UNIVERSITY OF SOUTHAMPTON

Cost-efficient Delay-Fault Sensors for Ageing Prediction

by

Gaole Sai

A thesis submitted for the
degree of Doctor of Philosophy

in the
Faculty of Physical Science and Engineering
School of Electronics and Computer Science

November 2018

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF PHYSICAL SCIENCE AND ENGINEERING
SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

by Gaole Sai

Aggressive technology scaling has accelerated the ageing of CMOS devices. Ageing refers to a slow progressive degradation in the performance of MOS transistors. Consequently, the speed of a chip can significantly degrade over time; this results in delay faults. Dynamic reliability management schemes have been proposed to assure an IC's lifetime reliability. Such schemes are typically based on the use of ageing sensors to predict a circuit's failure before the actual errors appear. Existing ageing sensors are usually placed on the circuit's longest delay paths, which are deemed to be the most vulnerable to delay faults. Such an approach is very costly and may be infeasible in today's complex designs that typically have a large number of long delay paths that need to be monitored. Existing ageing models are proposed to estimate the lifetime of an IC before it's fabrication. However, The result is inaccurate without considering the actual operating conditions of the circuit. Various ageing mitigation technique has been proposed to extend the lifetime of an IC. A trade-off between lifetime and performance usually achieves such approaches. Such sacrifices are reluctant.

We propose two sensors, Parity Check Circuit (PCC) and Differential Multiple Error Detection Sensor (DMEDS), for cost-efficient delay fault monitoring. Both of those two sensors can monitor multiple paths simultaneously, which reduces the number of sensors significantly. The PCC has been designed and verified in a 65nm technology. Our results indicate that using the proposed sensor for delay fault monitoring in a 32-bit MIPS can lead to a significant saving in the area and power overheads, compared to the use of canary flip-flops [40]: by two-thirds and one-third, respectively. The DMEDS has been designed at transistor level in a 32 nm and 90 nm CMOS technology and verified at the system level. Our results indicate that the use of the proposed sensor for delay fault monitoring across ten paths can lead to a significant saving in area overhead compared to Razor [29], and Canary [40]: 87.59%, 77.67%, respectively. We also propose an idea of lifetime prediction system, and it compares the data with reference to analysis the ageing of the device. Our results indicate that the use of the proposed system for lifetime prediction system can accurately estimate the lifetime of the IC compared with the data from the ageing model. The error is controlled within 5% with limited reference data.

Contents

Acknowledgements **xv**

1 Introduction **1**

1.1 Motivations For Research 1

1.2 Delay Fault Monitoring and PVTA Variations 2

1.2.1 Delay Fault Monitoring Techniques 2

1.2.2 PVT Variations 3

1.2.3 Ageing Variation 3

1.2.4 The limitation of state-of-the-art techniques 4

1.3 Objectives 4

1.4 Contributions 5

1.5 Thesis Structure 5

1.6 Publications 6

2 Background **7**

2.0.1 Static Timing Analysis and Delay Fault 7

2.1 Process, Voltage and Temperature Variations 8

2.1.1 Process Variations 10

2.1.2 Voltage Variations 10

2.1.3 Temperature Variation 12

2.1.4 Discussion 12

2.2 Negative-Bias Temperature Instability and Ageing Mechanisms 13

2.2.1 Negative-Bias Temperature Instability 14

2.2.2 Influence Factors of NBTI 17

2.2.3 HCI and PBTI 22

2.3 Ageing Mitigation and Lifetime Prediction Techniques 23

2.3.1 Ageing Mitigation 23

2.3.2 Lifetime Prediction 27

2.4 Ageing Sensors 28

2.4.1 Razor FF 28

2.4.2 Razor II 31

2.4.3 Canary FF 32

2.4.4 TDS FF 32

2.5 Concluding Remarks 35

3 Parity Check Circuitry **37**

3.1 Introduction 37

3.2 PCC Design Principles . . . 38
3.3 Path Selection for PCC . . . 43
3.3.1 Path Selection . . . 43
3.3.2 Potential Critical Path Identification . . . 45
3.4 PCC Implementation . . . 48
3.5 Verification and Comparative Analysis . . . 51
3.5.1 System level Simulation Results . . . 51
3.5.2 Area and Power Overheads Comparison . . . 53
3.6 Concluding Remarks . . . 55

4 Differential Multiple Error Detection Sensor 57
4.1 Introduction . . . 57
4.2 DMEDS Design Principles . . . 58
4.3 Operating Principles of SEDS . . . 59
4.4 Operating Principles of DMEDS . . . 59
4.5 Transistor Level Design of DMEDS . . . 63
4.6 Gate Level Design of DMEDS . . . 66
4.7 DMEDS Implementation . . . 67
4.8 Verification and Comparative Analysis . . . 68
4.8.1 Transistor Level Verification and Analysis in 90 nm . . . 69
4.8.1.1 Transistor Level Simulation Results in 90 nm . . . 69
4.8.1.2 Transistor Level Cost Comparison in 90 nm . . . 72
4.8.2 System Level Verification and Analysis in 90 nm . . . 73
4.8.2.1 DMEDS Synthesis and Simulation . . . 73
4.8.2.2 Comparison of Power and Area Overhead . . . 75
4.8.3 Transistor Level verification and cost Comparison in 32 nm . . . 77
4.8.3.1 Transistor Level verification in 32 nm . . . 77
4.8.3.2 Cost Comparison in 32 nm . . . 78
4.8.4 System Level verification in 32 nm . . . 79
4.9 Concluding Remarks . . . 79

5 DVS for Ageing Prediction 83
5.1 Introduction . . . 83
5.2 Ageing Prediction System Design Principle . . . 85
5.3 Reference Circuit Estimation . . . 91
5.3.1 Define the Margin . . . 92
5.3.2 Define the Characteristic of DVS . . . 97
5.3.3 Circuit Ageing Delay Estimation . . . 98
5.3.4 Reference Circuit Ageing Behaviour Estimation . . . 101
5.3.5 Discussion . . . 107
5.4 Verification and Analysis . . . 107
5.4.1 Lifetime Estimation and Comparison in Details . . . 107
5.4.2 Lifetime Estimation and Comparison with More than One Reference . . . 111
5.5 Concluding Remarks . . . 111

6 Conclusions 113

6.1 Conclusions 113
6.2 Future work 114

References **117**

List of Figures

2.1 The Architecture of a Pipelined MIPS, Reproduced from [48] 8

2.2 Margins, Reproduced from [57] 8

2.3 Technology node of the CPUs from Intel, Reproduced from [45, 61] 9

2.4 V_{th} Shift Random Due to the Process Variations, Reproduced from [52] . 10

2.5 Relation Between Supply Voltage and Logic Delay of a Multiplier Circuit in 65nm Technology, Reproduced from [85] 11

2.6 Within Die Temperature Variation, Reproduced from [14] 12

2.7 NBTI: Hydrogen Atoms Diffusion, Reproduced from [77] 14

2.8 Static NBTI, Reproduced from [10] 15

2.9 Dynamic NBTI, Reproduced from [4] 16

2.10 Dynamic NBTI for a minimum sized PMOS transistor of 90nm technology, Reproduced from [10] 17

2.11 Long Term Dynamic NBTI with Different Duty Cycle 18

2.12 Long Term Dynamic NBTI with Difference Temperature, Reproduced from [3] 19

2.13 Long Term Dynamic NBTI with Different Vgs, Reproduced from [62] . . . 20

2.14 Relationship between Duty Cycle, VDD and ΔV_{th}, Reproduced from [10] 21

2.15 NBTI: Hydrogen Atoms Diffusion, Reproduced from [77] 22

2.16 Ageing Mechanisms, Reproduced from [25, 90] 23

2.17 NBTI-aware Power Gating, Reproduced from [27] 24

2.18 The Frequency f vs. Supply Voltage VDD Shmoo Plot of the Traditional DVS Technique, Reproduced from [44] 25

2.19 Ageing mitigation by using DVS, Reproduced from [24] 26

2.20 Supply Voltage Scaling Flowchart, Reproduced from [17] 27

2.21 Lifetime Prediction Flow Chart, Reproduced from [22, 34] 29

2.22 The Razor FF and Timing Diagram, Reproduced from [35] 30

2.23 Razor II and Timing Diagram, Reproduced from [30] 31

2.24 The Canary FF and Timing Diagram, Reproduced from [40] 33

2.25 TDS FF, Reproduced from [71] 34

2.26 TDS FF Timing Diagram, Reproduced from [71] 34

3.1 Architecture of The Proposed Circuitry 39

3.2 Odd Number of Transitions 40

3.3 Even Number of Transitions 40

3.4 Unpredictable Error 41

3.5 PCC Implementation Process 42

3.6 Percentage of Transitions 44

3.7 Determining the Width of Detection Window 45

3.8 Potential BTI Critical Paths Estimation Process 47
3.9 Synthesised 4-bit signed full adder with a loose timing constraint 48
3.10 Synthesised 4-bit signed full adder with a tight timing constraint 48
3.11 Delay at Time Zero (Normal Process Variation) 49
3.12 Delay After 10 Years (Normal Process Variation) Random Signal Probability 49
3.13 Delay at Time Zero (The Worst Process Variation) 50
3.14 Delay After 10 Years (The Worst Process Variation) 50
3.15 DVS in a MIPS microprocessor with a pipeline 51
3.16 9 Input PCC Block Diagram after Synthesis 51
3.17 4-input PCC functional verification in a 32-bit MIPS 52
3.18 9 Input Canary FF Block Diagram after Synthesis 53
3.19 Area Overhead 54
3.20 Power Overhead 54

4.1 The Architecture and timing diagram of SEDS 60
4.2 The Architecture of DMEDS 61
4.3 MDU Timing Diagram 61
4.4 DMEDS Timing Diagram 62
4.5 The Adjustable Delay Element 64
4.6 The Stability Checker 64
4.7 6 Transistor XOR Gate 65
4.8 Previously Design of Stability Checker[66] 65
4.9 Stability Checker Timing Diagram 66
4.10 Gate Level DMEDS 66
4.11 Gate Level Transition Detection Timing Diagram 67
4.12 Ageing Prediction System in a MIPS microprocessor with a pipeline . . . 67
4.13 Timing Diagram of Buffer 68
4.14 Metastability Resistance Simulation result in Spice 69
4.15 Transistor Level Multiple Timing Error Detection Simulation result in Spice 70
4.16 Transistor Level Worst Case Timing Error Prediction Simulation result in Spice 71
4.17 Timing Report of the EX Stage in a 32-bit Pipelined MIPS 72
4.18 DMEDS Block Diagram after Synthesis 73
4.19 Verilog Simulation Result of Timing Error Prediction in Modelsim 74
4.20 Verilog Simulation Result of Worst Case Timing Error Prediction in Modelsim 75
4.21 Canary FF Block Diagram after Synthesis 75
4.22 Timing Report of the 32 bit MIPS 77
4.23 Multiple Timing Error Detection 78
4.24 Equivalent Model of DMEDS and Simulation Result 80
4.25 10-input DMEDS functionality verification in a 32-bit MIPS 80

5.1 The Average Supply Voltage of a Circuit with DVS Over the Years 85
5.2 Voltage Scaling 86
5.3 Relationship Between Circuit's Operating Time and Supply Voltage . . . 90
5.4 Lifetime Estimation Flow Chart 91

5.5 Reference Circuit Estimation Flow Chart 92
5.6 The worst and best case process variations of an inverter chain of 90 nm technology 93
5.7 The random process variations distribution for an inverter chain of 90 nm technology sample size (1000) 93
5.8 The Relation Between Supply Voltage and Delay 94
5.9 The Relation Between Temperature and Delay 95
5.10 The Relationship between Temperature Variation, Voltage Variation and Delay Degradation at time zero 96
5.11 The margin for inverter Chain 97
5.12 Long Term Dynamic NBTI with Different Duty Cycle 99
5.13 Long Term Dynamic NBTI with Different Temperature 100
5.14 Long Term Dynamic NBTI with Different Vgs 101
5.15 Circuit Ageing Behaviour Estimation Flow Chart 102
5.16 Circuit Lifetime Estimation When T=50°C 103
5.17 Lifetime Estimation of a Circuit With DVS When T=50°C 104
5.18 Average Vdd During the Circuit's Lifetime When T=50°C 105
5.19 Maximum Vdd During the Circuit's Lifetime When T=50°C 106
5.20 Circuit Lifetime Estimation when T=45°C 108
5.21 Lifetime Estimation of a Circuit with DVS when T=45°C 109
5.22 Average Vdd During the Circuit's Lifetime When T=45°C 110
5.23 Maximum Vdd During the Circuit's Lifetime When T=45°C 110

List of Tables

2.1 NBTI Vth Degradation Comparison in Different Operating Temperatures 19
2.2 Estimated Threshold Voltage and Delay Degradation of an Inverter 21

3.1 Percentage of Transitions from 4 paths 39

4.1 Truth Table of the RS Latch 66
4.2 Comparison with Razor and Canary FF 73
4.3 Comparison with Canary FF 76
4.4 Comparison with Canary FF in a 32 bit MIPS 76
4.5 Comparison with other designs 79

5.1 Delay Degradation Comparison in Different Temperatures 95
5.2 The Best and Worst Intrinsic Delay Degradation of a Long Delay Path . . 96
5.3 Intrinsic Delay with Different Vdd Under the Worst PVT variations . . . 97
5.4 NBTI Vth Degradation Comparison in Different Operating Temperatures 100
5.5 Weighted Average of Power Saving When Average T = 50°C 106
5.6 The Ratio of Average VDD $T_{avg} = 50°C$ 107
5.7 Weighted Average of Power Saving When Average T = 45°C 108
5.8 Lifetime Estimation When T=45°C 109
5.9 Reference Data When T= 55°C and 50°C 111
5.10 Target Data When T= 51°C - 54°C 112

Research Thesis: Declaration of Authorship

Print name:	Gaole Sai

Title of thesis:	Cost-efficient Delay-Fault Sensors for Ageing Prediction

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:

Sai G, Halak B, Zwolinski M. A cost-efficient delay-fault monitor[C]//Circuits and Systems (ISCAS), 2017 IEEE International Symposium on. IEEE, 2017: 1-4.

Sai G, Halak B, Zwolinski M. Multi-Path Aging Sensor for Cost-Efficient Delay Fault Prediction [J]. IEEE Transactions on Circuits and Systems II: Express Briefs, 2018, 65(4): 491-495.

Signature:		Date:	

Acknowledgements

I would like to thank my supervisor, Professor Mark Zwolinski, for his continuous support and guidance in This work. It has been an honour to be his PhD student. His patient and vision in approaching a problem is a great thing that I have learnt during my PhD. I am grateful all this contribution of ideas and time to make my PhD experience fruitful and inspiring.

I would like to acknowledge my co-supervisor, Dr Basel Halak, for his advice and valuable reviews on this research. I immensely appreciate that I had a chance to work with him.

I would also like to take this opportunity to thank my parents and my partner for their continuous support and encouragement over the years.

Finally, I would like to thank all my colleagues who have helped me during my PhD.

Chapter 1

Introduction

1.1 Motivations For Research

This research is motivated by the challenges arising from the reliability issues in modern technology node. The feature size of the technology node has been shrunk by more than 700 times, in the last five decades. The technology scaling results in a significant achievement regarding performance, power-efficiency and device density of the semiconductor devices. However, the side effect leads to the increasing of reliability issues. The Process, Voltage, Temperature variations and Ageing-induced (PVTA) device degradation are becoming major reliability concerns in modern semiconductor technologies. Those phenomena lead to performance degradation, and hence a timing error. To avoid delay fault induced failures, the integrated circuits are typically designed with large safety margins. This generally means a circuit is designed for worst-case operating conditions. Such an approach may limit system performance and lead to an increase in power consumption.

Dynamic reliability management schemes have been proposed to assure an IC's lifetime reliability. Such schemes are typically based on the use of sensors to predict circuit failures before the actual errors appear. The system can then adaptively scale its operating frequency and supply voltage according to the actual operating conditions to compensate for performance degradation. Various in-situ delay monitoring sensors have been proposed. These include delay fault detection and prediction techniques. Existing delay fault sensors are usually placed on the circuit's longest delay paths. However, the increasing complexity of ICs has led to a significant rise in the number of long paths and potential ageing-critical paths that may be vulnerable to timing errors. This means the cost of in-situ delay monitoring may be prohibitive.

Many research groups have been working on ageing models for years to estimate the lifetime of an IC before it's fabrication. However, the performance and lifetime degradation from ageing effect is depending on the circuit's operating temperature, supply voltage and stress workload. Each chip might be running in different workload and environment. Therefore the environment and workload of an IC are unpredictable in the design stage of the CMOS circuits. The lifetime prediction of each chip is when it is predicted by using such models. On the other hand, the intrinsic delay of each chip is different due to the process variations delay degradation. This means the ageing prediction approaches are becoming unattainable. On the other hand, various ageing mitigation technique has been proposed to extend the lifetime of an IC. A trade-off between lifetime and performance usually achieves such approaches. A faster operation speed always the pursuit of people. Therefore, performance is one of the primary concern of the IC design nowadays. The performance is the one of primary concern of the IC design nowadays. Such sacrifices are reluctant.

1.2 Delay Fault Monitoring and PVTA Variations

This Chapter briefly introduces the preliminary about delay fault monitoring and PVTA variations, and Section 1.2.1 introduces the delay fault monitoring techniques, Section 1.2.2 outlines PVT variations, Section 1.2.3 explains the PVT variations and 1.2.4 summarises the limitation of state-of-the-art techniques.

1.2.1 Delay Fault Monitoring Techniques

Delay fault occurs when a late signal arrives and crossed timing constraint in a sequential logic circuit. This may result in the failure of the circuit. Various in-situ delay monitoring sensors have been proposed. These include delay fault detection and prediction techniques. The delay fault detection sensors check data consistency after the rising clock edge and delay prediction sensors check data consistency after the rising clock edge. The in-situ delay monitoring sensors usually work with the Dynamic Voltage and Frequency Scaling (DVFS) system. The sensor triggers an error signal when a late arriving signal is captured. It then sends the error signal to the DVFS system. Then the DVFS will decide to scale up the supply voltage or scale down the operating frequency. However, existing sensors usually detect/predicts the delay fault from one path, with the increasing complexity of ICs has led to a significant rise in the number of long paths and potential ageing-critical paths, cost of in-situ delay monitoring has become prohibitive. Moreover, existing sensors usually replace the flip-flop of the original design; therefore it is hard to implement, and it influences the functionality of the original design. Moreover, flip-flop type sensors are usually suffering from the metastability issue which affects the reliability and power dissipation of the digital circuit.

1.2.2 PVT Variations

The process, voltage, and temperature (PVT) variations have negative impacts on the performance of a manufactured chip. The effect of PVT variations has been increased due to aggressive fabrication technology scaling. Process variations are occurred due to the variations in channel length or doping concentration which causes the threshold voltage shift at time zero it, therefore, affects the performance of the CMOS devices. Voltage variations are considered as there are unexpected voltage drops in the power supply networks and also the variations in the supply voltage itself. Temperature variations are raised due to temperature fluctuations of the CMOS circuits and the environmental impacts. The PVT-induced leads to delay variations which affect the propagation delays the synchronous circuits it, therefore, cause the timing errors. The PVT variations affect both clock signals and data paths. Hence the longest delay paths may fail to deliver its output data within a given clock period.

1.2.3 Ageing Variation

Negative Bias Temperature Instability (NBTI) is considered one of the most critical ageing reliability issues in PMOS transistors. NBTI occurs because interface traps at the gate oxide interface are generated when a negative gate bias – a negative potential difference between gate and source, is applied. It manifests itself as an increase in the threshold voltage (Vth) and a decrease in the drain current, thereby degrading timing. NBTI can cause more than 20% timing degradation in the worst-case operating conditions, which reduces the lifetime of an IC.

Positive BTI (PBTI) is an ageing reliability issue in NMOS transistors. The threshold voltage shift of PBTI has become significant in High-K Metal Gate technologies. The double effect of NBTI and PBTI exacerbates the timing degradation of ICs.

Hot-Carrier Injection (HCI) is a critical ageing issue in NMOS. HCI phenomenon occurs because the interface traps are generated at the drain end gate oxide interface when the current flow between the source and drain during the switching time. It manifests as the degradation of the switching characteristics, threshold voltage, drain-current and noise margin. Even worse, the HCI effect is unrecoverable. Therefore, although the HCI effect will not cause significant delay degradation like the NBTI effect, however, it accelerates the degradation of the device lifetime directly.

Existing ageing models estimate the lifetime of an IC before its fabrication without considering the actual operating conditions. The estimation result will be a range of lifetime which is not accurate.

1.2.4 The limitation of state-of-the-art techniques

The shortage of excising state-of-the-art techniques are summarised as follows:

1. Technology scaling arises the reliability issues of IC by the PVTA variations.
2. The cost of existing sensors implementation is very high, and the implementation usually replaces the original Flip Flop of the design as a sensor, which brings more reliability issues of ICs. The sensors are suffering from the metastability, and usually not suitable for ageing prediction.
3. Existing lifetime prediction techniques are either unreliable or not cost-efficient, and the state-of-the-art ageing mitigation techniques are usually sacrifices the performance of an IC.

The objectives of this research are raised by the limitation of state-of-the-art techniques, shown in the next Section.

1.3 Objectives

There are several research questions which are summarised by state-of-the-art techniques introduced in Section 1.2.

1. How do the PVTA variations impact the CMOS circuits, and what the PVTA variations arise reliability issues?
2. Is there a solution for cost-efficient delay-fault monitoring without replacing the original Flip Flop of the design, resistance to the metastability, and suitable for ageing prediction?
3. Can we propose a reliable lifetime prediction algorithm with the consideration of real operating conditions and also suitable for ageing mitigation without sacrifice the performance?

The first question is answered by the background study for this research shown in Chapter 2. The second question is answered in Chapter 3 and 4. Two cost-efficient sensors are proposed for delay-fault monitoring, Parity Check Circuit (PCC) and Differential Multiple Error Detection Sensor (DMEDS). Both sensors can monitor two or more paths synchronously and monitor the delay fault without replacing the original Flip Flop of the Design. Furthermore, the DMEDS is resistance to the metastability. The last question is answered in Chapter 5. An ageing prediction algorithm is proposed for real-time lifetime prediction. The ageing prediction algorithm predicts the lifetime of an IC by using the limited information which is provided by the sensors (PCC or DMEDS). Together with the Dynamic Voltage Scaling (DVS), the system can predict and extend the lifetime simultaneously. The contributions of this research are shown in the next Section.

1.4 Contributions

In Chapter 3.1, we propose a cost-efficient Parity Check Circuit (PCC) for delay fault prediction to mitigate the cost of in situ delay monitoring. PCC can monitor multiple paths simultaneously, which significantly reduces the number of sensors. The proposed sensor has been designed and verified in a 65nm technology. Our results indicate that using the proposed sensor for delay fault monitoring in a 32-bit MIPS can lead to a significant saving in the area and power overheads, compared to the use of canary flip-flops [40]: by two-thirds and one-third, respectively.

In Chapter 4, we propose a new Differential Multiple Error Detection Sensor (DMEDS) for timing errors. DMEDS can monitor multiple paths simultaneously, which significantly reduces the number of sensors needed to monitor ageing-induced delay faults. DMEDS has been designed at transistor level in a 32nm 90 nm CMOS technology and verified at the system level. Our results indicate that the use of the proposed sensor for delay fault monitoring across 10 paths can lead to a significant saving in area overhead compared to Razor [29], and Canary [40]: 87.59%, 77.67%, respectively.

At the end of this thesis, we propose an idea of lifetime prediction system, and the ageing prediction controller receives the error signal from the ageing prediction sensors such as DMEDS and PCC, it then compares the data with reference to analysis the ageing of the device. The ageing data of the reference comes from the existing ageing model. Our results indicate that the use of the proposed system for lifetime prediction system can accurately estimate the lifetime of the IC compared with the data from the ageing model. The error is controlled within 5% with limited reference data.

1.5 Thesis Structure

The rest of the thesis is organized as follows. Chapter 2 briefly describes related work, the Dynamic Voltage Scaling (DVS) system; ageing mechanisms; and existing sensors. Chapter 3 outlines the design of new ageing predicting sensor PCC. Chapter 4 outlines the design of new ageing predicting sensor DMEDS. A new lifetime predicting model is outlined in chapter 5. Finally, conclusions and future works are drawn in chapter 6.

1.6 Publications

The contributions presented in this thesis have been published in the flowing papers:

1. Sai, Gaole, Basel Halak, and Mark Zwolinski. ”A cost-efficient delay-fault monitor.” Circuits and Systems (ISCAS), 2017 IEEE International Symposium on. IEEE, 2017.
2. Sai, Gaole, Basel Halak, and Mark Zwolinski. ”Multi-Path Aging Sensor for Cost-Efficient Delay Fault Prediction.” IEEE Transactions on Circuits and Systems II: Express Briefs 65.4 (2018): 491-495.

Chapter 2

Background

This chapter gives an extensive overview of state-of-the-art existing technologies which are related to the research of this thesis. It introduces the relevant background of Chapter 3, 4 and 5 respectively.

2.0.1 Static Timing Analysis and Delay Fault

Static timing analysis (STA) is used to predict performance (the clock speed) of a fabricated digital circuit. It is widely used in high-performance circuit design, [51]. The critical path (the longest delay path) is defined as the path between an input and an output with the maximum delay after STA. Figure 2.1 shows the architecture of a basic pipelined MIPS. The pipelined divides the processor to several to improve the performance of the processor. The traditional processor runs the instructions one by one, and it usually takes a few clock cycles to run each instruction. The pipelined design is able to run more than one instruction simultaneously. The MIPS, as shown in 2.1, for example, is able to run 5 instruction at the same clock cycle. The delay fault occurs when a late signal arrives and crossed timing constraint in a sequential logic circuit. This may result in the failure of the circuit, [38, 70] .

The critical path of the design, as shown in Figure 2.1 is the longest delay path from those state, IF, ID EX MEM and WB after. The minimum clock cycle should be higher than the delay from the critical path; otherwise, it may result in a delay fault. The technology scaling results in a significant achievement regarding performance, power-efficiency and device density of the semiconductor devices. However, the side effect leads to the increasing of reliability issues. The Process, Voltage, Temperature variations and Aging-induced (PVTA) device degradation are becoming major reliability concerns in modern semiconductor technologies. Those phenomena lead to performance degradation, and hence a timing error. To avoid failures caused by delay fault, the integrated circuits are

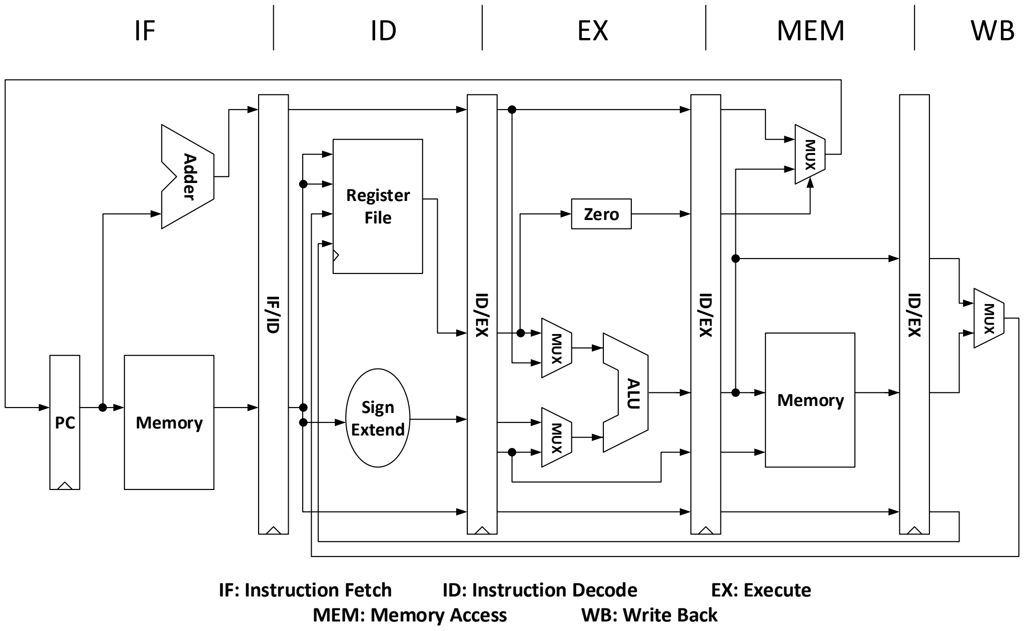


Figure 2.1: The Architecture of a Pipelined MIPS, Reproduced from [48]

typically designed with large safety margins. This generally means a circuit is designed for worst-case operating conditions, as shown in Figure 2.2.

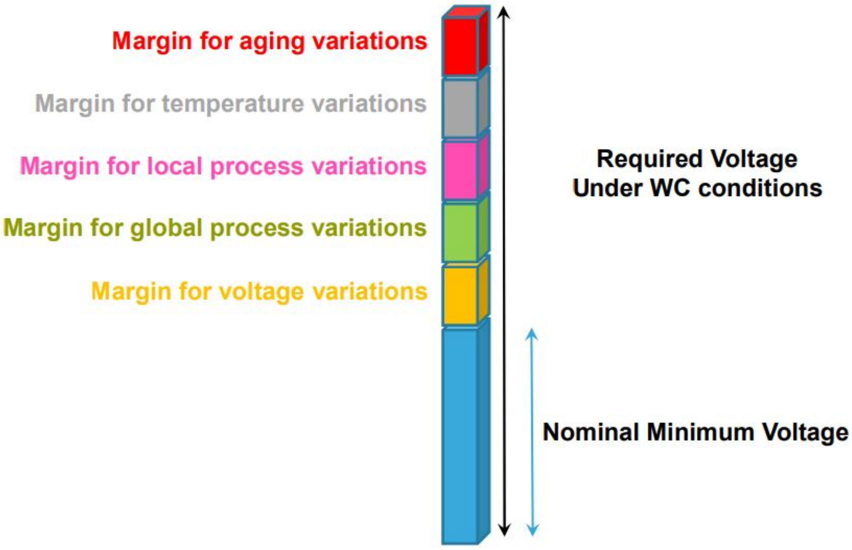

Margin for aging variations
Margin for temperature variations
Margin for local process variations
Margin for global process variations
Margin for voltage variations
Required Voltage
Under WC conditions
Nominal Minimum Voltage

Figure 2.2: Margins, Reproduced from [57]

2.1 Process, Voltage and Temperature Variations

The process, voltage, and temperature (PVT) variations have negative impacts on the performance of a manufactured chip. The effect of PVT variations has been increased due to aggressive fabrication technology scaling. Figure 2.3 shows the technology node of the CPUs from Intel over the last 46 years. As the Figure shows, the technology has been shrunk 714 times over the years.

