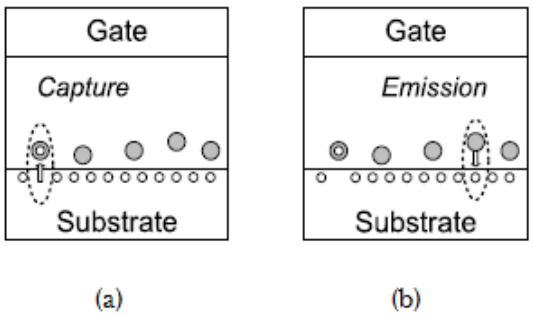


Figure 2.4: (a) Charge trapping component and (b) Charge detrapping component of T-D model [10]

Interface traps are electrically active defects that are present along the interface between the gate dielectric and the substrate. Figure 2.4(a) shows that the interface traps capture the charge carriers (holes) responsible for the flow of current between the source and drain of the pMOS transistor. This reduces the drain current because the number of holes in the channel decreases. Thus, the charge traps essentially increase the transistor threshold voltage. This corresponds to the stress phase of the R-D model. The trap occupation probability increases with gate bias and temperature, and is independent of the stress time [11]. On the other hand, when the trapped carriers are released due to positive V_{GS}, it results in recovery and leads to a decrease in the number of occupied traps [11]. Overall, the trapping-detrapping events lead to discrete shifts in the threshold voltage. Current research has revealed that the statistical probability of trapping depends on the time constants, the number of traps, the location of traps and the trap energies [11]. However, much effort is needed to develop an accurate, compact aging model based on this mechanism. Thus, the R-D model will be used for further modelling purposes as will be elaborated in Chapter 3.

2.4.2 Positive Bias Temperature Instability

NBTI occurs only in pMOS transistors, whereas Positive Bias Temperature Instability (PBTI) occurs only in nMOS transistors. When a positive gate-to-source voltage is applied to the nMOS transistor, it is in inversion and experiences PBTI. Similar to NBTI, PBTI effect can also be explained using the R-D and T-D models. PBTI was not very significant in earlier technology nodes. With the recent use of high-κ metal gates, it caused almost the same amount of V_{th} degradation as NBTI. Since further analysis is based on devices with SiO_2 as the gate dielectric, the effect of PBTI is neglected.

2.4.3 Hot Carrier Injection

This phenomenon affects the reliability of nMOS transistors primarily. Unlike the BTI effect which affects transistors when they are ON, the HCI effect is predominant during the switching phase.

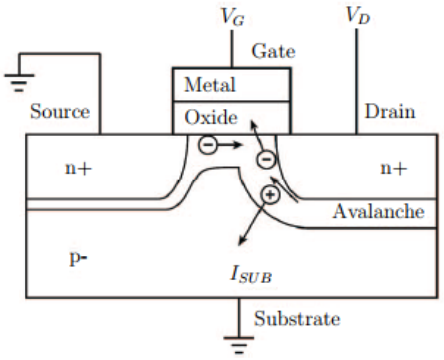


Figure 2.5: Channel hot carrier degradation in nMOSFET [11]

Channel Hot Carrier (CHC) effect is one of the mechanisms contributing to HCI. When an nMOS transistor conducts due to a positive V_{GS}, few carriers may gain tremendous energy due to large electric fields. As these carriers move along the channel from the source to drain, their energy gain becomes sufficient to surpass the potential barrier of the Si/SiO_2 interface, and leave the channel [11]. Such carriers are referred to as "hot carriers", because their energy in the conduction channel is higher than their energy in thermal equilibrium. As shown in Figure 2.5, the hot carriers may get lodged in the gate dielectric and damage it. This charge trapping leads to an increase in the threshold voltage, paving the way to increased variability in circuits. Unlike BTI, CHC is an asymmetric aging effect, since the damage is mostly toward the drain region of the transistor. An alternative explanation to the HCI effect is the Drain Avalanche Hot Carrier (DAHC) mechanism. Hot carriers collide with Si atoms in the interfacial layer, causing impact ionization and leading to the generation of electron-hole pairs [11]. The resulting carriers are accelerated and cause impact ionization again (avalanche effect). This results in a proliferation of high-energy carriers, some of which may get trapped in the oxide or damage the interface. Unlike BTI mechanisms, HCI is not recoverable.

2.4.4 Time-Dependent Dielectric Breakdown

Time-dependent dielectric breakdown, also referred to as soft oxide breakdown, affects both nMOS and pMOS transistors. This phenomenon is also attributed to trap formation associated with large electric fields across the gate oxide. Initially, the traps that are formed within the gate oxide are distributed and non-overlapping. This can be seen on the left portion of the transistor cross-sectional view in Figure 2.6. When a device experiences TDDB, its gate leakage current fluctuates, as the traps move randomly within the gate oxide [12]. This does not result in complete device failure, though its energy, delay and noise margins are affected. A gate-to-diffusion (source or drain) breakdown represents the worst-case scenario for TDDB.

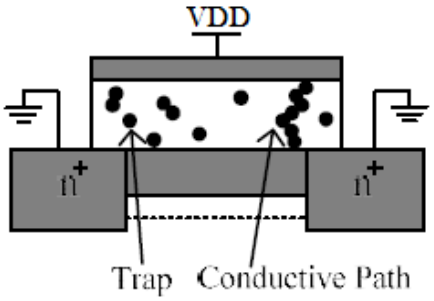


Figure 2.6: TDDB effect in nMOSFET [12]

When the traps accumulate over time, they may overlap, forming a resistive conducting path from gate to the channel. The conductive path is shown on the right portion the device in Figure 2.6. When the conduction path is formed, more traps are generated due to thermal damage [12]. The new traps widen the conduction path further, causing higher current to flow between the gate and channel and therefore, increasing the temperature. This vicious cycle leads to a catastrophic failure of the device, referred to as hard break down (HBD). Recent studies have shown that the rate of trap formation during TDDB increases with the permittivity of the dielectric (κ). TDDB is usually characterized at the transistor-level by modelling the increase in gate leakage current, as opposed to modelling threshold voltage in BTI and CHC.

2.4.5 Electromigration

Electromigration (EM) is caused by increased current densities on interconnects and on-chip contacts. The flow of electrons may force metal atoms to move in their direction, causing metal voids. When such metal voids form on interconnects, it results in open circuits or high resistive paths, which ultimately leads to reduced performance and circuit malfunction. Additionally, electromigration may also cause metal voids in the connections between metal contacts and silicon [13]. As the metal void grows, the aluminium can diffuse down to the silicon, forming metal spikes in the silicon region and shorting the p-n junctions [14]. Electromigration is a serious concern for reliability in the nanoscale regime because of the high current density.

2.5 Effects of Aging

Aging manifests differently at various level of the circuit. At the lowest level of abstraction, aging causes deterioration in the transistor device parameters, such as threshold voltage, leakage current etc. The extent of change depends on several factors, such as temperature, voltage, and activity factor. Gates that contain such degraded transistors experience an increase in the gate propagation delay. The V_D-I_D plot of a healthy gate (without aging) against an aged gate is shown in Figure 2.7.

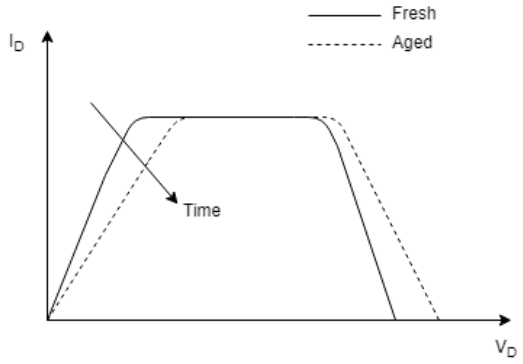


Figure 2.7: Aging effect in combinational gate

Figure 2.7 shows that the V-I curve of a fresh gate is sharp and takes up minimal delay. However, as the circuit experiences aging, the V-I response of the gate shows a gradual slump. Similar to combinational gates, sequential elements in the circuit (flip flops) also have increased propagation delay due to aging. At the circuit level, the critical path of the circuit dictates the timing requirements, i.e., the critical path delay determines the operating frequency of the circuit. The critical path of the circuit is that path that has the highest number of gates and exhibits maximum delay. As long as the delay of the critical path does not exceed the maximum permissible delay, the circuit's reliability is retained. As a result of aging, the critical path delay may increase over time, resulting in timing violations and circuit failure.

A more eminent reliability issue due to aging is that it introduces new critical paths over time. Though the actual critical path of the circuit is known before operation, it keeps changing as the circuit ages. A circuit may also contain near-critical paths, which exhibit a delay closer to the critical path delay. Sometimes, the critical path of the circuit may age slowly when compared to the near-critical paths. In such cases, timing violations may arise from near-critical paths, and thus, designers must take those paths into careful consideration too. Therefore, as time progresses, aging can turn a non-critical path into a potentially critical one.

Figure 2.8 shows the delay variation of ten critical paths in a 4-bit, 2-stage pipeline multiplier (PM4-2) circuit. The analysis is performed with a 65nm Predictive Technology Model (PTM)[15]. Among the ten probable critical paths, the three paths with the highest delay toward the end of a period of ten years are highlighted in figure 2.8. It can be seen that, even though critical path (CP) 1 is not likely to cause

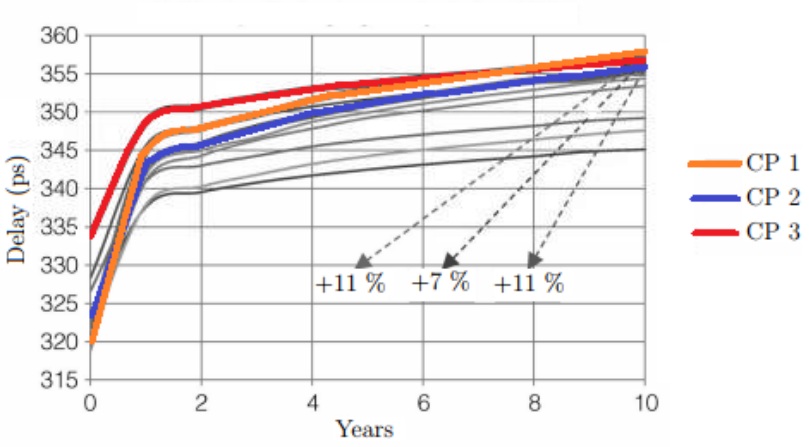


Figure 2.8: Critical paths' aging analysis for PM4-2 circuit - PTM 65nm, f_{clk} = 100 Hz, T = 110° C, V_{DD} = 1.1 V [15]

timing violations when the circuit begins operation (at t = 0 years), over the course of time, it becomes the most critical one due to aging. CPs 1, 2 and 3 undergo a delay degradation of 11%, 11% and 7% respectively, much higher than the remaining paths. Hence, aging-aware circuit analysis of the logic part of the circuit must focus on the near-critical paths, in addition to the actual critical path.

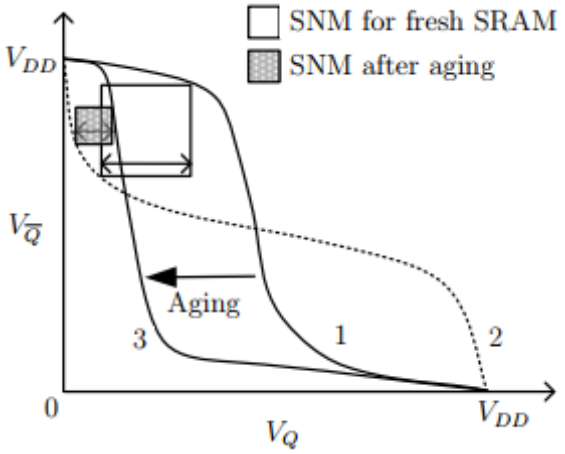


Figure 2.9: SRAM cell SNM degradation [16]

Aging mechanisms also affect CMOS memory devices, such as Static Random Access Memory (SRAM) elements. In the presence of aging, SRAM cache cells are exposed to degradation. Static Noise Margin (SNM) is an important parameter for the stability of a memory cell. SRAM cell aging may cause its SNM to be drastically reduced, causing read instability. SNM is the biggest noise voltage that the SRAM cell can tolerate [16]. In figure 2.9, V_Q and $V_{\overline{Q}}$ are the internal node voltages of a 6-transistor (6-T) SRAM cell. For a fresh SRAM cell, the V_Q-$V_{\overline{Q}}$ plot represents a well-balanced trend

known as the butterfly curve (1-2 curve pair in Figure 2.9). For such a healthy SRAM cell, the SNM region is quite wide, making it well resilient to noise spikes. As the cell ages, curve 1 moves further to the left, making the SNM progressively smaller. If the SNM degrades to a value smaller than the expected noise, the stored data might be flipped, causing a failure when data is read out [16]. Ultimately, this makes the SRAM cell non-functional. [17] shows that Read SNM (SNMR) undergoes higher degradation than the Write Trip Point (WTP) and Hold SNM (SNMH) metrics of an SRAM cell. Moreover, FinFET based SRAM cells are more sensitive to NBTI degradation than MOSFET based cells [17].

Apart from the drawbacks of aging, it has one significant benefit in logic and memories that implement low-power techniques. Aging mechanisms such as BTI, HCI increase device threshold voltages, substantially reducing the leakage sub-threshold current. This minimizes static power dissipation and makes CMOS logic very energy-efficient over time. The static power reduction depends on operating conditions, such as temperature and stress time. Static power dissipation reduces by more than 50% by the end of first year of operation, and by more than 74% at the end of five years. For a nominal life expectancy of ten years, the static power reduction becomes almost 80% [18]. This induces considerable energy savings in CMOS designs.

3 Aging Models

Extensive research has been conducted into the aging phenomenon, and the adverse effects it has on chip life. Predictive models have been introduced to determine the short-term and long-term effects of the various aging phenomena. Such models are very accurate in predicting the actual degradation of device parameters, so that designers are better equipped with lifetime expectations. The remaining part of this chapter details the short-term and long-term models used to describe transistor-level degradation, and the models describing aged behaviour of combinational cells, sequential cells and the aging sensor.

Various aging mechanisms and their effects were discussed in Chapter 2. This thesis primarily focuses on NBTI and CHC as the aging mechanisms for further modelling. Detailed theoretical analysis, Monte-Carlo simulation and experimental verification of the charge trapping component of BTI are performed in [19]. [20] and [21] aid in establishing a main difference between the R-D and T-D models of BTI. The level of degradation due to the T-D model follows a logarithmic dependence on time, whereas the R-D model predicts a power-law dependent degradation with time. The various reliability issues in nanoscale CMOS devices with dielectrics other than SiO_2 (SiON, high-κ dielectrics) are elaborated in [21]. [22] identifies that R-D model does not impose self-limiting recovery, i.e., it has been found that the amount of parametric shift induced by the stress cycle is almost nearly identical to that recovered during the relaxation cycle. Repeated stress and relaxation experiments have verified this in [22]. The parametric evolution after a fixed set of stress and recovery cycles is cyclic. This cyclic behavior is attributed to the same group of traps responding to a given set of experimental conditions. Additionally, the amount of recovery per cycle of the parameter of interest (threshold voltage shift) is shown to remain constant and independent of the number of stress and recovery cycles. [23] proposes an investigation of the T-D mechanism through analysis of the Gate-Induced Drain Leakage (GIDL) current, rather than the threshold voltage shift. Evidence from [24] suggests that HCI occurs when electrons gain energies higher than 3.7 eV. [12] proposes models to analyse the effect of TDDB in the performance degradation of combinational logic.

3.1 Integrated NBTI and CHC Model

Till date, research on NBTI and CHC has been progressively made within the device and reliability physics communities. This can be partially attributed to the complexity, lack of design knowledge and computer-aided design (CAD) tools for managing NBTI and CHC degradation [25]. Noted industrial companies develop their own models and tools to assist them in evaluating this effect. These tools are proprietary and customized to a certain technology level. Thus, generic models that can accurately

predict device degradation are of utmost use. A predictive model can be used to bridge the gap between the technical community and CAD tool developers.

[26] proposes a unified model that directly predicts the change of key transistor parameters under various process and design conditions for both NBTI and CHC. Additionally, the model in [26] has been verified using the 65-nm technology. Thus, all subsequent simulations will be done using the 65-nm Predictive Technology Model (PTM). Both NBTI and CHC result in threshold voltage (V_{th}) increase, which causes poor drive current, lower noise margin, and shorter circuit lifetime. Thus, this model analyses the degradation effects in terms of V_{th}. Normally, CHC is characterized using the substrate current (I_{SUB}) or gate current (I_G) induced by hot carriers. However, in the nanoscale region, this approach is problematic because of various leakage components dominating I_{SUB} such as, gate leakage current, junction current, and gate-induced drain leakage current [26]. This makes it difficult to differentiate the degradation in I_{SUB} due to aging. Instead of resorting to I_{SUB}, [26] unifies NBTI and CHC using the reaction-diffusion (R-D) model and directly develops degradation models for key transistor parameters, such as threshold voltage degradation (ΔV_{th} and mobility, μ). These integrated models have been comprehensively verified with industrial 65nm technology in [28]. This model can effectively capture the dependence of V_{th} degradation on key process and design parameters, such as V_{DD}, V_{th}, duty cycle, etc [27].

High-κ dielectric stack consists of two layers, high-κ dielectric layer and interfacial SiO_2 layer. The fast stress and recovery component most likely associates with the defects in the SiO_2 layer, which is induced by the overlaying high-κ film. Thus, stess-induced changes in threshold voltage (ΔV_{th}) of high-κ stacks must be modelled by both R-D and T-D model. Fast stress and recovery mechanism can be modelled by the hole trapping theory, while slow stress and recovery can be explained by the reaction-diffusion model. Experimental data for 65nm technology without high-κ stacks does not show hole trapping/detrapping [28]. Thus, reaction-diffusion is the dominant mechanism that can effectively model ΔV_{th} in transistors with SiO_2 dielectric.

3.2 NBTI Degradation Model

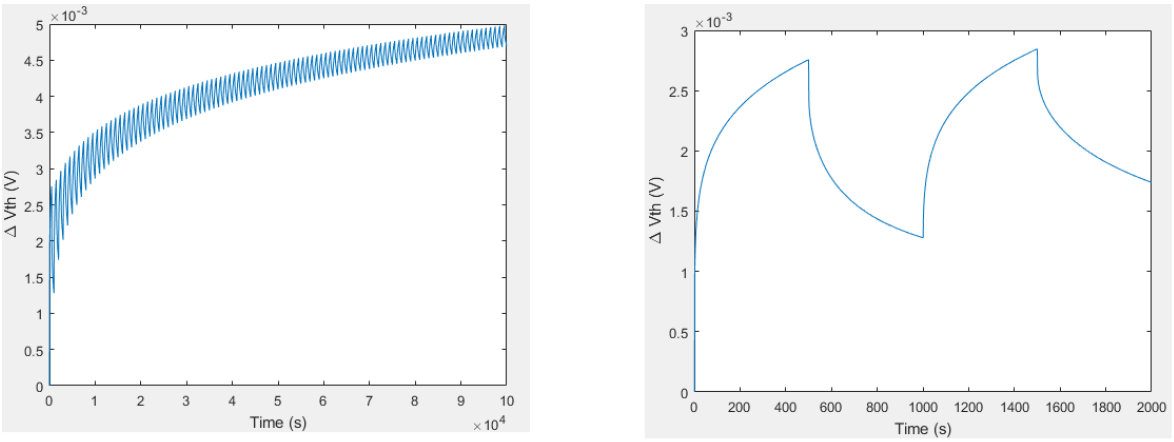
The short-term model which depicts V_{th} degradation in a pMOS transistor is given by [28] as

$$\text{Stress} : \Delta V_{th}(t) = \left(K_v (t - t_0)^{1/2} + \sqrt[2n]{\Delta V_{th}(t_0)} \right)^{2n} \tag{3.1}$$

$$\text{Recovery} : \Delta V_{th}(t) = \Delta V_{th}(t_1) \left(1 - \frac{2\xi_1 t_e + \sqrt{\xi_2 C (t - t_1)}}{(1 + \delta) t_{ox} + \sqrt{Ct}} \right) \tag{3.2}$$

t_0 and t_1 correspond to the start times of the stress and recovery phases respectively. This model has two primary advantages. It is easily scalable with key process and design

parameters [28], such as t_{ox}, V_{gs}, V_{th}, effective gate length (L_{eff}) and temperature T. Secondly, this model can be easily characterized for any predictive technology model and implemented in a simulation environment. Threshold voltage degradation due to NBTI over a period of 10^5 seconds is shown in Figure 3.1(a). The graph is plotted by a MATLAB script that calculates the V_{th} increase using (3.1) and (3.2). Table B.1 lists the values for the various terms in the equations, which were obtained from [26] and 65nm CMOS process data sheet from Fujitsu [36].