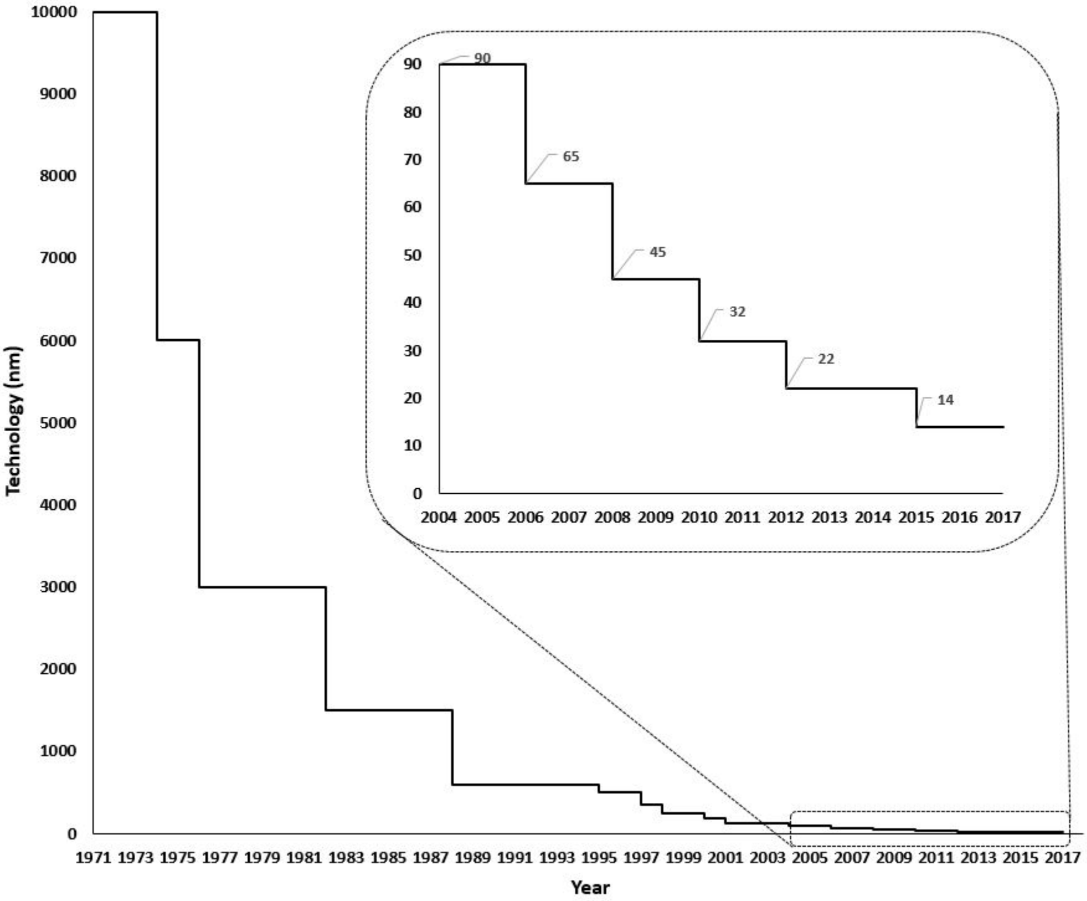


FIGURE 2.3: Technology node of the CPUs from Intel, Reproduced from [45, 61]

Process variations are occurred due to the variations in channel length or doping concentration which cause the threshold voltage shift at time zero it, therefore, degrades the performance of the CMOS devices. Voltage variations are considered as there are unexpected voltage drops in the power supply networks and also the variations in the supply voltage itself. Temperature variations are raised due to temperature fluctuations of the CMOS circuits and the environmental impacts. The PVT variations lead to the delay degradation which affect the propagation delays the synchronous circuits it, therefore, cause the timing errors. The PVT variations affect both clock signals and data paths. Hence the longest delay paths may fail to deliver its output data within a given clock period. Section 2.1.1, 2.1.2 and 2.1.3 describes the process, voltage and temperature variations respectively. Section 2.1.4 concludes performance degradation due to the PVT variations.

2.1.1 Process Variations

Process variation is a naturally occurring variation in the semiconductor circuit. It is embodied in the variations in length, widths, oxide thickness of transistors during the circuits fabrication. The technology scale saves the area significantly. However, the process variations have become more significant due to the aggressive fabrication technology scaling with the percentage of the variations in length, widths, oxide thickness.

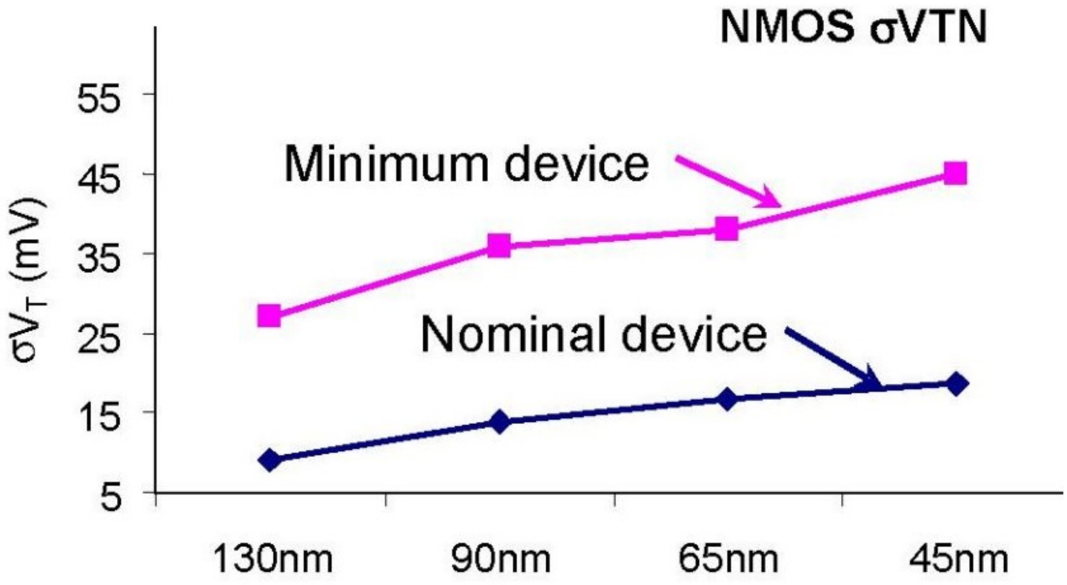


FIGURE 2.4: V_{th} Shift Random Due to the Process Variations, Reproduced from [52]

Same as other critical areas in the semiconductor processing, meeting Moores law scaling with variation challenges require the combining of innovations to improve the strategies, [13]. Figure 2.4 shows the V_{th} shift of NMOS transistor due to process variations over the technology generation. As the Figure shows, the V_{th} shift at 45nm technology node has almost doubled compared with the V_{th} shift at 130nm technology, and the percentage of V_{th} shift is even worst. The value of V_{th} decreases over the technology generation, from 0.34V to 0.29V. This means the technologies scaling has increased the uncertainty of the intrinsic delay after fabrication. The process variation has become a new challenge associated with advanced CMOS technologies for the IC, industry, [13, 53].

2.1.2 Voltage Variations

Voltage variations are mainly occurred by the IR drop and di/dt noise. Both effects will not only decrease the supply voltage but also increase the supply voltage in an IC. The IR drop is caused by current flow over the parasitic resistance of the integrated circuit supply rail. The parasitic inductance, resistance and capacitance of the integrated circuit supply rail and the package causes the di/dt noise, [85]. Therefore, the voltage variations

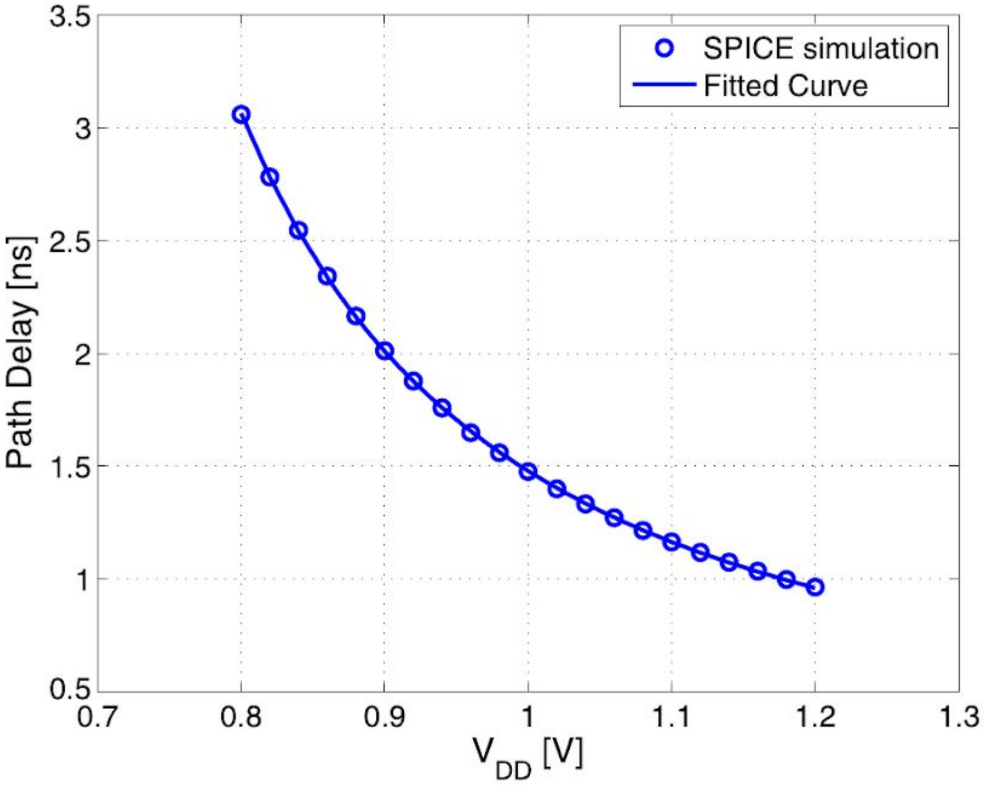


FIGURE 2.5: Relation Between Supply Voltage and Logic Delay of a Multiplier Circuit in 65nm Technology, Reproduced from [85]

follow the Ohm's Law and the Inductance Effect, as shown in equation (2.1) and (2.2). With the aggressive technology scaling of CMOS circuits well into the nanometer regime and the increasing demand for the ultra-low power/low voltage systems, the reliability due to voltage variations become more and more impotent in CMOS circuits and systems, [23, 88].

$$\Delta V_{IR\_Drop} = IR_{parasitic} \tag{2.1}$$

$$\Delta V_{di/dt} = L_{parasitic}\frac{di}{dt} \tag{2.1}$$

$$T_{pd} = \frac{V_{DD}}{b(V_{DD} - V_{th})^{a}}[85] \tag{2.3}$$

Figure 2.5 shows the relation between the supply voltage and logic delay of a multiplier circuit in 65nm Technology. The simulation result is accurately matched with equation (2.1), where V_{DD} is the supply voltage, V_{th} is the threshold voltage, a and b are the specific gate fitting parameters, [85]. As the figure shows, the correlation between the

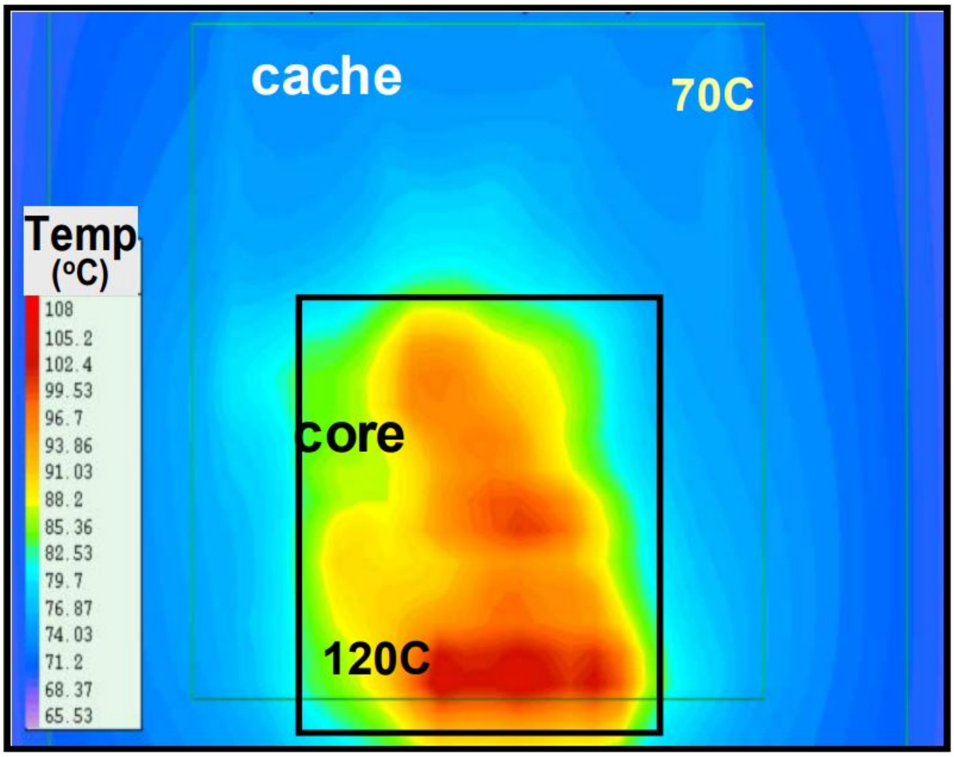


Figure 2.6: Within Die Temperature Variation, Reproduced from [14]

supply voltage and propagation delay are negative. Hence, means the performance of a digital circuit decreases when the voltage goes down.

2.1.3 Temperature Variation

Temperature changes the mobility of transistors. The mobility goes down when the temperature increases, and this causes a decrease in the drain current[9]. It, therefore, increases the propagation delay of the CMOS circuits. Temperature variation has become more and more critical as the increase of power density due to technology scaling, [86]. Figure 2.6 shows the thermal image of a microprocessor. Within-die temperature fluctuations have become a significant performance and packaging challenge for many years. The High temperature causes performance degradation. Moreover, the temperature variation may cause the performance mismatches, therefore, resulting in failures of a system, [14]. As the Figure shows, the internal temperature of a chip can be higher than 100℃.

2.1.4 Discussion

As section 2.1.1, 2.1.2 and 2.1.3 shows, the PVT variations not only reduces the propagation delay but also enhances the propagation delay of a digital circuit. The increase of propagation delay leads to significant performance degradation, especially in the synchronous circuit system. The temperature variation causes the most reduction of the performance compared with others. Furthermore, ageing variations are becoming significant due to the technology shirking, see the next Section.

2.2 Negative-Bias Temperature Instability and Ageing Mechanisms

With aggressive scaling down of the feature size of the IC, the ageing effect has become one of the most important critical issues of MOS devices. Especially in many timing-dependent wear-out mechanisms such as BTI, HCI and TDDB refer to a slow progressive degradation in the performance of MOS transistors. Consequently, the speed of a chip can significantly degrade over time; this results in delay faults.

Many research groups and companies have carried out extensive research on the ageing of MOS devices:

1. The Electronic Systems and Devices research group at the University of Southampton is focusing on the reliability of memory and PUF, ageing monitoring and ageing modelling, [1, 33, 59, 60, 63, 67, 80].
2. The Nano-scale Integration and Modelling research group at the Arizona State University is focusing on the NBTI and HCI modelling, [10, 58, 77].
3. The Device Modelling screech Group research group in the University of Glasgow is focusing on reliability problems of SRAM due to device wear out and process variation, [6, 31].
4. A research group from the University of Algarve is focusing on delay-faults prediction Ageing sensors, [56, 66].
5. A research group from the University of Athens is focusing on delay-faults detection Ageing sensors, [71, 72, 73].
6. Cisco systems ltd and IBM Technology and Qualification are focusing on the lifetime of DRAM under HCI effect, [7, 8, 15, 84].
7. The U.S. Patent and Trademark Office on 11^{th} March 2014 issued Apple a patent describing a method of monitoring the ageing of a device's electronics, [39].

This section describes the specific detail of the NBTI effect and the Reaction-Diffusion (RD) model, which is the most prevalent NBTI model. It then states the main influence factors of NBTI and NBTI mitigation. A brief overview of PBTI and HCI effect due to ageing is stated at the end of this chapter.

Section 2.1.1, 2.1.2 and 2.1.3 describes the process, voltage and temperature variations respectively. Section 2.1.4 concludes performance degradation due to the PVT variations.

2.2.1 Negative-Bias Temperature Instability

Negative-Bias Temperature Instability (NBTI) was discovered in 1966, [65]. However, It is considered as one of the most important reliability issues in integrated circuits during late few years, as a result of technology scaling and the increase of device operating temperature. NBTI is an ageing reliability issue in P-channel Metal-Oxide-Semiconductor field-effect (PMOS) transistors, [11, 20]. NBTI occurs because interface traps at the gate oxide interface are generated when a negative gate bias, a negative potential difference between gate and source (Vgs), is applied, [60]. It manifests itself as an increase in the threshold voltage (Vth) and a decrease in the drain current. The Vth is the minimum voltage to conducting the path between the source and drain of a transistor. Thereby, the shift of the Vth produces delay degradation of CMOS circuits. NBTI can produce more than 20% timing degradation in the worst-case operating conditions[58], which means the performance degradation especially of the synchronous circuit system is becoming more intense.

ReactionDiffusion Model

To estimate the behaviour of an IC due to NBTI effects, many NBTI models have been proposed. The Reaction-Diffusion (R-D) model is the most widely used NBTI model to estimate the Vth degradation due to NBTI, [10, 77].

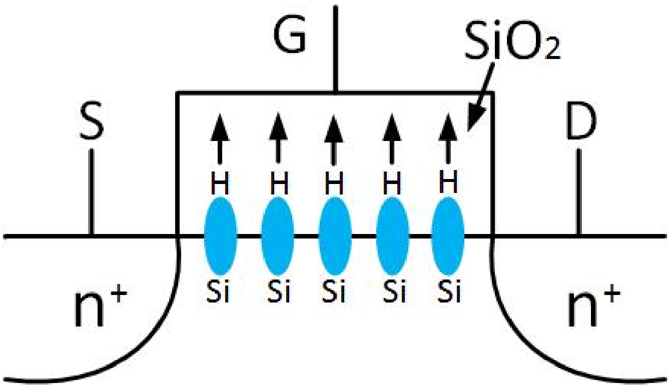


FIGURE 2.7: NBTI: Hydrogen Atoms Diffusion, Reproduced from [77]

As shown in Figure 2.7, the Si-H bonds at the oxide gate interface will be broken when a PMOS transistor is under electrical stress (the reaction phase) by creating positive holes at the gate oxide interface. The hydrogen atoms which are generated by the reaction then diffuse away from the interface to the gate (diffusion). The NBTI effect generates the interface traps uniformly in the channel, [77]. Models for threshold voltage degradation (ΔV_{th}) due to NBTI have been proposed for different levels of abstraction, [10, 33, 77].The behaviour of a CMOS circuit can be predicted by integrating the ΔV_{th} estimation result into circuit simulation tools such as Spice. There are two types of NBTI effect: static NBTI and Dynamic NBTI, shown as follows:

Static and Dynamic NBTI

In this section, The R-D model from one of those research groups is applied to estimate the NBTI effect as shown in Figure 2.10 and Figure 2.8, [10, 58, 77].

A. Static NBTI

Static NBTI is a situation when a PMOS transistor is always under stress (a constant logic 0 input). In this case, the PMOS transistor suffers from the most serious wear out due to the NBTI effect.

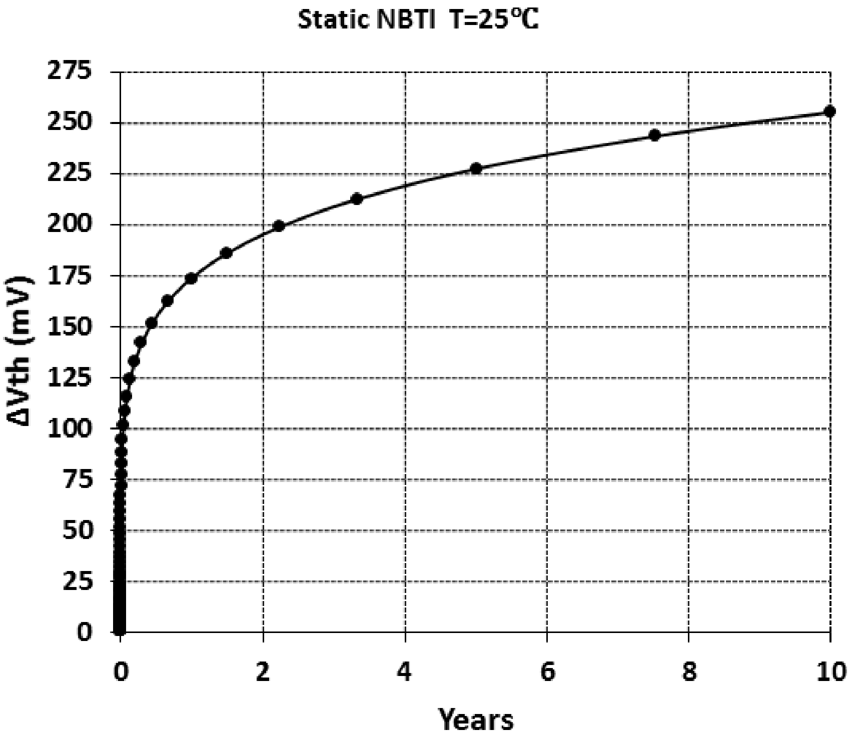


Figure 2.8: Static NBTI, Reproduced from [10]

Figure 2.8 shows the threshold voltage degradation of static NBTI for a minimum sized PMOS transistor of 90nm (Vth at time zero 0.276V) technology in 25 °C and Vgs=1.2V over ten years. As the figure shows, The threshold voltage degradation is increased 255mV after ten years, which is 92.4% threshold voltage shift compared with time zero. This threshold voltage shift leads to significant performance degradation of a circuit. The static become an issue if the when a logic gate has a constant logic 0 input. Such as the data stored in memory.

B. Dynamic NBTI

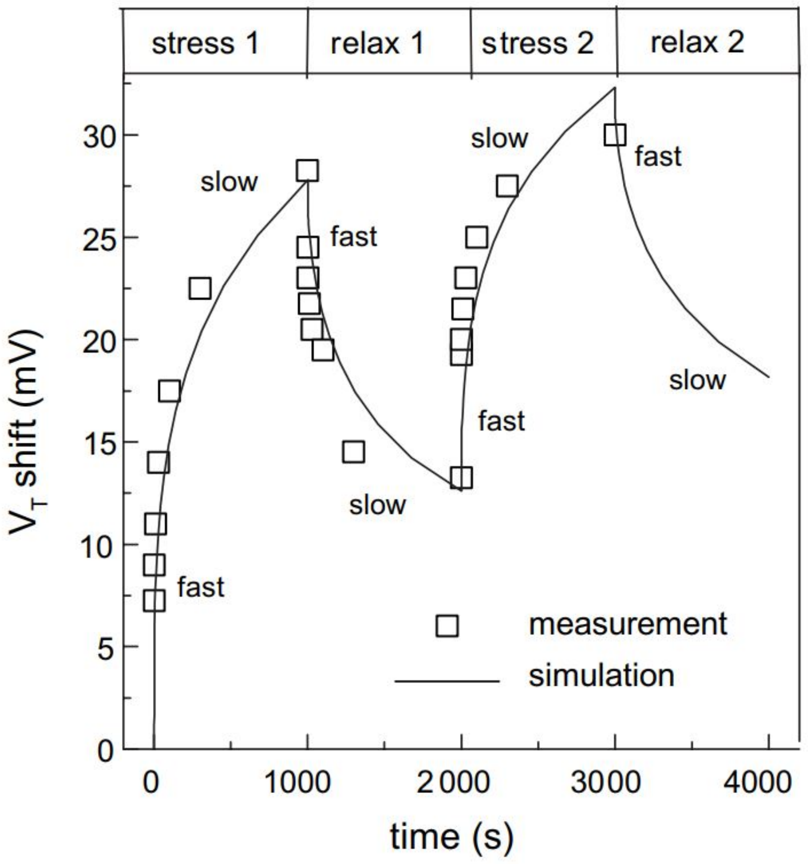


Figure 2.9: Dynamic NBTI, Reproduced from [4]

Unlike the static NBTI, the PMOS transistor will not have a constant input over time in the case of dynamic NBTI. The input of PMOS transistor circulates between stress period (a logic 0 input) and recovery period (a logic 1 input). As the name suggests, Vth increases during the stress period, and it decreases during the recovery period over time. Figure 2.9 shows the threshold voltage degradation during two stress and recovery process. As the threshold voltage shift during the stress time cannot be fully recovered during the relax time, ΔV_{th} increases after each stress and relax period. Therefore, it causes the threshold voltage degradation over a long period.

Figure 2.10 shows the threshold voltage degradation of dynamic NBTI for a minimum sized PMOS transistor of 90nm technology at 0.5 stress duty cycle, 100MHz switching activity, 125 °C and 1.2V Vgs over the years. As the figure shows, the threshold voltage is slowly accumulating during its operating. The Vth shifts about one-third compared with time zero over ten years. In spite of that, the Vth degradation of dynamic NBTI is not as substantial as the Vth degradation of static NBTI; however, it leads to significant performance degradation of a long delay path in complex circuitry(see the delay degradation over time in Chapter 5).

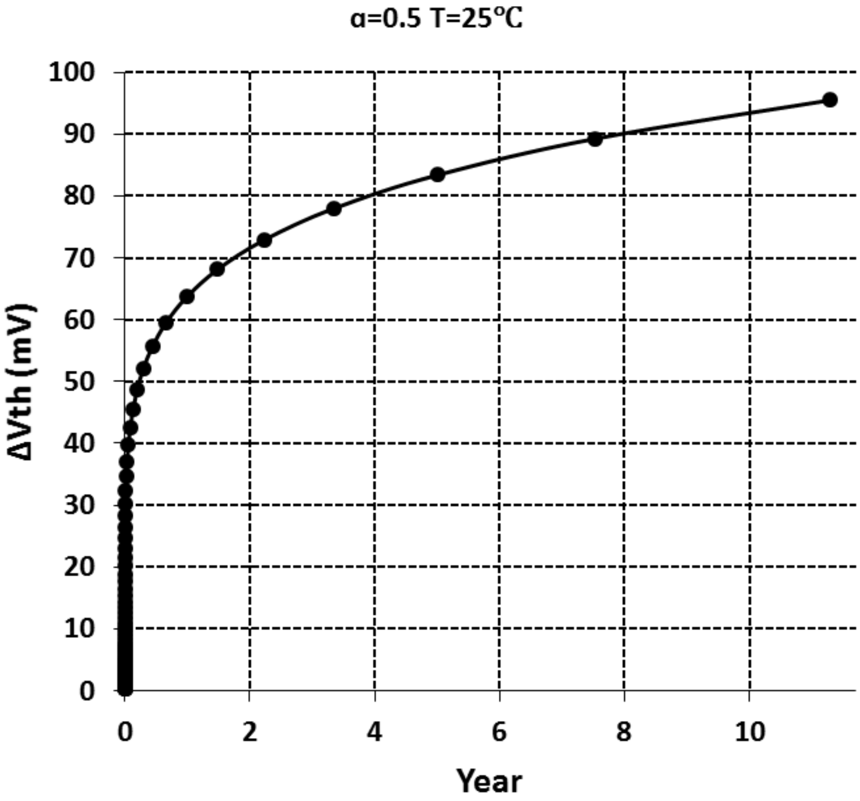


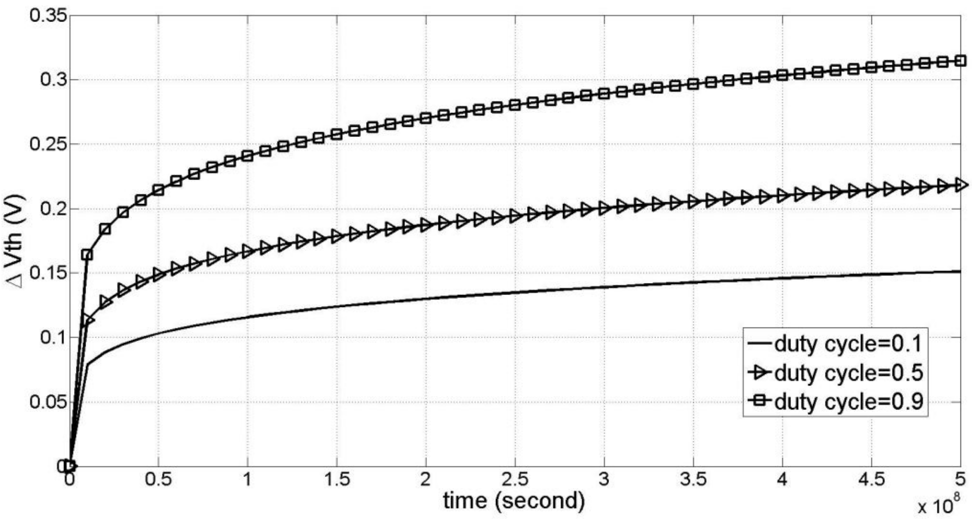
Figure 2.10: Dynamic NBTI for a minimum sized PMOS transistor of 90nm technology, Reproduced from [10]

2.2.2 Influence Factors of NBTI

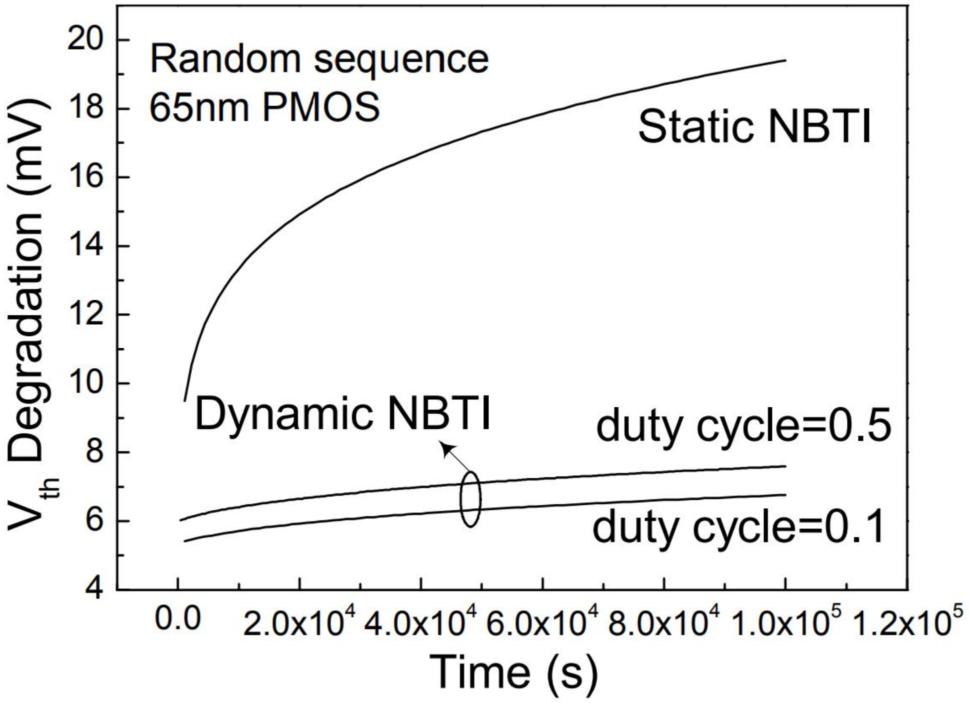
Research groups have been working on the R-D model of NBTI for years. NBTI effect is operation time, temperature, stress duty cycle and V_{gs} dependent, [10, 77, 82, 83].