(a) (b)

Figure 3.1: V_{th} degradation due to dynamic NBTI

This plot is made for nominal voltage, $V_{DD} = 1.3$ V and temperature, $T = 27^o$ C, with a stress duration of 500 seconds. It can be observed that the difference in threshold voltage with its nominal value (ΔV_{th}) increases progressively over time. Figure 3.1(b) depicts two stress-recovery cycles. For $t \in [0, 500s]$ (first stress cycle), ΔV_{th} rises to almost 2.8 mV. During the subsequent relaxation phase, i.e., $t \in [500s, 1000s]$, the V_{th} degradation reverses partially, in accordance with the R-D model. Over time, the recovered voltage moves further away from the initial point, as illustrated in Figure 3.1(a).

The long-term degradation model depicting NBTI is given by [28] as,

$$\Delta V_{th}(t) = \left(\frac{\sqrt{{K_v}^2 \alpha T_{clk} / min(\alpha, 1-\alpha)}}{1 - {\beta_m}^{1/2n}} \right)^{2n} \tag{3.3}$$

where,

$$\beta_m = \left(1 - \frac{2\xi_1 t_e + \sqrt{\xi_2 C(1-\alpha) T_{clk}}}{(1+\delta) t_{ox} + \sqrt{C T_{clk}}} \right)$$

α is the duty cycle, which is the ratio of the stress time to the total cycle time, and T_{clk} is the duration of one stress-recovery cycle. A.1 elaborates the derivation of the short-term and long-term models from [28].

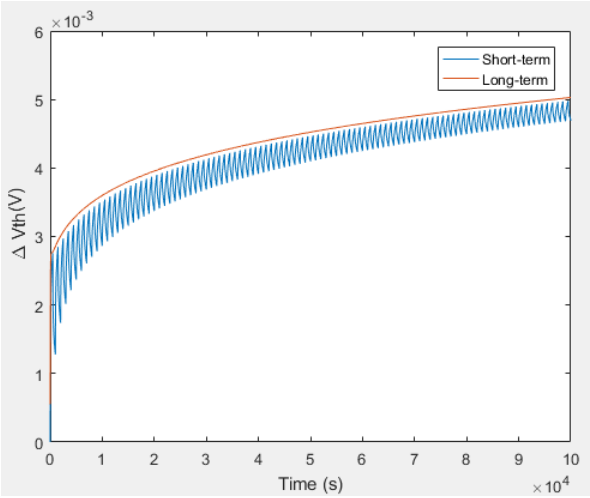


Figure 3.2: Long-term vs short-term V_{th} degradation due to NBTI

Figure 3.2 shows the trend of long-term degradation over time. The long-term model forms a clear upper bound to the ΔV_{th} degradation predicted by the short-term model. The maximum error between the long-term and short-term simulation results is within 4%.

Characteristics of NBTI Degradation

It is necessary to understand the degradation dependencies on factors such as temperature, supply voltage, process elements, etc. The amount of degradation in a low-V_{th} transistor is higher than that of a high-V_{th} transistor. Figure 3.3 shows the degradation variations of low, nominal, and high V_{th} devices.

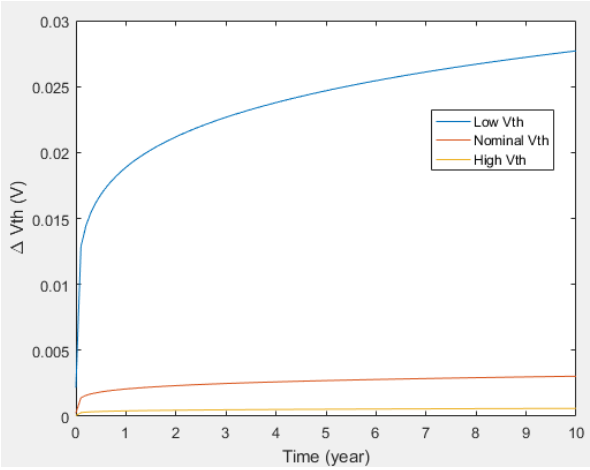


Figure 3.3: Comparison of NBTI effect in low, nominal, and high V_{th} transistors (V_{DD} = 1.3 V, T = 300.15 K

pMOS transistors with low V_{th} switch ON quickly. This implies that low threshold voltage transistors remain in stress for a longer duration compared to ones with higher

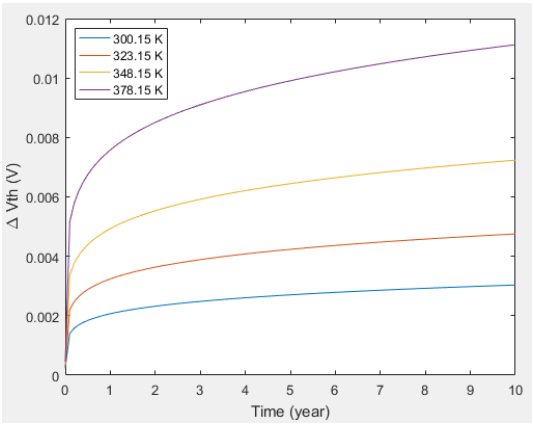


Figure 3.4: Variation of V_{th} degradation with temperature (V_{DD} = 1.3 V)

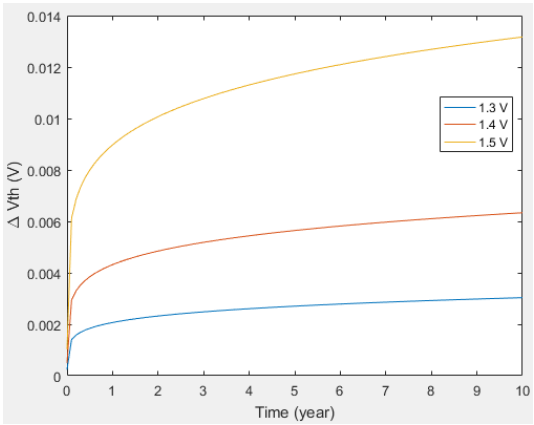


Figure 3.5: Variation of V_{th} degradation with V_{DD} (T = 300.15 K)

V_{th}, for the same cycle time. Since the overall span of stress is longer, the low V_{th} transistor is subjected to higher vertical electrical field across its oxide (E_{ox}), causing more Si-H bond dissociation. Ultimately, this translates to a greater threshold voltage shift. The variation of V_{th} with temperature and supply voltage is shown in Figures 3.4 and 3.5 respectively. It can be seen that ΔV_{th} increases with temperature and supply voltage. This is attributed to an accelerated Si-H dissociation process due to high temperatures and greater E_{ox} across the oxide (with increasing V_{DD}). Thus, the worst NBTI induced degradation occurs at high temperature, high V_{DD}, and high duty cycle (α).

3.3 CHC Transistor Degradation Model

CHC is primarily caused by "hot carriers", as a result of heating inside the channel due to large lateral electric fields [28]. These high energy carriers may get injected into the gate oxide, changing transistor properties such as threshold voltage, transconductance, etc. NBTI occurs in standby mode, whereas CHC occurs during the switching phase of operation. The V_{th} degradation caused by CHC is given by [28] as

$$\Delta V_{th}(t) = \frac{q}{C_{ox}} K_2 \sqrt{Q_i} exp\Big(\frac{E_{ox}}{E_{o2}}\Big) exp\Big(-\frac{\varphi_{it}}{q\lambda E_m}\Big) t^{n'} \tag{3.4}$$

where, E_{ox} is the electric field across the oxide (vertical), E_m is the maximum lateral electric field (between source and drain), and n' is the time exponential.

CHC does not include any recovery phase. Figure 3.6 shows the variation of threshold voltage over a period of 10^5 seconds. The factors on which CHC degradation depends on are analyzed. Figure 3.7 shows the degradation level in different types of V_{th} devices. Contrary to the trend in NBTI, high V_{th} transistors undergo the most CHC degradation. In such transistors, higher lateral electric fields (between source and drain) are required to make the device conduct, which are maximum near the drain end. More Si-O or Si-H bonds break near the drain end and get lodged in the oxide,

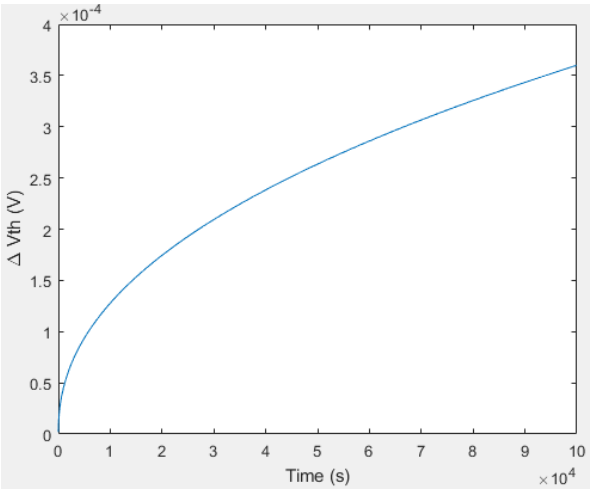


Figure 3.6: V_{th} degradation of due to CHC

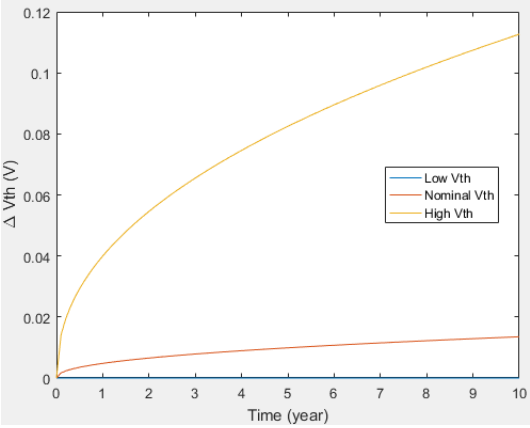


Figure 3.7: Comparison of CHC effect in low, nominal, and high V_{th} transistors (V_{DD} = 1.3 V, T = 300.15 K

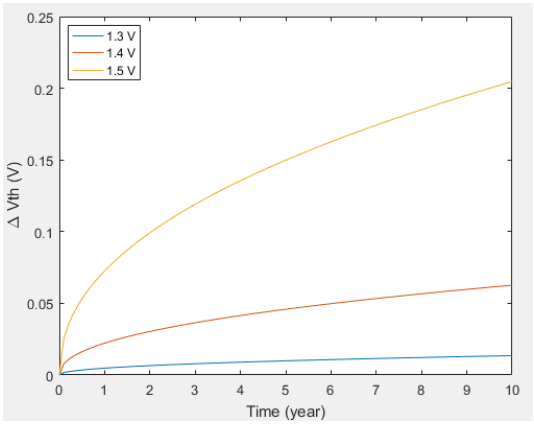


Figure 3.8: Variation of V_{th} degradation with V_{DD} (T = 300.15 K)

leading to greater elevation in the threshold voltage [28]. The same phenomena causes more degradation with higher V_{DD}, as shown in Figure 3.8.

3.4 Combinational Circuit Aging

Timing analysis is required to determine the maximum frequency a circuit can operate at, to perform circuit optimization during synthesis and layout, and verify that timing constraints are not violated due to local modifications. Timing analysis of a circuit is performed at the gate-level. Static timing analysis approaches have two primary advantages compared to simulating complete circuits to obtain timing characteristics. It is significantly faster, especially, when a simplified gate and interconnect model are used. Secondly, no input vectors are required to cover all worst-case timing errors. For timing analysis, a gate model is needed to compute gate propagation

delays. Two propagation delays are typical for a gate - rising and falling. Rising propagation delay (t_{rpd}) is defined as the time interval between input and rising output waveforms crossing 50% of their maximum value (i.e., $V_{DD}/2$). Similarly, falling propagation delay (t_{fpd}) is the difference between time instants, when input and falling output cross 50% of V_{DD}. To illustrate aging effects at the gate-level, let us consider a simple inverter (INV) gate. Inverter gate schematic is shown in Figure 3.9.

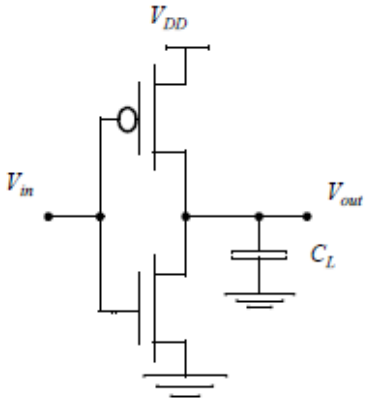


Figure 3.9: Inverter gate schematic

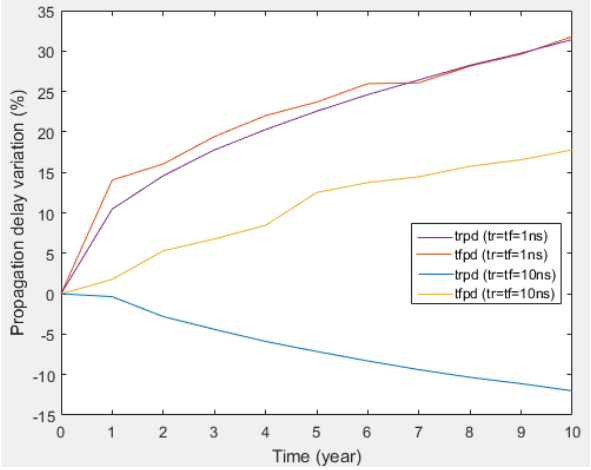


Figure 3.10: Propagation delay variation for different rise and fall times (V_{DD} = 1.3 V, T = 300.15 K)

V_{in} and V_{out} are the input and output voltages, and C_L is the load capacitance. The output rises when pMOS transistor, p, is turned ON (i.e., $V_{in} = 0$). The ON transistor acts as a closed switch, connecting V_{out} directly with V_{DD} and causing the output to rise. When $V_{in} = V_{DD}$, n turns ON and p turns OFF. The charge stored in C_L is discharged via transistor n to GND, causing the output to fall. Thus, the pMOS transistor is along the charging path and the nMOS transistor is along the discharging path between the input and the output. This implies that NBTI, which increases the threshold voltage of pMOS transistors, causes a subsequent increase in the rising propagation delay of the gate. Correspondingly, CHC will lead to an increase in the falling propagation delay. This is true, only if, the input rise and fall times are within 2 ns. Combined simulation of NBTI and CHC effects result in different propagation delay trends depending on the input transition times, as shown in Figure 3.10. It can be seen that when the rise (and fall) times are small, both rising and falling propagation delay increase with aging. For longer input transition times, the rising propagation delay decreases, while the falling delay increases.

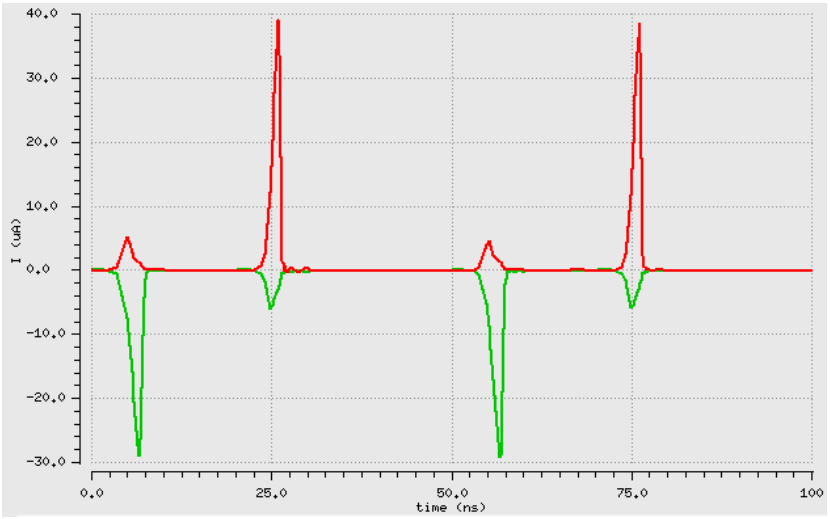


Figure 3.11: Drain current of p and n when rise time = fall time = 10 ns

Figure 3.11 shows the drain current (I_D) variations (observed using Cadence Virtuoso simulations) in transistors p and n when the rise and fall times are equal to 10 ns. There are four spikes in I_D corresponding to transistors p and n turning ON (positive spikes - n turns ON, negative spikes - p turns ON). Figure 3.11 shows a small current spike in the complementary transistor, when the main spike occurs. This indicates that for a very small interval in gate operation, both n and p transistors conduct. In such a case, both transistors influence gate delay collectively, such that the delay associated with the most prominent effect increases, reducing the impact of aging in its counterpart. Hence, further analysis and modelling is conducted for short rise and fall times (typically, 1 ns), so that it is closer to replicating actual circuit signals and uniform behavior can be observed in t_{rpd} and t_{fpd}. [29] shows that depending on the switching activity (α), NBTI can cause up to 19% gate delay increase due to the degradation in its component transistors and up to 4.8% additional delay increase due to transistor degradation in the adjacent gates. [30] shows that NBTI causes 2.3× more degradation in pMOS transistors than PBTI in nMOS transistors. This leads to a higher delay degradation in NOR gates due to NBTI and in NAND gates due to PBTI. Furthermore, [30] presents the dependence of gate delay increase due to BTI on several factors such as, duty cycle (α), frequency, and the location of the stressed transistor.

3.4.1 Gate Delay Models

Several models accurately describe the aged behaviour of gates. A long-time industry standard is the Non-Linear Delay Model (NLDM), which represents gate delay and output slew rate in 2-D look-up tables (LUTs) as a function of input slew rate (S_{in}) and effective load capacitance (C_L). NLDM-based timing analysis is very fast, as delay computation involves linear interpolation from LUTs based on S_{in} and C_L. A few of the other notable delay models are Current Source Models (CSMs), Composite Current Source (CCS) models, multiple-port CSMs, voltage dependent CSMs and transistor-level gate delay models. Specifically, multiple port CSMs consider physical

effects (e.g. Miller effects) such as high interconnect resistance and noise propagation, are compatible with arbitrary waveform shapes, include parasitic resistance and coupling effects in interconnects, and capture electrical effects. Such models can be incorporated in SPICE-based model simulations, where accurate noise and power analysis is needed. [31] proposes a delay model for NBTI induced degradation, which takes threshold voltage shift and hole mobility degradation under temperature variations into account. [32] incorporates a gate delay model which shows that the gate output transition time increases by 8.56% over 10 years. For straightforward timing analysis requirements, the following gate delay model is incorporated.