Duty cycle Dependence

The threshold voltage degradation from dynamic NBTI of a PMOS transistor depends on the duty cycle of the input signal. Figure 2.11 shows the threshold voltage degradation of static NBTI and dynamic NBTI over the time. As Figure 2.11 (a) shows, the Vth degradation is stronger with higher duty cycle. As the results show in Figure 2.11 (b), ΔV_{th} grows over time, and the Vth degradation of static NBTI is much stronger than dynamic NBTI. Because the recovery of ΔV_{th} is rapid during the beginning of



(a) NBTI with Different Duty Cycle, Reproduced from [32]



(b) Compared with The Static NBTI, Reproduced from [79]

Figure 2.11: Long Term Dynamic NBTI with Different Duty Cycle

recovery time, as shown in Figure 2.9. Compare to static NBTI, ΔV_{th} of dynamic NBTI is decreased significantly even though the duty cycle is close to one. Because of the correlation between duty cycle and ΔV_{th} are positive, the input vector control can mitigate the performance degradation from NBTI, as shown in [83].

Temperature Dependence

The threshold voltage degradation from dynamic NBTI of a PMOS transistor depends on the operating temperature. Figure 2.12 shows the threshold voltage degradation due to dynamic NBTI in temperature over time. As the estimate ΔV_{th}, as shown in Figure 2.12, the correlation between temperature and ΔV_{th} are positive. As the temperature increases, the degradation of the Vth increases sharply.

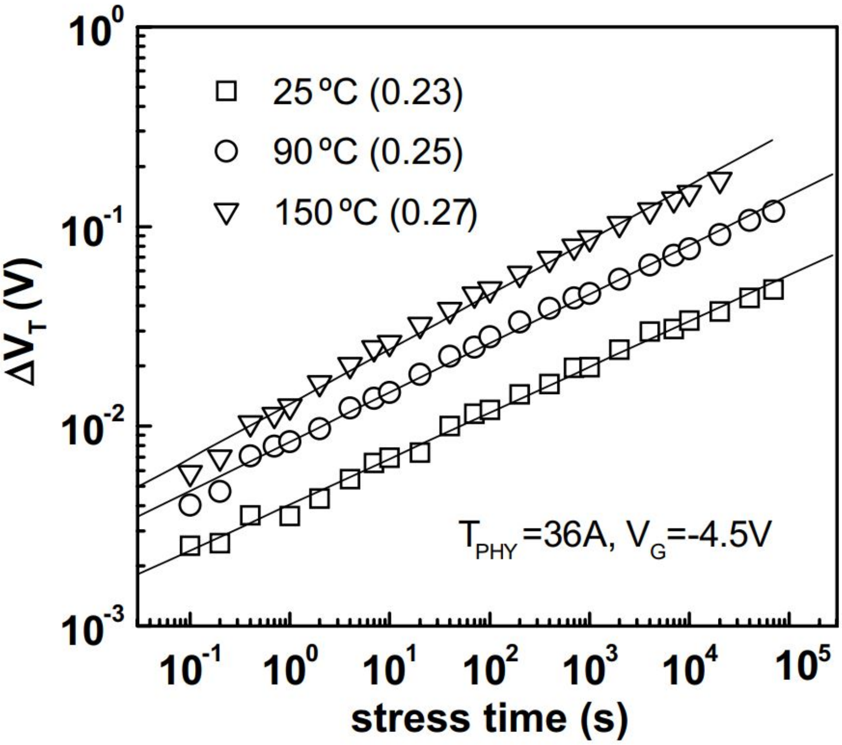


FIGURE 2.12: Long Term Dynamic NBTI with Difference Temperature, Reproduced from [3]

Temperature(°C)	Delay (ns)	Delay degradation
25	82.99	30.00%
30	91.58	33.18%
35	100.65	36.47%
40	110.19	39.93%
45	120.19	43.55%
50	130.63	47.32%

TABLE 2.1: NBTI Vth Degradation Comparison in Different Operating Temperatures

The Vth degradations are estimated and compared with relatively smaller temperature difference, shown in Table 2.1, to show the influences from the temperature in detail. According to the estimation results, shown in the Table, each 5-degree increase in the temperature leads to a 3% degradation of Vth. The internal temperature of a chip increases rapidly when the system is running at the high performing stage. Combined with the results, as shown in Section 2.1.3, a good cooling system will not only let the IC circuit running at higher performance but also extends the lifetime of the IC circuit.

Vgs Dependence

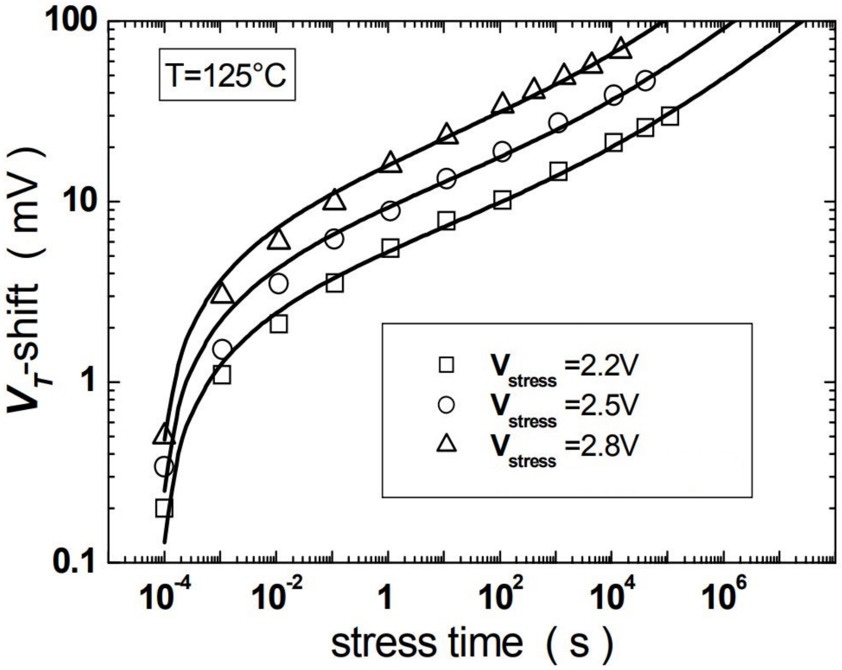


FIGURE 2.13: Long Term Dynamic NBTI with Different Vgs, Reproduced from [62]

Figure 2.13 shows the threshold voltage degradation of dynamic NBTI Vgs after time. As the estimated results shown, the correlation between Vgs and ΔV_{th} is positive.

Discussion

As stated above, the NBTI effect is dependent on time, temperature, and stress duty cycle and Vgs. Figure 2.14 shows the relationship between Duty Cycle, VDD and ΔV_{th}. Previous research groups have also stated that the relationship between ΔV_{th} and operating frequency (switching activity) is de-correlated or ΔV_{th} weak frequency dependent at high-frequency,[3, 10, 77]. However, this discounts the relationship between switching activity and temperature. The Power dissipation of a CMOS is consists of static power dissipation, dynamic power dissipation and short-circuit power dissipation, [43, 75]. Dynamic power dissipation and short-circuit power dissipation are depended on switching activity, and $\Delta T = T_0 + \theta$ P [69] , where θ is the thermal resistance. Therefore, the temperature is switching activity dependent. Hence, the NBTI effect is non-frequency dependent if and only if there is a cooling system which keeps the temperature of the device to be constant.

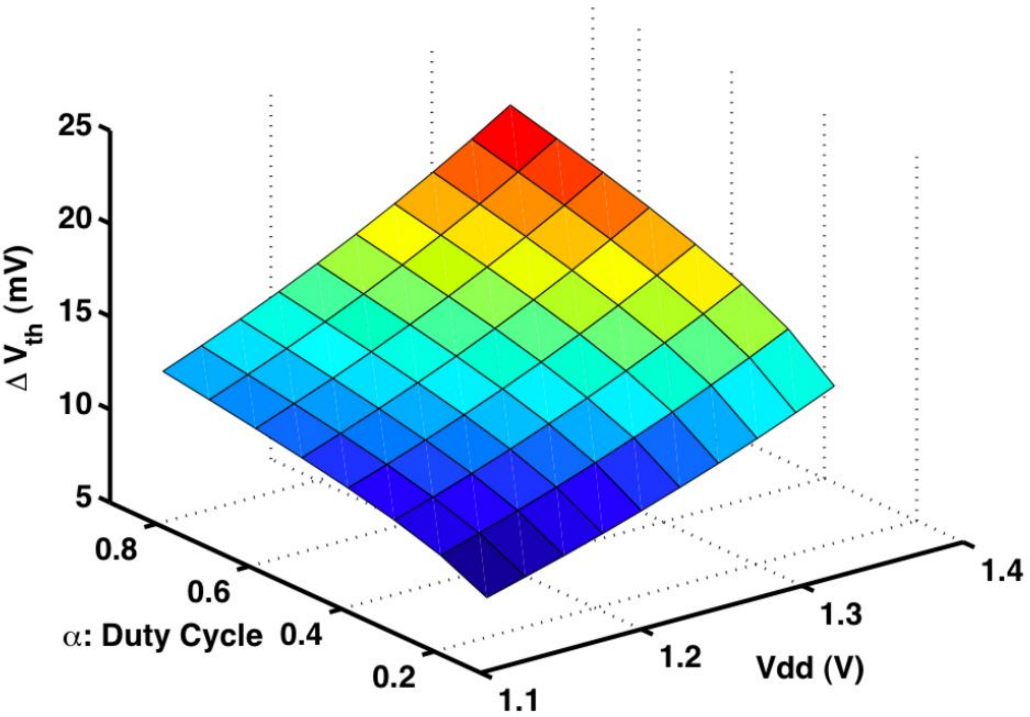


FIGURE 2.14: Relationship between Duty Cycle, VDD and ΔV_{th}, Reproduced from [10]

Once the threshold voltage degradation (ΔV_{th}) is estimated, the NBTI effect can be integrated into the circuit simulation tools such as Spice. The threshold voltage and delay degradation are estimated by using the R-D model at the 90nm technique is shown in table 2.2. As the table shows, the delay degradation changes in different time, duty cycle, temperature and Vgs.

Years	Duty Cycle	Vgs(V)	Temperature °C	ΔV_{th}(mV)	Delay(ps)	Δt
0	0.5	2.0	50	0	16.40	0%
1	0.5	2.0	50	112.15	18.14	10.61%
5	0.5	2.0	50	146.65	18.73	14.21%
5	0.9	2.0	50	167.77	19.09	16.40%
5	0.5	2.0	75	212.39	19.91	21.40%
5	0.5	2.5	50	162.25	19.00	15.85%

TABLE 2.2: Estimated Threshold Voltage and Delay Degradation of an Inverter

2.2.3 HCI and PBTI

Hot-Carrier Injection

Hot-Carrier Injection (HCI) is a critical ageing issue in n-channel Metal-Oxide-Semiconductor field-effect transistor (NMOS), [28, 42, 49, 87]. HCI phenomenon occurs because the interface traps are generated at the drain end gate oxide interface when the current flow between the source and drain during the switching time. It manifests as the degradation of the switching characteristics, threshold voltage, drain-current and noise margin. Even worse, the HCI effect is unrecoverable, [77, 81]. Therefore, although the HCI effect will not cause significant delay degradation like the NBTI effect[77], however, it accelerates the degradation of the device lifetime directly.

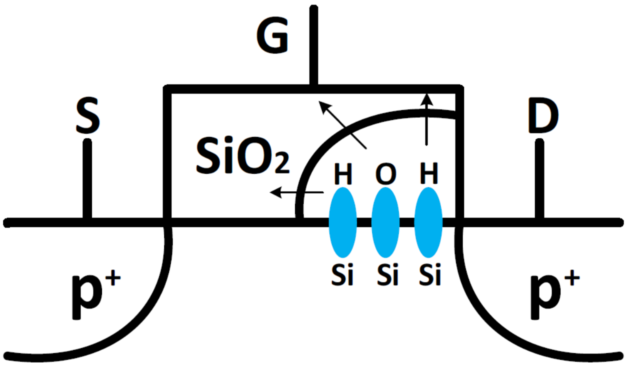


FIGURE 2.15: NBTI: Hydrogen Atoms Diffusion, Reproduced from [77]

The R-D model can describe the HCI effect. The Si-H or Si-O bonds at the drain end of gate oxide interface might be broken (reaction) when the PMOS transistor is switching, and this generates the interface traps at the drain end gate oxide interface permanently. Those hydrogen atoms generated by reaction then diffuse away from the interface to the gate (diffusion), [64, 77, 81].

Positive-Bias Temperature Instability

Positive BTI (PBTI) is an ageing reliability issue in N-channel MOS (NMOS) transistors. Unlike the NBTI, PBTI occurs when a positive gate bias, a positive potential difference between gate and source (Vgs), is applied, [74]. It has been considered negligible because the degradation due to PBTI was inconsequential compared with other ageing phenomena; however, the threshold voltage shift of PBTI has become significant in High-K Metal Gate technologies, [74]. The double effect of NBTI and PBTI exacerbates the timing degradation of ICs.

In this section, we introduced three ageing mechanisms, NBTI, HCI and PBTI, shown in Figure 2.16. Those three mechanisms lead to the lifetime degradation of CMOS circuits.

The ranking of critical paths might change due to those wear out mechanisms. Vth degradation can be estimated by using the existing ageing model. The ΔV_{th} can be incorporated into the simulation tools to estimate the propagation delay after ageing of the design. A method of potential NBTI critical path identification is introduced in Section 3.3.

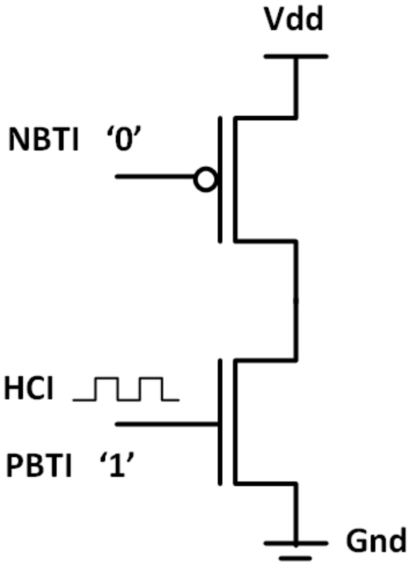


FIGURE 2.16: Ageing Mechanisms, Reproduced from [25, 90]

2.3 Ageing Mitigation and Lifetime Prediction Techniques

2.3.1 Ageing Mitigation

As Section 2.2.1 shows, the ageing of an IC degrades the IC's performance and reduces the IC's lifetime significantly over time. Research groups proposed many ageing mitigation techniques. Such as Input Vector control [1, 37, 83], Ageing-Aware synthesis[33, 67], Dynamic Voltage Scaling [58], and Power Gating [18, 27]. This Section introduces two efficient ageing mitigation techniques for NBTI effect mitigation.

Power Gating

Power Gating is widely used to reduce the leakage power when a digital IC is at the idle periods, as the percentage of leakage power has increased significantly due to the technology note scaling, [24, 27]. The power gating techniques can also be used for the NBTI mitigation. Figure 2.17 shows the structure of a header-based power gating sleep transistor. As the figure shows, a High-Vth PMOS transistor is inserted between the power rail and the logic gates as a sleep transistor. The virtual Vdd replaces the Vdd of the original circuit. The size of the transistors is usually carefully designed to reduce the leakage current and mitigate the NBTI simultaneously, [24].

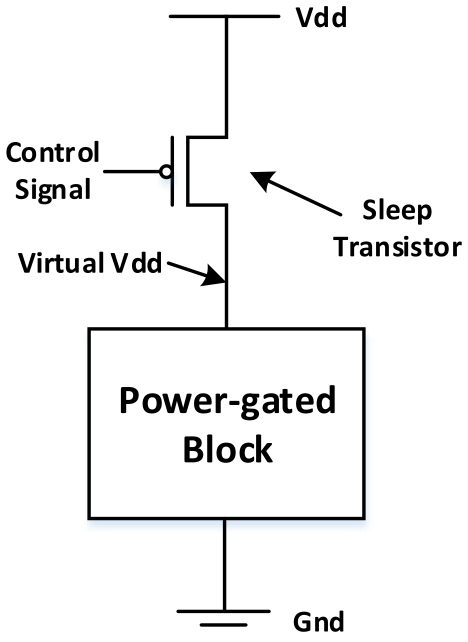


Figure 2.17: NBTI-aware Power Gating, Reproduced from [27]

The power gating technique causes the delay, area and wake-up time overhead as the insertion of the PMOS transistor. According to the result stated in Section 2.2.2, the NBTI is Vgs dependent. The 'Control signal', as shown in Figure 2.17, is set as a log '1' while the IC is at the idle period. The virtual Vdd is tending to the grand gradually as the leakage current of the Power-gated Block is relatively higher than the sleep transistor. It, therefore, reduced the stress of the PMOS transistors in the Power-gated Block. The power gating technique can extend the lifetime effectively, [18, 24]. However, it causes the reduction of performance and additional area overhead simultaneously, which are the primarily concerned of the IC design nowadays.

Dynamic Voltage Scaling

Dynamic Voltage Scaling (DVS) techniques are a widely used technique to reduce power consumption. There are two types of DVS techniques, Traditional DVS and On-line DVS.

The traditional DVS scales the scales the supply voltage and the operating frequency simultaneously. Figure 2.18 shows the frequency f vs. supply voltage VDD shmoo plot of the ARM M4F combined with the traditional DVS technique. As the figure shows, the system is running under a specific operating frequency with a corresponding supply voltage, because of the propagation delay of a CMOS device decreases with the lower supply voltage.

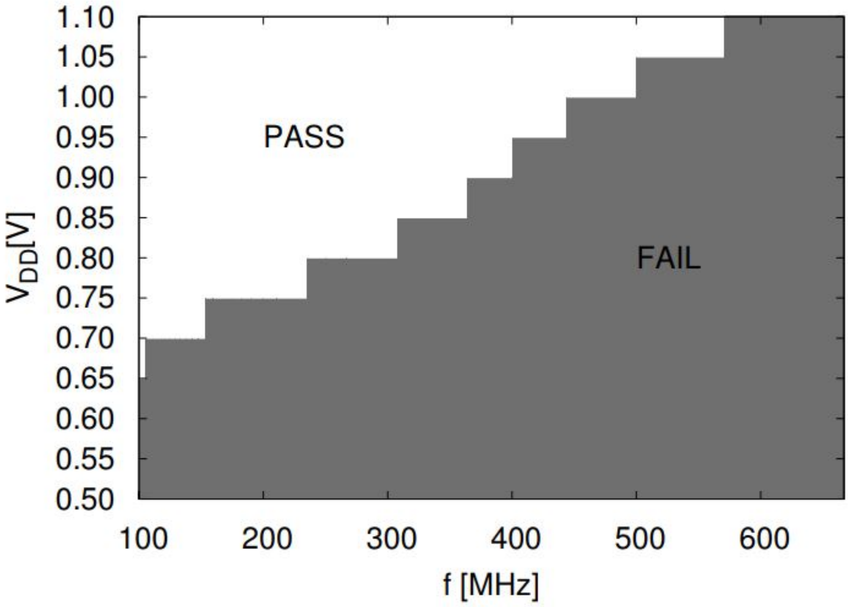
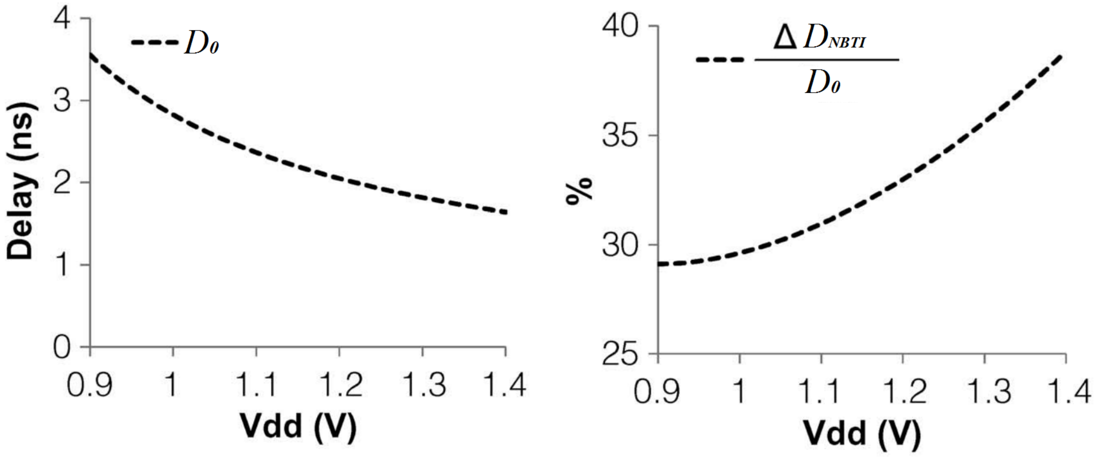


FIGURE 2.18: The Frequency f vs. Supply Voltage VDD Shmoo Plot of the Traditional DVS Technique, Reproduced from [44]

The On-line DVS widely used technique to reduce power consumption and compensate for PVTA variations. Process variations and ageing-induced device degradation are becoming major reliability concerns in modern semiconductor technologies. Both phenomena lead to performance degradation, and hence timing errors. The integrated circuits are typically designed with large safety margins to avoid delay fault induced failures, [2]. This generally means a circuit is designed for worst-case operating conditions. Such an approach may limit system performance and lead to an increase in power consumption, [58]. The on-line DVS has been proposed to assure an IC's lifetime reliability. Such schemes are typically based on the use of sensors to predict circuit failures before actual errors appear. The system can then adaptively scale its operating frequency and supply voltage according to the actual operating conditions to compensate for performance degradation, [2]. Unlike the traditional DVS, the on-line DVS removes the large safety margin for PVTA variations to reduce power consumption further and improve the performance of an IC.

Various in situ delay monitoring sensors have been proposed. These include delay fault detection and prediction techniques, [29, 40]. Existing delay fault sensors are usually placed on the circuit's longest delay paths. However, the increasing complexity of ICs has led to a significant rise in the number of long paths and potential ageing-critical paths that may be vulnerable to timing errors, [19, 33, 67], and this means the cost of in situ delay monitoring may be prohibitive.



(a) The relationship between intrinsic delay and supply voltage

(b) The relationship between percentage intrinsic delay degradation and supply voltage

FIGURE 2.19: Ageing mitigation by using DVS, Reproduced from [24]

Figure 2.19 (a) shows the relationship between intrinsic delay and supply voltage. As the Figure shows, the intrinsic delay decreases with the increase of the supply voltage. Figure 2.19 (b) shows the relationship between intrinsic delay degradation and supply voltage. As the Figure shows, the intrinsic delay degradation increases with the decrease of the supply voltage.

DVS is the adjustment of power and speed settings on a computing devices various processors, controller chips and peripheral devices to optimise resource allotment for tasks and maximise power saving when those resources are not needed. DVS allows devices to perform needed tasks with the minimum amount of required power. The technology is used in almost all modern computer hardware to maximise power savings, battery life and longevity of devices while still maintaining ready compute performance availability. Figure 2.20 shows the supply voltage scaling flowchart. Where N_{loop} is the number of the clock cycles that the circuit has been monitored in one voltage scaling period, N_{set} defines the voltage scaling period, n is the number of error that occurred in voltage scaling period, n_{set} is the threshold for voltage scaling .

The DVS system costs more area overhead due to the sensors insertion, however, it reduces the power and extends the lifetime without sacrifice the performance of ICs, which is an advantage over the power gating techniques.

As Figure 2.20 shows, the principle of the DVS system is to keep the supply voltage at the lower state while the number of error is tolerable. Compared with leaving a safety margin for timing degradation from the Process, Voltage, Temperature and Ageing (PVTA) variations, implementing a DVS system for the devices not only saves power

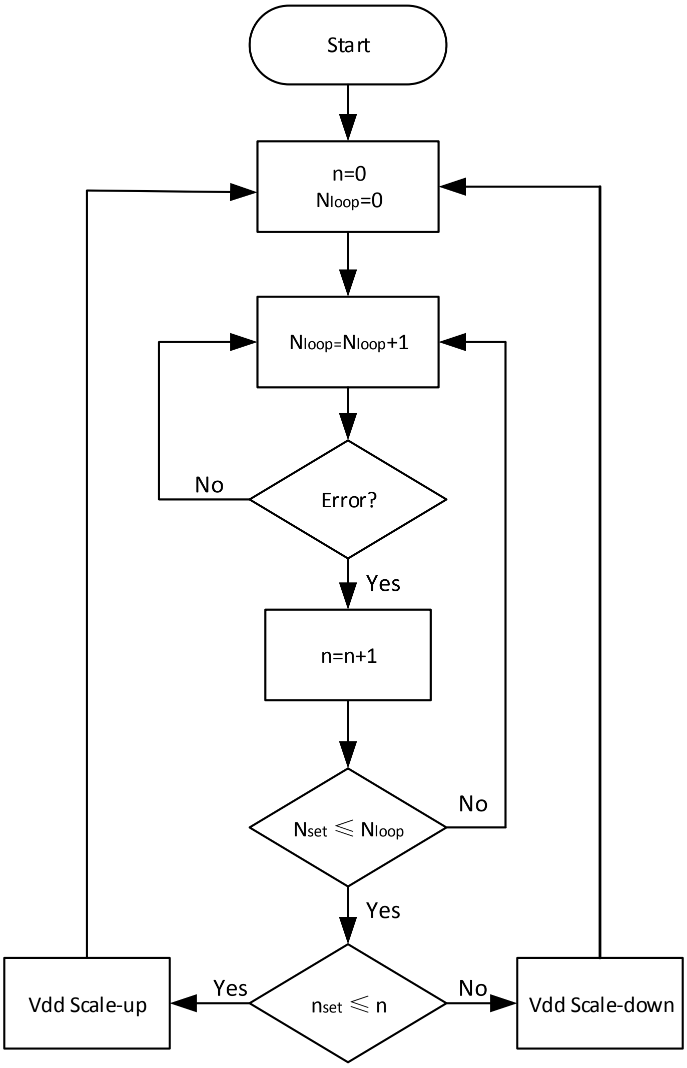


Figure 2.20: Supply Voltage Scaling Flowchart, Reproduced from [17]

during the lifetime but also slows down the power supply voltage dependent wear out mechanisms.

2.3.2 Lifetime Prediction

The RD model can predict the behaviour of an IC, [33, 67], therefore predicts the lifetime of ICs before its fabrication. Equation (2.4) shows a simplified NBTI model, [33].

$$\triangle D_{nbti} = K.D_0.\alpha^{0.16}.t^{0.16} \tag{2.4}$$

Where $\triangle D_{nbti}$ is the delay degradation for a given set of operating time and stress duty cycles. K is a fitting parameter dependent on the operating temperature, supply voltage and the technology node; α is the stress duty cycle; t is the operational time of the circuit. As the equation shows, the delay degradation is depending on stress duty cycle during the circuits operating time. In another word, the workload affects the lifetime of a circuit. The Equation can also be derived as follows:

$$t = \sqrt[0.16]{\frac{\triangle D_{nbti}}{K.D_0.\alpha^{0.16}}} \tag{2.5}$$

In modern IC design, a timing margin is inserted for the ageing variation. The lifetime of an IC can be estimated by substitute the timing margin into the equation as D_{nbti}. Figure 2.21 shows the lifetime prediction flowchart. Where N is the current number of cycles, N_{size} is the sample size of input pattern random probability, T_{est} is the lifetime estimation result from a single calculation, T_{max} and T_{min} are the maximum and the minimum lifetime estimation which was calculated so far.

As Figure 2.21 shows, the first step in lifetime prediction is to estimate the signal probability (duty cycle), from input patterns and logic circuit net-list. Then estimate lifetime (T_{est}) by using the RD model (equation(2.5)) and the timing margin. Finally, the T_{est} will be compared with the maximum and the minimum lifetime estimation (T_{max} and T_{min}). T_{max} will be saved as T_{est} if T_{est} is greater than or equal to $T_{max}$and T_{mmin} will be saved as T_{est} if T_{est} is less than or equal to T_{max}. The loop will end until N is greater than or equal to N_{size}. The lifetime in specific operating temperature, supply voltage can be predicted precisely if the sample size is large enough.

HSPICE MOSRA is a model for transistor-level ageing behaviour estimation , [46]. Unlike the RD model, as shown in equation 2.4, MOSRA is more suitable for small circuitries. It invokes ageing mechanisms such as HCI and BTI to estimate the behaviour of a circuit after ageing. Ageing mechanisms depend on the workload, the circuit's operating conditions, the circuit's operating time, technology node and the structure of circuitry. MOSRA combines those parameters and predicts circuit's behaviour after ageing. Therefore, the HSPICE MOSRA is useful in the design phase of a small circuit when the circuit's ageing is taking into account.