The relational model [11] calculates the delay change with respect to $\Delta\ V_{th}$. Input slew rates and load capacitances impact delay shifts, and this model is independent of them [11]. Based on the drain current of a short channel device, the delay of a digital gate, t_d, is expressed as [37]

$$t_d \propto \frac{C_L V_{DD}}{V_{DD} - V_{th}} \tag{3.5}$$

[11] expresses the change in gate delay in terms of change in V_{th} (Δt_{dVth}) as,

$$\Delta t_{dVth} = \frac{K \Delta V_{th}}{V_{DD} - V_{th}} . t_d \tag{3.6}$$

Thus, using (3.6), the change in propagation delay can be obtained in terms of change in V_{th}. Another behavioral delay model is the α-power law model [37]. However, it relies heavily on the proportionality constants, which vary with different temperatures, V_{DD}, the assessment time (for e.g., different values for computing delay in the first year and the tenth year), the type of gate, etc. Storing proportionality constants for each and every scenario will result in large LUTs. The expression in (3.6) overcomes this problem by expressing delay change as a function of V_{th} change. The proposed model in (3.6) predicts the shift in gate delay (rising and falling) in case of inverter and NAND gates, where one pMOS and nMOS transistor exists between input and output. In case of NOR gates, the situation is quite different because there are multiple transistors between switching input and output. In such a case, the change in gate delay due to V_{th} degradation is given by,

$$\Delta t_{dVth}(\text{rising}) = \frac{(k_1 \Delta V_{th1} + k_2 \Delta V_{th2}) t_d}{V_{dd} - V_{th}} \tag{3.7}$$

where, k_1 and k_2 indicate the contribution of the two pMOS transistors in series toward the rising propagation delay [11]. Falling propagation delay of a NOR gate can be expressed using (3.6). Typical values of k_1 and k_2 are in the range 0.5-1. With the correct choice of K, k_1 and k_2 in (3.6) and (3.7), the model accuracy can be increased.

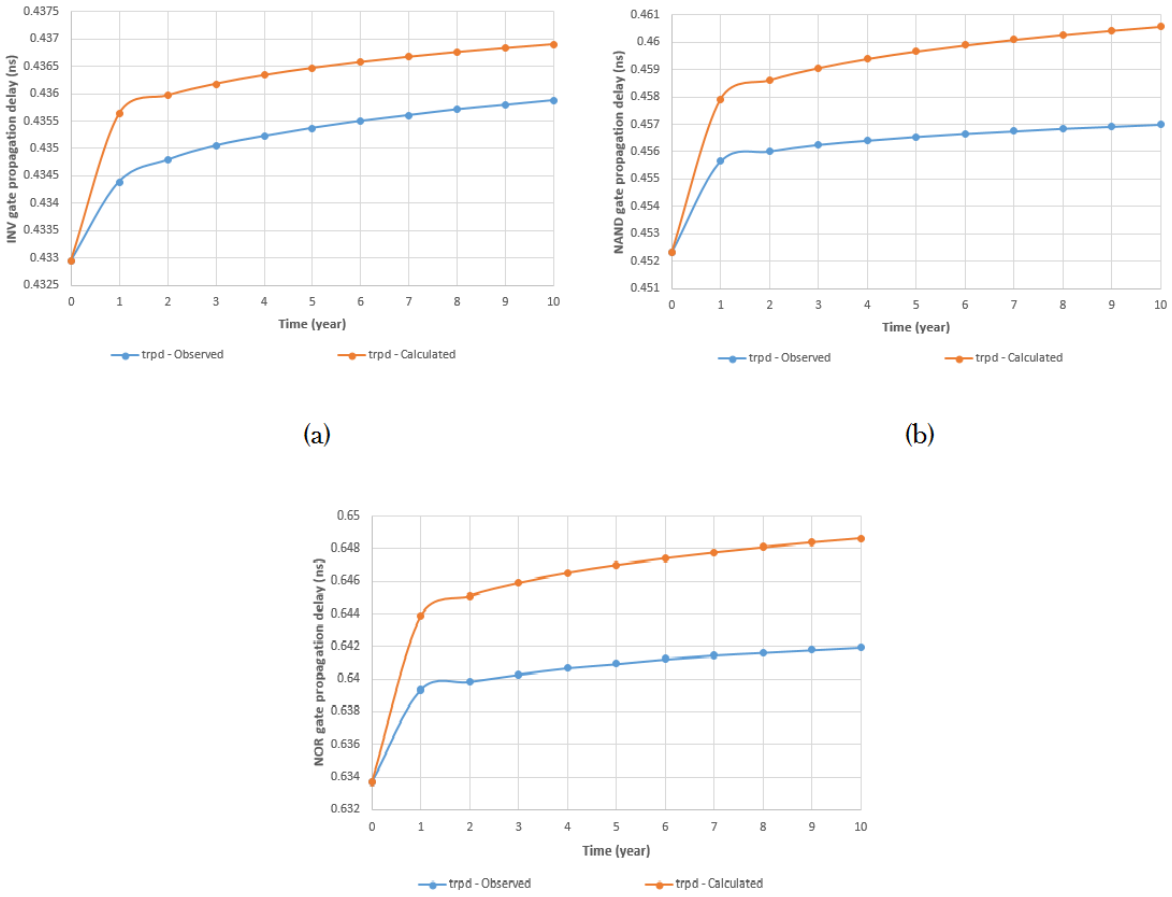


Figure 3.12: Rising propagation delay degradation of (a) Inverter gate (b) 2-input NAND gate (c) 2-input NOR gate (V_{DD} = 1.3 V, T = 300.15 K)

The process of determining the fitting constants to depict aged behaviour is called characterization. The characterization process performed for all combinational and sequential cells is uniform. Firstly, the gate is simulated in Cadence Virtuoso. The inputs that provide the best-case and worst-case delays are identified. Reverse biasing the transistors by using a voltage source between the source and bulk can manually alter their threshold voltages. Thus, the elevated V_{th} values, which were calculated using MATLAB, can be simulated in Cadence. Transient simulations are run, and rising and falling propagation delays can be calculated from the resulting input/output waveforms. These are the observed delay values. Using (3.6) and (3.7), the models can be tuned to reduce the error between observed and calculated values. The characterization process is repeated for the range of supply voltages and temperatures that are of interest.

Figures 3.12(a), (b) and (c) show the actual and calculated degradation in rising delays of inverter, 2-input NAND and 2-input NOR gates. As shown in the graphs, the calculated figures are very accurate in expressing gate delays, with a mean error less than 1.5% compared to observed values. The falling propagation delays can be calcu-

lated in a similar way. Apart from these basic gates, delays of various combinational cells such as transmission gate, AND gate, OR gate, XOR gate, etc. can be modelled using (3.6) and (3.7).

3.4.2 Factors Influencing Gate Delay

Several elements affect gate propoagation delay, such as transistor size, input slew rate, load capacitance, size of fan-in/fan-out, etc. Increasing gate size corresponds to increasing the size of its transistors. Larger gates experience smaller gate delay. In simple terms, gate delay is the time taken to charge/discharge the load capacitance, C_L. When the transistor size is large, its resistance decreases, leading to a stronger drive current at C_L. This hastens the charging/discharging process, thereby reducing gate delay. However, increasing gate size will ultimately increase the load capacitance of the previous gate, leading to a higher delay in the previous gate. The transistors of all gates have constant width (W) and length (L) equal to $2\mu m$ and $240nm$, respectively.

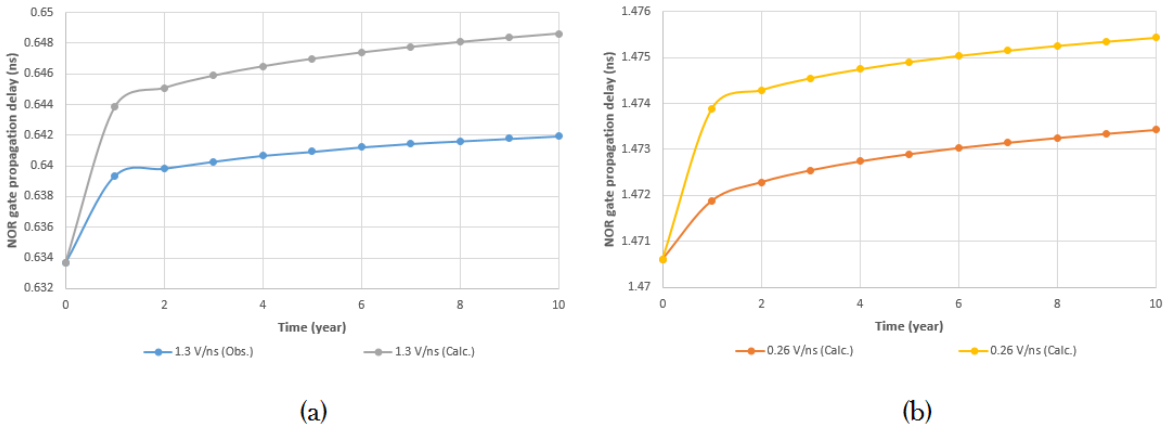
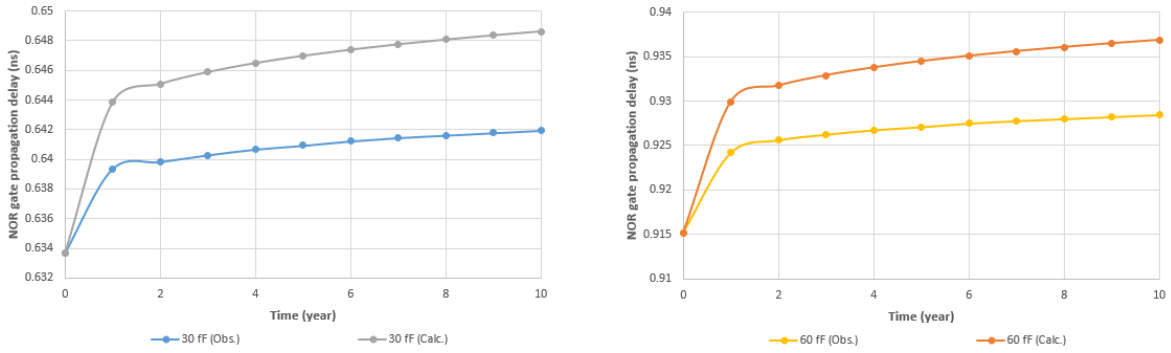


Figure 3.13: Rising propagation delay degradation of 2-input NOR gate for (a) input slew rate = 1.3 V/ns (b) input slew rate = 0.26 V/ns (V_{DD} = 1.3 V, T = 300.15 K)

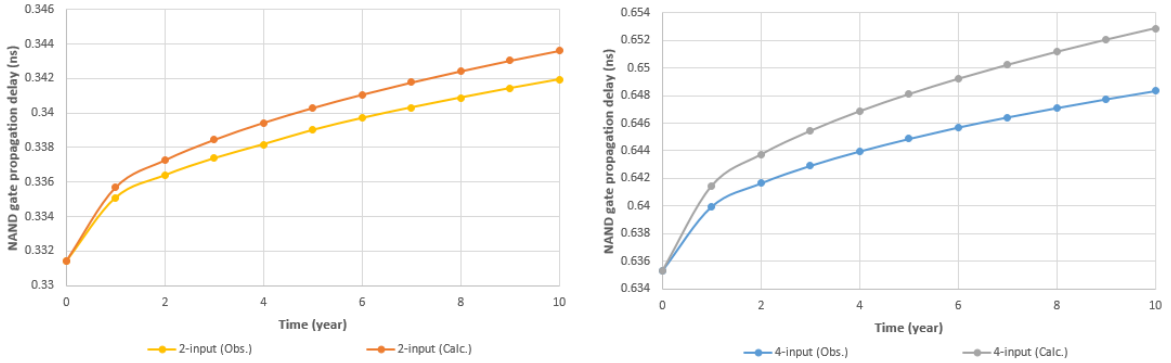
Input slew rate is defined as the rate of change of input voltage from 10% to 90% of its maximum value. Higher slew rate indicates a faster transition from 0 to V_{DD}, and vice versa. When the input voltage has a slower transition (low slew rate), its delay is more. Figures 3.13(a) and (b) show the rise in delay for a 2-input NOR gate over a span of ten years, for two different input slew rates. It can be seen that the initial gate delay (at $t = 0$) increases for smaller input slew rates (Figure 3.13). Load capacitance, C_L, is the capacitance seen by the device driving the load. Increasing C_L leads to a higher propagation delay. As it can be seen in Figures 3.13(a), 3.13(b), 3.14(a) and 3.14(b), irrespective of the input slew rate and load capacitance, (3.6) and (3.7) capture the gate delay with good accuracy, within 10 ps.

Fan-in is the number of digital inputs to a gate, and fan-out is the maximum number of digital inputs that the output of a single logic gate can feed. The propagation delay increases rapidly as the fan-in/fan-out size increases. The large number of transistors



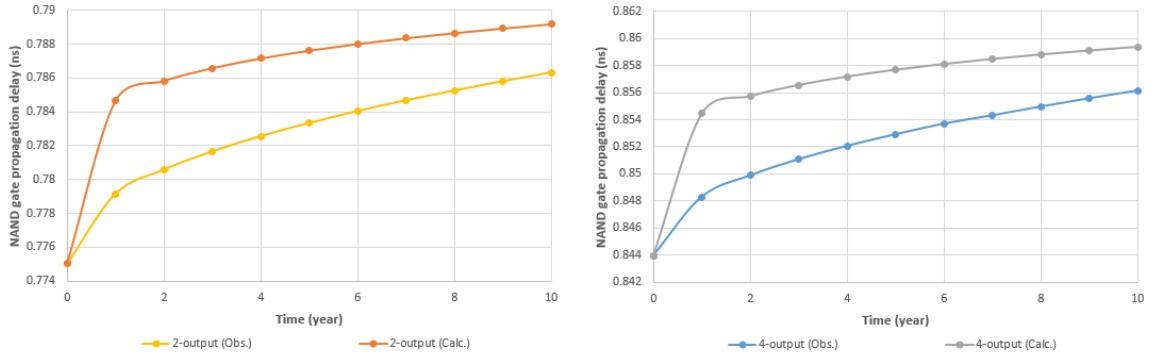
(a) (b)

Figure 3.14: Rising propagation delay degradation of 2-input NOR gate for (a) load capacitance = 30 fF (b) load capacitance = 60 fF (V_{DD} = 1.3 V, T = 300.15 K)



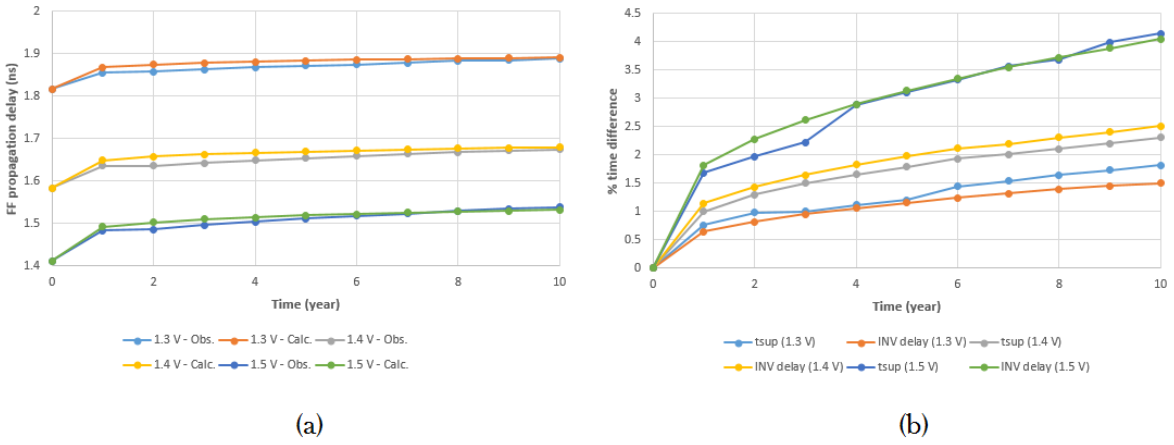
(a) (b)

Figure 3.15: Falling propagation delay degradation of (a) 2-input NAND gate (b) 4-input NAND gate (V_{DD} = 1.3 V, T = 300.15 K)



(a) (b)

Figure 3.16: Rising propagation delay degradation of 2-input NAND gate with (a) fan-out of 2 gates (b) fan-out of 4 gates (V_{DD} = 1.3 V, T = 300.15 K)



(a) (b)

Figure 3.18: Flip-flop characterization for different V_{DD} (a) delay degradation variation (b) plot for setup time sensitivity

associated with higher fan-in/fan-out increases the overall capacitance of the gate, leading to a higher propagation delay. Figures 3.15(a) and (b) show the increase in falling propagation delay as the number of fan-in signals increases. Alternately, a rise in the initial delay with an increase in the fan-out can be seen in Figures 3.16(a) and (b). In all these scenarios, the gate models depict the propagation delays very precisely.

3.5 Sequential Circuit Aging

Sequential circuits consist of logic gates and storage elements. Majority of sequential circuits are synchronous and use edge-triggered flip-flops. The schematic of the flip-flop (FF) chosen for analysis is shown in Figure 3.17 [38]. Flip-flop delay is the time interval between the input data edge and the clock edge when they cross 50% of their maximum magnitude (V_{DD}). This is also referred to as clock-to-Q delay ($t_{clk-to-Q}$). Apart from the FF delay, two other important timing constraints are the setup time t_{SUP} and hold time t_{HLD}. Setup time is time before the active edge of the clock that the input data must be valid. Hold time represents the time that the input must be held stable after the rising edge of the clock.

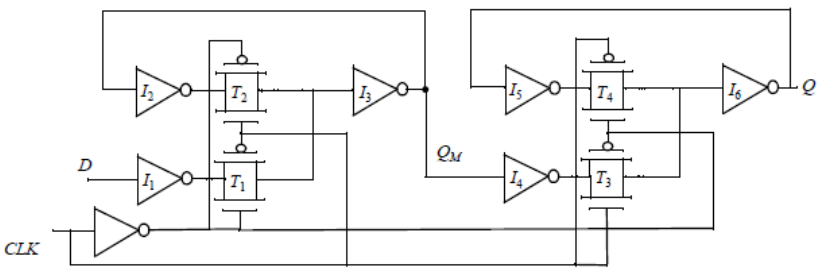


Figure 3.17: Master-slave positive edge-triggered register using multiplexers [38]

Characterization of t_{SUP} and t_{HLD} cannot be done directly. Most commonly, they

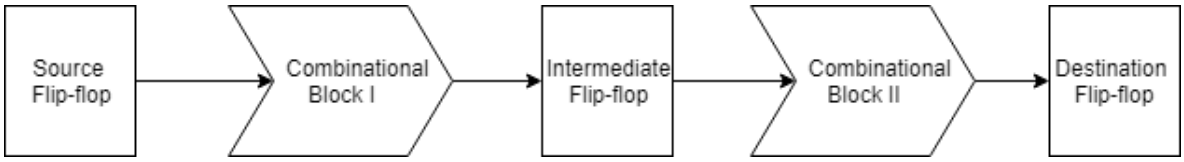


Figure 3.19: Circuit block schematic

are obtained by solving an optimization problem - optimize the time difference between data signal edge and clock edge such that $t_{clk-to-Q}$ is 110% of its relaxed value. [39] presents an analytical model to characterize interdependence between t_{SUP} and t_{HLD} in a fast and accurate manner. However, FF delay is the sum of setup time and clock-to-Q delay. Thus, the optimization approach is deemed sufficient for characterizing FFs. During FF characterization, it was observed that the transient response was faulty. Hence, the inverter size was increased, such that W and L dimensions are 4 μm and 720 nm. This modification was done to facilitate faster operation of the pull-down network. Figure 3.18(a) shows the degradation of $t_{clk-to-Q}$, when all its transistors have the same V_{th} drift.

The delay and setup time can be expressed using the delays of the inverters and transmission gates in the FF [38], depending on the path taken by signal to propagate from input to output. Hence, by identifying an appropriate value for K through characterization, (3.6) can be used to model $t_{clk-to-Q}$ delay precisely. Figure 3.18(b) compares the percentage variation of inverter gate delay and the percentage change in setup time with aging. The variation is very similiar, and this invariably shows that the same delay model can be used to depict setup time changes too.

3.6 Aging Sensor

Typically, IC lifetime requirements are determined using worst-case assumptions, leading to highly conservative margins and under utilization of technology potential. To make better use of the available technology improvement, pessimistic assumptions must be relaxed and combined with a dynamic reliability management framework that relies on sensors to measure IC aging. Accurate and efficient degradation measurements are vital for dynamic reliability management in aging aware systems. The fundamental idea in using an aging sensor is to detect timing violations that may occur in the circuit. Complex circuits are split into smaller combinational blocks, interspersed with flip-flops, as shown in Figure 3.19. The final output of the combinational path will be latched onto the destination flip-flop. When the delays of the individual combinational and sequential cells increases, the path delay rises ultimately. Circuit functionailty is preserved as long as the input signal to the destination FF arrives well before its required minimum setup time. As aging increases the path delay, data input to the destination FF may arrive late, giving insufficient time for the data to be latched by the FF (i.e., setup time violations occur). An aging sensor must be capable of predicting such delay errors in advance.

Numerous efforts are made into developing robust aging sensors. Three typical aging sensors are analyzed. The schematic of a Razor flip-flop [40] is shown in Figure

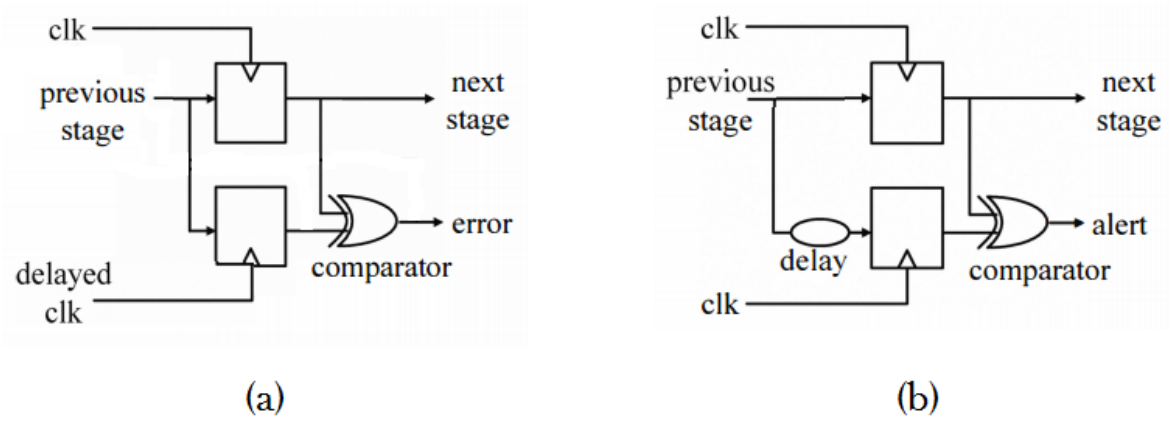


Figure 3.20: (a) Razor flip-flop (b) Canary flip-flop

3.20(a), which consists of a main FF and a shadow FF (labelled M and S in the figure). The output of the previous combinational stage is input to the main and shadow FF. However, the clock input to the shadow latch is delayed. When the path delay increases, it may cause setup time violation in the main FF, but not in the output of the shadow FF, as its clock input is delayed. The correct output from the shadow latch is compared with the faulty output from the main latch using an XOR gate (comparator). Any difference in the outputs from the main and shadow FFs will issue an error.

Canary FF [41] has a minor change compared to the Razor FF. Instead of delaying clock input to the shadow FF, its data input is delayed, as shown in Figure 3.20(b). This delay can be introduced by a delay element, which is typically an even number of inverter gates connected in series. Deferring the input to the shadow FF will cause a setup violation in the shadow latch first (in the razor FF, the main latch will first experience setup time violation). The outputs of the main and shadow latches are compared to issue an error signal. Since the shadow latch fails first, canary FF is ideal for predicting delay errors, whereas razor FF is ideal for error detection. Multiple comparator outputs can be combined to form a single error signal.

The Adaptive Error Prediction flip-flop (AEP-FF) [7] can detect late transitions at the flip-flop data input. The sensor schematic is presented in Figure 3.21(a). These sensors are placed at key flip-flops where synchronization errors are most likely to occur. These sensors must be integrated in FFs terminating in critical or near-critical paths (CP). Contrary to the razor and canary FFs, AEP-FF does not require a separate shadow latch. The master-slave configuration of the FF, along with a delay element (DE) and stability checker (SC) suffice. DE is a simple buffer that introduces a delay to create virtual guard-band (t_g) to identify late transitions at the FF input. In high performance circuits, DE is usually two inverter gates in series. The RESET signal facilitates continuous on-line monitoring, as well as sensor activation at certain time intervals. Stability checker determines whether the output of the master latch changes well before the observation interval (t_g) and issues an error if late transitions occur. The output of the aging sensor (AS_OUT) will be high if a delay error is predicted. Table 3.1 presents a comparison between the three aging sensors.

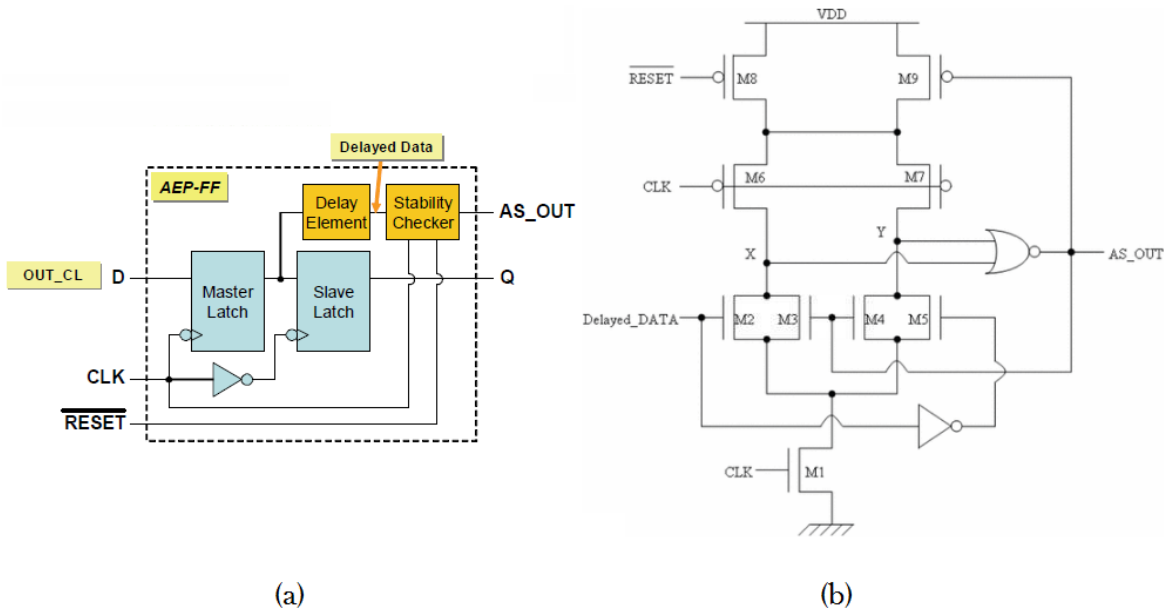
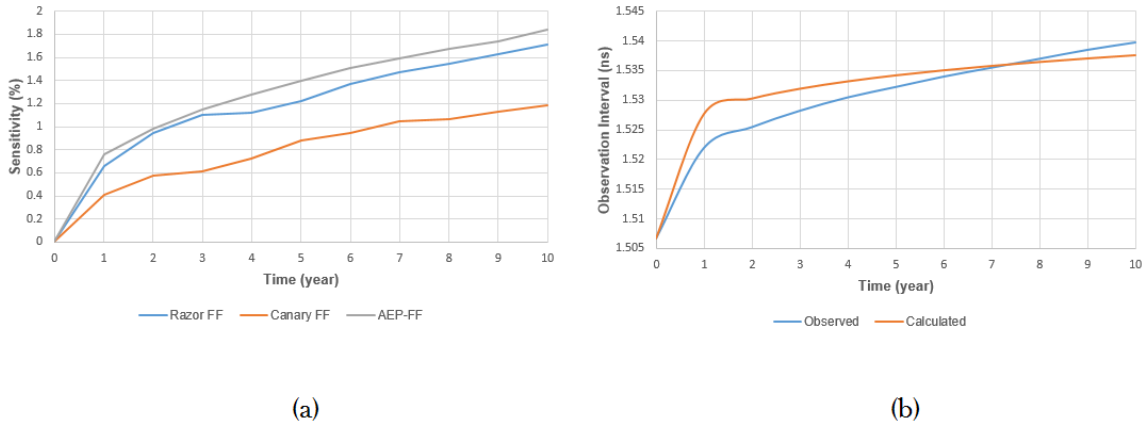


Figure 3.21: (a) Adaptive error prediction flip-flop topology (b) Stability checker architecture with on-retention logic [7]

	Razor FF	Canary FF	AEP-FF
Error prediction	No	Yes	Yes
Observation interval	Moderate response to aging	Minimal response to aging	Adapts best to aging
Monitoring process	Constant	Constant	Controlled through RESET
Hardware requirements (# of transistors)	48	48	39

Table 3.1: Aging sensor comparison

It can be seen from Table 3.1 that AEP-FF provides better sensing performance than razor and canary FFs. The observation interval is the effective guard-band in which delay errors can be identified. A good sensor has an observation interval that responds best to aging, i.e., as the sensor ages, its sensitivity must be enhanced. Figure 3.22(a) shows the sensitivity comparison of the three sensors. As it can be seen from the figure, the observation interval of AEP-FF increases well with aging, followed by the razor FF. Canary FF experiences the least adaptability to aging. AEP-FF is chosen for characterization because of its advantages listed in Table 3.1. AEP-FF sensor is composed of combinational and sequential elements which were characterized previously. Hence, the relational equations are applicable for modelling the observation interval as well. Figure 3.22(b) shows the accuracy of the model in depicting actual values. The error in the modelling the observation interval is less than 1.2%. AEP-FF is not placed at the end of every path in a circuit. Instead, the sensor is placed using a



(a) (b)

Figure 3.22: (a) Sensitivity comparison of aging sensors (T = 300.15 K, V_{DD} = 1.3 V) (b) Comparison of AEP-FF observation interval with calculated values

particular procedure. The critical and near-critical paths of the circuit can be identified using a tool such as PrimeTime. The delays of all the paths are determined, and the paths are ranked by their delay. The user specifies a parameter, γ, which will ensure monitoring effectiveness. The terminating flip-flops along any path that has a delay above $\gamma\times$ critical path delay will be identified as a critical memory element (CME) [7]. Lastly, the CMEs are replaced with AEP-FFs. A low value of γ will ensure a good coverage, but increase the hardware cost since many AEP-FFs will need to be placed.

4 System Aging Management

This chapter details the methodology used to evaluate the long-term aging effects at the system-level and the corresponding measures to counteract premature IC failure. The system overview is presented, the mitigation measures are explained, and architecture implementation in System-C AMS is detailed.

4.1 Prior Work

Performance degradation in circuits has been evaluated using conventional reliability tools such as BERT, RelXpert, etc. [42]. Berkeley reliability tool (BERT) determines the performance degradation caused by HCI. Additionally, it also computes the probability that a circuit fails due to TDDB and EM. Firstly, BERT determines the drain current ($I_D(t)$), the gate current ($I_G(t)$) and the substrate current (I_{SUB} (t)). Subsequently, these three parameters are used to determine an AGE value for every transistor, which quantifies the amount of degradation [43]. However, it is not possible to simulate the circuit for its entire lifetime. Instead, the circuit is simulated for a shorter duration, and the AGE values are extrapolated. Commercial reliability simulators, like RelXpert from Cadence, HSPICE from Synposys and ELDO [44] provide integrated reliability analysis. Two primary disadvantages of these tools are that numerous SPICE parameters are required, and several iterations are needed to determine circuit lifetime. Reliability simulators on circuit-level can be very accurate. However, it is quite time consuming and realistic input vectors are required to bring out the best-case and worst-case circuit behaviour.

[45] proposes a path-based circuit aging analysis scheme, which makes several lookup tables (LUTs) for different conditions like lifetime, temperature, etc. Values in between the stored values of the LUTs are obtained by interpolation. An advantage of the LUT-based models is that their accuracy can easily be increased by performing characterization at additional supporting points. However, LUT-based approaches require numerous tables to cover all possible operating conditions. [32] presents a design technique for reliability improvement, where the increase in transition time is converted to an increment in the voltage with a sensitivity of 0.50 mV/ps, and a biasing voltage is applied for NBTI mitigation subsequently. All these tools require knowledge of the underlying circuit dynamics, such as creating the appropriate circuit elements, inducing aging by varying V_{th} or I_{SUB} manually, etc.

The existing system aging simulation methodologies have the following major drawbacks. System aging assessment through current circuit-level simulation approaches are severely time-intensive. Additionally, determining system reliability using the tools mentioned previously requires various SPICE parameters and lookup tables. This framework is devoid these restrictions. Furthermore, the proposed aging assessment

prototype abstracts the lower level aging models to the system-level, making the assessment process faster, without requiring any knowledge of underlying circuit dynamics.

4.2 System-Level Aging Assessment

The next higher level of abstraction deals with determining aging at the system level. The low-level aging effects described so far must be translated to the system-level, facilitating high-level analysis. Typically, a system consists of a processor, memory unit, logic unit, and a bus system to interconnect them, as shown in Figure 4.1. In addition to these basic components, the proposed aging-aware system consists of a Management Unit (MU).

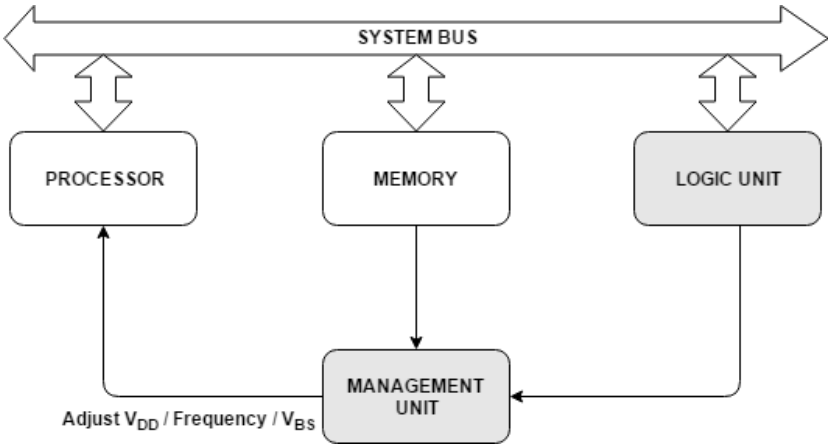


Figure 4.1: High-level overview of an aging-aware system

The MU performs two important tasks - actively monitors the path for timing violations and minimzes the impact of system aging by employing Dynamic Voltage Scaling (DVS), Dynamic Frequency Scaling (DFS) and Adaptive Body Bias (ABB). Without loss of generality, the system-level aging assessment involves logic unit and the management unit, as highlighted in Figure 4.1. Logic unit consists of many combinational and sequential circuit elements, forming various paths between the input and output. The path which exhibits the longest delay in the circuit is called the critical path. The speed of operation is limited by the critical path delay. As a result, when timing violations occur in the critical path, circuit functionality is compromised. Near-critical paths experience delays closer to the critical path, and can potentially become the most crucial one because of aging. Thus, system-level aging assessment can be performed by monitoring the circuit's critical/near-critical paths (CP) to determine its lifetime under given best-case and worst-case operating conditions. This provides the user an estimate of expected lifetime from a circuit. The behavioral models, which identify long-term aged performance of gates and FFs, can be used to simulate long-term critical path operation. The critical path delay is determined based on incremental timing analysis (path-based), where the arrival time of the input signal at a particular node is the sum of arrival time at the previous node (gate/FF) and delay of previous node. Following

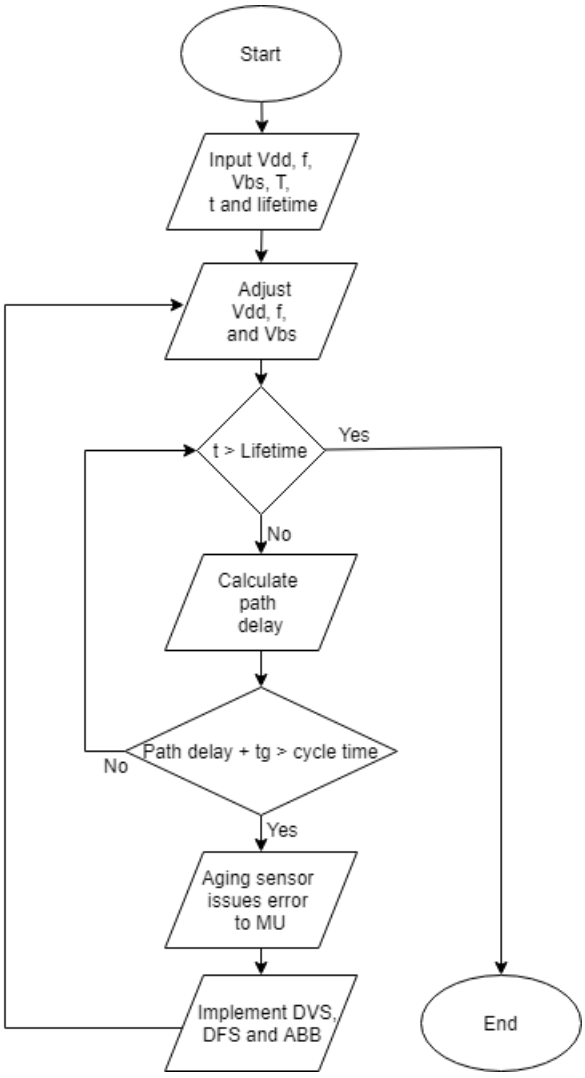


Figure 4.2: Flowchart to perform system lifetime assessment

this method, the sum of individual circuit element delays provides the path delay. The flowchart that describes system aging assessment process is shown in Figure 4.2.

The system is initialized by providing the operating conditions such as supply voltage (V_{DD}), frequency (f), temperature (T), bulk-to-source voltage (V_{BS}), time (t) and lifetime. Additionally, the cycle time is predetermined for the target circuit. When the assessment procedure starts, the degraded threshold voltages are calculated for both pMOS and nMOS transistors in every circuit element. Using the characterized delay models, the propagation delays (rising and falling) of all circuit elements are computed to determine the path delay. If the sum of the path delay and the guard-band of the aging sensor exceed the set cycle time, the aging sensor will issue an error signal to the MU, which takes the necessary corrective action by implementing DVS, DFS, and ABB. This assessment process is repeated till the required lifetime is met.