2.4 Ageing Sensors

2.4.1 Razor FF

Razor FF [35, 50] is a delay-fault detection sensor, which is a flip-flop-type sensor using the double sampling technique. It detects the consistency of the output signal from

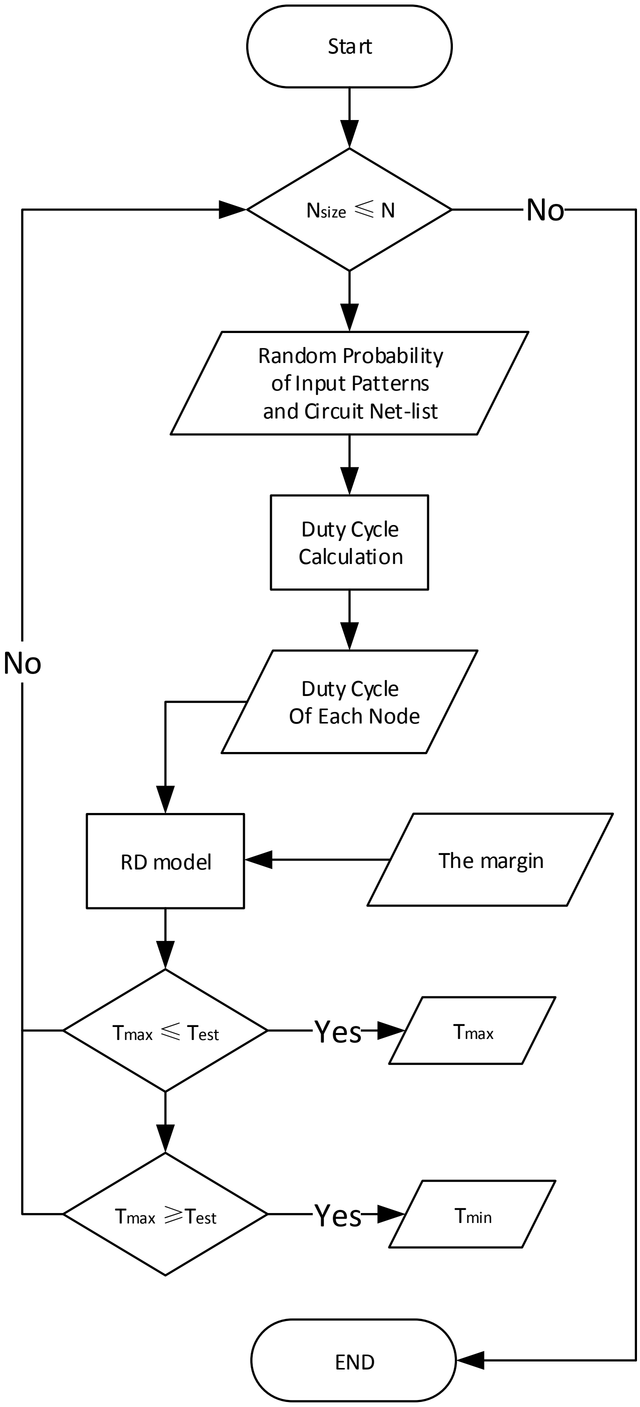
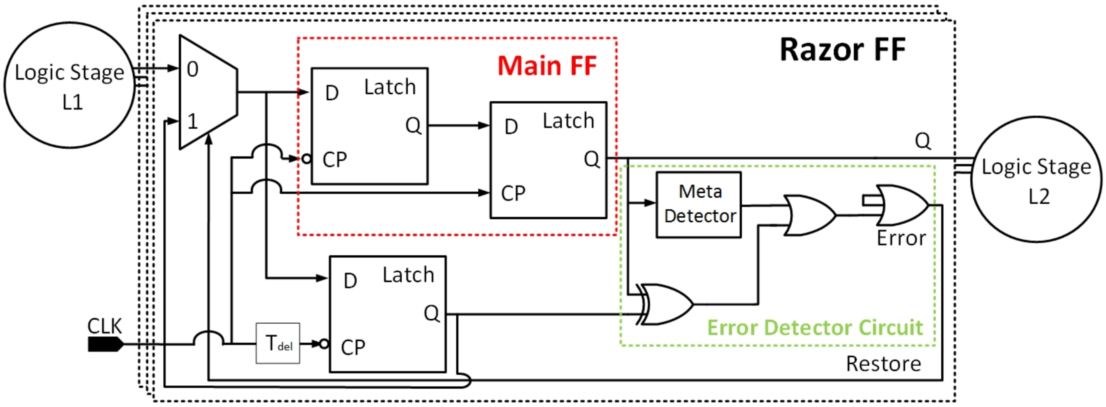
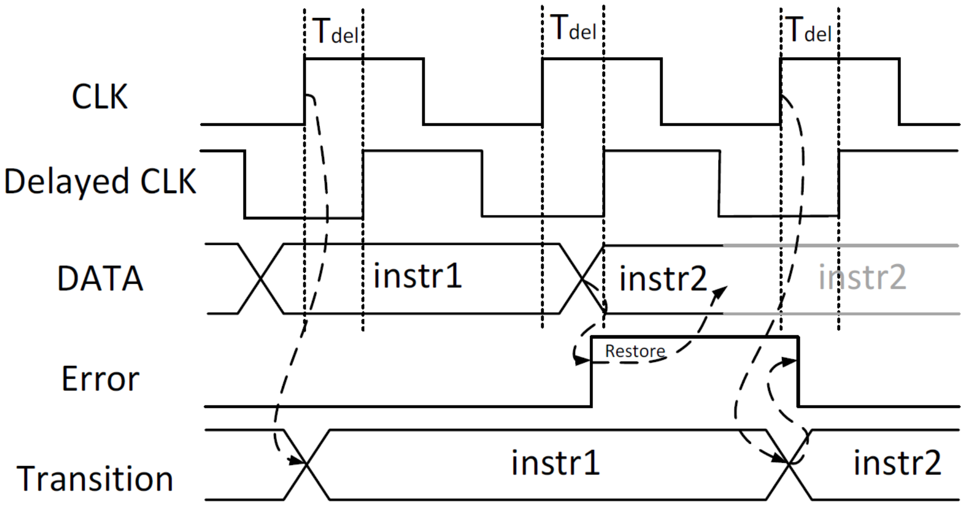


FIGURE 2.21: Lifetime Prediction Flow Chart, Reproduced from [22, 34]

a near critical path after the rising clock edge. As Figure 2.22 (a) shows, Razor FF consists of a main Flip-Flop, a shadow latch, a multiplexer, one delay element and an Error Detector Circuit. The shadow latch receives data as a reference which is sampled after clock rising edge to help the detection circuit detecting the delay-faults. The error signal is generated by comparing the data between the main FF and shadow latch. Razor FF is a fault tolerant sensor which can restore the data when the sample is missed. The delay element generates a detection window, Tdel, shown in Figure 2.22 (b). Transitions in this detection window, Tdel, causes the shadow latch to sample a different signal with the main FF, therefore, triggers the error signal as the inconsistency between the samples from the main FF and the shadow latch. Usually, there are two methods to correct this error: Suspend the process for one clock cycle then restore the data in the shadow latch via the multiplexer in the next clock cycle shown in Figure 2.22 (b) or replay the instruction at the system level, [40].



(a) The Razor FF



(b) Timing Diagram of The Razor FF

FIGURE 2.22: The Razor FF and Timing Diagram, Reproduced from [35]

Razor FFs are typically implemented on the critical paths of logic designs, but it only detects a timing error after it occurs. Therefore they cannot be used for predicting ageing-induced delay-faults unless the design is modified accordingly, which would further increase its area overheads. Razor FF can only detect the delay-faults in one path, and the area cost for multi-path monitoring using Razor can be prohibitive. The timing degradation of paths increases with circuit ageing, the slack between the rising clock edge and the transition of the input signal decreases. Hence, the input signal first violates the setup time of the main FF. Meta-stability occurs in the main FF when the clock and input change about the same time. Therefore, the checking circuitry is required to detect meta-stability in the main FF, as shown in Figure 2.22 (a). Moreover, as a delay fault detection sensor, Razor is not suitable for ageing prediction. Because of the upcoming timing error is unpredictable in this case.

2.4.2 Razor II

A number of latch-type delay fault monitoring sensors have been proposed in recent years, [16, 89]. These replace the main FF with a latch to solve the metastability issue and to decrease the transition time from D to Q. Razor II was proposed as a latch-type version of Razor FF, which can address the timing issues more effectively. It replaces the Flip-Flop by a single latch, which decreased transition time from D to Q.

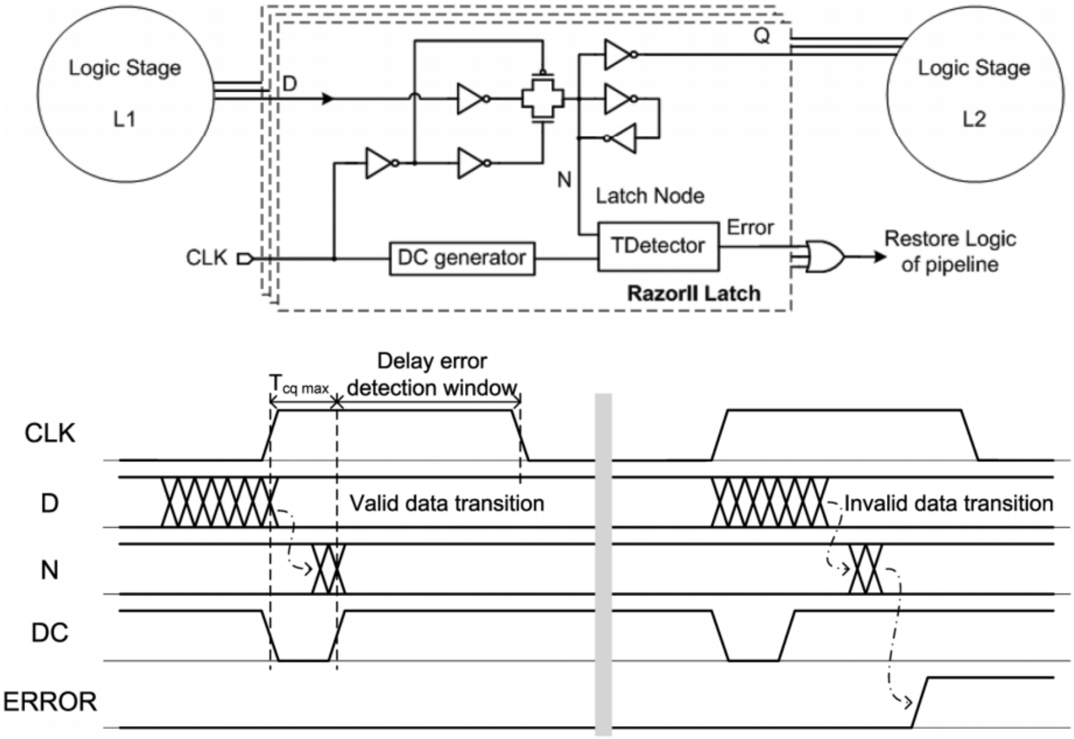


Figure 2.23: Razor II and Timing Diagram, Reproduced from [30]

Razor II consists of a latch, a Detection Clock (DC) generator and a Transition Detector. The detection window is between the rising edge of the DC signal and the falling edge of the clock signal. The signal DC generates a slack which ensures the error signal is not triggered during the transition time between Q and D after the clock rising edge. The Transition Detector detects the stability of the input signal within the detection window. As Figure 2.23 shown, transitions during this detection window trigger the error signal. As Razor II is a latch, which passes the signal through from D to Q during clock is logic

’1’, metastability does not occur when the transition is approaching the clock rising edge. A metastability checker is not required in Razor II. However, there might be a misjudgement when Razor II is operating as a delay fault detection sensor, any glitches during the operation of the circuits while the clock is logic ’1’ is considered as a timing error. Same as Razor FF, Razor II only detect a timing error after it occurs. Therefore it cannot be used for predicting ageing-induced delay-faults, and it can only detect the timing error for one path. Therefore the area cost for multiple path monitoring might be prohibitive. As a delay fault detection sensor, Razor II is not suitable for ageing prediction. Because of the upcoming timing error is unpredictable in this case.

2.4.3 Canary FF

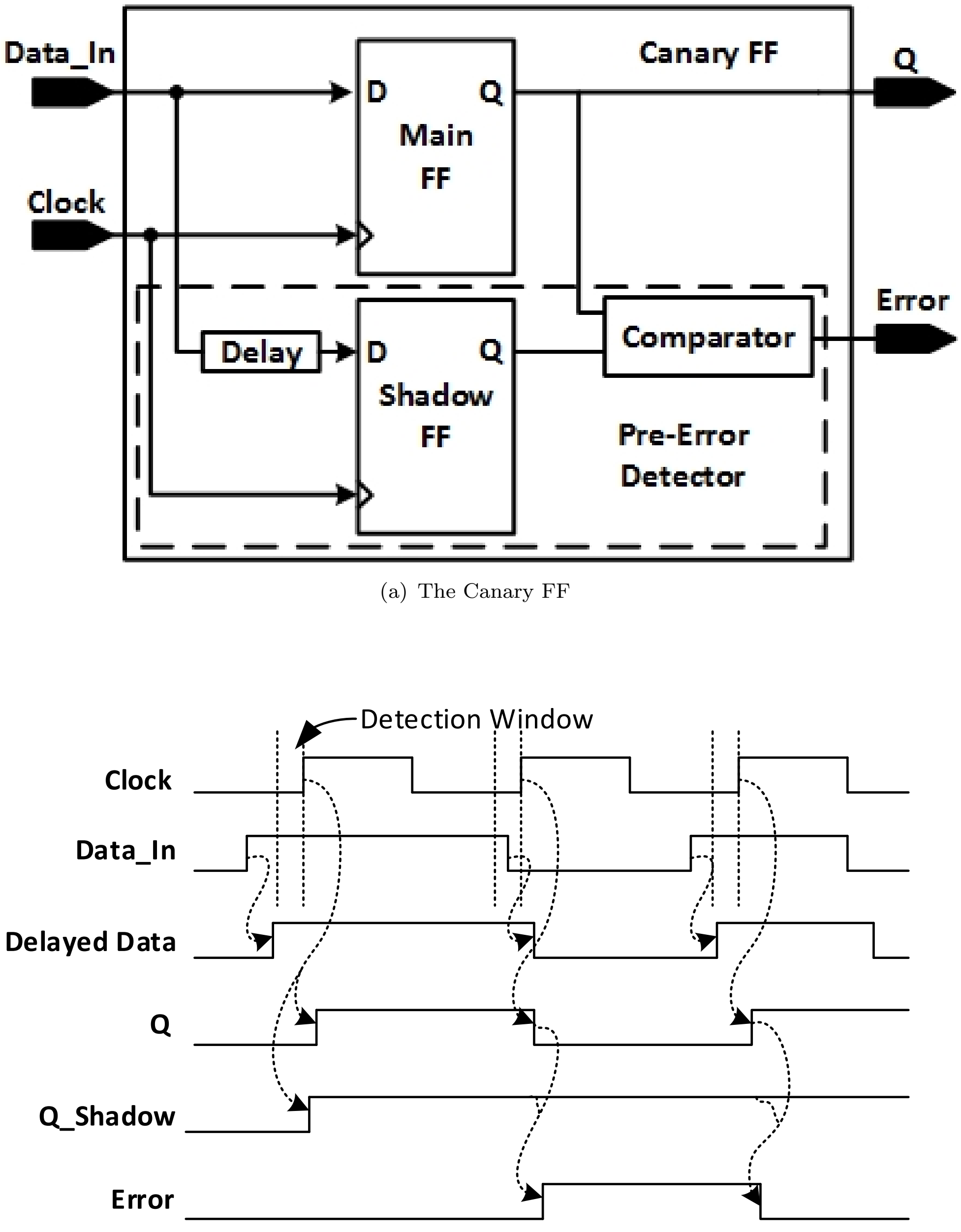
Unlike delay fault detection sensors, like Razor FF[29], the Canary FF[40], shown in Figure 2.24, is a delay fault prediction sensor that checks data consistency before the rising clock edge. Canary FF consists of a main Flip-Flop, a shadow Flip-Flop, one delay element and a comparator. As shown in Figure 2.24, the shadow FF receives delayed data as a reference and compares it with the data from the main FF. As the path ages, the Error signal will be triggered if the delayed data violates the setup time of the shadow FF.

Canary FF has a small safety margin to detect if a delay-fault is about to occur. As a delay-fault prediction sensor, Canary FF is more suitable for ageing detection compared to Razor. In this case, the input signal may violate the setup time of the shadow FF as a result of circuit ageing, not the main FF. Therefore it does not require any circuitry for timing error recovery nor meta-stability detection for the main FF such as Razor FF. However, meta-stability occurring in the shadow FF causes errors. On the other hand, implementing a shadow FF of the same size causes a relatively large area overhead. As a delay fault prediction sensor, Canary is suitable for ageing prediction. However, a Canary FF can only detect the timing error for one path. Therefore the area cost for multiple path monitoring might be prohibitive.

2.4.4 TDS FF

TDS FF [71] is a delay-fault detection and fault tolerant sensor, which is achieved by the transition detecting technique. Unlike Canary and Razor, The main advantage of TDS FF over other sensors is multi-detection. Which means TDS can detect the delay-fault from more than one paths simultaneously.

Figure 2.25 shows the design of the TDS FF, as the figure shows, the TDS FF consists of a main Flip-Flop, an Error Flip-Flop, a multiplexer, an XOR gate, an OR gate and one delay element for the Mem_CLK. The XOR gate detects the transitions between D



(a) The Canary FF

(b) Timing Diagram of The Canary FF

Figure 2.24: The Canary FF and Timing Diagram, Reproduced from [40]

and Q of the main FF. The Mem_CLK signal is a delayed clock signal, and it generates a detection window, Tdel, shown in Figure 2.26. The signal Error_L will be triggered when the data between D and Q are different. It then triggers the signal Error_R. Signal data from Error_R will be sampled on the ring edge of the Mem_CLK signal, the transition from D to Q is considered as an error if Once the timing error occurs, the data will be latched by the multiplexer and the process will be suspended for one clock cycle to restore the data.

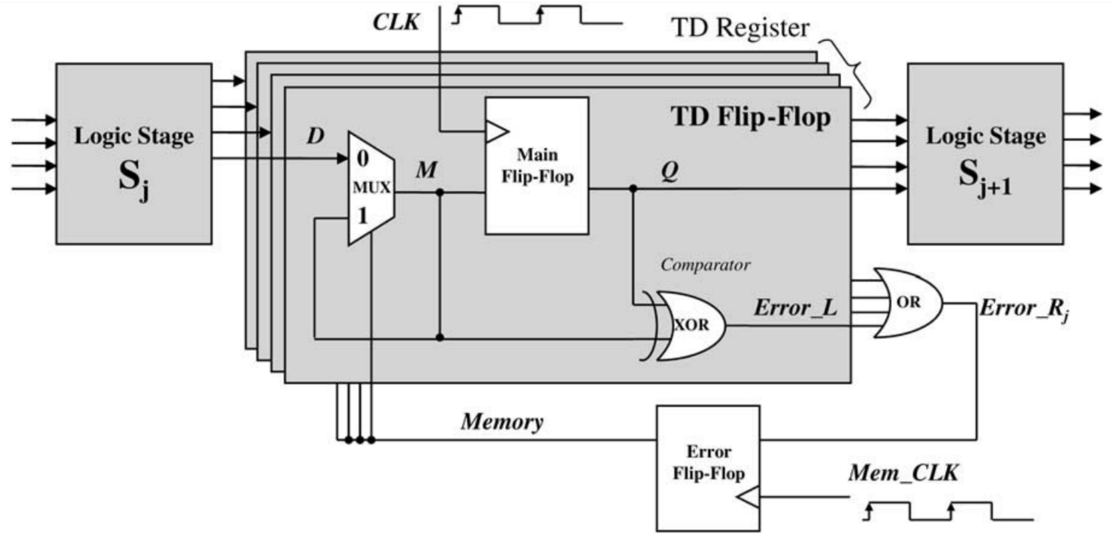


Figure 2.25: TDS FF, Reproduced from [71]

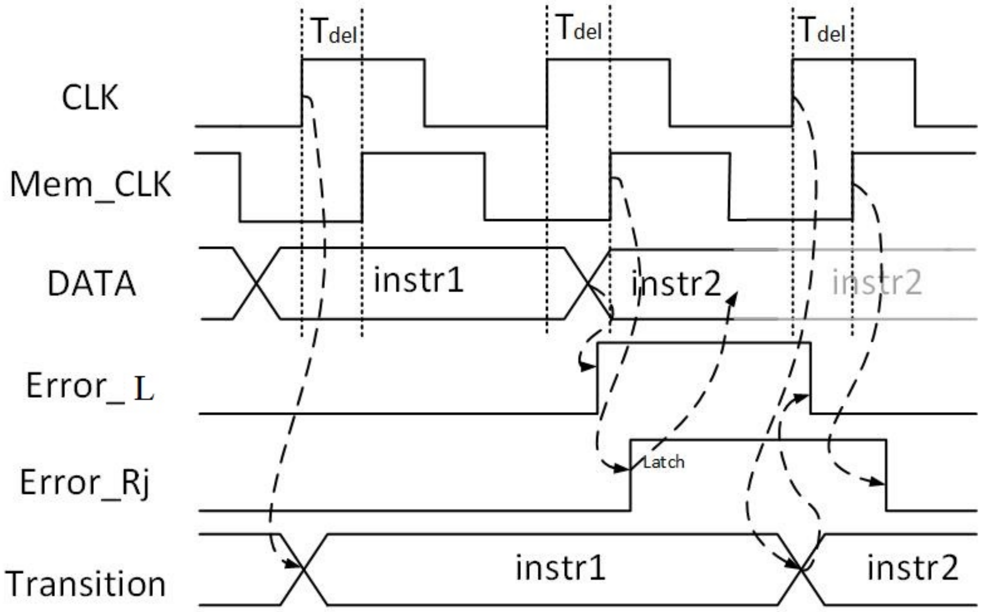


Figure 2.26: TDS FF Timing Diagram, Reproduced from [71]

With PVTA variations, the input signal will violate the setup time of the main FF before the error occurs. Meta-stability will occur in the main FF when the clock and input change about the same time. Therefore, checking circuitry is also required to detect meta-stability in the TDS FF. On the other hand, the timing error will not occur for a long period of time since time zero. Therefore the PMOS transistor in the OR gate, Error FF and multiplexer will undergo serious ageing due to NBTI, a constant logic '0' input. The noise margin of the Error FF will change due to NBTI, and it may cause the bit flipping after a certain period of time. As a delay fault detection sensor, TDS is not suitable for ageing prediction. Because the upcoming timing error is unpredictable in this case.

2.5 Concluding Remarks

The literature reviewed in this chapter suggests that the traditional dynamic reliability management schemes used in delay fault predicting are prohibitive as the one sensor can only monitoring the delay fault from one path. On the other hand, existing sensors replace flip-flops of the original design. Therefore it is hard to implement, and it might influence the functionality of the original design. Most of the sensors are suffering from the metastability, which affects the reliability and power dissipation of the digital circuit. Ageing models are proposed to estimate the lifetime of an IC before it's fabrication. However, the performance and lifetime degradation from ageing effect is depending on the circuit's operating temperature, supply voltage and stress workload. Each individual chip might be running in different workload and environment. Therefore the environment and workload of an IC are unpredictable in the design stage of the CMOS circuits. The lifetime prediction of each individual chip is certain when it is predicted by using such models. Various ageing mitigation technique has been proposed to extend the lifetime of an IC. The trade-off between lifetime and performance usually achieves such an approach. A faster operation speed always the pursuit of people. Therefore, performance is one of the primary concern of the IC design nowadays. Such sacrifices are reluctant, and this leads to the abandonment of the application of such approaches in practice. Therefore, the cost-effective delay fault monitoring, accurate lifetime estimation and affordable ageing mitigation are the significant challenges for modern commodity microprocessor design.

Chapter 3

Parity Check Circuitry

Delay-fault monitoring sensors are widely used for Dynamic Voltage and Frequency Scaling (DVFS) to compensate for intrinsic Process, Voltage, Temperature and Ageing (PVTA) variations. Such techniques are generally based on monitoring the circuit's critical paths. Shrinking IC technology has enhanced the timing dependence process and ageing variations, which varies the ranking of the critical path. Consequently, such an approach have a large number of long delay paths that need to be monitored. This means the cost of delay-fault monitoring is becoming exorbitant. This chapter presents a new delay-fault monitoring circuit, which is able to monitor multiple paths simultaneously. The proposed circuitry has been designed and verified in a 32 bit MIPS processor using a 65nm technology. Our results indicate that the use of the proposed sensor for delay monitoring can lead to a significant saving in area and power overheads of two-thirds and one-third, respectively, compared to a canary flip-flop.

3.1 Introduction

Integrated circuits are typically designed with a safety margin as the performance varies due to unavoidable PVTA variations, [17, 26]. This means the circuits are generally designed for a combination of worst-case conditions, which limits the system performance and leads to an increase in power consumption during the system's lifetime. DVFS schemes have been proposed to dynamically eliminate the unused safety margin to improve the power-efficiency [47]. Such schemes use in situ sensors to monitor the delay-fault from the longest delay path[17, 36, 40, 55]. The system can then dynamically scale the supply voltage and the operating frequency to compensate for the impact from PVTA variations [2, 76].

Generally, the in situ delay monitoring sensors aim to detect data consistency before or after the clock rising edge to predict or detect delay faults, [35, 40]. This can be done either by double sampling or by stability checking techniques, [17, 66]. Delay fault detection sensors trigger an error signal when an actual error occurs. The advantage of using this is that the supply voltage will always remain at the lowest value, [35]. However, it requires an error recovery scheme and buffers to be inserted on short paths to disambiguate between early transitions in those short paths and actual errors, [16]. Delay fault prediction sensors trigger an error signal before an error actually occurs. This may lead to a relatively higher power consumption, compared with delay fault detection sensors, but the buffers and the error recovery scheme are not required in this case, [40]. The sensors are usually implemented at the end of near-critical paths. Each sensor only monitors delay faults from the paths that share the same end. Technology shrinking has exacerbated the timing-dependent process and ageing variations, which may change the ranking of critical paths. This will lead to a significant rise in the number of ageing and process variation Potential Critical Paths (PCPs) that are deemed to be vulnerable to delay faults, [41, 58]. The cost of conventional in situ delay monitoring is thus becoming prohibitive.

In this chapter, we propose a cost-efficient Parity Check Circuit (PCC) for delay fault prediction to mitigate the cost of in situ delay monitoring. PCC is able to monitor multiple paths simultaneously, which significantly reduces the number of sensors. The proposed sensor has been designed and verified in a 65nm technology. Our results indicate that using the proposed sensor for delay fault monitoring in a 32-bit MIPS can lead to a significant saving in the area and power overheads, compared to the use of canary flip-flops [40]: by two-thirds and one-third, respectively.

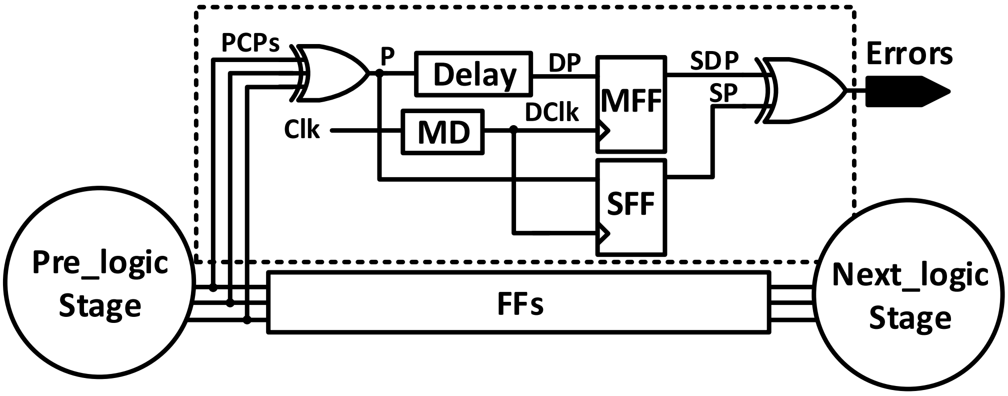
The rest of the chapter is organized as follows: Section 3.2 outlines the design principles of the PCC. Verification results and a cost analysis are discussed in section 3.5. Finally, conclusions are drawn in section 3.6.

3.2 PCC Design Principles

This section outlines the operating principles of the proposed circuit. The main advantage over existing sensors is that the PCC can be used for multiple path delay fault prediction; hence it requires less area overhead. Furthermore, compared with latch-type sensors, implementing PCC does not influence the functionality of the original design, as it does not need to replace the FFs on the monitored paths nor add buffers to the short paths of the original design.

Assume that a group of data from monitored PCPs is handled as a single number. Transitions on a PCP will change the parity of that number, from even parity to odd parity or from odd parity to even parity. A delay fault can be predicted when this change is captured before a clock rising edge. The architecture of the PCC is shown in Figure 3.1. The PCC consists of one multiple-input XOR gate, a delay element, a matched delay element, one main FF, one shadow FF and a 2-input XOR gate. The multiple-input XOR gate checks the parity of the input from the PCPs. The output signal ‘P’ is not able to represent the parity at the current time because of the propagation delay of the multiple-input XOR gate. This will cause a phase shift in the detection window. The matched delay element matches the delay of the multiple-input XOR gate to compensate for this phase shift.

Compared with Canary, which detects the data consists of a single PCP, the PCC checks the parity consistency to monitor multiple PCPs simultaneously. Hence, with an increase in the number of PCPs, the PCC implementation will not lead to a rapid growth in overheads. However, the parity of two sample point will remain the same if an even number of transitions from PCPs occur within the detection window and the delay fault behaviour will be unpredictable in this case. Nevertheless, PCC predicts the delay faults when transitions from PCPs approach the clock rising edge. Therefore it is not necessary to predict every single delay fault on the paths that the PCC is monitoring.



MD: Matched Delay element
MFF: Main FF, SFF: Shadow FF

Figure 3.1: Architecture of The Proposed Circuitry

Table 3.1: Percentage of Transitions from 4 paths

Activity Rate	α	10%
One Paths Transition	$C(1,4) \times \alpha \times (1-\alpha)^3$	29.16%
Two Paths Transitions	$C(2,4) \times \alpha^2 \times (1-\alpha)^2$	4.86%
Three Paths Transitions	$C(3,4) \times \alpha^3 \times (1-\alpha)^1$	0.36%
Four Paths Transitions	$C(4,4) \times \alpha^4$	0.01%
No Transition	$C(4,4) \times (1-\alpha)^4$	65.61%

In reality, it very unlikely that the transitions from different PCPs are 100% correlated with each other. Table 3.1 shows the percentage of transitions from 4 monitoring points with a 10% activity rate, [54]. As the Table shows, there is a 29.52% probability that an odd number of transitions occurs and 4.87% probability that an even number of transitions occurs during the system operating time. The delay fault prediction rate is 95.13% when a single PCC monitors four paths simultaneously. Moreover, the delay faults will be unpredictable if and only if an even number of transitions occur in the detection window, as shown in Figure 3.2 and 3.3. Therefore the actual delay-fault prediction rate would be higher than 95.13%. A delay fault will be eventually detected by the PCC when an odd number of transitions occurs.

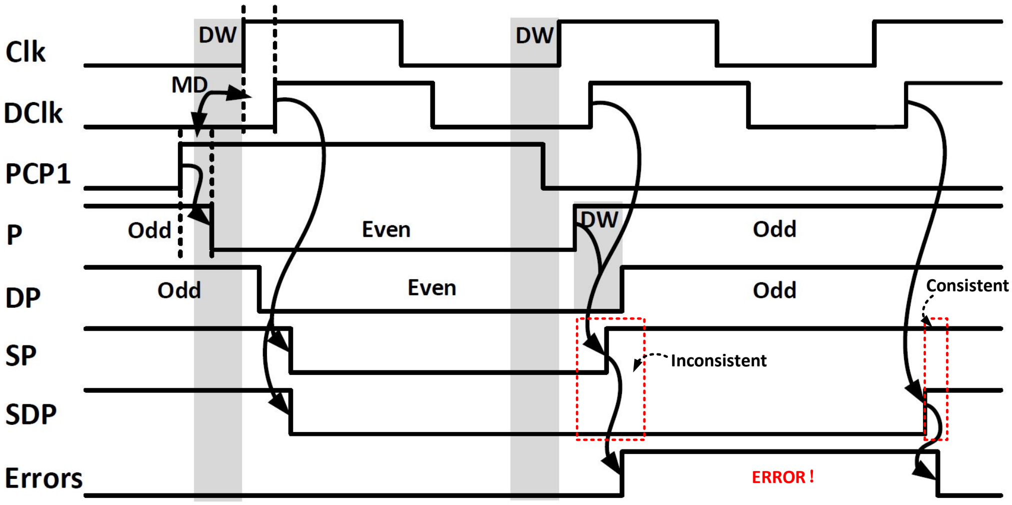


FIGURE 3.2: Odd Number of Transitions

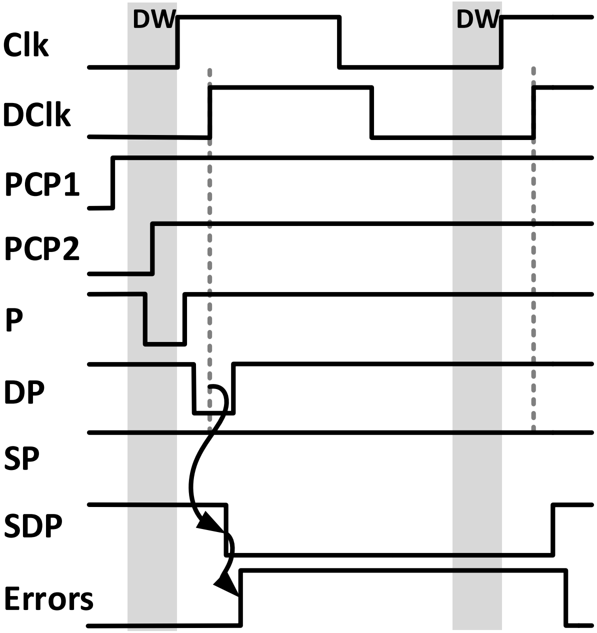


FIGURE 3.3: Even Number of Transitions

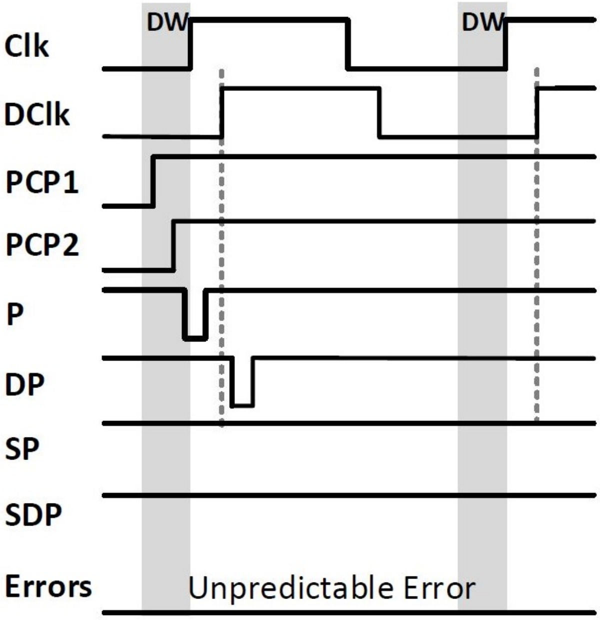


FIGURE 3.4: Unpredictable Error

Figure 3.2, Figure 3.3 and Figure 3.4 shows the timing diagram of the PCC when it monitors a group of paths simultaneously, where CLK is the clock signal of the system, $DClk$ is the output signal of the matched delay element, $PCP1$ and $PCP2$ are the output signals from two different PCPs, P is the parity status of the output signals from the PCPs ('0' for even parity, '1' for odd parity), DP is the delayed parity status (which generates the detection window (DW)), SP and SDP are the output signals of the main FF and the shadow FF, MD is the delay generated by the matched delay element, and $Errors$ is the delay fault prediction flag of the PCC. There are three typical operating cases during the PCC operation time, shown in Figure 3.2,Figure 3.3 and Figure 3.4 , respectively.