4.3 System Architecture

The detailed architecture, shown in Figure 4.3, determines the lifetime bounds of a system. The critical path can be completely modelled using long-term behavioral models of gates and FFs. Coupling the critical path performance with the aging sensor, logic unit aging can be determined. The system architecture consists of four major components -

- Testbench
- Logic Unit
- Management Unit
- Auxiliary Unit

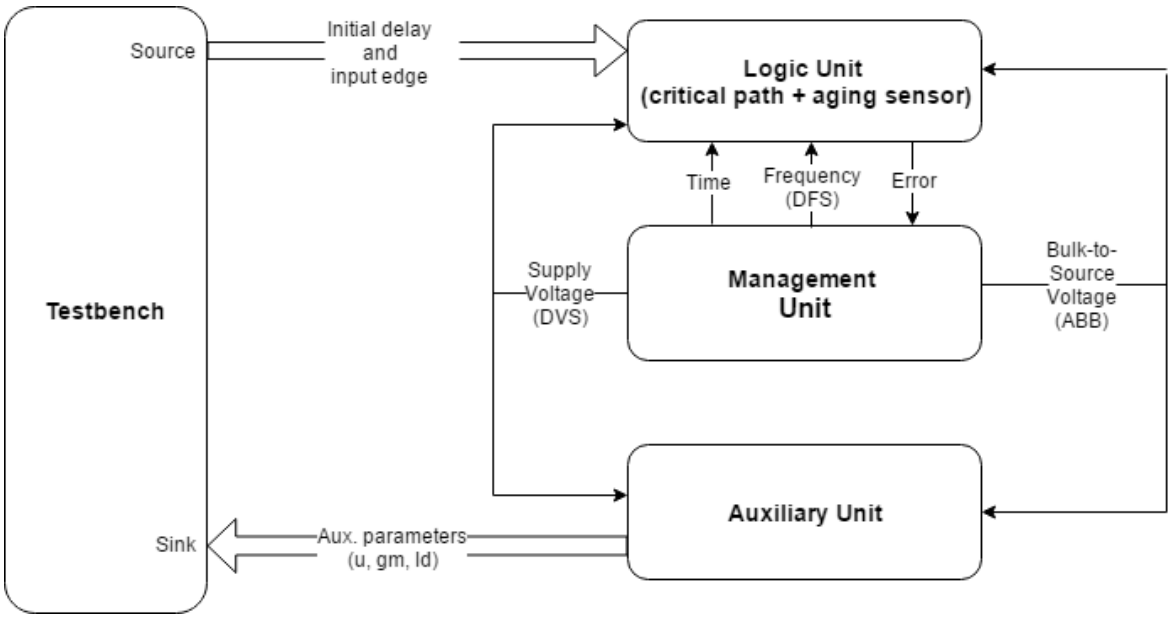


Figure 4.3: System architecture to evaluate high-level aging effects and implementation of corresponding mitigation measures

Testbench: Testbench is composed of a source and a sink. Source supplies the primary inputs to the path and sink receives the outputs from the path for observation purposes. The source initializes the delay value to the circuit path, and specifies whether the input to the first element in the circuit path will be a rising or a falling transition to capture rising and falling degradation accurately.

Logic Unit: The logic unit includes combinational and sequential logic, and aging sensor. Every component in the logic unit (node) receives a set of values from its previous node, performs required computations and provides values to the subsequent node. Supply voltage (V_{DD}), bulk-to-source voltage (V_{BS}), frequency (f), sum of

path delays of previous gates (t_{pd}), time (t), and type of incoming edge are the values input to every node. Every node performs V_{th} degradation computations based on [26] and gate/FF delay values are also calculated. The sum of delay values from the preceding gates are added to the delay of the current gate. This value is passed along to the succeeding gate along with V_{DD}, V_{BS}, f, t_{pd}, t, and type of outgoing edge. As the delays get accumulated further along the path, the output values from the last node contain the path delay. By explicitly including time as a communication signal between the nodes, higher analysis resolution can be achieved.

Management Unit: The Management Unit initializes the assessment procedure by providing the default V_{DD}, V_{BS} and f, continuously monitors the logic unit for timing violations, and corrects aging using DVS, DFS and ABB. Additionally, MU is the source point of the time signal, so that the aging assessment process is synchronized. If a delay error is predicted at time t, the MU counteracts at the same instant, making the detection and mitigation processes concurrent.

Dynamic Voltage Scaling is one of commonly used power management schemes. However, it can also be implemented to change the timing characteristics of the circuit. DVS is the process of adjusting the supply voltage to satisfy certain constraints (power and cycle time). As it can be seen in (3.40) and (3.41), the change in propagation delay is inversely proportional to the supply voltage. Stepping up V_{DD} can, therefore, help minimize the circuit delay. Dynamic Frequency Scaling is a technique where the circuit frequency (consequently, its cycle time) is relaxed to accommodate the increase in critical/near-critical path delays. This is effective for applications where high speed is not very crucial. Adaptive Body Biasing is the final mitigation measure. This method is also usually employed for leakage power reduction in high-performance processors. To mitigate system-level aging effects, ABB can be employed to forward bias the transistors, i.e., maintain aged V_{th} near zero-bias V_{th}. ABB is used to dynamically tune transistor threshold voltages by adjusting the voltage difference between bulk and source (V_{BS}). The relation between V_{th} and V_{BS} can be expressed as

$$V_{th} = V_{th0} + \gamma(\sqrt{\phi_s + V_{BS}} - \sqrt{\phi_s}) \tag{4.1}$$

where, V_{th0} is the nominal threshold voltage, γ is the body effect coefficient, and ϕ_s is the surface potential. When the body is not forward biased (i.e., $V_{BS} = 0$), the threshold voltage becomes equal to the nominal V_{th}. As threshold voltage increases due to aging, applying a small positive V_{BS} will revert the elevated V_{th} back to a value close to its nominal V_{th}.

Auxiliary Unit: The auxiliary unit (AU) provides additional information such as mobility (μ), transconductance (g_m) and drain current (I_d). These parameters depend on the threshold voltage, which can be expressed as follows

$$\mu = \frac{\mu_{eff}}{(1 + \alpha_{mob} N_{IT})^m} \tag{4.2}$$

$$I_d = 2 \times \mu \times C_{ox} \times \frac{W}{L} \times (V_{DD} - V_{th})^2 \tag{4.3}$$

$$g_m = \frac{I_d}{V_{DD}} \tag{4.4}$$

where, μ_{eff} is the effective mobility, N_{IT} is the number of interface traps given by $\Delta V_{th} C_{ox}/q$, α_{mob} and m are technology dependent parameters, C_{ox} is the oxide capacitance, and W and L are the transistor width and length respectively. Including these signals as a part of the circuit elements of the logic unit is redundant. Hence, consolidating them in a single auxiliary unit makes the architecture more modular. The AU is connected to the testbench sink, so that the degradation of μ, g_m, and I_d due to aging can be observed.

4.4 System-C AMS Implementation

The System-C AMS implementation of the system architecture (Figure 4.3) is shown in Figure 4.4, along with its significant ports and signals.

Source: The output ports in the source portion of the testbench are `path_out` and `edge_out`. They are declared as follows:

```
sca_tdf::sca_out<double> path_out;
sca_tdf::sca_out<bool> edge_out;
```

The `path_out` port provides initial delay value to the circuit path. The `edge_out` port is of type `bool` and indicates whether the input to the first element will be a rising (HIGH/1) or a falling one (LOW/0). These functions are performed within the `processing()` function of the source node.

```
void processing
{
    path_out.write(0);
    edge_out.write(1);
}
```

Logic Unit: The structure of gate/FF node is shown in Figure 4.5(a). For every gate/FF node in the system, the input `Vdd_in` port receives the supply voltage signal from the previous node, and issues an output signal to the next node through the `Vdd_out` port. Similarly, the `Vbs_in/out` and `f_in/out` port pairs are used to communicate bulk-source voltage (V_{BS}) and frequency (f) signals. The `path_in/out` input-output port pair handles the path delay. The pseudocode of every node in the logic unit is as follows.

```
MODULE: GATE/FF
{
    1: Initialize values of variables required for Vth and delay
       calculations
```

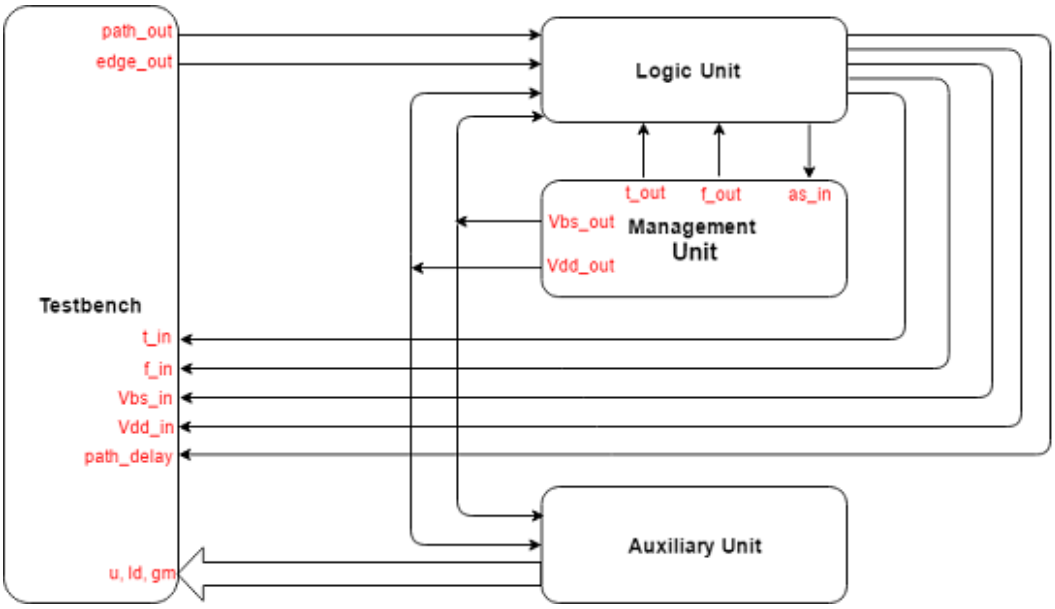


Figure 4.4: System-C AMS implementation of system-level aging assessment platform

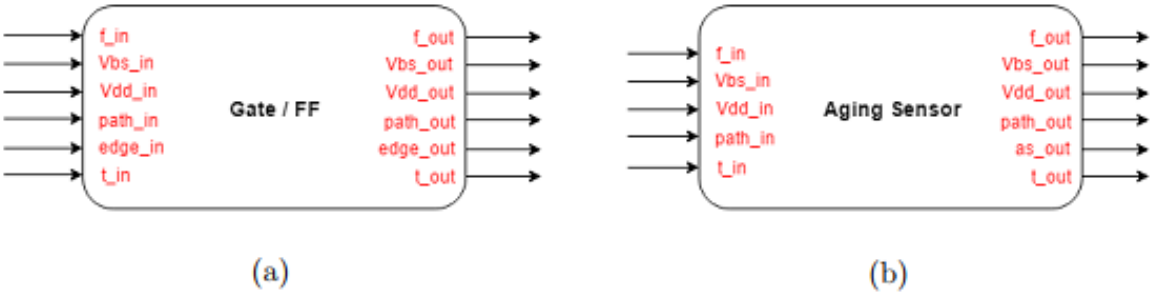


Figure 4.5: (a) High-level gate/FF architecture (b) High-level aging sensor architecture

```
    2: Assign initial values of Vth and delay observed from
       gate/FF characterization
    3: Receive input port values
    4: If num_of_iteration = 1 or MU implements DVS, DFS and ABB
    5:      Compute preliminary values needed to perform Vth
            degradation calculation
    6: Calculate del_Vth_p and del_Vth_n
    7: Perform trpd and tfpd calculations
    8: Issue updated values to corresponding output ports
}
```

Firstly, the values for the various parameters used in V_{th} and delay calculations are assigned. As per the second step of the pseudocode, the initial values of V_{th} and delays for the particular type of gate are assigned in the `processing()` function. These primary values have been obtained from the characterization process. Afterwards, the node performs the necessary initial calculations (E_{ox}, K_v, etc.) to determine ΔV_{th} when $t = 0$ or when MU alters the value of the control parameters (V_{DD}, f and V_{BS}). Subse-

quently, the values of ΔV_{th} for pMOS and nMOS transistors are calculated (`del_Vth_p` and `del_Vth_n` respectively), and the change in gate delay is computed. The `path_in` port receives the incoming delay value from the previous node, adds it to the propagation delay of the current node, and assigns the sum to the outgoing port, `path_out`. As the delays get accumulated further along the path, the value from the `path_out` port of the last node corresponds to the path delay. The `edge_in/out` ports represent the incoming and outgoing edges, and the `t_in/out` ports are used to propagate time signals along the path. By explicitly including time as a signal, better analysis resolution can be achieved. For example, aging assessment can be performed yearly, monthly, or daily, by adjusting the time signal accordingly. The FF node also has the same structure and internal functionality. The following pseudocode describes the aging sensor node:

```
MODULE: AGING SENSOR
{
    1: Initialize values of variables required for Vth and tg
       calculations
    2: Assign initial values of Vth and tg observed from
       sensor characterization
    3: Receive input port values
    4: If num_of_iteration = 1 or MU implements DVS, DFS and ABB
    5:      Compute preliminary values needed to perform Vth
            degradation calculation
    6: Calculate del_Vth_p and del_Vth_n
    7: Determine tg
    8: Cycle time = path delay + calculated tg
    9: If cycle time >= calibrated value
    10:     Issue error to MU
        else
    11:     No error issued
    12: Write updated values to corresponding output ports
}
```

The aging sensor shown in Figure 4.5(b) has an architecture similar to that of a gate, and an additional output port, `as_out`, which communicates to the MU if it predicts a delay error. The first part of the node functionality is almost the same as that of a gate/FF. Once V_{th} degradation is determined, the guard-band interval (t_g) of the sensor is calculated. If the sum of the path delay and the sensor's guard-band interval exceeds the predetermined cycle time value, `as_out` writes a HIGH value to its output.

Management Unit: The Management Unit has five output ports, namely `Vdd_out`, `Vbs_out`, `f_out`, `t_out` and `as_in`. The first four ports are used to supply the initial values of supply voltage (`V1`), bulk-to-source voltage (`Vb1`), frequency (`f1`) and time. The last port receives `bool` input from the aging sensor (the `as_out` port), whose value is HIGH if the sensor detects an impending timing violation. The pseudocode depicting the functionality of the MU is shown below.

```
MODULE: MANAGEMENT UNIT
{
    1: Initialize values of Vdd (V1), f (f1), Vbs (Vbs1), and t
    2: Receive input port values
```

```
    3: If as_in = TRUE
    4:      Adjust t to implement countermeasure from the same time
            instant
    5: If mitigation method = DVS
    6:      If current Vdd = V1 and as_in = TRUE or
            If current Vdd = V2 and as_in = FALSE
    7:          Vdd <- V2
    8:      If current Vdd = V2 and as_in = TRUE or
            If current Vdd = V3 and as_in = FALSE
    9:          Vdd <- V3
    10: If mitigation method = DFS
    11:     If current f > f3 and current f < f1
    12:         f <- f2
    13:     Else if current f < f2
    14:         f <- f3
    15. If mitigation method = ABB
    16:     If current Vbs = Vb1 and as_in = TRUE or
            If current Vbs = Vb2 and as_in = FALSE
    17:         Vbs <- Vb2
    18:     If current Vbs = Vb2 and as_in = TRUE or
            If current Vbs = Vb3 and as_in = FALSE
    19:         Vbs <- Vb3
    20:     If current Vbs = Vb3 and as_in = TRUE or
            If current Vbs = Vb4 and as_in = FALSE
    21:         Vbs <- Vb4
    22: Write updated values to corresponding output ports
}
```

Initially, the MU issues the preliminary values of the operating parameters, such as V_{DD}, f, V_{BS} and t (time). If the aging sensor predicts a delay error and `as_in` goes HIGH, the MU takes corrective action at the same time instant. Then, the MU alters the value of V_{DD}, f and V_{BS} through the corresponding output ports. When an error is predicted and the current value of V_{DD} is equal to its initial value (V1), DVS steps up the operating voltage to V2. When system operation sustains at V2 without any errors (i.e., `as_in.read() == 0`), the system continues to operate at the same voltage. The same principle applies for DFS and ABB, with a difference in the control parameters (f and V_{BS} respectively).

Auxiliary Unit: An outline of the AU implementation is as follows:

```
MODULE: AUXILIARY UNIT
{
    1: Initialize values of variables required for Vth, mobility (u),
       transconductance (gm) and drain current (Id) calculations
    2: Assign initial values of Vth of pMOS and nMOS transistors
    3: Perform long-term Vth, transconductance, mobility and drain
       current calculations of pMOS transistor
    4: Perform long-term Vth, transconductance, mobility and drain
       current calculations of nMOS transistor
    5: Issue computed parameter values to corresponding output ports
}
```

Inside the `processing()` function of the auxiliary unit, the ΔV_{th} values of pMOS and nMOS transistors are determined, and additional parameters such as mobility, transconductance, and drain current are evaluated according to (4.2), (4.3) and (4.4). The AU is connected to the input ports of the testbench sink, namely `u_in`, `gm_in`, and `Id_in` so as to observe the changes in mobility, transconductance and drain current due to aging.

5 Results and Discussion

This chapter describes the analysis results of three circuits that are subjected to system-aging assessment. The simulations verify if the lifetime requirements are met, and execute necessary attenuation steps to mitigate system-level aging.

5.1 Experimental Results

The models, which represent the aged behaviour of combinational circuits, sequential circuits and aging sensor, are created in System-C AMS, considering 65-nm Predictive Technology Model (PTM). Every individual circuit element (gate, FF, aging sensor) is modelled as a System-C AMS class, with its corresponding input and output ports. These classes compute V_{th} and delay degradation internally, communicating the necessary values through their ports. The command line requires four input arguments, which are

```
./Executable_Name Temperature Required_Lifetime Analysis_Resolution
Mitigation_Measure
```

Analysis resolution is the minimum interval between two successive aging assessments. The last input parameter refers to the type of counteractive measure chosen. Simulation results are presented for three CUTs (circuit under test). A nominal V_{DD} of 1.3 V is considered. Temperature (T) variations are restricted from 27°C to 105°C. Worst-case (WC) conditions are $\alpha = 0.9$ (stress duty cycle) and $T = 105$°C, and best-case (BC) operating conditions are $\alpha = 0.1$ and $T = 27$°C. For a complete and thorough analysis, three circuits are considered.

5.2 CUT I - FIR5 Filter

A 5^{th} order finite impulse response (FIR) filter circuit is primarily used in signal processing applications for digital communication purposes. A few of its critical applications include noise suppression in medical imaging devices, and signal storage in media. The circuit schematic is shown in Figure B.1 of Appendix A. The filter circuit has a critical path with 23 gates, one FF (source) and one aging sensor. The critical path composition of FIR5 filter is shown in Table 5.1. Figure 5.1 shows the simulated best-case and worst-case path delay variations of the circuit using the system-level prototype and compares it with the circuit-level simulations.

As it can be seen from Figure 5.1, the critical path delay increases by almost 1.8 ns from the start of its fresh operation to the point it reaches its required lifetime under persistent WC conditions. Best-case circuit performance deterioration is almost half its

Type of cell	# of cells
2-input AND	5
2-input OR	2
2-input XOR	8
AO21TTF	5
MUX	2
DFF	1

Table 5.1: CUT I critical path composition

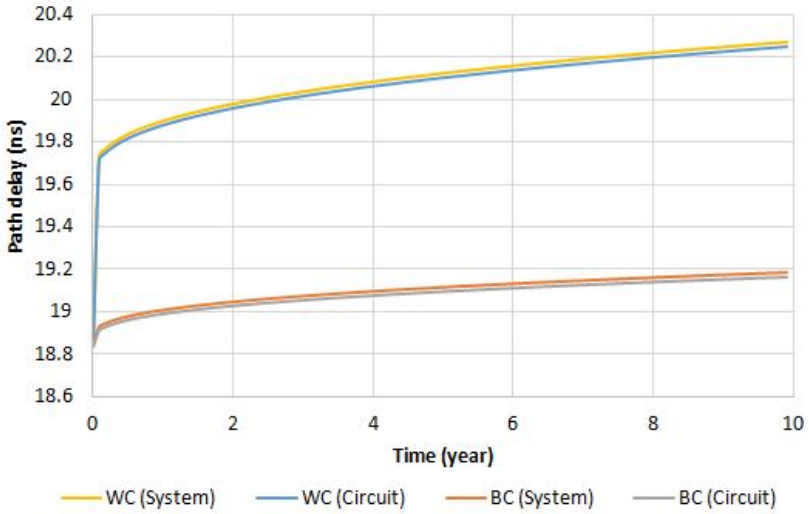


Figure 5.1: Best-case (BC) and worst-case (WC) path delay variations of CUT I (V_{DD} = 1.3 V)

worst-case value. These trend lines give bounds for the circuit lifetime estimates. This circuit is calibrated to operate at a frequency of 46.7 MHz, i.e., with a cycle time of 21.4 ns (sum of path delay and guard-band of aging sensor). It is evident from Figure 5.1, the set lifetime will not be met under WC conditions. Therefore, it is vital to employ mitigation measures.