Figure 3.2 The PCPs might share the same ends with other short paths. In the first clock cycle, a short path exists and $PCP1$ changes from '0' to '1' before the detection window (the other PCPs remains the same). This conversion causes the signal P to switch from odd parity to even parity. The error signal is not triggered as both P and DP switched to '0' before DClk rises. In the second clock cycle, a PCP is asserted and $PCP1$ switches within the detection window, which results in the signal DP changing from even parity to odd parity after the $DClk$ rising edge. The error signal is triggered due to inconsistent sampling between the main FF and the shadow FF. This error signal is then cleared in the next clock cycle.

Figure 3.3 In this case, two PCPs are asserted in the same clock cycle. $PCP2$ reaches the detection window before $PCP1$. Transitions from those two paths trigger a pulse in signals P and DP. Signal P switches back to odd parity before $DClk$ rises and signal

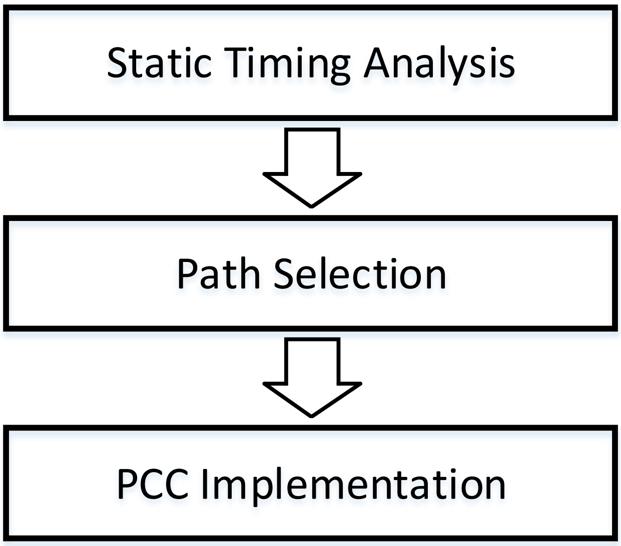


Figure 3.5: PCC Implementation Process

DP remains at even parity when *DClk* is rising. The error signal is then triggered as the inconsistency is captured by the main FF and the shadow FF.

Figure 3.4 When the transitions from two PCPs occur in the detection window, pulses on signals *P* and *DP* will be generated before and after the *DClk* rising edge. The inconsistency will not be captured by the main FF and the shadow FF, thus there is an unpredictable error. This is a rare situation, which will not arise every time from a group of PCPs are monitored. The error will be eventually predicted when an odd number of PCPs are asserted (see Table 3.1).

In practice, the routing area overhead should also be considered. Compared with existing sensors such as Canary, there is more routing on the input side of the PCC as the multiple-input XOR gate needs to be connected to the output signals of the PCPs. However, existing sensors will have more routing on the output side to manage error signals. The implementation of PCC will not lose the advantage in this case.

Figure. 3.5 shows the PCC implementation process. As the figure shows, the first step of the process is the static timing analysis for a well-synthesised design. The second step is path selection, which identifies the PCPs. The final step is the PCC implementation. The details of path selection and PCC implementation is shown in Section 3.3 and 3.4.

3.3 Path Selection for PCC

3.3.1 Path Selection

To present the path selection in a PCC implementation, we have considered a 32-bit pipelined MIPS. The PCPs were identified by performing a detailed timing analysis using a 65nm technology. The PCPs can be classified as follows:

I. Data is written back to the specific bits of different addresses. The data will not be written to different addresses in the same clock cycle, thus is 100% de-correlated.

II. Data is written back to the specific bits of the same address or read from the register file. The critical path will be active while the data changes from '1' to '0' or '0' to '1'. The PCPs switching rate will be 0.25 if the signal probability is 50%. Figure 3.6 shows the transition probability of PCPs, where α is the PCPs switching rate and n is the number of PCPs, according to Equations (1), odd transitions, (2), even, and (3), none, respectively. As Figure 3.6 shows, the probability of odd transitions is generally higher than even transitions. The correlation between PCPs increases with the number of PCPs. In practice, the correlation rate over a certain amount of time determines the effectiveness of the delay fault prediction. In the worst case, the probabilities of even and odd transitions are 50% with a 99% confidence level and ±4% interval in every 1000 clock cycles. Therefore, the number of PCPs which are detected by a single PCC is dependent on the accuracy constraints and error sampling size of the design.

$$\sum_{k=0}^{i} C(2k+1,n)\alpha^{2k+1}(1-\alpha)^{n-(2k+1)}, (i \leq \frac{n-1}{2}) \tag{1}$$

$$\sum_{k=1}^{i} C(2k,n)\alpha^{2k}(1-\alpha)^{n-2k}, (i \leq \frac{n}{2}) \tag{2}$$

$$(1-\alpha)^{n} \tag{3}$$

III. Data is assigned to the carry chain in the ALU. This might cause a few inputs of the FFs at the end of EX stage to change in the same clock cycle. The correlation rate between transitions will be high in this case, as those paths share the carry chain. However, path sharing means the same work load, and hence the transistor stress of those paths will be about the same. The ranking of those paths is most likely not affected by ageing, which means that only one of them needs to be monitored. In some particular cases, the correlated paths should be monitored by different PCC to ensure decorrelation of the signals.

IV. A critical path under monitoring is never activated under a specific workload; Some transistors from this path are under static NBTI, which means the path will be aged

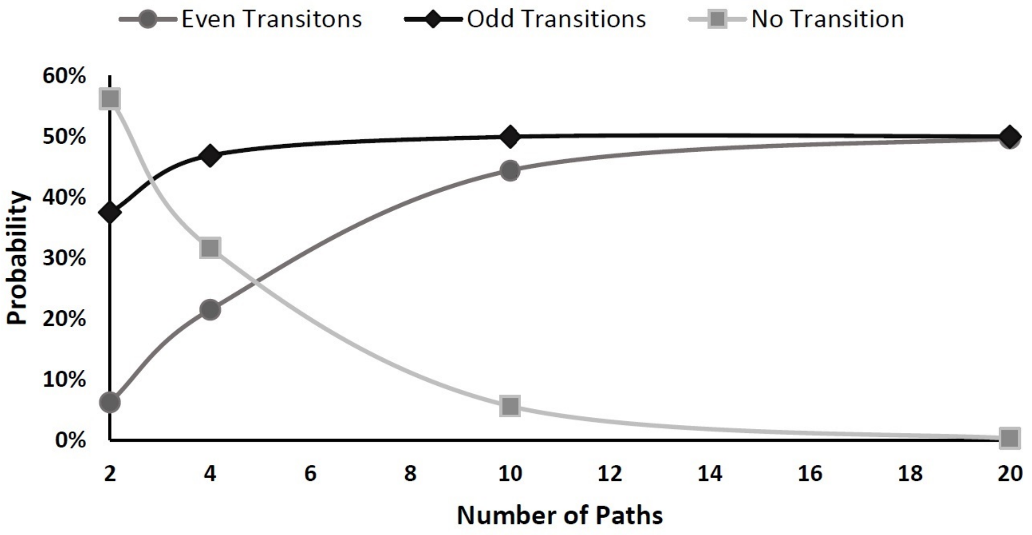


Figure 3.6: Percentage of Transitions

significantly. By changing the workload, the path is activated and produces actual errors, before the PCC has the chance to predict the failure. The delay degradation from ageing increases slowly over the years, and it can be 20% over ten years, [33]. Assuming that the operating frequency of an IC is 500MHz and a path is not activated in one day. In this case, the switching activity of this path will be less than 2.31×10^{-12}, and a circuit will not be aged in one day. Therefore, this scenario is almost impossible unless the design of this IC is immature.

To monitor N paths, the compare unit has N-1 2-input XOR gates and the data propagates through $\lceil \log_2 N \rceil$ XOR gates. The increase in N will lead to an increase in the prediction margin, therefore the prediction margin can be larger than the original safety margin, reserved for PVTA variations, if N is large enough. Hence, the number of paths monitored by one PCC should be limited according to the application. In Dynamic Voltage Scaling(DVS) approaches, the prediction margin should be slightly larger than the delay degradation of a minimum voltage drop, which ensures the true error will not occur after the voltage scaling and guarantees the system is running at the lowest voltage level. In this case, limiting the number of paths monitored by same PCC or inserting a clock buffer for the stability checkers is required. Therefore more than one PCC is required for delay fault prediction. The routing area overhead might increase with the increase of N if the distances between the monitoring points are too large. Therefore, the PCC should choose a group of close paths for the delay fault monitoring (e.g.one PCC only monitors the critical paths from the same stage in the pipelined MIPS).

The delay element determines the width of DW, the minimum scaling the supply voltage (VDD) unit is an essential index for the determination of the DW. Figure 3.7 shows the timing diagram before and after the VDD scaling down. As the figure shows, the width of DW should be wide enough to ensure the actual error will not occur after the VDD scales

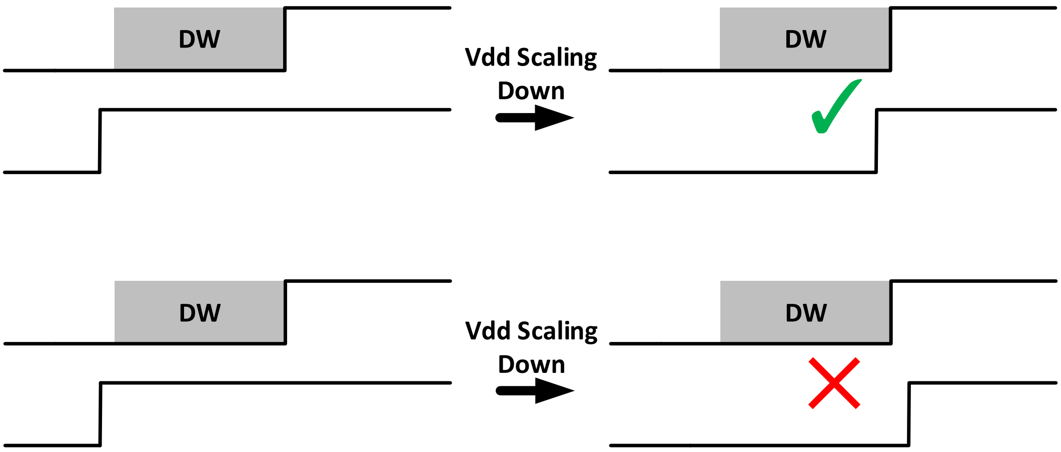


FIGURE 3.7: Determining the Width of Detection Window

down. However, the DVS will lose its advantage if the DW is too wide as the advantage of the DVS is to remove the unused margin for PVTA variations. Furthermore, the ageing effect will increase the width of the DW. Therefore, the width of the DW should be controlled as small as possible under the premise that the actual error will not occur after the VDD scales down.

The delay fault monitoring sensors would not be implemented on every path. A limited set of paths should be selected, [17]. Timing-dependent ageing and process variations should be considered as they may change the ranking of the critical paths. A PCP may become the critical path after fabrication due to process variations, or after a certain time because of ageing. Various ageing models are available for timing dependent ageing variations [58]; the potential critical paths after ageing can be identified using those models. The range of behaviours due to process variations can be estimated by applying the data provided for different technologies using the worst and best case models, see the details in the next Section.

As an external sensor, PCC would be implemented on the inputs of FFs on the PCPs. Therefore the load capacity of the last gate on the potential critical path may require adjustment. The safety margins need to be defined according to the variability analysis of each sensor implementation. The delay from different inputs of the multiple-input XOR gate should be balanced before implementation.

3.3.2 Potential Critical Path Identification

The first step in any delay-fault prediction technique is to identify the long delay paths to be monitored. These refer to the critical paths of the circuit under consideration, in addition to all long delay paths that may cause timing errors due to ageing and process variation induced delay degradation. This Section shows the result of potential ageing critical paths identification. An ageing model [32, 33] in 65nm was applied to estimate the potential critical paths after considering the ageing variations (BTI). From previous work, a timing monitoring sensor would not be implemented on every path. A limited set of paths should be identified, [17, 66]. In order to minimise power and area overheads, the sensors are implemented on the critical path and near-critical paths. A specific timing analysis is required for critical path identification. Static Timing Analysis tools can generate the timing report.

On the other hand, timing dependent ageing degradations and process variations may change the ranking of the critical paths. The range of behaviour due to process variations can be estimated by Spice using worst and best case models. Timing dependence ageing degradations, such as BTI, manifests itself as the degradation of the threshold voltage and a decrease of drain current. This degradation increases the processing time of logic gates. A more accurate way to identify the potential NBTI critical path is to estimate the threshold voltage degradation at a specific time of each component, then integrate that data with the design at time 0. All of the influencing factors need to be considered such as temperature, stress duty cycle and Vgs. Figure 3.8 shows the potential BTI critical paths estimation process.

As Figure 3.8 shows, the first step in lifetime prediction is to estimate the potential critical paths after process variation, as a near critical might become the critical path after fractionation. The second step is to estimate the signal probability (duty cycle), from input patterns and logic circuit net-list. Then estimate lifetime using the RD model. Finally, the estimation will be compared with the ranking with the other estimated results. A new potential critical path will be added when a near critical becomes the critical path after ageing.

Figure 3.9 and Figure 3.10 shows the schematic of 4-bit signed full adder with a loose and tight timing constraint respectively. As the Figure shows, the circuit is synthesised differently with different timing constraint. In practice, the timing constraint should be set as small as possible to improve the performance of an IC. As the Figure shows, there are 5 monitoring point in this circuit. To identify how many paths are considered to be the potential critical paths, timing analyses at time zero, with different process variations and after ageing are required. Figures 3.11 to 3.14 shows that the potential critical paths identification by applying the ageing models in different circumstances.

Figure 3.11 shows the ranking of the critical path after static timing analysis with normal process variation. As the Figure shows, 'out[3]' is the critical paths at time zero. Figure 3.12 shows the ranking of the critical path after 10 years ageing (the same circuit with Figure 3.11). As the Figure shows, the ranking is changed after ten years. 'out[1]' becomes the circuital path during its ageing.

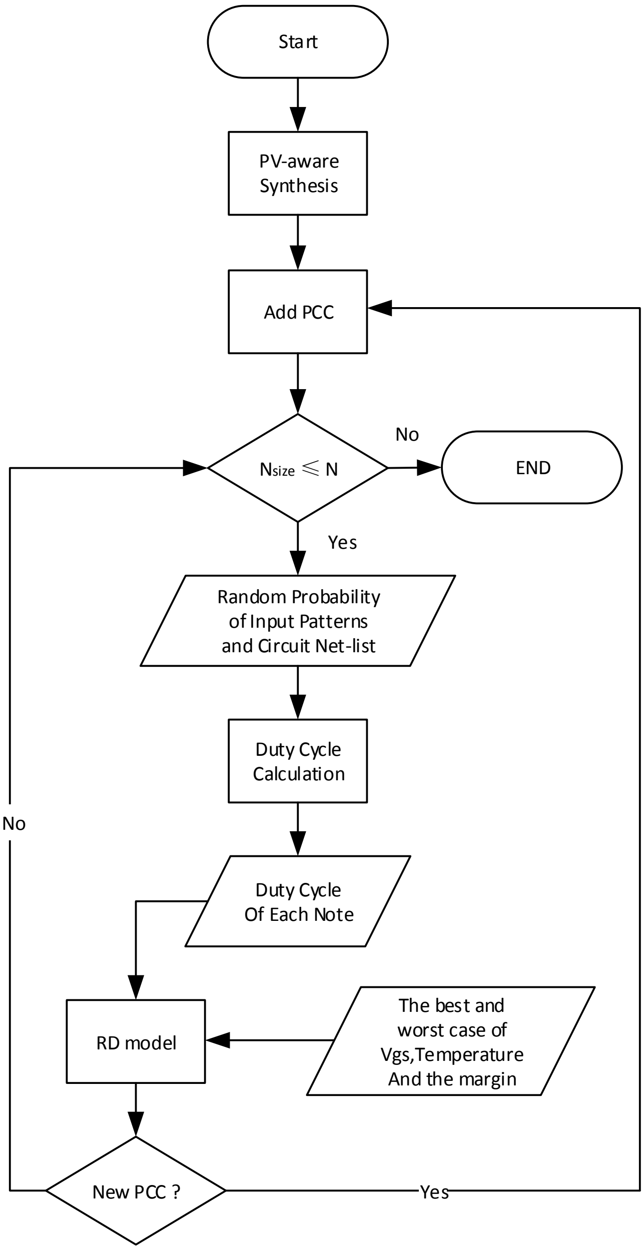

Start
PV-aware Synthesis
Add PCC
Nsize ⩽ N
No
END
Yes
Random Probability of Input Patterns and Circuit Net-list
Duty Cycle Calculation
No
Duty Cycle Of Each Note
RD model
The best and worst case of Vgs,Temperature And the margin
New PCC ?
Yes

Figure 3.8: Potential BTI Critical Paths Estimation Process

Figure 3.13 shows the ranking of the critical path in the worst case of process variation. As the Figure shows, the ranking is changed after worst case of process variation compared with the normal process variation as shown in Figure 3.11. As the Figure shows, the ranking is changed because of the process variation, however, 'out[3]' is still the critical paths at time zero. Figure 3.14 shows the ranking of the critical path after 10 years (the same circuit with Figure 3.13). As the Figure shows, the ranking is changed after ten years, however 'out[3]' is still the critical paths after ageing.

It is therefore concluded, the potential critical paths identification of the four-bit full

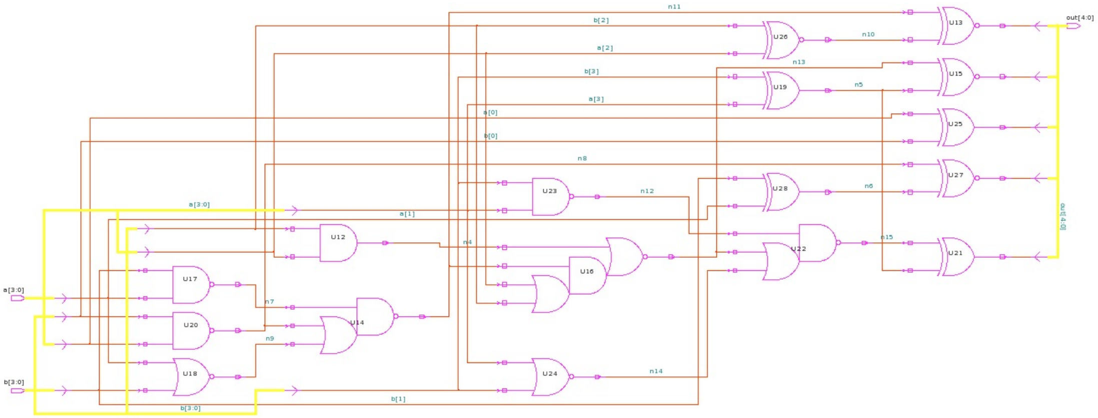


FIGURE 3.9: Synthesised 4-bit signed full adder with a loose timing constraint

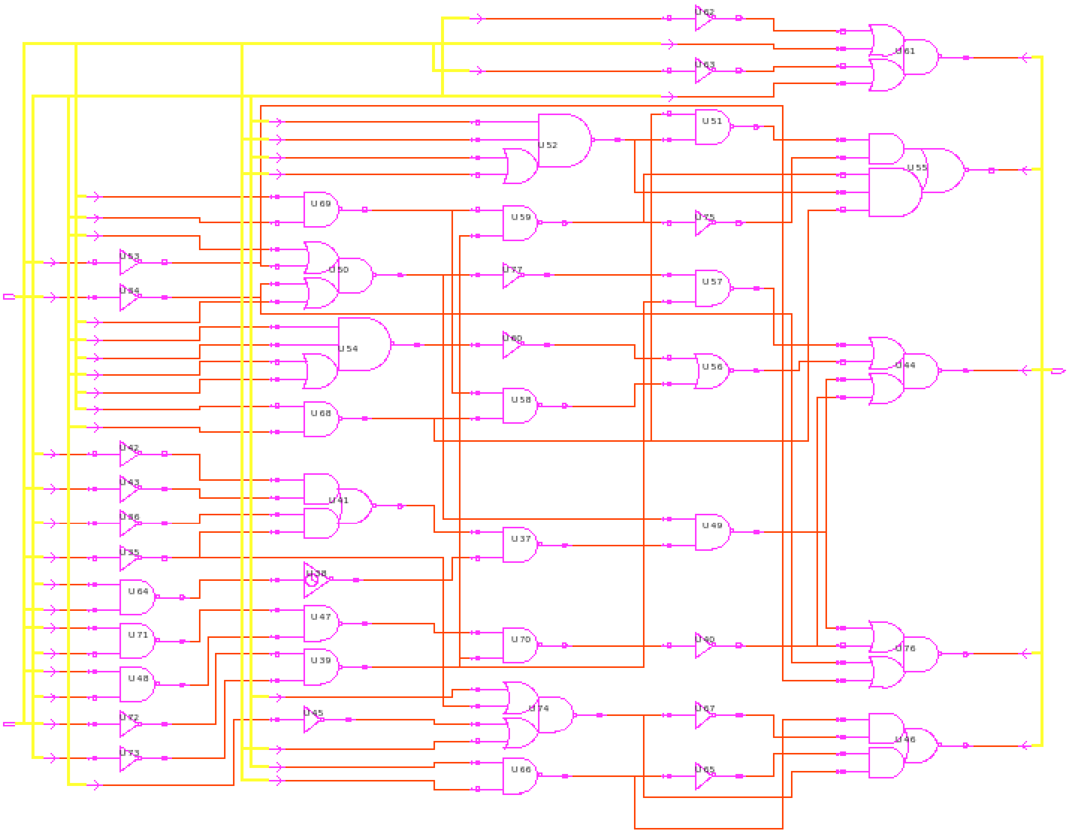


FIGURE 3.10: Synthesised 4-bit signed full adder with a tight timing constraint

adder is identified as 'out[1]' and 'out[3]'. Those two paths (2 out of 5) need to be monitored in this case.

3.4 PCC Implementation

In a stage-by-stage (pipelined) architecture, the transmission of error signals is also pipelined, shown in Fig. 3.15, [17]. The error signal is transported stage by stage until

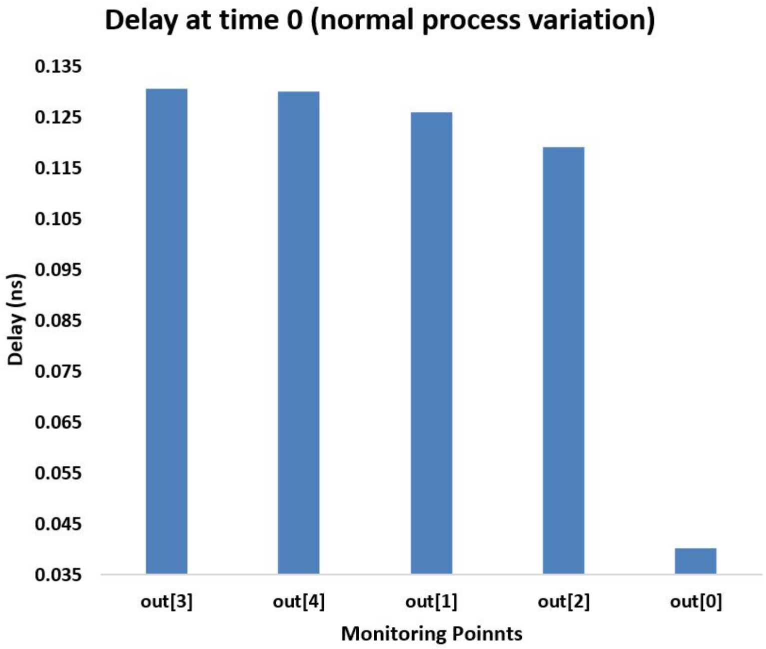


FIGURE 3.11: Delay at Time Zero (Normal Process Variation)

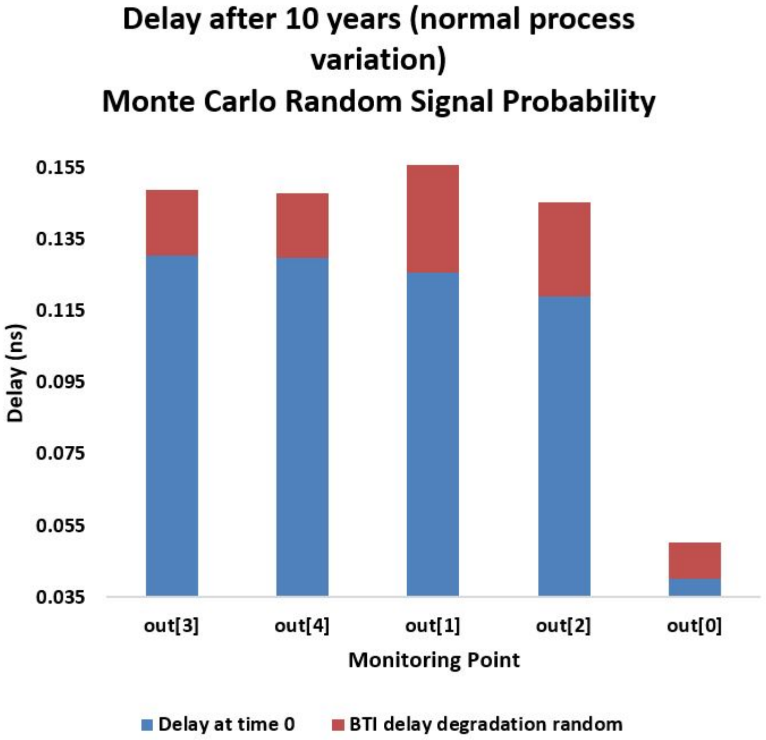


FIGURE 3.12: Delay After 10 Years (Normal Process Variation) Random Signal Probability

the DVS revives the error. The DVS then scale the supply voltage (VDD) according to the circuit's operating condition. Fig. 3.15 shows the DVS system in a pipelined MIPS architecture microprocessor (see more details in Chapter 5). The FFs in the system is

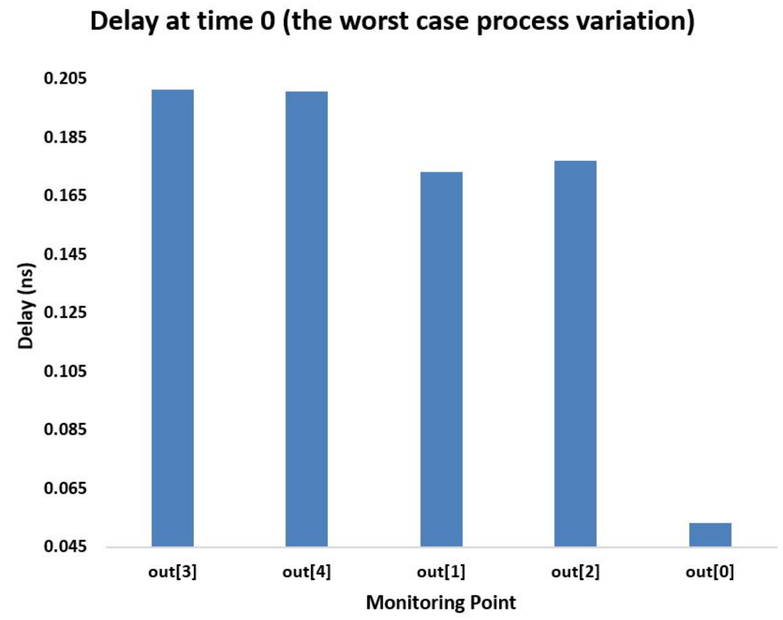


FIGURE 3.13: Delay at Time Zero (The Worst Process Variation)

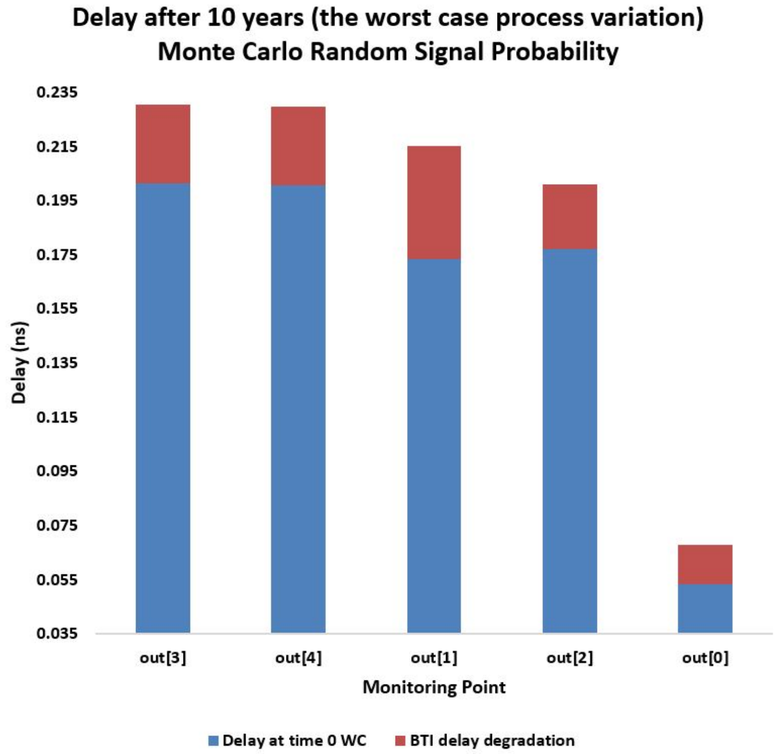


FIGURE 3.14: Delay After 10 Years (The Worst Process Variation)

designed as a clock rising edge sensitive FF. The DVS receives the global error signal at the end of the process. The error signal arrives at the same time with the data which triggers the error signal. Hence, instruction accesses can be addressed by the DVS.