From Figure 5.2(a), it can be seen that the path delay increases very rapidly and causes a delay violation even before the circuit completes a year of continuous operation. When aging causes the path delay to breach the predetermined cycle time, the MU counteracts by stepping up the voltage to 1.4 V (steps of 0.1 V from 1.3 V). Increasing the supply voltage has an adverse effect on the propagation delay, and it can be observed that the path delay immediately reduces well below the critical level. It can be observed that MU scales V_{DD} at the same instant when aging sensor predicts an error, without delaying action any further.
Figure 5.2(b) shows the dynamic frequency scaling measure to mitigate aging. Dynamic frequency scaling can be achieved by tuning the circuit frequency. Unlike the DVS mechanism, DFS does not have a direct impact on the propagation delay. Rather, it increases the cycle time (by reducing the frequency in steps of 0.5 MHz) of the circuit.

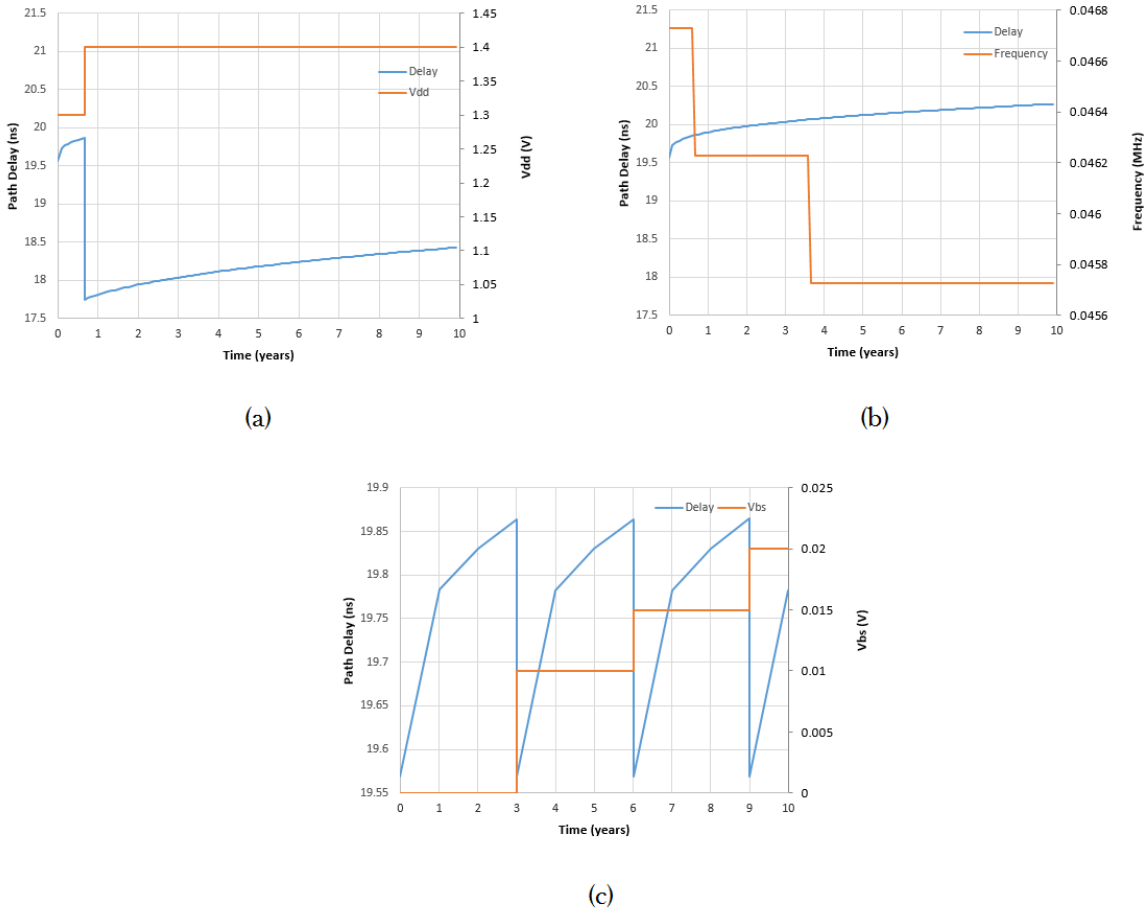


Figure 5.2: CUT I operation and corrective measure by MU using (a) dynamic voltage scaling (b) dynamic frequency scaling (c) adaptive body biasing

This measure can be implemented in specific applications where the speed of circuit operation is not very important. ABB forward biases the transistors to maintain aged V_{th} near zero-bias V_{th}. From Figure 5.2(c), it can be observed that as the propagation delay increases, MU adjusts V_{BS} (in steps of 5 mV) such that it brings back the delay closer to its initial value.

5.3 CUT II - ISCAS 74181

This circuit is an ISCAS benchmark, with 14 inputs, 8 outputs, and 61 gates. Most commonly used arithmetic and logic units (ALU) are based on this 4-bit benchmark. This circuit can perform all the traditional add, subtract, decrement operations with or without carry. It can also perform AND, NAND, OR, NOR, XOR and shift operations. Multiplication and division is not provided, but can be performed in multiple steps using shift and add/subtract operations. The circuit schematic is shown in Figure 5.3, with the critical path highlighted.

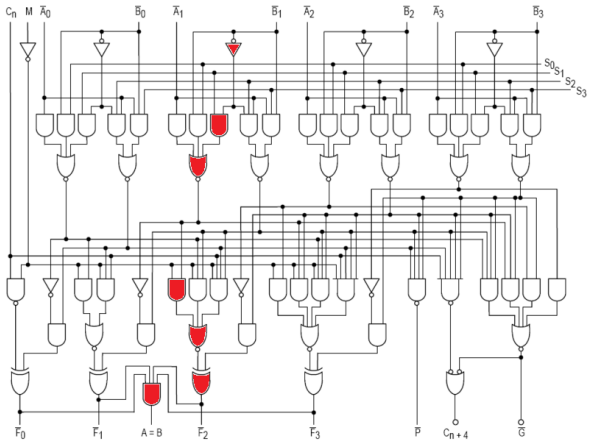


Figure 5.3: CUT II circuit schematic

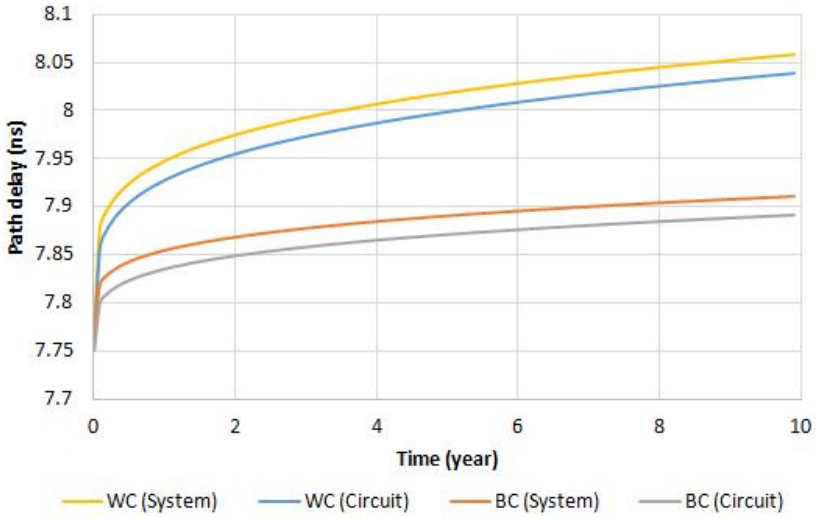


Figure 5.4: Best-case (BC) and worst-case (WC) path delay variations of CUT II ($V_{DD} = 1.3$ V)

ALU is a typical part of many processors, hence, its reliability must be ensured. The critical path has 7 gates from input to output. Figure 5.4 shows the best-case and worst-case delay of 74181 IC. The initial path delay is 7.77 ns. Including margins for guard-band, the operating frequency is fixed at 105.263 MHz. This corresponds to a cycle time of 9.5 ns. Figures 5.5(a), (b) and (c) show the response of MU to aging through DVS, DFS and ABB. DVS and ABB cause variations in the path delay throughout the circuit's lifetime. DFS tolerates the rise in path delay by adjusting the circuit frequency accordingly. As 74181 is an integral component is modern day processors, DVS can be incorporated during the first half lifetime, and ABB for the subsequent half lifetime.

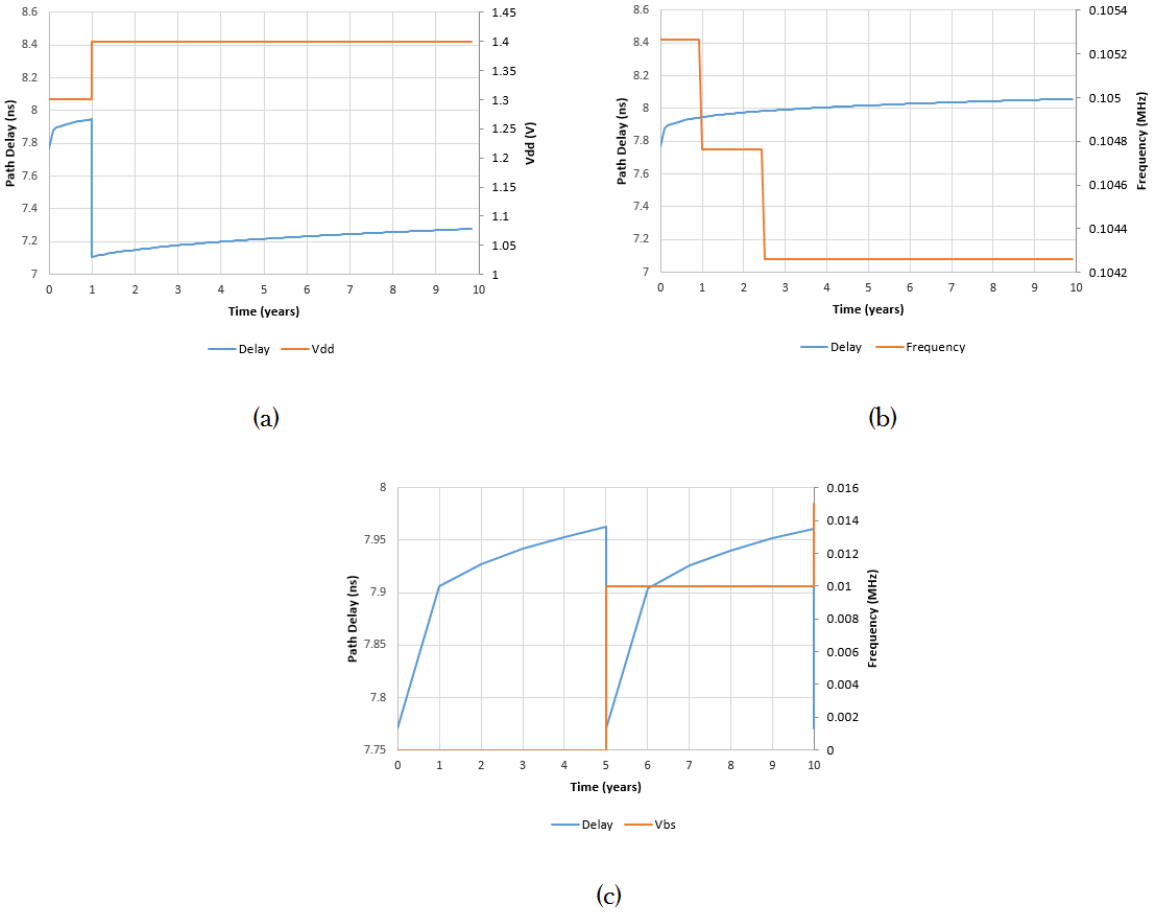


Figure 5.5: CUT II operation and corrective measure by MU using (a) dynamic voltage scaling (b) dynamic frequency scaling (c) adaptive body biasing

5.4 CUT III - Divider FSM

This circuit is an 8-bit divider that uses a finite state machine (FSM). FSM circuits are important in systems that function real-time. Predictability (next state must be definitely known) and reliability (hard real-time deadlines must be met) are the main characteristics of any real-time system. In such applications, ensuring circuit reliability takes priority.

Figure 5.6 shows the circuit layout and the critical path, which has 7 gates. The path delay bounds are shown in Figure 5.7, where the delay can be expected to vary between 12.13 ns and 12.6 ns over its lifetime. Critical path delay increases by 50 ps in 10 years. The frequency and clock cycle time are calibrated at 71.994 MHz and 13.89 ns, respectively.

When the time resolution specified is smaller, fine granularity in the analysis is achieved. In Figure 5.8(a), the cycle time exceeds the set value at 0.4 years, causing the MU to ramp up the supply voltage. Figure 5.8(b) shows that two levels of frequency tuning is required in DFS, at the 0.4 year mark and at the 1.6 year mark. The rate of

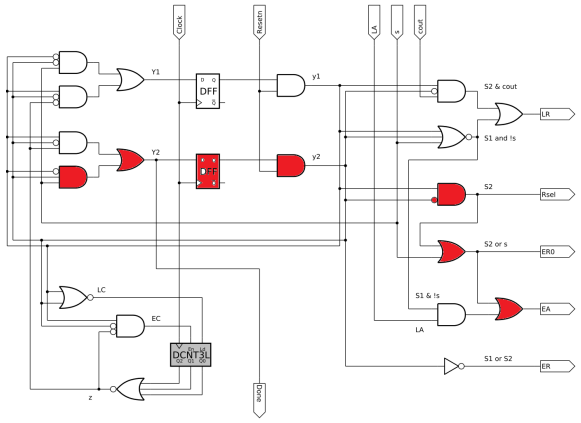


Figure 5.6: CUT III circuit schematic

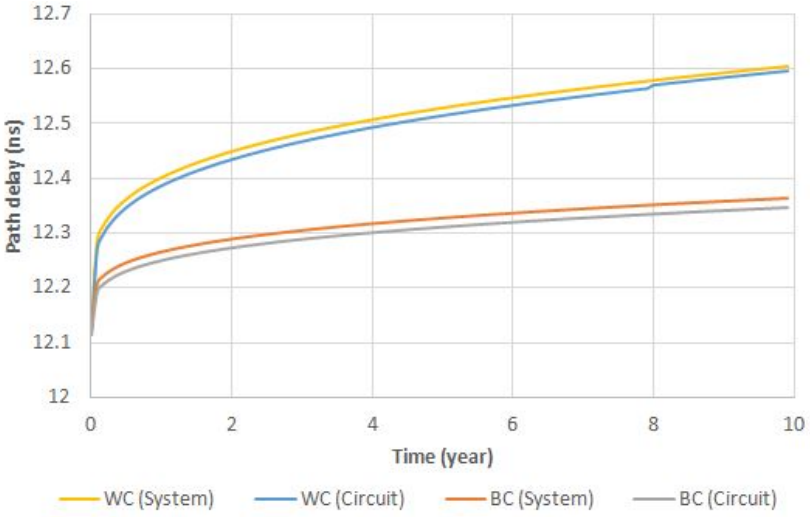


Figure 5.7: Best-case (BC) and worst-case (WC) path delay variations of CUT III (V_{DD} = 1.3 V)

delay increase is prominent during the first few years of circuit operation. As the change in path delay gradually reduces, no more adjustments in the frequency is required. Figure 5.8(c) shows that V_{BS} is varied in steps of 5 mV, which causes the path delay to drop, as soon as aging causes a timing violation. The rise and fall in path delay is almost periodic, when same operating conditions prevail. Figures 5.9(a), (b) and (c) show the degradation in mobility, drain current, and transconductance at WC conditions.

It can be noted that for the same lifetime requirements, ABB requires more levels than DFS, which in turn, requires more levels than DVS. The expected lifetimes of the three CUTs under best-case and worst-case conditions are shown in Table 5.2. The lifetime of every CUT barely exceeds 1 year under persistent WC conditions. The simulated MU prototype helps maintain system reliability in such cases. Table 5.3 shows a speedup comparison between Cadence NCSim and the System-C AMS simulation platform. Simulations were carried out on a 4-core Intel i5-4690 processor

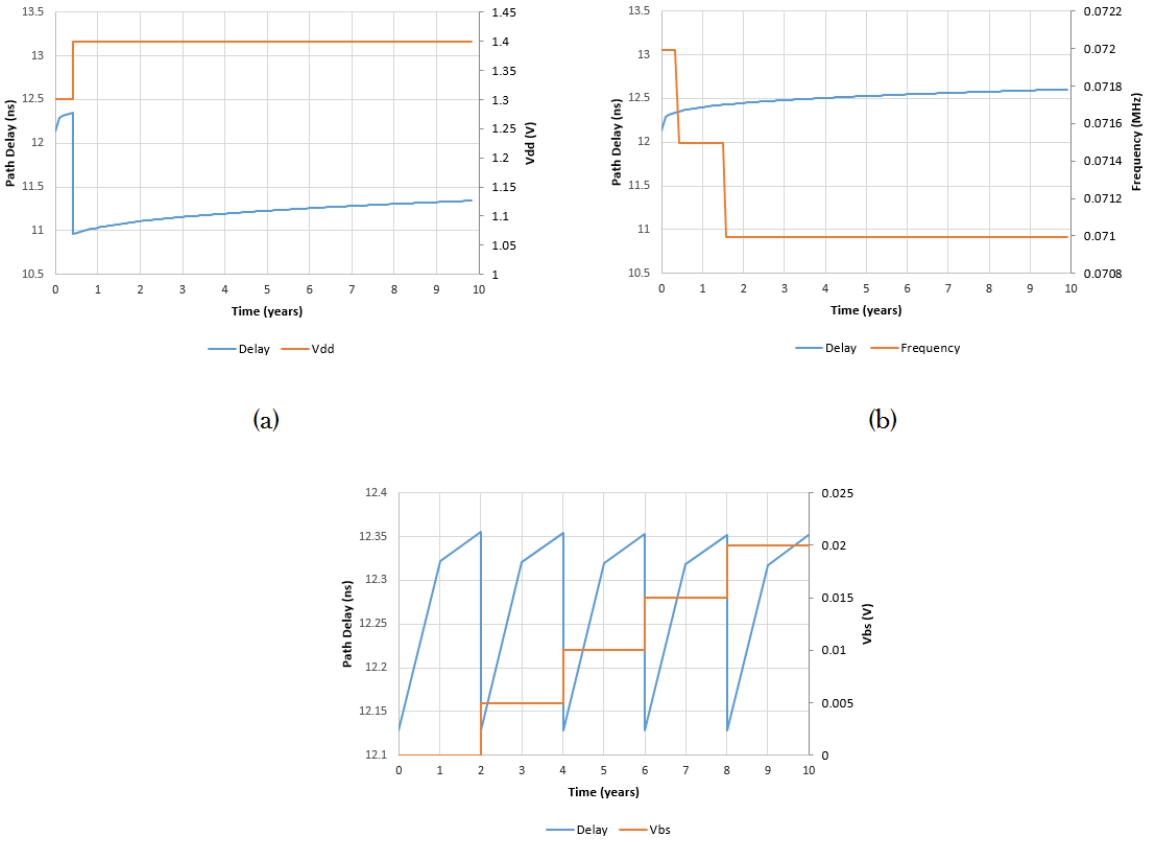


Figure 5.8: CUT III operation under WC conditions and corrective measure by MU using (a) dynamic voltage scaling (b) dynamic frequency scaling (c) adaptive body biasing

Circuit	Speed (GHz)	WC lifetime (years)	BC lifetime (years)
CUT I	0.0467	0.6	8
CUT II	0.105263	0.8	10
CUT III	0.071994	0.4	10

Table 5.2: CUT lifetime assessment

with 8 GB RAM, running at a frequency of 3.5 GHz. As shown in Table 5.3, the System-C AMS platform is faster, with an average speedup of almost 15 times over Cadence. Furthermore, additional aging effects such as TDDB can be included in this aging assessment method, making this framework very flexible for analysis across aging mechanisms and process technologies. Remarkably, every other reliability tool requires design of the actual system, in order to view the timing report and assess the lifetime. Such simulations are complex and time-consuming. Through the proposed

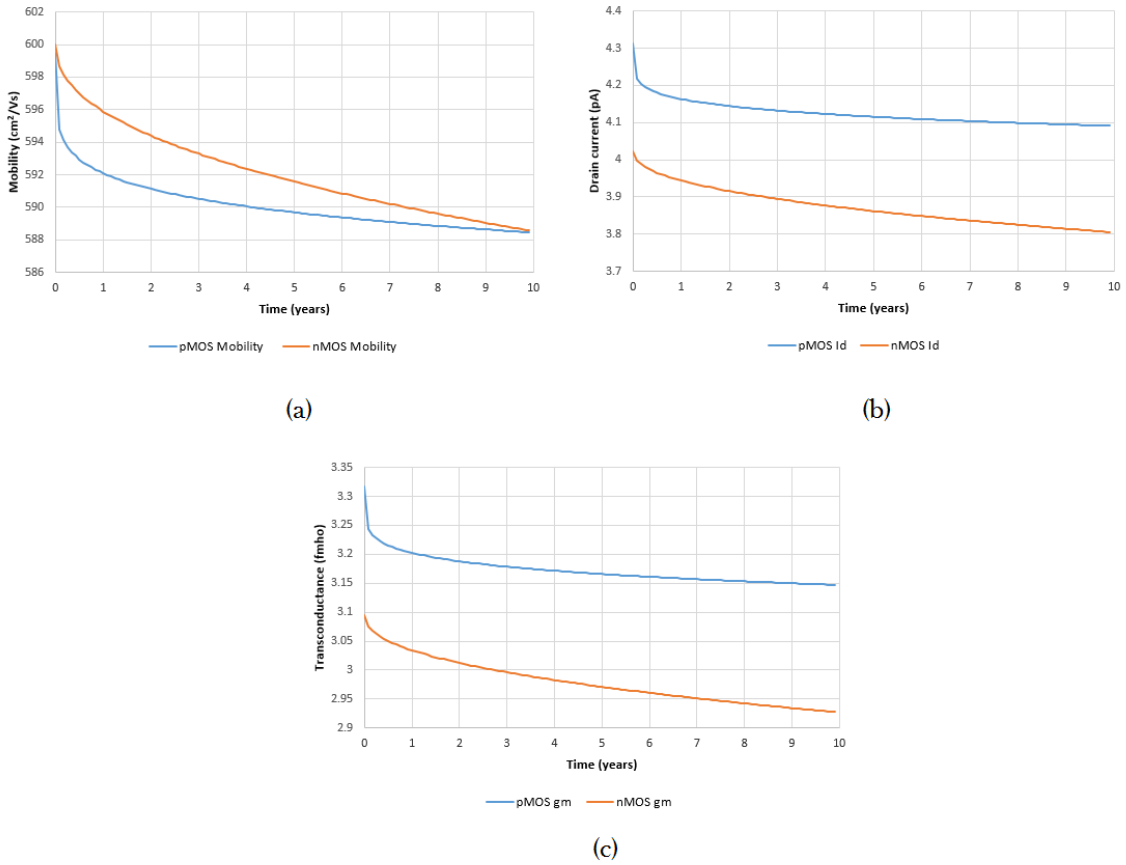


Figure 5.9: (a) Mobility (b) Drain current (c) Transconductance variation of pMOS and nMOS transistors with aging

Circuit	Execution Time (μs)		Speedup
	Cadence	System-C AMS	
CUT I	0.68	0.031	21.93
CUT II	0.44	0.032	13.75
CUT III	0.39	0.033	11.81

Table 5.3: Speedup comparison

aging assessment framework, this process can be simplified.

6 Conclusion and Future Work

This chapter provides a summary of this thesis and highlights the goals that were achieved. The scope for future work is also presented.

6.1 Summary

Aging causes severe reliability issues in nanoscale circuits, which ultimately shortens IC lifetime. Typically, designers include safety margins for circuit operation to accommodate these effects. However, such a conservative feature fails to exploit the high speeds that these nanoscale devices have to offer. Using circuit simulations to determine the long-term aging effects of a typical SoC is very complicated. Moreover, the designer has to rely on the lower-level circuit foundations (transistors, gates, FFs, interconnections) to perform aged circuit simulations. This approach is not practical in simulating a system to perform aging analysis. Thus, a methodology, which abstracts low-level aging models to the system-level, expediting system aging assessment process without requiring any knowledge of underlying circuit dynamics was presented.

A brief summary of the thesis is presented as follows.

- The goal of this thesis to abstract device aging models to the system level is stated. The related objectives to perform system-level aging assessment are identified and the main contributions are specified. A concise introduction to each chapter was provided.
- Various root causes of aging that cause CMOS device failure are discussed. The reliability-aware framework to be used is introduced. Following a bottom-up approach, several aging phenomena at the transistor level are listed and their processes are elaborated. Aging effects cause threshold voltage increase at the transistor-level, propagation delay increase at the gate-level, and cumulatively lead to path delay errors at the circuit-level.
- An integrated model for NBTI and CHC aging effects based on the Reaction-Diffusion (R-D) model is explained. The NBTI shows that the threshold voltage (V_{th}) degradation in a pMOS transistor is strongly dependent on supply voltage, temperature, and stress duty cycle. On the other hand, the CHC model shows that the amount of V_{th} increase in nMOS transistors is influenced by the supply voltage. MATLAB plots of these models are presented, which show the long-term V_{th} increase in p- and n-MOSFETS.
- Aging impact at the next higher stage, i.e., in combinational and sequential circuit elements is analyzed. Gate delay models that can replicate propagation delays accurately are discussed. The relational model is chosen for good accuracy (error

less than 2%) and lack of large LUTs to store proportionality constants. Similarly, sequential cell (flip-flop) delay is also modelled. The required gates are classified based on their type and characterized to replicate their aged behaviour for the required operating conditions (V_{DD} and T). The Adaptive Error Prediction flip-flop is chosen as the aging sensor to detect timing violations, and its observation interval is characterized.

- The system architecture used to determine SoC lifetime is proposed. The concept of a Management Unit (MU), which continuously monitors the path for timing errors and overcomes aging by employing DVS, DFS and ABB, is introduced. System-C AMS implementation is described, wherein, the circuit elements are black-box models with input/output ports for communication purposes (values). This does not require any knowledge of the underlying circuit dynamics. For a given circuit, the critical path/near-critical path can be simulated using the characterized circuit elements.

- Analysis results are presented for three CUTs. Under persistent worst-case conditions, the lifetime of each circuit is very limited. Simulations show that the MU increases the life expectancy of the circuit. The System-C AMS prototype platform has an average speedup of 15 times over other tools. Moreover, additional effects such as PBTI, TDDB can be incorporated in this analysis since behavioural models are used.

In a working system, the critical path may vary over time, depending on its workload and other factors (environmental, temperature, etc). Thus, aging might cause a particular path, which was considered to be non-critical, to become the most pivotal one in determining system reliability. This framework assumes that the critical path of the system remains the most critical one throughout its lifetime, which may not be the case. Additionally, the framework performs reliability assessment for worst-case corners, taking a homogeneous stress duty cycle value (0.1 for best-case and 0.9 for worst-case). In reality, this varies over time, and therefore, it is crucial to impart this characteristic of α in the framework.

6.2 Future Work

Research on several aspects of aging (from transistor-level models to system performance assessment) is still being carried out. The scope of the proposed high-level aging assessment scheme can be improved to include the following:

1. In this thesis, the critical path is assumed to be identified beforehand. In cases where this information is not available sooner, modelling procedures must be modified accordingly.

2. In addition to the R-D model, several other models, such as the T-D model, I_{SUB}-based model etc. are available. An analysis can be made if these behavioral models can enhance the accuracy of aging prediction.

3. Characterizations can be extended to include whole technology libraries. This would increase the simulation coverage of a wide range of circuits. Based on gate characterizations performed at certain base conditions (V_{DD} and T), the proportionality constants for operating conditions in between can be obtained by interpolation.

4. Although best- and worst-case aging analysis is performed by assuming homogeneous stress-duty cycle, it may vary in reality. Such variations can be considered for performing characterization. Additional aging effects can be included to complete all the device-level aging phenomena.

5. The proposed methodology assumes that the system lifetime is solely dependent on the critical path. Multiple path simulation is required to incorporate near-critical paths along with the critical path. Dynamic variation of α that causes a near-CP to become the most critical one is required to be accounted for.

Appendix A

A.1 NBTI Transistor Degradation Model

A.1.1 Stress Phase

As per the R-D mechanism, NBTI can be physically described as the generation of charges in the region close to the interfacial layer, i.e., the Si-SiO_2 surface [26]. During the stress phase of NBTI, interface charges (H^+) are induced due to the breaking of Si-H bonds. Let N_0 be the initial concentration of the Si-H bonds and P be the concentration of inversion holes. The generation rate of the interface traps is, thus, given by [28]

$$\frac{dN_{IT}}{dt} = k_F(N_0 - N_{IT})P - k_R N_H N_{IT} \tag{A.1}$$

where, k_F and k_R are the reaction rates of the forward and reverse reactions, and N_H is the hydrogen density at the Si-SiO_2 interface. When the stress phase starts, the trap generation is quite slow [33]. Thus, $dN_{IT}/dt \approx 0$ [28] and $N_{IT} \ll N_0$, and (A.1) becomes

$$N_H N_{IT} \approx \frac{k_F}{k_R} P.N_0 \tag{A.2}$$

As Hydrogen (H) atoms are continually generated at the interface during the forward reaction, two H atoms may combine to form a more stable H_2 molecule. The concentration of H_2 molecules (N_{H_2}) is related to the concentration of H atoms [28] through

$$N_{H_2} = k_H {N_H}^2 \tag{A.3}$$

Driven by the increasing density of generated H_2, the H_2 current diffuses into the oxide and then into the poly-Si (the gate). This process influences the balance of the diffusion process [28] by

$$\frac{dN_H}{dt} = D_{H_2}\frac{d^2 N_H}{dx^2} \tag{A.4}$$

D_{H_2} represents the diffusion constant of the H_2 molecule [28]. The solution of (A.4) exhibits a power-law dependence on time. The value of the power-law index depends on the type of diffusion species, i.e., based on whether H or H_2 diffuses toward the gate. At any point in time t, the diffusion front moves forward by a distance [28]

$$x_{DF}(t) = \sqrt{D_{H_2} t} \tag{A.5}$$

Let t_0 be the time taken by H_2 to reach the SiO_2/poly-Si interface. Once the diffusing species reaches the poly-Si, the diffusion front moves further in the poly-Si [28] as

$$x_{DF}(t) = \sqrt{D_{H_2}(t - t_0)} \tag{A.6}$$

Since the diffusion rate of H_2 in oxide is much higher than its diffusion rate in poly-Si [27], the time taken by H_2 to reach the SiO_2-polySi interface (i.e., inside the oxide substrate) is very small, or $t \gg t_0$ [28]. Thus, after a certain amount of time t, the diffusion front can be approximated to be at a distance of $\sqrt{D_{H_2}t} + t_{ox}$ from the Si-SiO_2 interface, where t_{ox} indicates the thickness of the oxide layer. The total number of interface charges produced after time t is twice the number of H_2 molecules generated during this time [28], since two Hydrogen atoms combine to form a Hydrogen molecule. Thus, from [28],

$$N_{IT} = 2\int_0^{x_{DF}(t)} N_{H_2}(x)dx \tag{A.7}$$

The total hydrogen can be divided into two categories - 1) hydrogen in the oxide and, 2) hydrogen in the poly-Si. There exists a very small difference between the concentration of H_2 at the Si-SiO_2 interface and the SiO_2-polySi interface. A fractional parameter δ is introduced to represent the fractional drop in the concentration of H_2 at the SiO_2-polySi interface [28]. Thus, (A.7) can be rewritten as [28],

$$N_{IT} = 2\int_0^{t_{ox}} N_{H_2}(x)dx + 2\int_{tox}^{\sqrt{D_{H_2}t}+t_{ox}} N_{H_2}(x)dx \tag{A.8}$$

$$\approx 2\left(\frac{1}{2}(1+\delta).N_{H_2}(0).t_{ox} + \frac{1}{2}N_{H_2}(0).\sqrt{D_{H_2}t}\right) \tag{A.9}$$

where, $N_{H_2}(0)$ is the concentration of H_2 at the Si-SiO_2 interface and $\delta N_{H_2}(0)$ represents the density of the N_{H_2} at the SiO_2-polySi interface. Replacing $N_{H_2}(0)$ from (A.3) in terms of $N_H(0)$ [28], we get

$$N_H(0) = \left(\frac{N_{IT}}{k_H((1+\delta)t_{ox} + \sqrt{D_{H_2}t}}\right)^{\frac{1}{2}} \tag{A.10}$$

Using (A.2) and (A.10), N_{IT} [28] can be represented as

$$N_{IT} = \left(\frac{\sqrt{k_H}k_F N_0 P}{k_R}\right)^{\frac{2}{3}}\left((1+\delta)t_{ox} + \sqrt{D_{H_2}t}\right)^{\frac{1}{3}} \tag{A.11}$$

where, $k_F N_0/k_R$ and inversion hole density ($P = C_{ox}(V_{gs} - V_{th})$) are proportional to the vertical electrical field ($E_{ox} = (V_{gs} - V_{th})/t_{ox}$) [28]. The change in threshold voltage ΔV_{th} due to the interface charges is given by [28] as

$$\Delta V_{th} = \frac{Q_{IT}}{C_{ox}} = \frac{qN_{IT}}{C_{ox}} \tag{A.12}$$

where, Q_{IT} is the total interface charge, q is the electron charge, N_{IT} is the number of interface charges and C_{ox} is the oxide capacitance per unit area. Substituting (A.11) in (A.12), we obtain the general form of V_{th} degradation [28] as

$$\Delta V_{th}(t) = A\left((1+\delta)t_{ox} + \sqrt{Ct}\right)^{2n} \tag{A.13}$$

where,

$$A = \left(\frac{qt_{ox}}{\epsilon_{ox}}\right)^{1/2n}\sqrt{K^2 C_{ox}(V_{gs} - Vth)\left(exp\left(\frac{E_{ox}}{E_0}\right)\right)^2}, \tag{A.14}$$

C is the diffusion constant having a temperature dependence as $C = {T_o}^{-1}exp(-E_a/kT)$ [28], k is the Boltzmann constant and T_o is another constant. For a H_2 based diffusion model, $n = 1/6$ and for a H based model, $n = 1/4$. The initial concentration of Si-H bonds, N_0, and the quality of oxide will wary with different processes. Higher N_0 and poor oxide quality can cause higher V_{th} degradation [28]. In the model described in (A.14), the parameter K is used to adjust the V_{th} degradation rate due to different processes.

A.1.2 Recovery Phase

In the recovery phase, there is no net generation of interface traps due to the absence of holes. Thus, the forward reaction term in (A.1) becomes zero [35]. The hydrogen species that were generated during the stress phase continue to diffuse away from the interface toward the poly-Si gate [28]. However, the reverse reaction term will not remain zero during the recovery phase. This indicates that, some of the hydrogen species that are closer to the interface diffuse back and re-passivate the broken Si^+ bonds [28]. This results in the reduction of H_2 density, $N^A_{H_2}$.

Let t_1 be the time at which recovery starts and $N_{IT}(t_1)$ be the total number of interface charges at the end of the stress cycle. Let $N^A_{IT}(t)$ be the number of annealed charges at time t. Therefore, the number of interface charges at time t is given by [28] as,

$$N_{IT}(t) = N_{IT}(t_1) - N^A_{IT}(t) \tag{A.15}$$

Substituting (A.9) in (A.15), the number of interface charges at time t is given by [28] as

$$N_{IT}(t) = \left((1+\delta)t_{ox} + \sqrt{Ct}\right)N_{H_2}(0) \tag{A.16}$$

Recovery is a two-step process - the fast step of the recovery process involves H_2 back diffusion (diffusion of H_2 back to Si-SiO_2 interface) abd the slow recovery of H_2

relates to back diffusion from poly-Si. Thus, the number of annealed traps can be due to two parts [28] and is given by

$$N_{IT}^{A}(t) = 2\Big(\xi_1 t_e + \frac{1}{2}\sqrt{\xi_2 C(t-t_1)}\Big)N_{H_2}(0) \tag{A.17}$$

where, the first term is approximation of the diffusion profile of H_2 in oxide and the second term is approximation of the diffusion profile of H_2 in poly-Si. ξ_1 and ξ_2 are the corresponding back diffusion constants. Depending on the recovery duration $(t-t_1)$, the effective oxide thickness (t_e) equals to either t_{ox} or the diffusion distance of hydrogen in the initial stage of recovery [28]. Let t' be the time at which all hydrogen species has recombined with the interface traps. Specifically, this is the time taken by the diffusing species to diffuse to a distance of t_{ox}. Typical values of t' are 2.5 ms for H_2 and 0.14 μs for oxide with thickness 1.2 nm [28]. If the recovery duration exceeds t', t_e is equal to t_{ox} [28]. If not, t_e equals the exact diffusion distance of hydrogen in the oxide. From (A.16) and (A.17), we get

$$N_{IT}^{A}(t) = N_{IT}(t)\left(\frac{2\xi_1 t_e + \sqrt{\xi_2 C(t-t_1)}}{(1+\delta)t_{ox} + \sqrt{Ct}}\right) \tag{A.18}$$

Substituting this value of $N_{IT}^{A}(t)$ from (A.18) in (A.15), simplifying using the approximation $1/(1+x) \approx 1$ - x, and using (A.12) [28], we obtain

$$\Delta V_{th}(t) = \Delta V_{th}(t_1)\left(1 - \frac{2\xi_1 t_e + \sqrt{\xi_2 C(t-t_1)}}{(1+\delta)t_{ox} + \sqrt{Ct}}\right) \tag{A.19}$$

where, $V_{th}(t_1)$ is the threshold voltage at the end of the stress phase.