In predicted, the DVS does not scale the Vdd at every clock cycles. There is a counter

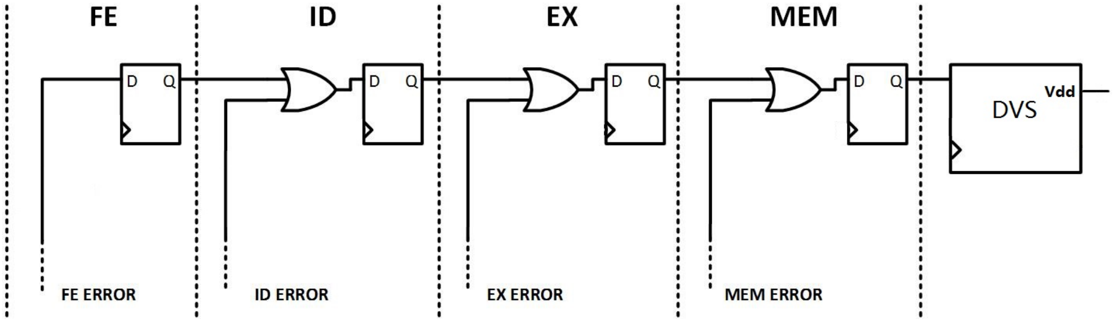


Figure 3.15: DVS in a MIPS microprocessor with a pipeline

for the error signal statistics. The Vdd is scaled up or down if the number of errors is higher or lower than a threshold in a certain number of clock cycles (e.g. 1000 clock cycles). The threshold for the PCC should be lower than the other sensors, such as Canary, as the detection rate of the PCC is not 100%. The threshold should of PCC be set as 50 out of 1000 for 20 paths monitoring if the threshold of Canary is 100 out of 1000 because of the detection rate for 20 paths monitoring is about 50%.

3.5 Verification and Comparative Analysis

This section first presents the results of functional verification at the system level, and then we summarise the cost of the proposed delay fault prediction technique when implemented in a 32-bit pipelined MIPS.

3.5.1 System level Simulation Results

The design of PCC is programmed in the Hardware Description Language, System Verilog and then synthesised by a logic synthesis tool, Design Vision, which takes HDL designs and synthesise them to gate-level HDL net-lists and generates estimated area, timing and power report of the design. Figure 3.16 shows a 9 input PCC Block Diagram after Synthesis.

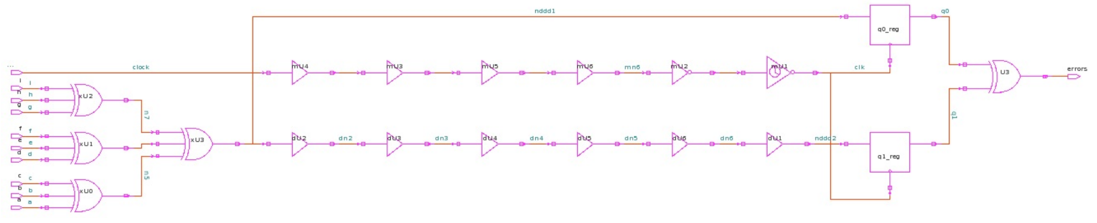


Figure 3.16: 9 Input PCC Block Diagram after Synthesis

In this example, four paths from the PCPs were selected after the timing analysis for a 32-bit pipelined MIPS in a 65nm technology. The PCC was used to monitor the 4 PCPs simultaneously for verification and evaluation.

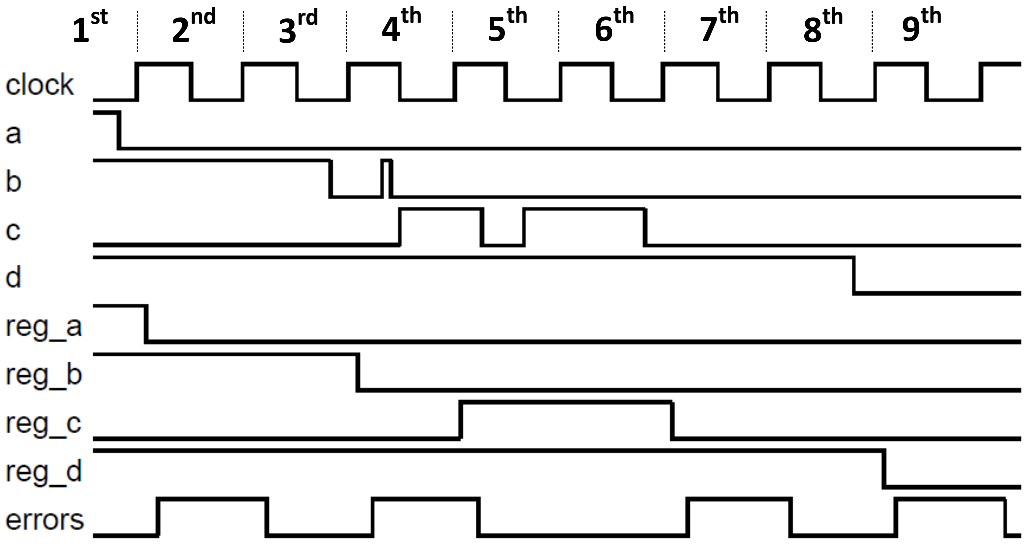


FIGURE 3.17: 4-input PCC functional verification in a 32-bit MIPS

Figure 3.17 shows the system level simulation results when a PCC monitored 4 PCPs simultaneously. Those four PCPs will be assigned when the MWB stage writes data '0' back to four different register files. The signals a, b, c and d are the output signals of those four PCPs and the signals *reg_a*, *reg_b*, *reg_c* and *reg_d* are the data which is stored in the register files. As the figure shows, transitions occur in the first, third, fourth, fifth, sixth and eighth clock cycles. Those transitions can be divided into three categories (I), (II) and (III).

(I) The PCPs were asserted in the first, third, sixth and eighth clock cycles which triggers the late transitions at monitoring points a, b, c and d, respectively. The error signals are generated when the transitions are detected by the PCC without an actual error.

(II) In the fourth clock cycle, a transition is triggered at monitoring point c by a short path. The error signal is not triggered as the data settles before the detection window.

(III) The pulse signals on monitoring points b and c in the fourth and the fifth clock cycles occur due to the competition risk between logic circuits. The the error signal is not triggered as the pulse signals occur before the detection window.

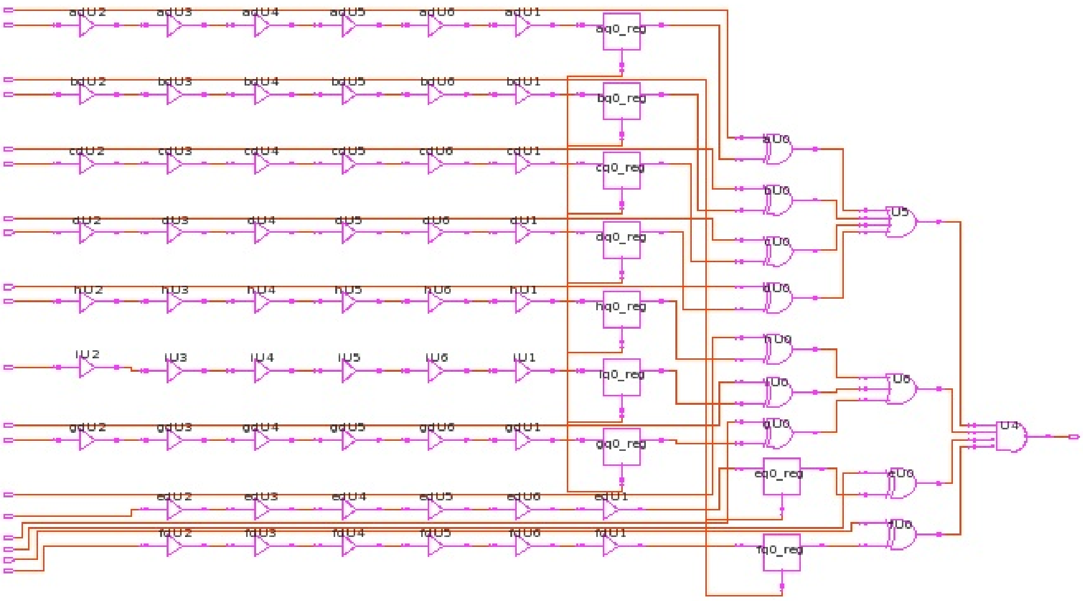


FIGURE 3.18: 9 Input Canary FF Block Diagram after Synthesis

3.5.2 Area and Power Overheads Comparison

To compare the area overhead of the proposed design PCC with Canary FF, we have considered a 32-bit pipelined MIPS. The PCC and Canary FF are designed by using exactly the same double sampling circuitry with the same width of detection window to produce an equitable comparison. The PCC and Canary FF were applied to monitor the same group of PCPs. The power and area overheads were estimated from Design Compiler using a 65 nm technology. Figure 3.18 shows the Block Diagram of Canary FF after Synthesis. Compared Figure 3.16, the Canary will cost more area than the PCC.

Figure 3.19 and Figure 3.20 shows the trends of the area and power overheads for 4, 6 and 9 path monitoring when Canary and PCC were applied to the MIPS respectively at the highest operating frequency (800 MHz) with 1.05V supply voltage. As the Figure 3.19 shows, the area overhead of PCC is generally smaller than that of the Canary FF. Compared with Canary FF, the growth in area overhead is more than 6 times slower when the PCC was applied to the MIPS. The PCC has a higher power overhead compared with the Canary FF when fewer than 6 paths are monitored, shown in Figure 3.20. This is due to the dynamic power overhead produced by the matched delay element, as the matched delay element is connected to the clock signal. However, the power overhead of Canary might be underestimated as Canary FF requires more clock buffers after place and route. PCC saves two-thirds area overhead and one-third power overhead compared with Canary FF when PCC monitors 9 paths simultaneously.

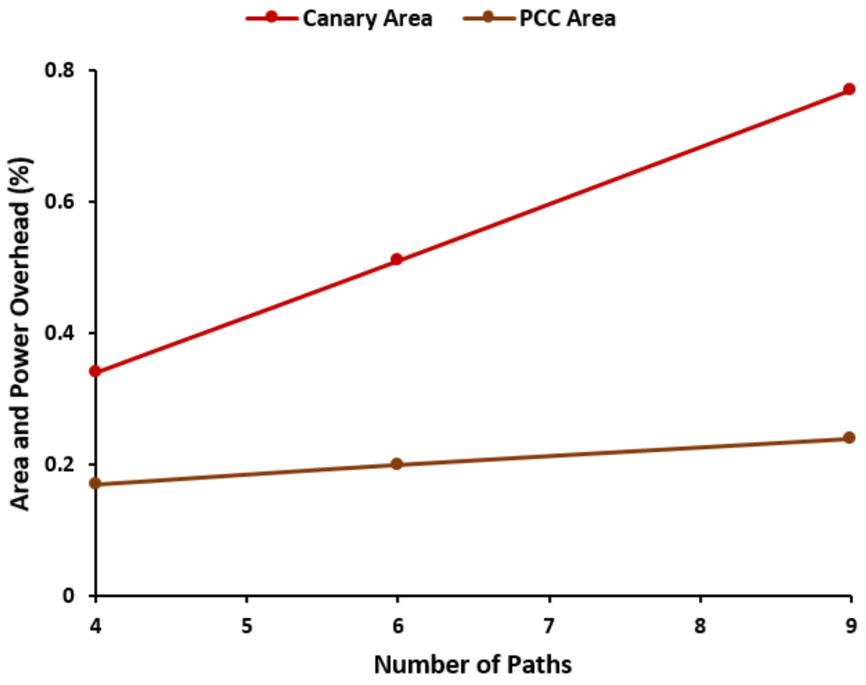


Figure 3.19: Area Overhead

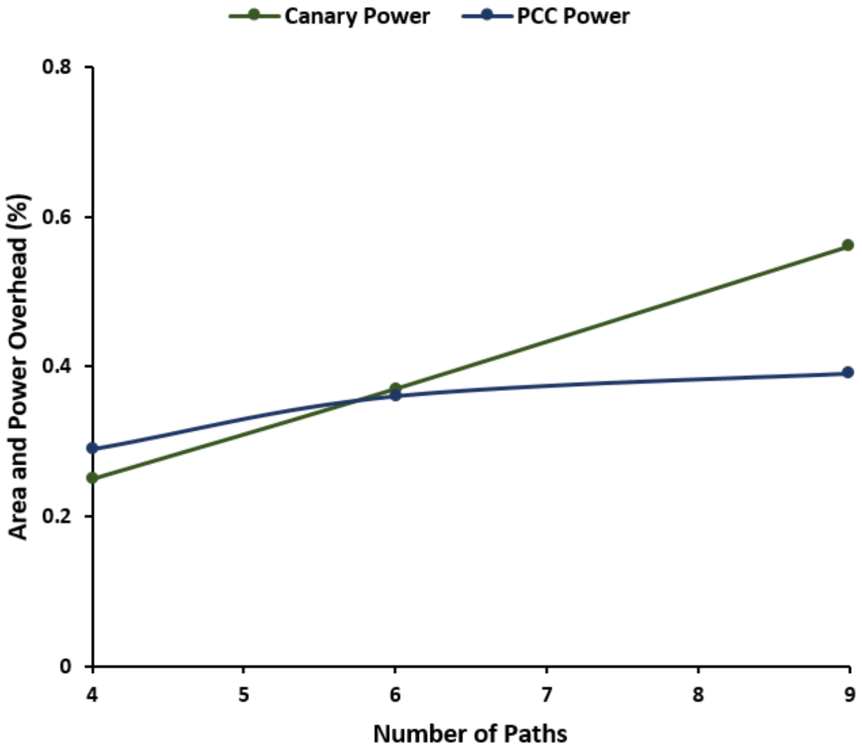


Figure 3.20: Power Overhead

3.6 Concluding Remarks

In this chapter, we have a new delay-fault prediction circuitry, named PCC. PCC is a multiple delay-fault predictor which improves the cost and energy efficiency. Compared with Canary FF, PCC saves two-thirds and one-third of area and power overheads respectively in a 32-bit MIPS. The design was implemented and verified in a 32-bit MIPS in a 65 nm technology. The PCC is a gate level design of the multiple error prediction sensor. The cost-efficiency can be improved if the sensor was designed at the transistor level.

On the other hand, an error will not be predicted when even number of transitions are asserted in the detection window. The prediction rate can be improved if the sensor was designed by using the transition detection technique. A better version of the multiple delay fault monitoring sensor will be shown in the next Chapter.

Chapter 4

Differential Multiple Error Detection Sensor

This Chapter proposes a new ageing sensor, named Differential Multiple Error Detection Sensor (DMEDS), which capable of monitoring multiple paths concurrently. The main advantage of the DMEDS over the PCC are: The DMEDS is designed at the transistor level. Therefore, it further improved the cost-efficiency of the delay fault monitoring; The DMEDS has a higher detection rate over the PCC. DMEDS is a semi-digital and semi-analogue circuit. In order to verify the functionality of the DMEDS, the proposed sensor has been designed at transistor level using a 32 nm and a 90 nm technology nodes and applied to a 32-bit MIPS to monitor ten paths concurrently. Our results indicate that using the proposed sensor for monitoring ten paths can save 87.59% and 77.67% in area overheads compared to Razor and Canary, respectively.

4.1 Introduction

Process variations and ageing-induced device degradation are becoming major reliability concerns in modern semiconductor technologies. Both phenomena lead to performance degradation, and hence timing errors. To avoid delay fault induced failures, integrated circuits are typically designed with large safety margins, [2]. This generally means a circuit is designed for worst-case operating conditions. Such an approach may limit system performance and lead to an increase in power consumption, [58]. Dynamic reliability management schemes have been proposed to assure an IC's lifetime reliability. Such schemes are typically based on the use of sensors to predict circuit failures before errors actually appear. The system can then adaptively scale its operating frequency and supply voltage according to the actual operating conditions to compensate for performance degradation, [2].

Various in situ delay monitoring sensors have been proposed. These include delay fault detection and prediction techniques, [29, 40]. Existing delay fault sensors are usually placed on the circuit's longest delay paths. However, the increasing complexity of ICs has led to a significant rise in the number of long paths and potential ageing-critical paths that may be vulnerable to timing errors, [19]. This means the cost of in situ delay monitoring may be prohibitive. The objective of this work is to minimize the area overhead with an increase in potential ageing-critical paths.

In this chapter, we propose a new Differential Multiple Error Detection Sensor (DMEDS) for timing errors. DMEDS is able to monitor multiple paths simultaneously, which significantly reduces the number of sensors needed to monitor ageing-induced delay faults. DMEDS has been designed at transistor level in a 32nm 90 nm CMOS technology and verified at the system level. Our results indicate that the use of the proposed sensor for delay fault monitoring across ten paths can lead to a significant saving in area overhead compared to Razor [29], and Canary [40]: 87.59%, 77.67%, respectively.

This chapter is organized as follows. Section 4.2 outlines the design principles of DMEDS. Verification results and cost analysis are discussed in section 4.8. Finally, conclusions are drawn in section 4.9.

4.2 DMEDS Design Principles

The disadvantages of existing sensors are:

1. Usually placed on the circuit's longest delay paths however, the relentless increase in the complexity of silicon chips has led to a significant rise in the number of long paths which are deemed to be vulnerable to timing error. This means the cost of conventional in-situ delay monitoring approaches is becoming prohibitive.
2. Replace the FF of the original design, therefore it is hard to implement and it will influence the functionality of the original design.
3. Data from main Flip-Flop or sensor itself may suffer with the metastable state which will effete the reliability and power dissipation of the digital circuit.
4. Widely used in adaptive voltage scaling system which might bring more reliability problem during the scaling of the supply voltage of devices.

This chapter outlines the operating principles of the proposed sensor. Two versions of the design are presented: the first is a Single Error Detection Sensor (SEDS). Its main advantages over the Canary FF are that it requires less area overhead and unlike Canary FF, it does not suffer from the meta-stability error problem explained above.

The second version of the design is called DMEDS; the latter can be used for multiple path delay-fault prediction. A new implement process for Ageing Prediction System and an idea of Device Failure Warning System are also proposed at the end of this chapter.

4.3 Operating Principles of SEDS

The architecture of SEDS is shown in Figure 4.1. Compared with existing delay-fault predictive sensors such as Canary, SEDS retains the original main FF and predicts delay-faults at the same time, [40, 56, 66]. The SEDS FF consists of one delay element and a stability checker. The stability checker receives $Data\_In$ via a delay element, T_pre, which is a minimum sized inverter chain. The stability checker checks the stability of the delayed signal while the clock is '1'. Thereby the detection window of SEDS starts from T_pre before the rising clock edge until T_pre before the falling clock edge. The error signal is triggered by transitions of Data_In during the Detection Window (DW). As Figure 4.1 shows, initially the propagation delay of the path, monitored by SEDS, is shorter than $T\_clock\_cycle - T\_pre$, and no error occurs. After serious ageing, the propagation delay of the path increases due to ageing until the remaining slack is shorter than T_pre, which triggers the error signal. Meta-stability of the main FF is avoided as the error signal triggers before the setup time of the FFs is violated.

4.4 Operating Principles of DMEDS

The increase in the number of long delay paths in today's complex systems makes it infeasible to insert a delay sensor on every path. DMEDS is capable of monitoring multiple paths for delay faults. This can significantly improve the cost-efficiency. As shown in Figure 4.2, DMEDS consists of a Multiple Detection Unit (MDU) and a stability checker. The MDU monitors two or more of the potential critical paths simultaneously.

As Figure 4.3 shows, any transitions in the data will trigger a transition of the MDU output signal (from '0' to '1' or '1' to '0'). This transition will be captured if it occurs during the detection period of the stability checker, thus signalling a delay-fault. The MDU has a small safety margin from its own delay. The small safety margin ensures the signal error triggers before the real delay-fault occurs, thus enabling delay-fault prediction.

In reality, it is very unlikely that the data from two or more different paths converge at the exact same time. The MDU assumes that there will be timing differences between the transitions from different paths. The XOR gate in the MDU compares the data monitored by the DMEDS and each transition from the input data will trigger a transition at the MDU output. However, the sensitivity of the MDU circuit is not sufficiently

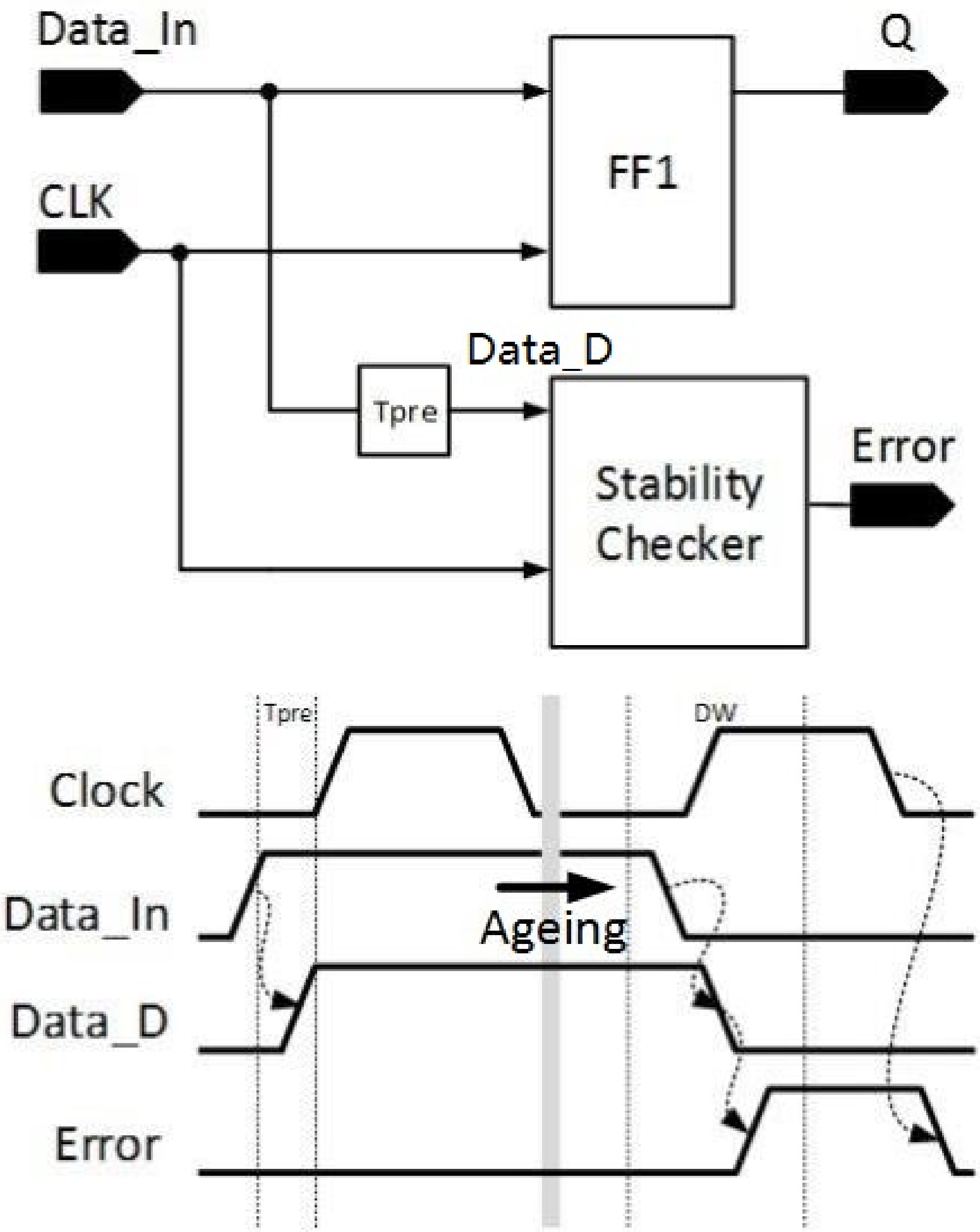


FIGURE 4.1: The Architecture and timing diagram of SEDS

high to detect an extremely small difference between the transitions of the input data. Transitions become undetectable when an even number of signals change at about the same time. In order to generate a measurable time difference between the transitions, adjustable delay elements are embedded in the MDU. The delay of the adjustable delay element can be increased or decreased by scaling the control signal 'S'. As shown in Figure 4.4 (b), the control signal 'S' will be scaled down when the delay times of two potentially critical paths become close, due to process or ageing variations, such that the difference is shorter than the measurable margin of the MDU. Actually, the data from different paths is unlikely to change in the same clock cycle every time. Without the adjustable delay elements, the MDU is able to detect timing errors when an odd

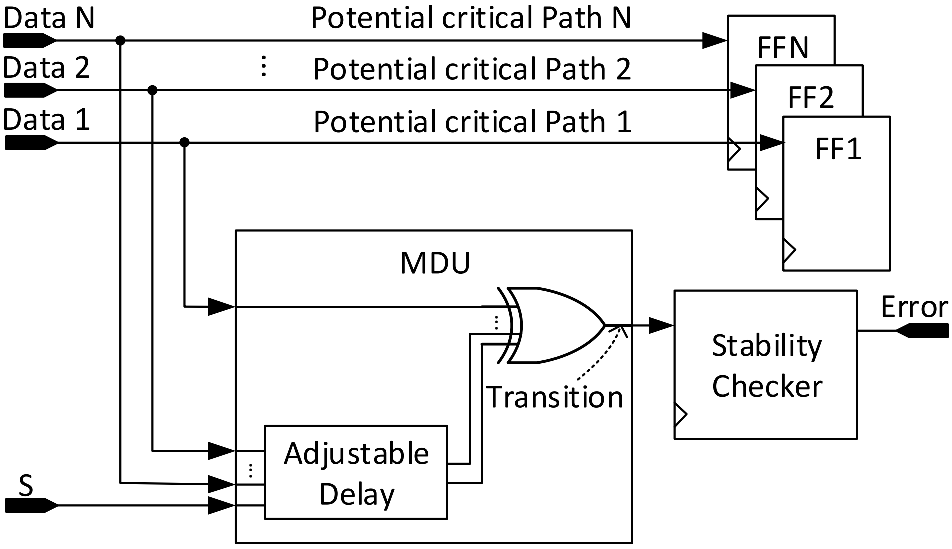


Figure 4.2: The Architecture of DMEDS

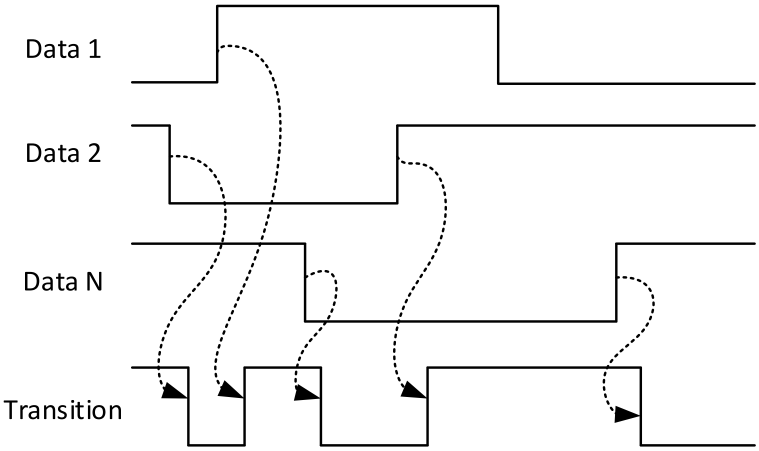
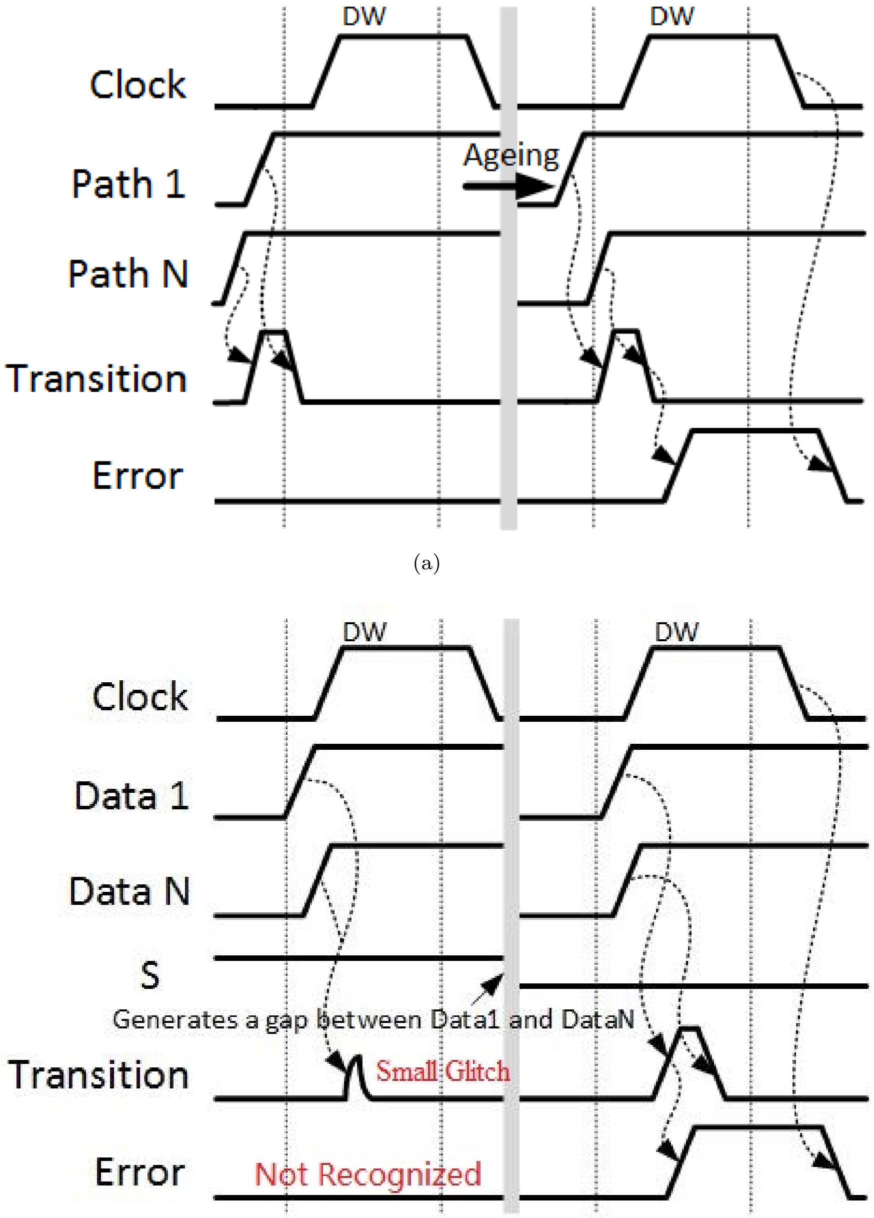


Figure 4.3: MDU Timing Diagram

number of transitions occur during the same clock cycle. In the worst case, with the adjustable delay element, the DMEDS is able to predict the delay fault if the transitions between two paths are 100% correlated, hows in Section 3.3, by Scaling the signal 'S'. (see the simulation result shown in Figure 4.16). However, it will increase the power consumption and the area overhead especially the routing area overhead of the system if the adjustable delay elements are implemented. The DMEDS will be still functional if the high-correlated paths are monitored by different DMEDS, which will further reduce the power and area overhead. The probability of undetectable errors increases with the number of potentially critical paths that the MDU is monitoring in this case. The stability checker checks the stability of the output signal from the MDU. It captures transitions of the MDU output signal during the checking period.

The checking period of the stability checker starts from the rising edge until the falling edge of the clock signal. As shown in Figure 4.4, output signal 'Error' will be triggered if the signal 'Transition' changes during the checking period and is cleared after the clock falling edge. The width of the detection window is half a clock cycle, starting from *Dmdu* before the clock rising edge, where *Dmdu* is the propagation delay of the MDU. There are two typical operating cases during the DMEDS operation time, shown in Figure 4.4 (a) and (b).



(a)

(b)

FIGURE 4.4: DMEDS Timing Diagram

(a) When the DMEDS is monitoring two or more paths simultaneously, the potential critical path, 'Path N', is ageing much faster than the critical path at time zero, 'Path 1', due to different operating conditions. Initially, the propagation delay of 'Path 1' is longer than that of 'Path N' (and other paths if more than two paths are monitored). The difference between paths triggers a pulse of the output signal 'Transition'. As the remaining slacks in any of those paths do not reach the DW yet, the error signal is not triggered. After a certain period of time, both 'Path 1' and 'Path N' age, but as 'Path N', is ageing faster, the remaining slack in 'Path N' shrinks to the detection window duration before those in 'Path 1' (and other paths). The output signal of MDU is activated by the transition in 'Path N' within the detection period of the stability checker. This conversion is then captured and triggers the error signal.

(b) In this case, the potential critical path, 'Path N', is ageing only slightly faster than the critical path at time zero, so the difference between the changes from those paths becomes closer and closer during the operation time. After a certain period of time, the remaining slack of both two paths approaches the DW, and the difference between 'Path 1' and 'Path N' is shorter than the measurable margin. The sensitivity of MDU is not sufficient to trigger a strong pulse. A small glitch is generated when 'Path 1' and 'Path N' change at about the same time. The glitch is too small to be captured by the stability checker, and a glitch is not strong enough to be recognised by the stability checker. The sensitivity of the stability checker can be improved by tuning transistor sizes, [5]. The delay-fault of both paths becomes unpredictable. By scaling the voltage of signal 'S', a more significant difference is generated between the paths. As Figure 4.4 (b) shows, a pulse is generated by the MDU even though the data changes at the same time. This pulse is then captured by the stability checker and triggers the error signal.

In practice, it is not necessary to check delay faults during the whole lifetime of the IC. A schedule can be set up for the prediction circuitry (e.g. once a day or once a week). During the checking period, the system scales the voltage of signal 'S' to monitor the ageing conditions of critical and potential critical paths. Hence, the DMEDS is also suitable for delay-fault prediction for DVS.

4.5 Transistor Level Design of DMEDS

The circuit schematics of the adjustable delay element, the stability checker and the MDU are shown in Figure 4.5, Figure 4.6 and Figure 4.7, respectively. The adjustable delay element is able to scale the delay by scaling the input signal 'S'. Unlike general delay elements consisting of an inverter chain, the adjustable delay element consists of two inverters with adjustable input NMOS transistors. The propagation delay of the adjustable delay element increases with decreasing voltage at 'S'.

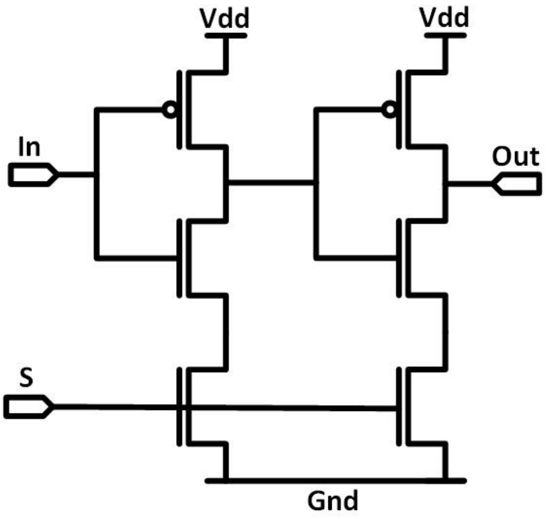


FIGURE 4.5: The Adjustable Delay Element

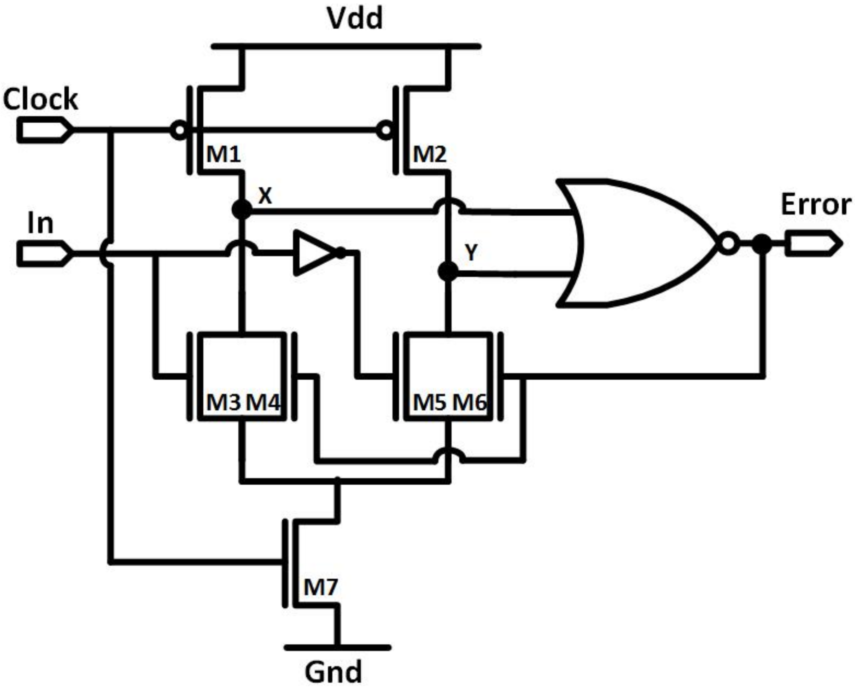


FIGURE 4.6: The Stability Checker

The stability checker, shown in Figure 4.6, is simplified with respect to the stability checkers proposed previously shown in 4.8, [56, 66]. As the name suggests, the stability checker detects whether the input signal is stable. X and Y are the input signals of a NOR gate. Figure 4.9 shows the timing diagram of the stability checker. As the figure shows, during the clock high state, the checking period, node X is pulled down when the input signal is logic '1' and node Y is pulled down when the input signal is logic '0'. The error signal will be triggered when a transition takes place at the input signal. Both nodes X and Y will be pulled up during the clock low state, which clears the error

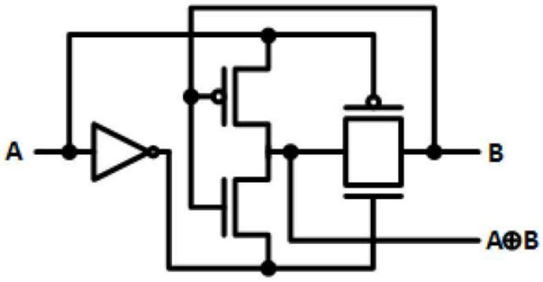


Figure 4.7: 6 Transistor XOR Gate

signal. The error signal needs to be stored as the stability checker cannot latch the error signal after the falling edge of the clock signal. Figure 4.7 shows a 2-input MDU, which is a 6 transistor XOR gate.

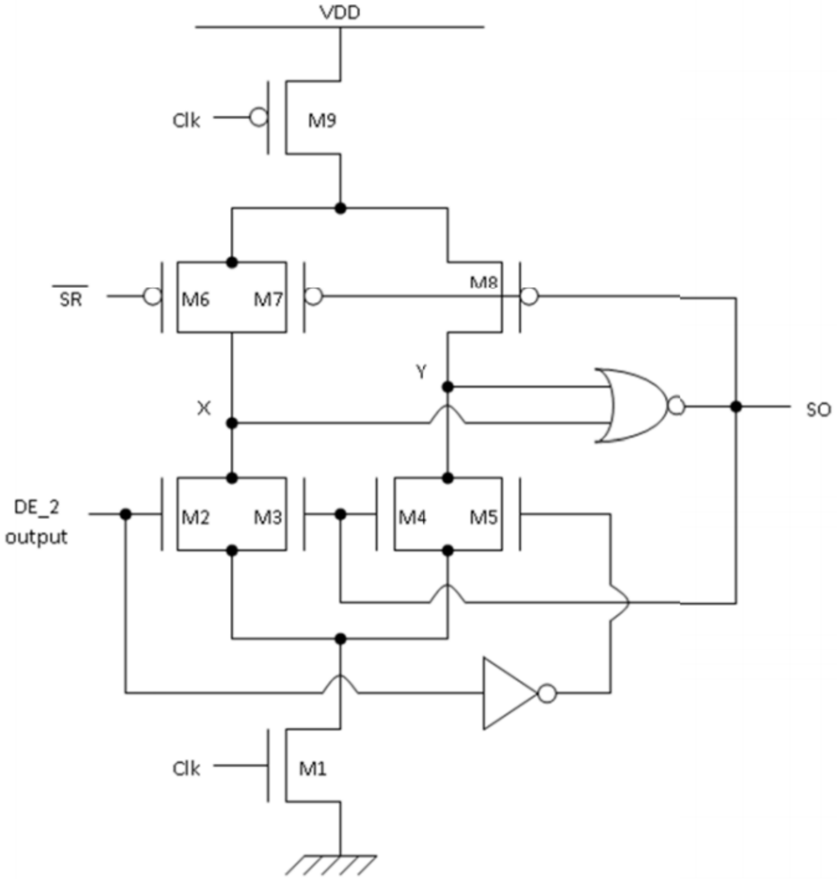


Figure 4.8: Previously Design of Stability Checker[66]

The stability checker is metastability resistant. Metastability occurs with two or more signals changes about the same time in a digital circuit. The output signal will remain between logic '1' and logic '0', an unknown signal. When output signal Error is between logic '1' and logic '0', M4 and M6 will work in a linear region, which will pull down nodes X and Y. Then the Error signal will become a strong logic '1'.

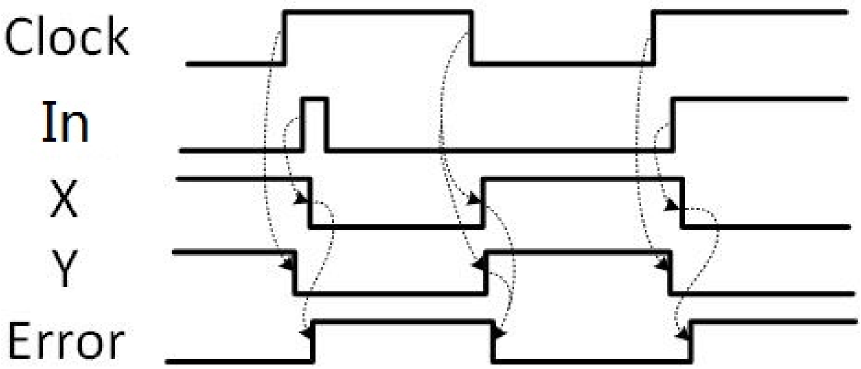


FIGURE 4.9: Stability Checker Timing Diagram

4.6 Gate Level Design of DMEDS

The Stability Checker shown in Figure 4.6 does not consist of standard logic cells. A cell library design in layout level is required to implement the DMEDS into a processor. This will take a lot of time for the layout level design and simulation for different techniques. An equivalent circuit of the stability checker is proposed to make the DMEDS sensor implementation more efficient.

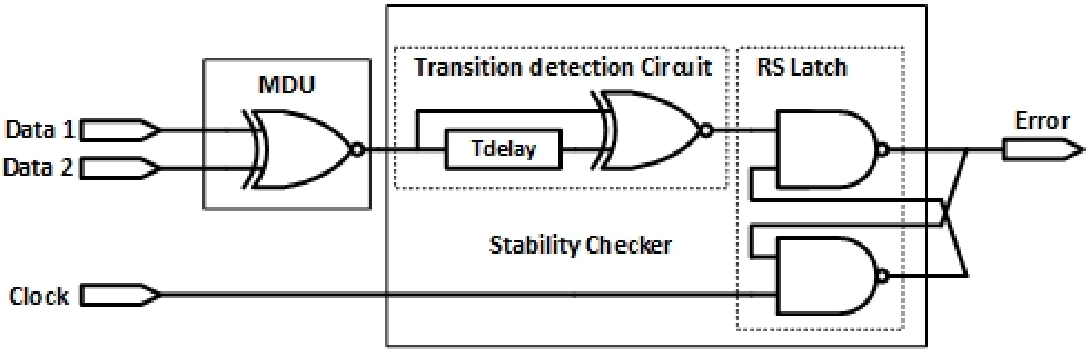


FIGURE 4.10: Gate Level DMEDS

As shown in Figure 4.10, the gate level stability checker consists of a transition detection circuit and an RS Latch. The transition detection circuit captures the transition from the output of MDU and generates a pulse signal shown in Fig 4.11. The RS latch latches the error signal during clock is logic '1', and the error signal will be cleared after the clock falling edge shown in table 4.1.

R	S	Error
0	0	1
0	1	1
1	0	0
1	1	memory

TABLE 4.1: Truth Table of the RS Latch

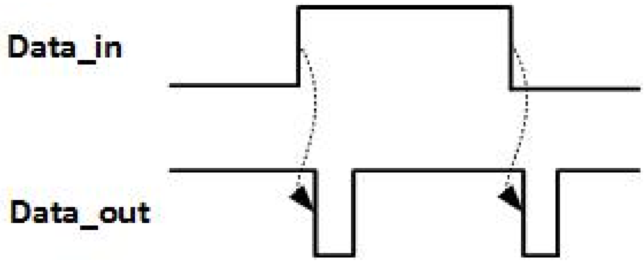


FIGURE 4.11: Gate Level Transition Detection Timing Diagram

The adjustment delay element is ignored in the gate level design of DMEDS. For an ageing prediction system, it is not necessary to detect every signal delay fault signal during the lifetime. The transitions between different paths are de-correlated. The error signal will be captured in a different combination of input.

Compared with the design shown in section 4.5, the design at gate level sacrifices the area overhead (more transistors are required in the design of gate level), but the sensor implementation process is simplified.

4.7 DMEDS Implementation

In a stage-by-stage (pipelined) architecture, usually, the transmission of error signals is also pipelined, shown in Fig. 4.12, [17]. As the error will be cleared after the clock falling edge, the Flip-Flops (FF) which are used to transmit the error signals need to be designed as falling edge sensitive FFs.

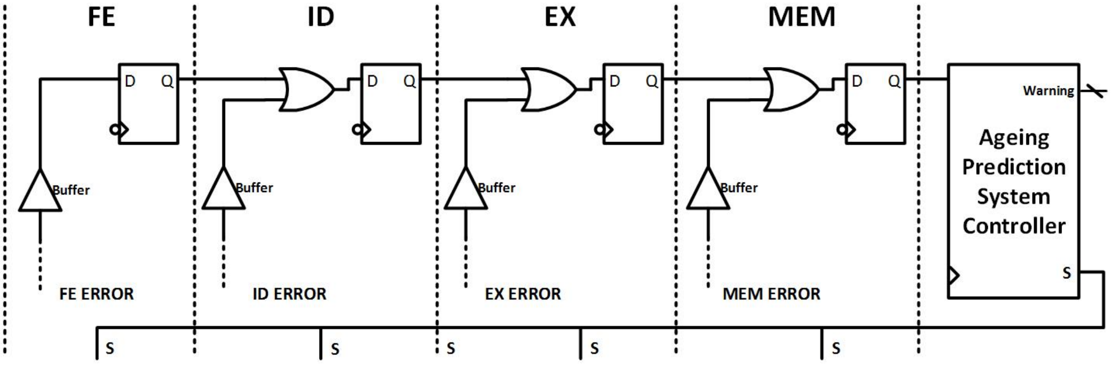


FIGURE 4.12: Ageing Prediction System in a MIPS microprocessor with a pipeline

Fig. 4.12 shows the Ageing Prediction System(APS) system in a pipelined MIPS architecture microprocessor (see more details in Chapter 5). The FFs in the system is designed as a clock falling edge sensitive FF which stores the error signal at the clock falling edge. The global error signal will be received by ageing prediction system controller at the end of the process. The error signal arrives at the same time with the data which triggers the error signal. Hence, instruction accesses can be addressed by the APS. A signal S is generated by the controller to scale the delay of the adjustable delay element in the comparator. The situation of unpredictable timing error is avoided, by scaling the adjustable delay element randomly. The warning signal will be sent the predicted lifetime to the user after the lifetime is predicted (See more detail about ageing prediction in the next Chapter). A buffer is implemented after the timing error sensor. The role of the buffer is to give the FFs enough time, setup time and hold time, store the error before it goes away. The timing diagram of the buffer is shown below.

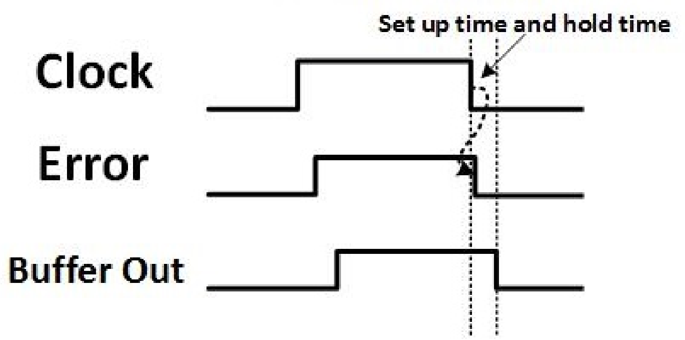


Figure 4.13: Timing Diagram of Buffer

As an external sensor, DMEDS is implemented on the inputs of FFs on the potential critical paths. Therefore the load capacity of the last gate on the potential critical path may require an adjustment. The safety margins need to be defined according to the variations analysis in each sensor implementation. Hence, the delay of the MDU also needs an adjustment in different cases.

It is therefore concluded that the delay fault detection and prediction sensors which were proposed previously need to replace the FF of the original design, which is not easy to implement and it will not influence the functionality of the original design. All of them can only detect the timing error for one path. Therefore the area cost for multiple path monitoring might be prohibitive.

4.8 Verification and Comparative Analysis

This chapter first presents the results of functional verification, and then we summarise the cost analysis of the proposed delay fault monitoring technique when implemented in a 32-bit pipelined MIPS.

DMEDS is a semi-digital and semi-analogue circuit. A semi-analogue circuit might work differently in different technology node as the drain current and capacitance are different when feature size changes. The proposed sensor has been designed at the transistor level using a 32 nm and a 90 nm technology nodes to verify the functionality of the DMEDS. Moreover, The DMEDS is also applied to a 32-bit MIPS to monitor ten paths concurrently.

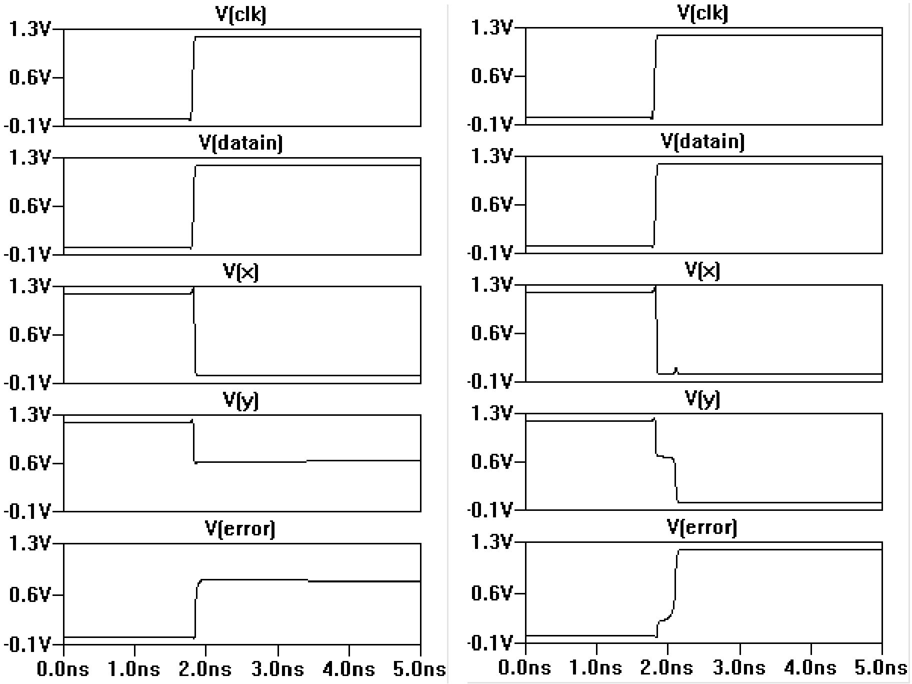
4.8.1 Transistor Level Verification and Analysis in 90 nm

This section presents the transistor level functional verification and comparative analysis at 90nm technology. The transistor level design of DMEDS was implemented to predict the delay fault from two paths simultaneously. The cost analysis calculated the amount of the transistor when the DMEDS were monitoring two paths on the EX Stage in a 32-bit pipelined MIPS. The design is described by transistor-level net-list and simulated by Spice.

4.8.1.1 Transistor Level Simulation Results in 90 nm

The DMEDS was implemented in a 90nm technology for verification and evaluation. Figure 4.15 and Figure 4.16 show the Spice simulation results in different situations when the inputs of two conventional FFs were monitored simultaneously.

Section 4.5 stated that the stability checker shown in 4.6 is metastability resistant. Transistor M4 and M6 will pull down nodes X and Y to make the output signal error a strong logic '1'.



(a) (b)

FIGURE 4.14: Metastability Resistance Simulation result in Spice

Figure 4.14 (a) the simulation result of transistor-level stability checker design without the pull-down transistors M4 and M6. When the input signal 'datain' and 'clock' changes about the same time, the metastability occurs as node Y remains between logic '1' (1.2v) and logic '0' (0V). In Figure 4.14 (b) the simulation result of transistor-level stability checker design without the pull-down transistors M4 and M6. When the input signal 'datain' and 'clock' changes about the same time, node Y remains in the metastable state momently, then pulled down by transistor M6.

Figure 4.15 shows the situation when the delay between potential critical paths is still greater than the measurable margin, and both transitions from the input of FF switch a little before the clock rising edge and the setup time. 'clk' is the clock signal of FFs and the DMEDS. 'din1' and 'din2' are the data input signal of the two FFs respectively. 'q1' and 'q2' are the data output signal of the two FFs. 'cout' is the output of the MDU. 'error' is the output signal of the stability checker.

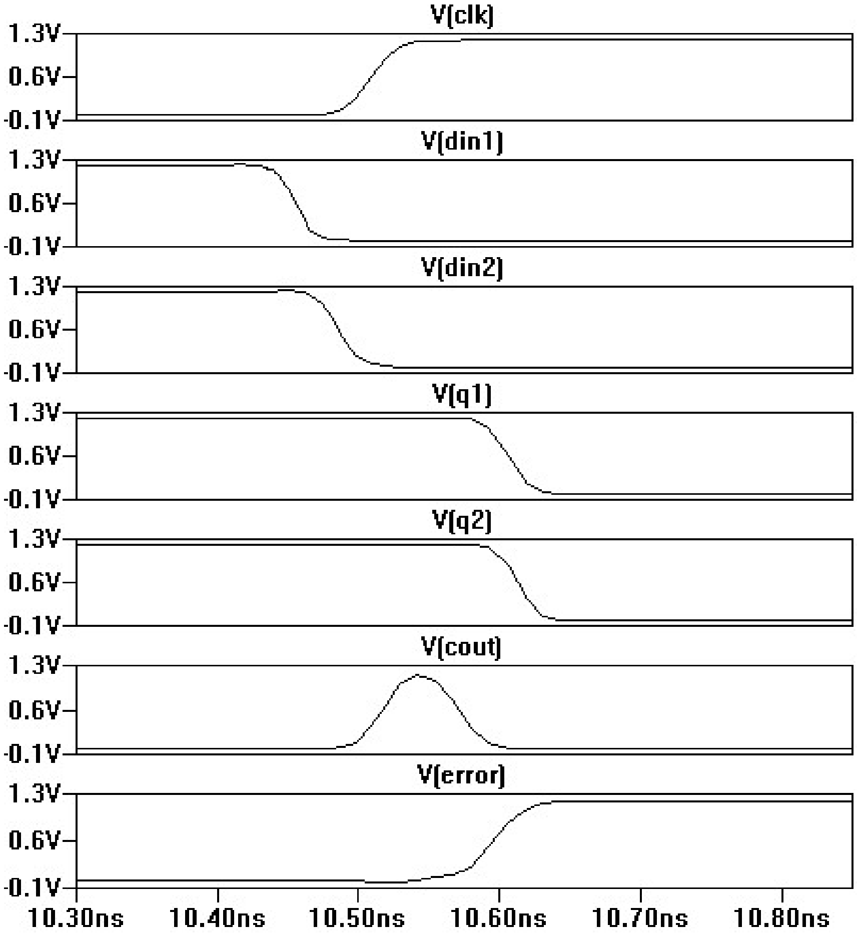


FIGURE 4.15: Transistor Level Multiple Timing Error Detection Simulation result in Spice

As Figure 4.15 shows, 'din1' and 'din2' switch a little before the rising clock edge. The MDU generates a pulse during the stability checking period by the difference between ‘din1’ and ‘din2’. The stability checker captures the pulse and triggers the error signal. As FFs receive ’din1’ and ’din2’ after the rising clock edge, there is no timing error.

Figure 4.16 shows the situation when the delay between potential critical paths is still less than the measurable margin and both transitions from the input of FF switch a little before the clock rising edge and the setup time. ’clk’ is the clock signal of FFs and the DMEDS. ’din1’ and ’din2’ are the data input signal of two FF respectively. ’q1’ and ’q2’ are the data output signal of two FF. ’S’ is the input the adjustable delay element. ’cout’ is the output of the MDU. ’error’ is the output signal of the stability checker.

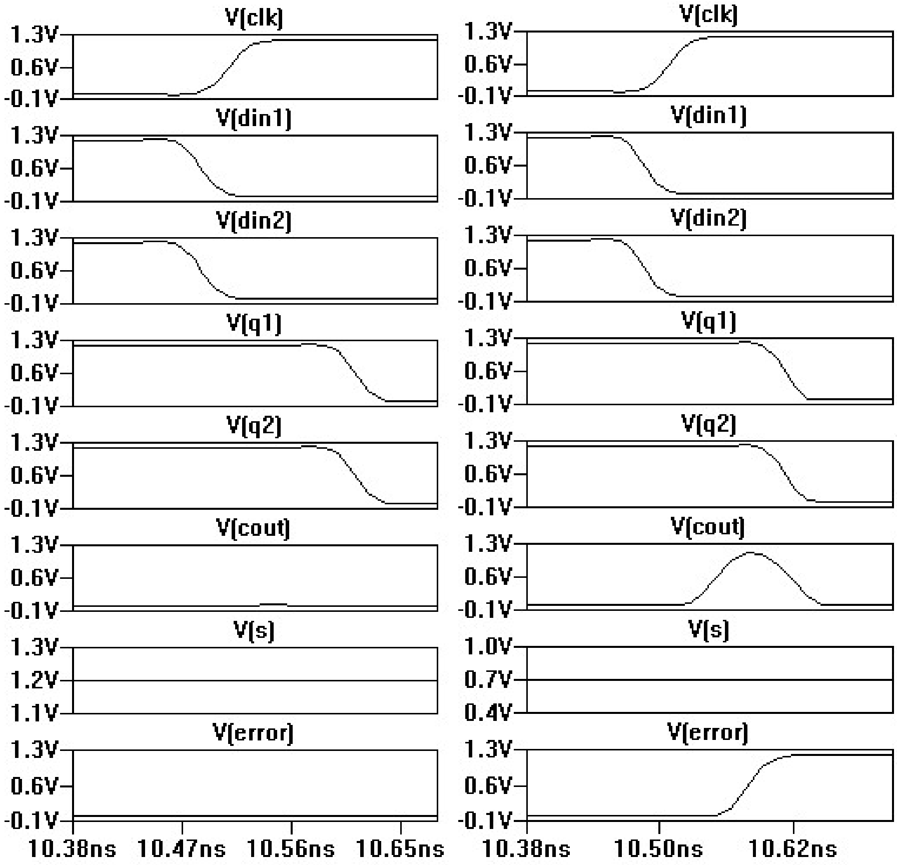


FIGURE 4.16: Transistor Level Worst Case Timing Error Prediction Simulation result in Spice

Figure 4.16 (a) shows the worst case circumstance, and two data signals are switching at about the same time, the switching is by the MDU in this case. In Figure 4.16 (b), the voltage of signal ‘S’ is scaled down from 1.2 V to 0.7 V. With the differential between ‘din1’ and ‘din2’ increasing, a pulse is generated by the MDU during the stability checking period. This pulse is then captured by the stability checker and generates an error signal.

4.8.1.2 Transistor Level Cost Comparison in 90 nm

The first step in any ageing prediction technique is to identify the long delay paths to be monitored. These refer to the critical paths of the circuit under consideration, in addition to all long delay paths which may cause timing errors due to ageing induced delay degradation. NBTI can cause more than 20% timing degradation in the worst-case operating conditions, [78]. Therefore in this work, we consider all paths whose delays violate the timing budget after 20% degradation to be vulnerable to ageing-induced delay faults, hence needing monitoring.

To compare the area overhead of the proposed design DMEDS with Razor FF and Canary FF, we have considered the EX stage of a 32-bit pipelined MIPS. First, we identified the paths which need monitoring, by performing a detailed timing analysis for a 90nm technology. Figure 4.17 shows the delay of the longest 40 paths in this circuit. To identify which of these paths are likely to cause timing errors, we need to take into consideration a potential increase of 3% and 20% from the nominal delay of each path, due to process variation and ageing, respectively [78], as shown in Figure 4.17. Hence, in the example under consideration, we have identified twenty-four paths in the EX stage of the MIPS which need to be monitored with DMEDS sensors for ageing-induced timing failures. Based on this analysis, we have estimated the area overhead of inserting ageing prediction sensors using our design, compared to Razor FF and Canary FF. The results are shown in Table 4.2.