A.1.3 Dynamic NBTI

When a pMOS transistor experiences dynamic NBTI, it undergoes alternate stress and relaxation cycles. Let T be one complete cycle with one stress (t_s) and recovery (t_r) duration, i.e., T = $t_s + t_r$. The V_{th} degradation during the first stress interval [0, t_s] is given by (A.13). Thus, at t = t_s [28],

$$\Delta V_{th}(t_s) = A\Big((1+\delta)t_{ox} + \sqrt{Ct_s}\Big)^{2n} \tag{A.20}$$

The V_{th} degradation during the first recovery [t_s, T] is given by (A.19). Therefore, at t = T [28],

$$\begin{aligned}\Delta V_{th}(T) &= \Delta V_{th}(t_s)\left(1 - \frac{2\xi_1 t_e + \sqrt{\xi_2 C(T-t_s)}}{(1+\delta)t_{ox} + \sqrt{CT}}\right) \\ &= A\Big((1+\delta)t_{ox} + \sqrt{Ct_s}\Big)^{2n}.\theta\end{aligned} \tag{A.21}$$

where,

$$\theta = \left(1 - \frac{2\xi_1 t_e + \sqrt{\xi_2 C(T - t_s)}}{(1+\delta)t_{ox} + \sqrt{CT}}\right) \tag{A.22}$$

$\Delta V_{th}(t)$ for $t \in [T, T + t_s]$ is given by (A.13) [28] as

$$\Delta V_{th}(t) = A\Big((1+\delta)t_{ox} + \sqrt{C(t-T)} + s\Big)^{2n} \tag{A.23}$$

At t = T, (A.23) becomes

$$\Delta V_{th}(T) = A\Big((1+\delta)t_{ox} + s\Big)^{2n} \tag{A.24}$$

In order for the threshold voltage degradation to be continuous, (A.21) and (A.24) must be equal [28]. Therefore, the value of s becomes

$$s = ((1+\delta)t_{ox} + \sqrt{Ct_s})\theta^{1/2n} - (1+\delta)t_{ox} \tag{A.25}$$

Substituting (A.25) in (A.24), we get

$$\Delta V_{th}(t) = (A^{1/2n}\sqrt{C(t-T)} + \sqrt[2n]{\Delta V_{th}(T)})^{2n} \tag{A.26}$$

Extending this for the (m+1)th stress cycle, i.e., for time $t \in [mT, mT + t_s]$ [28], we obtain

$$\Delta V_{th}(t) = (A^{1/2n}\sqrt{C(t-mT)} + \sqrt[2n]{\Delta V_{th}(mT)})^{2n} \tag{A.27}$$

(A.27) can be simplified to get the short-term V_{th} degradation equations for a stress-recovery cycle [28].

$$\text{Stress} : \Delta V_{th}(t) = \left(K_v(t-t_0)^{1/2} + \sqrt[2n]{\Delta V_{th}(t_0)}\right)^{2n} \tag{A.28}$$

$$\text{Recovery} : \Delta V_{th}(t) = \Delta V_{th}(t_1)\left(1 - \frac{2\xi_1 t_e + \sqrt{\xi_2 C(t-t_1)}}{(1+\delta)t_{ox} + \sqrt{Ct}}\right) \tag{A.29}$$

A.2 CHC Transistor Degradation Model

The interface trap generation can be written similar to that in NBTI as a balance between the dissociation and annealing rates of Si-H bonds [28]

$$\frac{dN_{IT}}{dt} = k_F(N_0 - N_{IT})P - k_R N_H N_{IT} \tag{A.30}$$

In the initial part of the stress phase, trap generation is slow [26]. Hence, $dN_{IT}/dt \approx 0$ and $N_{IT} \ll N_0$ [28]. (A.32) becomes

$$N_H N_{IT} \approx \frac{k_F}{k_R} P.N_0 \tag{A.31}$$

The diffusion profile in CHC is conical, since it is concentrated toward the drain end. Thus, the number of interface traps [28] is given by

$$N_{IT} = \frac{1}{3}\pi(\sqrt{D_H t})^2 N_H(0) \tag{A.32}$$

Integrating (3.33) and (3.34) gives

$$N_{IT} = \sqrt{\frac{1}{3}\pi\frac{k_F N}{k_R} N_0(\sqrt{D_H t})} \tag{A.33}$$

where, N is the electron concentration. k_R depends on both the lateral and vertical electric field, contrary to NBTI. The V_{th} degradation caused by CHC [28] is given by

$$\Delta V_{th}(t) = \frac{q}{C_{ox}} K_2 \sqrt{Q_i} exp\Big(\frac{E_{ox}}{E_{o2}}\Big) exp\Big(-\frac{\varphi_{it}}{q\lambda E_m}\Big) t^{n'} \tag{A.34}$$

Appendix B

B.1 Model Parameters for 65-nm Technology

The terms used in the determination of V_{th} degradation due to NBTI and CHC are as follows:

Parameter	**Unit**	**Value**	**Parameter**	**Unit**	**Value**
K_1	$C^{-0.5}nm^{-2.5}$	7.5	q	C	1.602 $\times 10^1 9$
δ	no unit	0.5	t_{ox}	nm	1.2
ξ_1	no unit	0.9	ε_{air}	$m^{-3}kg^{-1}s^4A^2$	8.854 $\times 10^{-21}$
K_2	$nmC^{-0.5}$	1.7 $\times 10^8$	ε_r	no unit	3.9
φ_{it}	eV	3.7	k	eV	8.6173303 $\times 10^{-5}$
V_t	V	0.0259	n	no unit	1/6
E_a	eV	0.49	T_{clk}	μs	100
ξ_2	no unit	0.5	W	μm	2
l	nm	1.7	L	nm	240
A_{bulk}	no unit	0.005	n'	no unit	0.45
m	no unit	1.6	T_0	s/nm^2	10^{-8}
E_{01}	V/nm	0.08	E_{02}	V/nm	0.8
E_{sat}	V/nm	0.011	λ	nm	7.8
α_{mob}	no unit	5	t_e	nm	1.2

Table B.1: 65-nm technology values for long-term performance degradation assessment due to NBTI and CHC

B.2 CUT I - FIR5 Filter

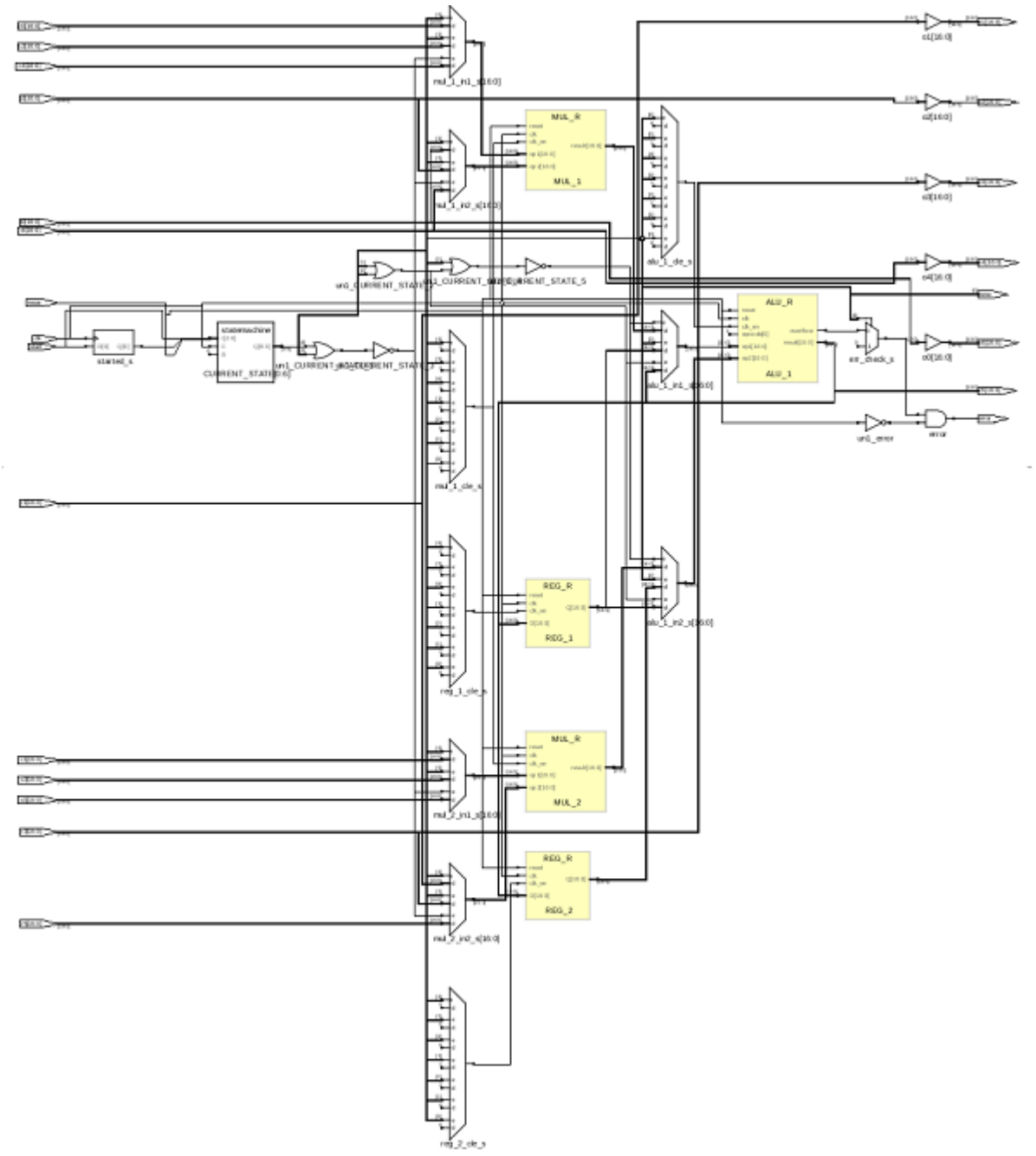


Figure B.1: CUT I - 5th order FIR filter circuit schematic

Bibliography

[1] "The scaling of MOSFETs, Moores law, and ITRS." [Online]. Available: http://userweb.eng.gla.ac.uk/fikru.adamu-lema/Chapter_02.pdf

[2] R. Dennard, F. Gaensslen, V. Rideout, E. Bassous, and A. LeBlanc, "Design of Ion-Implanted MOSFETs with Very Small Physical Dimensions," *IEEE Journal of Solid-State Circuits*, pp. 256 – 268, 1974.

[3] J. McPherson, "Reliability Trends with Advanced CMOS Scaling and The Implications for Design," *IEEE Custom Integrated Circuits Conference*, pp. 405 – 412, 2007.

[4] S. Hamdioui, D. Gizopoulos, G. Guido, M. Nicolaidis, A. Grasset, and P. Bonnot, "Reliability Challenges of Real-Time Systems in Forthcoming Technology Nodes," *Design, Automation Test in Europe Conference Exhibition (DATE)*, pp. 129 – 134, 2013.

[5] Y. Wang, "Aging Assessment and Reliability Aware Computing Platforms," 2013.

[6] M. Ohring, "Reliability and Failure of Electronic Materials and Devices," 1998.

[7] C. V. Martins, J. Semio, J. C. Vazquez, V. Champac, M. Santos, I. C. Teixeira, and J. P. Teixeira, "Adaptive Error-Prediction Flip-flop for Performance Failure Prediction with Aging Sensors," *VLSI Test Symposium*, pp. 203 – 208, 2011.

[8] H. Kkner, S. Khan, P. Weckx, P. Raghavan, S. Hamdioui, B. Kaczer, F. Catthoor, L. V. der Perre, R. Lauwereins, and G. Groeseneken, "Comparison of Reaction-Diffusion and Atomistic Trap-Based BTI Models for Logic Gates," *IEEE Transactions on Device and Materials Reliability*, vol. 14, no. 1, pp. 182 – 193, 2014.

[9] A. W. Strong, E. Y. Wu, R. P. Vollertsen, J. Sunea, G. L. Rosa, T. D. Sullivan, and S. E. Rauch, "Reliability Wearout Mechanisms in Advanced CMOS Technologies," 2009.

[10] J. B. Velamala, K. B. Sutaria, H. Shimizu, H. Awano, T. Sato, G. Wirth, and Y. Cao, "Compact Modeling of Statistical BTI under Trapping/Detrapping," *IEEE Transactions on Electron Devices*, vol. 60, no. 11, pp. 3645 – 3654, 2013.

[11] J. B. Velamala, "Compact Modeling and Simulation for Digital Circuit Aging," 2012.

[12] M. Choudhury, V. Chandra, K. Mohanram, and R. Aitken, "Analytical Model for TDDB-based Performance Degradation in Combinational Logic," *Design, Automation Test in Europe*, pp. 423 – 428, 2010.

[13] B. Mesgarzadeh, I. S. Saab, and A. Alvandpour, "Reliability Challenges in Avionics due to Silicon Aging," *IEEE International Symposium on Design and Diagnostics of Electronic Circuits Systems*, pp. 342 – 347, 2012.

[14] R. F. Orsagh, D. W. Brown, A. J. Hess, and T. Dabney, "Prognostic Health Management for Avionic Systems," *IEEE Aerospace Conference*, pp. 534 – 541, 2006.

[15] J. Pachito, C. V. Martins, J. Semio, M. Santos, I. C. Teixeira, and J. P. Teixeira, "The Influence of Clock-Gating on NBTI-Induced Delay Degradation," *IEEE International On-Line Testing Symposium*, pp. 61 – 66, 2012.

[16] W. Wang, C. Kim, K. Guerin, and D. Kim, "Practical Implementations of Cell Flipping in Instruction Caches for NBTI Lifetime Optimization," *IEEE Transactions on VLSI Systems*, pp. 396 – 402, 2010.

[17] S. Khan, I. Agbo, S. Hamdioui, H. Kukner, B. Kaczer, P. Raghavan, and F. Catthoor, "Bias Temperature Instability analysis of FinFET based SRAM cells," *Design, Automation Test in Europe Conference Exhibition (DATE)*, pp. 1 – 6, 2014.

[18] D. Rossi, V. Tenentes, S. Yang, S. Khursheed, and B. M. Al-Hashimi, "Aging Benefits in Nanometer CMOS Designs," *IEEE Transactions on Circuits and Systems*, vol. 64, no. 3, pp. 324 – 328, 2017.

[19] G. I. Wirth, R. da Silva, and B. Kaczer, "Statistical Model for MOSFET Bias Temperature Instability Component Due to Charge Trapping," *IEEE Transactions on Electron Devices*, vol. 58, no. 8, pp. 2743 – 2751, 2011.

[20] J. B. Velamala, K. B. Sutaria, T. Sato, and Y. Cao, "Aging Statistics based on Trapping/Detrapping: Silicon Evidence, Modeling and Long-term Prediction," *IEEE International Reliability Physics Symposium*, pp. 2F.2.1 – 2F.2.5, 2012.

[21] G. Ribes, M. Rafik, D. Roy, and J. M. Roux, "Reliability Issues for Nano-scale CMOS Dielectrics: From Transistors to Product Reliability -From SiON to High-K dielectrics," *IEEE International Conference on Integrated Circuit Design and Technology*, pp. 91 – 96, 2008.

[22] D. S. Ang, Z. Q. Teo, T. J. J. Ho, and C. M. Ng, "Reassessing the Mechanisms of Negative-Bias Temperature Instability by Repetitive Stress/Relaxation Experiments," *IEEE Transactions on Device and Materials Reliability*, vol. 11, no. 1, pp. 19 – 34, 2011.

[23] T. Wang, T. E. Chang, L. P. Chiang, C. H. Wang, N. K. Zous, and C. Huang, "Investigation of Oxide Charge Trapping and Detrapping in a MOSFET by Using a GIDL Current Technique," *IEEE Transactions on Electron Devices*, vol. 45, no. 7, pp. 1511 – 1517, 1998.

[24] C. Hu, S. C. Tam, F. C. Hsu, P. K. Ko, T. Y. Chan, and K. W. Terrill, "Hot-Electron-Induced IMOSFET Degradation - Model, Monitor, and Improvement," *IEEE Transactions on Electron Devices*, vol. 32, no. 2, pp. 375 – 385, 1985.

[25] A. S. Goda and G. Kapila, "Design for Degradation: CAD Tools for Managing Transistor Degradation Mechanisms," *International Symposium on Quality Electronic Design*, pp. 416 – 420, 2005.

[26] W. Wang, V. Reddy, A. T. Krishnan, R. Vattikonda, S. Krishnan, and Y. Cao, "Compact Modeling and Simulation of Circuit Reliability for 65-nm CMOS Technology," *IEEE Transactions on Device and Materials Reliability*, vol. 7, no. 4, pp. 509 – 517, 2007.

[27] A. Chatterjee, J. Yoon, S. Zhao, S. Tang, and K. S. *et al.*, "A 65-nm CMOS technology for mobile and digital signal processing applications," *IEEE International Electron Devices Meeting*, pp. 665 – 668, 2004.

[28] W. Wang, "Circuit Aging in Scaled CMOS Design : Modelling, Simulation and Prediction," 2008.

[29] S. Khan and S. Hamdioui, "Modeling and Mitigating NBTI in Nanoscale Circuits," *IEEE International On-Line Testing Symposium*, pp. 1 – 6, 2011.

[30] S. Khan, S. Hamdioui, H. Kukner, P. Raghavan, and F. Catthoor, "BTI Impact on Logical Gates in Nano-scale CMOS Technology," *IEEE International Symposium on Design and Diagnostics of Electronic Circuits Systems (DDECS)*, pp. 348 – 353, 2012.

[31] S. Khan and S. Hamdioui, "Temperature Dependence of NBTI Induced Delay," *IEEE International On-Line Testing Symposium*, pp. 15 – 20, 2010.

[32] S. Khan, N. Haron, S. Hamdioui, and F. Catthoor, "NBTI Monitoring and Design for Reliability in Nanoscale Circuits," *IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems*, pp. 68 – 76, 2011.

[33] M. A. Alam, "A Critical Examination of the Mechanics of Dynamic NBTI for PMOSFETs," *IEEE International Electron Devices Meeting*, pp. 14.4.1 – 14.4.4, 2003.

[34] A. T. Krishnan, C. Chancellor, S. Chakravarthi, P. E. Nicollian, V. Reddy, A. Varghese, R. B. Khamankar, and S. Krishnan, "Material Dependence of Hydrogen Diffusion: Implications for NBTI Degradation," *IEEE International Electron Devices Meeting*, pp. 684 – 691, 2005.

[35] S. Mahapatra, P. B. Kumar, and M. A. Alam, "Investigation and Modeling of Interface and Bulk Trap Generation during Negative Bias Temperature Instability of p-MOSFETs," *IEEE Transactions on Electron Devices*, vol. 51, no. 9, pp. 1371 – 1379, 2004.

[36] "65nm CMOS Technology, CS200 / CS200A Datasheet." [Online]. Available: http://www.fujitsu.com/cn/Images/65nm_CMOS.pdf

[37] T. Sakurai and A. R. Newton, "Alpha-Power Law Mosfet Model and its Application to CMOS Inverter Delay and Other Formulas," *IEEE Journal on Solid-State Circuits*, vol. 25, no. 2, p. 584 594, 1990.

[38] J. M. Rabaey, A. Chandrakasan, and B. Nikolic, "Digital Integrated Circuits - A Design Perspective - 2nd Edition," 1995.

[39] H. A. Balef, H. Jiao, J. P. de Gyvez, and K. Goossens, "An Analytical Model for Interdependent Setup/Hold-time Characterization of Flip-flops," *International Symposium on Quality Electronic Design*, p. 209 214, 2017.

[40] D. Ernst, N. S. Kim, S. Das, S. Pant, R. Rao, and t. T. Pham, "Razor: A Low-Power Pipeline Based on Circuit-Level Timing Speculation," *IEEE/ACM International Symposium on Microarchitecture*, p. 7 18, 2003.

[41] Sai, Gaole, Halak, Basel, Zwolinski, and Mark, "Multi-Path Ageing Sensor for Cost-efficient Delay-Fault Prediction," *Workshop on Early Reliability Modeling for Aging and Variability in Silicon Systems*, 2016.

[42] "Reliability Simulation in Integrated Circuit Design." [Online]. Available: http://www.cadence.com/

[43] D. Lorenz, "Aging Analysis of Digital Integrated Circuits," 2012.

[44] M. Karam, W. Fikry, H. Haddara, and H. Ragai, "Implementation of hot-carrier reliability simulation in ELDO," *IEEE International Symposium on Circuits and Systems*, vol. 5, pp. 515 – 518, 2001.

[45] L. C. Chen, S. K. Gupta, and M. A. Breuer, "A New Gate Delay Model for Simultaneous Switching and its Applications," *ACM/IEEE Design Automation Conference*, pp. 289 – 294, 2001.