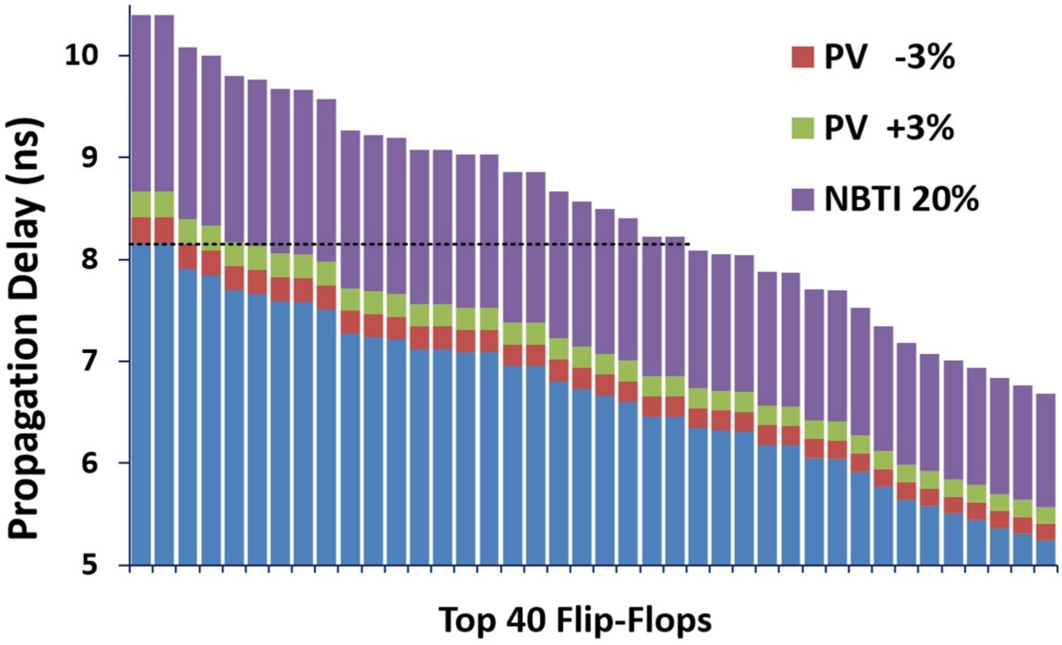


FIGURE 4.17: Timing Report of the EX Stage in a 32-bit Pipelined MIPS

	Canary FF	Razor	DMEDS
Area overhead (Extra Transistors Required)	720	1296	151

Table 4.2: Comparison with Razor and Canary FF

To monitor twenty-four paths, DMEDS requires 151 extra transistors when DMEDS monitors twenty-four paths simultaneously (twenty-three 2-input XOR and one stability Checker shared). Razor FF and Canary FF require 1296 and 720 extra transistors respectively. Compared with Razor and Canary FF DMEDS saves 88.33% and 79.00% of the transistors respectively, which significantly reduces the area.

4.8.2 System Level Verification and Analysis in 90 nm

This section presents the System level functional verification and comparative analysis in 32nm technique. The DMEDS was implemented to predict the delay fault from two or more paths simultaneously. The cost analysis calculated the area overhead and power overhead when the DMEDS were monitoring two or more path in a 32-bit pipelined MIPS.

4.8.2.1 DMEDS Synthesis and Simulation

The gate level design of DMEDS is programmed in the Hardware Description Language, System Verilog and then synthesised by a logic synthesis tool, Design Vision, which takes HDL designs and synthesise them to gate-level HDL net-lists and generates estimated area, timing and power report of the design. Figure 4.18 shows the DMEDS Block Diagram after Synthesis.

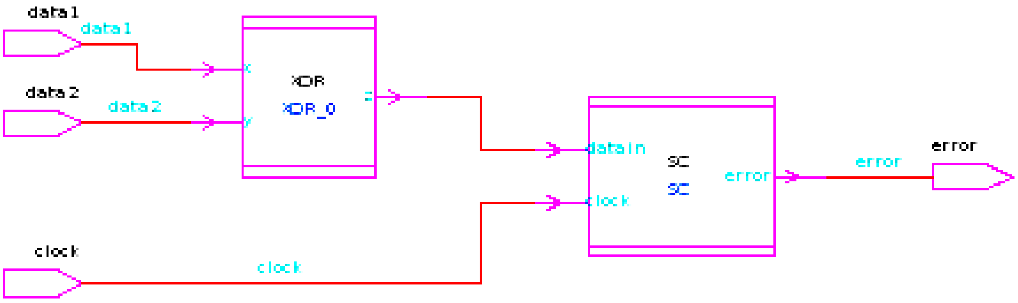


Figure 4.18: DMEDS Block Diagram after Synthesis

Figure 4.19 shows the Verilog simulation result when only one of the input data changes within the detection window and both transitions from the input of the FF switch before the clock rising edge and the setup time. ’Clock’ is the clock signal of FFs and the DMEDS. ’Data1’ and ’Data2’ are the data input signal of two FF respectively. ’Q1’ and ’Q2’ are the data output signal of two FF. ’error’ is the output signal of the stability checker.

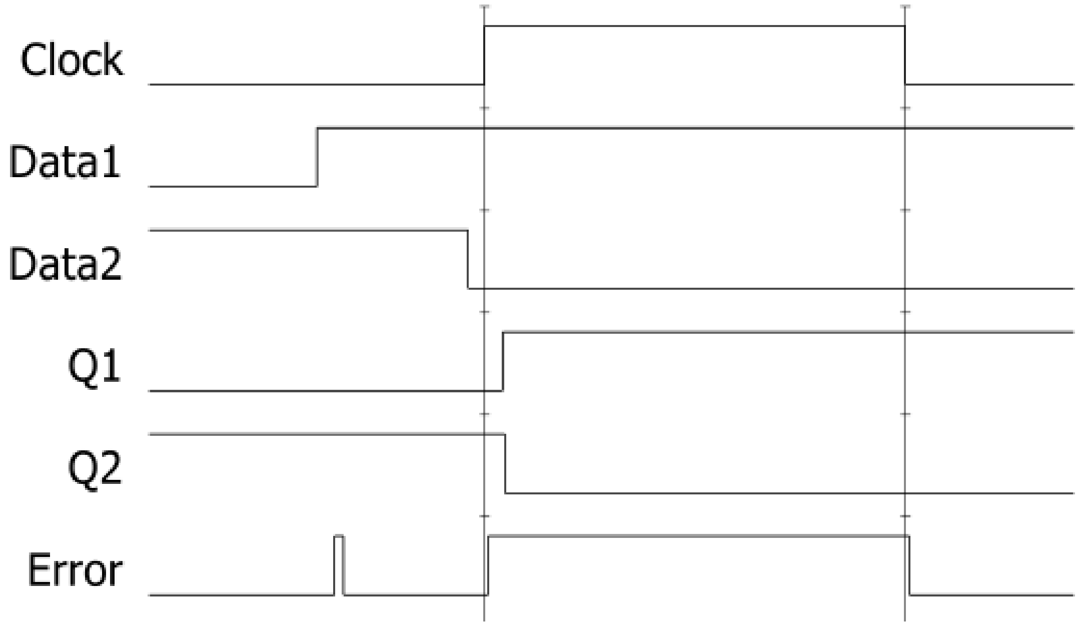


Figure 4.19: Verilog Simulation Result of Timing Error Prediction in Modelsim

As Figure 4.19 shows, ’Data1’ switch out of the detection window which triggers a short pulse on the ’Error’ signal. As the pulse was triggered during the ’Clock’ signal is at the low stage, the RS latch did not latch the transition in the stability checker. ’Data2’ changes within the detection window and before the setup time of the FF, the stability checker captures the transition and latches the error signal. As FFs receive ’data1’ and ’data2’ before the rising clock edge, there is no timing error.

Figure 4.20 shows the Verilog simulation result when both two input signal switch within the detection window about the same time. ’Clock’ is the clock signal of FFs and the DMEDS. ’Data1’ and ’Data2’ are the data input signal of two FF respectively. ’Q1’ and ’Q2’ are the data output signal of two FF. ’error’ is the output signal of the stability checker.

As Figure 4.20 shows the worst case circumstance, both ’Data1’ and ’Data2’ are switching at about the same time, the MDU does not recognise the switching in this case. As there is no adjustment delay element of the gate level design of DMEDS the delay fault become undetectable when two input data change about the same time. However, the transitions of two paths are de-correlated, the data of two different paths will not convert simultaneously every time, with the different combination, the input data will switch individually, it then triggers the error signal. The DMEDS is a delay fault prediction

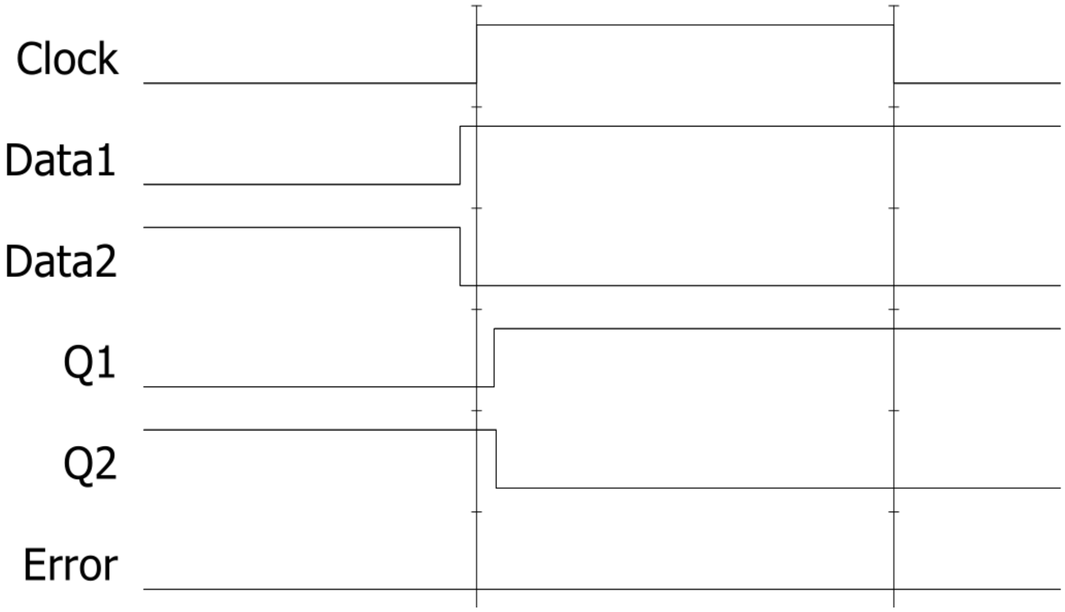


Figure 4.20: Verilog Simulation Result of Worst Case Timing Error Prediction in Modelsim

sensor, and it is not necessary to detect every signal delay fault which does not truly occurred.

4.8.2.2 Comparison of Power and Area Overhead

This section compares the power and area overhead between two delay fault prediction sensors, DMEDS and Canary FF. In section A, the comparison is based on the sensor itself without implementing the sensor in any circuit. Section B compares the power and area overhead when the sensors are implemented in the EX-stage of a 32 bit pipelined MIPS processor. Figure 4.21 shows the Block Diagram of Canary FF after Synthesis.

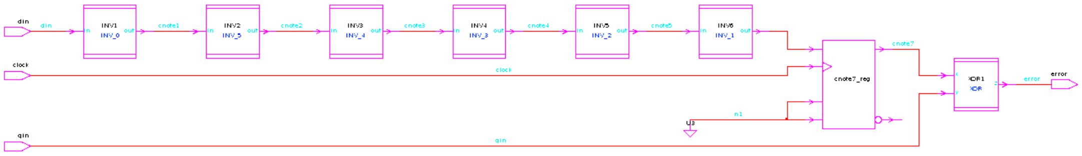


Figure 4.21: Canary FF Block Diagram after Synthesis

A.Comparison of Power and Area Overhead of DMEDS

DMEDS is able to predict the delay fault from two or more path simultaneously. To ensure the reliability of the sequential circuits, the size of the FF is relatively large compared with others. The delay fault sensors based on the double sampling technique achieve the functionality by adding a same sized FF to sample a reference data of the main FF. In practice the shadow FF can be half sized FF compared with the main FF without the consideration of ageing. However, the smaller shadow FF will age differently with the main FF. This will lead to the miss-match between those two FFs. The detection window of the Canary FF will move forward if a smaller sized shadow FF is inserted as a small component will ageing faster than the large ones, [32]. Hence the double sampling sensors will cost more area than the comparison result stated in the last section. The leakage current is dependent on the size of the design. The DMEDS also has the advantage in terms of power dissipation. The cost and power comparison of one Canary FF, compared to 2-input DMEDS, 3-input DMEDS and 4-input DMEDS. The results are shown in Table 4.3.

	Canary FF (for 1 path)	DMEDS-2 (for 2 paths)	DMEDS-3 (for 3 path)	DMEDS-4 (for 4 paths)
Cost (μm^2)	33.00	24.92	27.36	31.77
Power (μw)	51.75	31.10	43.35	53.50

Table 4.3: Comparison with Canary FF

The cost of one Canary FF, 2-input DMEDS, 3-input DMEDS, 4-input DMEDS are 33.00μm^2, 24.92μm^2, 27.36μm^2 and 31.77μm^2 respectively. As DMEDS monitors 2 to 4 path simultaneously, for the same amount of path monitoring, DMEDS saves 62.25%, 72.36% and 75.93% area compared with Canary FF with respect to monitor the same number of paths. In the case of power, DMEDS saves 69.99%, 72.07% and 82.4% power compared with Canary FF with respect to monitor the same number of paths.

B.Comparison of Power and Area Overhead in a Processor

Figure 4.22 shows the timing report of top 1000 critical paths of a 32 bit MIPS in 32 nm technique with 4.2 ns clock period generated by Design Vision. Assuming that the potential delay degradation of each path is the same with the degradation stated in section 4.8.1.2, there will be 738 paths potential critical paths which share 164 FFs at the end of those paths.

Based on the timing analysis of potential critical paths, we have estimated the area overhead of inserting ageing prediction sensors using our design without considering the routing area overhead (place and route), compared to Canary FF. The results are shown in Table 4.4.

	Canary FF	DMEDS-2	DMEDS-3	DMEDS-4
Area Overhead (%)	20.5%	8.1%	6.0%	5.1%
Power Overhead(%)	19.6%	5.8%	5.4%	5.1%

Table 4.4: Comparison with Canary FF in a 32 bit MIPS

As table 4.4 shows, the area overhead of implementing Canary FF, 2-input DMEDS, 3-input DMEDS, 4-input DMEDS on the potential critical paths are 3 20.5%, 8.1%,

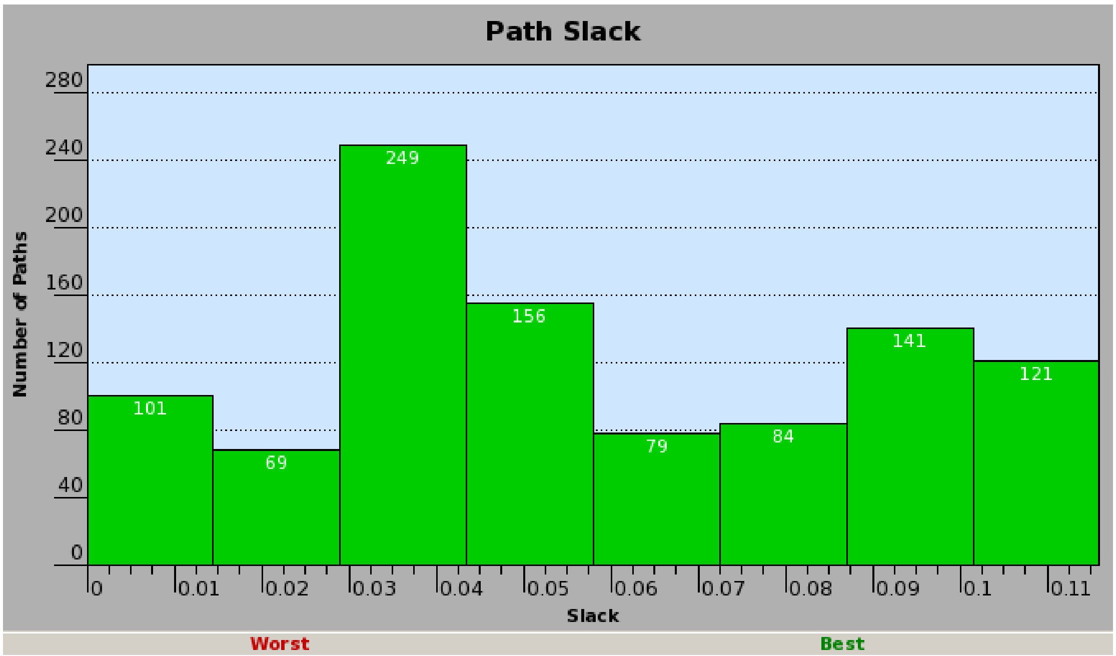


Figure 4.22: Timing Report of the 32 bit MIPS

6.0% and 5.1% respectively and the power overhead are 19.6%, 5.8%, 5.4% and 5.1% respectively.

4.8.3 Transistor Level verification and cost Comparison in 32 nm

4.8.3.1 Transistor Level verification in 32 nm

DMEDS was designed using a 32nm CMOS technology for verification and evaluation. Figure 4.23 shows the Spice simulation results when the inputs of two conventional FFs are monitored simultaneously at time zero and after 10 years. The ageing degradations were estimated using Synopsys HSPICE MOSRA. As Figure 4.23 (a) shows, the difference between 'data1' and 'data2' triggers a pulse of the output signal 'Transition' at time zero. As the remaining slacks in any of those paths do not reach the DW yet, the error signal is not triggered. After 10 years ageing, the minimum measurement resolution between two input has degraded from 21ps to 39ps. The width of the T_{MDU} and the FF's set-up time are increased by 24.5% and 126%, respectively, due to the ageing of the MDU and FF. The simulation results for DMEDS after 10 years' ageing is shown in Figure 4.23 (b). Both 'data1' and 'data2' switch at the same times as in Figure 4.23 (a). A pulse is generated by the MDU during the stability checking period caused by the difference between 'data1' and 'data2'. The stability checker captures the pulse and triggers the error signal due to the ageing degradations of the DW. As 'data1' and 'data2' are propagated to the output signals of FFs, 'q1' and 'q2', after the rising

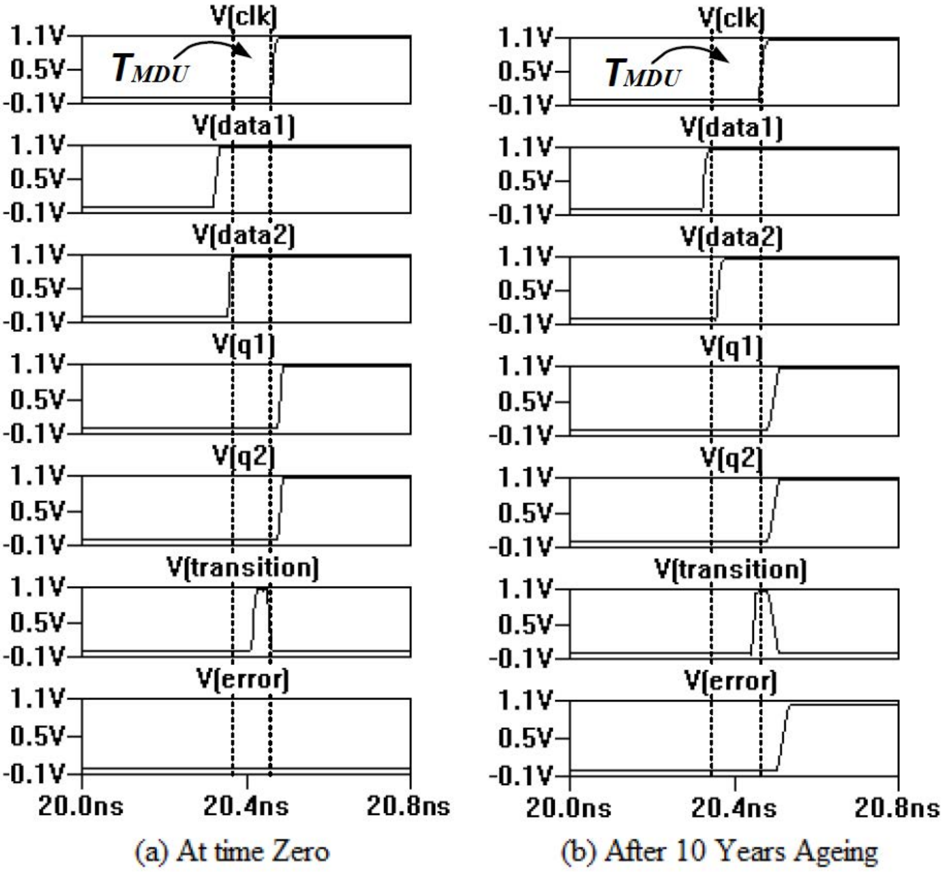


FIGURE 4.23: Multiple Timing Error Detection

clock edge, there is no timing error. According to the worst/best case process variation simulation results, the minimum measurement resolution, T_{MDU} and FF's set-up time vary by +0.4%/-5.0%, +3.2%/-2.2% and +2.3%/-5.8% at time zero, respectively.

4.8.3.2 Cost Comparison in 32 nm

Table 4.5 gives a cost and performance comparison with different delay fault detection or prediction sensors. The overhead of FF type sensors is compared with the standard 24 transistor FF. Compared with Razor and Canary, DMEDS saves 87.59% and 77.67% of the transistors respectively for 10 path delay fault monitoring (10 paths are monitored by one 10-input DMEDS simultaneously). DMEDS is also able to predict delay faults in a latch-type sensor. The overhead of a latch-type sensor is compared with the standard 16 transistor latch, Table 4.5. Compared with DSTB and iRazor, DMEDS saves 80.29% and 29.18% of the transistors respectively for 10 path delay fault monitoring, which significantly reduces the area overhead. The cost of MDU adjustment is not included in this comparison, as existing sensors will also require adjustments to ensure their functionality after serious ageing such as the delay chain of Canary FF [40] and the local CLK Generation of iRazor, [89]. The routing area overhead has been carefully

Table 4.5: Comparison with other designs

FF Type	Razor[29]	Canary[40]	DMEDS (II)
Extra number of Transistors (I)	54 225%	30 125%	6.7 27.9%
Metastability	Yes	Yes	No
Error Recovery Required	Yes	No	No
Replace FFs	Yes	No	No
Latch Type	DSTB [16]	iRazor[89]	DMEDS (II)
Extra number of Transistors (III)	34 212.5%	9.46 59.1%	6.7 41.9%
Metastability	No	No	No
Error Recovery Required	Yes	Yes	No
Replace FFs	Yes	Yes	No

(I) Compared to standard 24T FF (excluding the delay chain)
(II) 10-input DMEDS monitors 10 path simultaneously (nine 2-input XOR (54T) and one stability checker (13T) shared)
(III) Compared to standard 16T Latch

considered in this comparison. Compared with other designs, DMEDS has more routing area overhead on the input side. However, there is only one error signal on the output side, while there is a higher routing area overhead on the output side for the other designs. Furthermore, more circuitry will be required to manage the error signals when those signal delay fault sensors are implemented. Thus, the overhead of the other sensors is underestimated compared with DMEDS.

4.8.4 System Level verification in 32 nm

An equivalent model of DMEDS, written in a hardware description language has been used to verify the functionality of the DMEDS at the system level, as shown in Figure 4.24. The behaviour of the model is identical to that shown in Figure 4.4.

We have used a 32-bit pipelined MIPS to verify the functionality of DMEDS. Figure 4.25 shows the MWB stage writing data back to ten different register files. The unbalanced addressing will cause these ten paths to age differently. Transitions in those ten paths occur successively because they are triggered by the same instructions but by different addresses (100% de-correlated). The error signal is triggered when late transitions are detected by the 10-input DMEDS.

4.9 Concluding Remarks

This Chapter has introduced the timing dependent ageing mechanisms. It then presents four typical delay fault detection and prediction sensors. The ageing mechanisms might

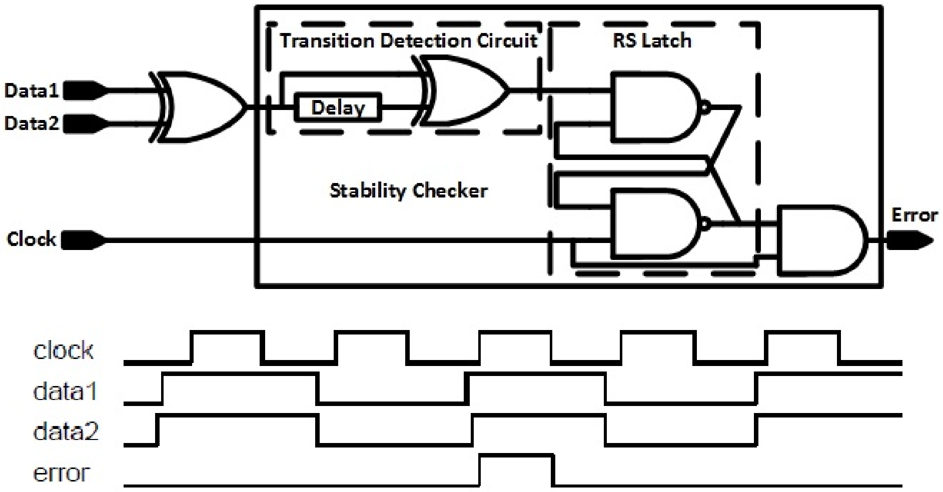


FIGURE 4.24: Equivalent Model of DMEDS and Simulation Result

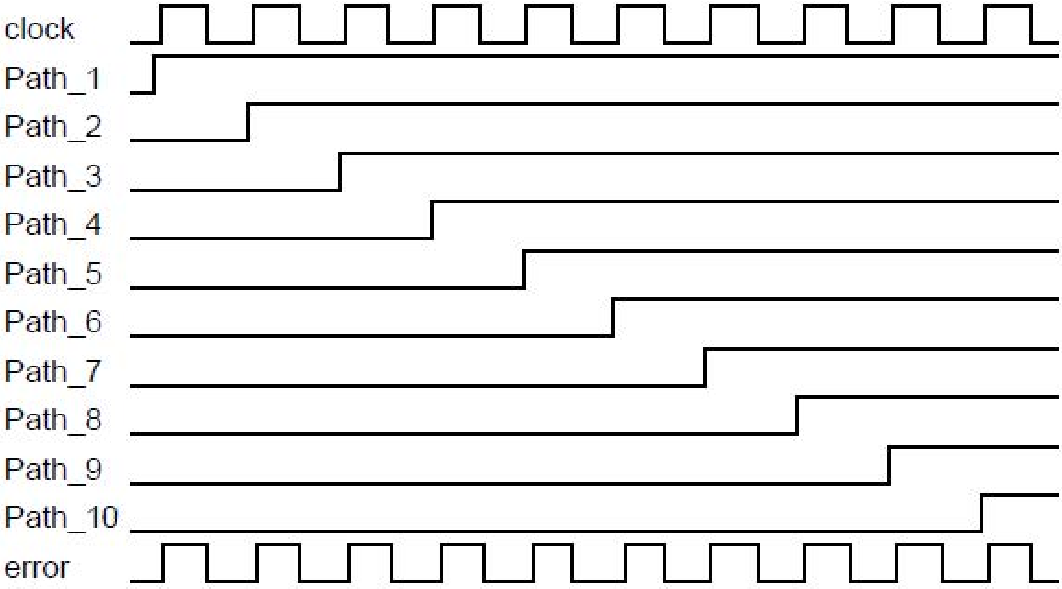


FIGURE 4.25: 10-input DMEDS functionality verification in a 32-bit MIPS

changes the ranking of the potential critical paths, which lead to area cost for multiple path monitoring might be prohibitive. A new sensor, named DMEDS, is proposed in this chapter for predicting timing errors. The advantages of DMEDS compared with existing approaches are:

1. DMEDS is a multiple error detector which improves the cost-efficiency. Compared with Razor and Canary FF, DMEDS saves 87.59% and 77.67 % of the transistors respectively for delay-fault monitoring in the EX stage of a 32-bit MIPS. In system level implementation, DMEDS saves 62.25%, 72.36% and 75.93% area overhead for delay-fault monitoring in a 32-bit MIPS when it is monitoring 2, 3 and 4 paths respectively compared with Canary FF.

2. DMEDS is an external sensor, it does not need to replace the FF of the original design, therefore it is easy to implement and it will not influence the functionality of the original design.

3. DMEDS is an error prediction sensor; a small safety margin is generated by the MDU. DMEDS is able to predict the delay fault before it occurs.

4. The design of stability checker in DMEDS is metastability resistant, metastability will not occur on both the main FF and the DMEDS.

This Chapter and Chapter 3 present two cost-efficient delay fault prediction sensors which reduce the area overhead significantly. It, therefore, brings the DVS system back to the consideration as an ageing mitigation technique. The delay fault prediction sensors with the DVS can provide useful information for lifetime prediction as shown in the next Chapter.

Chapter 5

DVS for Ageing Prediction

Aggressive technology shrinking has caused reliability issues of CMOS devices such as time-dependent process and ageing variations which will limit the lifetime of CMOS design. Lifetime mitigation approaches will sacrifice area, power and performance, which are the primary concern during the design of a CMOS circuit, it, therefore, kills the motivation of such approaches. Existing ageing prediction techniques predict the lifetime by one-time ageing analysis or real-time ageing analysis. The former one assumes a circuit is running under the certain condition which will overestimate or underestimate the lifetime degradation immoderately. The latter one requires acquiring various real-time information to assist the lifetime prediction which is complicated and expensive. This chapter presents an Dynamic Voltage Scaling (DVS) system, together with real-time ageing prediction algorithm, for lifetime prediction and time-dependent ageing effects mitigation. The cost of the DVS system is becoming prohibitive as technology shrinking and design complexity in the last few years. Two cost-efficient delay fault prediction sensors are presented in Chapter 3 and 4 which reduce the area overhead severely. It, therefore, brings the DVS system back to the scope of consideration as an ageing mitigation technique. The real-time ageing prediction algorithm compares the operating voltage scaling time with reference to estimate the lifetime of a CMOS device. Our result indicates that lifetime is accurately estimated with less than 5% of error, which is a lot more precise compared with existing ageing prediction models, [10, 77].

5.1 Introduction

This chapter outlined the DVS system, together with real-time ageing prediction algorithm, for lifetime prediction and time dependent ageing mitigation. We focus on the delay degradation induced by Negative Bias Temperature Instability (NBTI), which is one of the most common time-dependent ageing mechanism. NBTI increases the absolute threshold voltage which causes the degradation of mobility, drain current and threshold voltage transconductance of PMOS transistors. The delay degradation caused by NBTI is becoming more and more severe as the aggressive technology scaling over the past few decades, especially for sub-65nm technology. It, therefore, reduced performance and lifetime of integrated circuits bitterly. The performance degradation can be more than 20% depending on the operating conditions of a design. Time-dependent ageing mechanisms depend on the common factors: technology, the total device operating time, workload, temperature and supply voltage. Therefore, other ageing mechanisms can be expended into this system, such as Positive Bias Temperature Instability (PBTI), Hot Carrier Injection (HCI), although NBTI is the primary concern in this work.

As the last paragraph stated, ageing mechanisms, such as NBTI, depend on technology, the total device operating time, workload, temperature and supply voltage. Existing ageing mitigation approaches control those factors to extend the lifetime CMOS devices. However, such approaches will sacrifice area, power and performance which are the primary concern during the design of a CMOS circuit. DVS system is one of the most popular power-saving approaches. Usually, the DVS works with delay fault monitoring sensors, which detect or predict timing errors on the critical path of a CMOS device. However, modern IC designs are particularly demanding for power, area, and performance. The length of critical paths is synthesised into similar length. On the other hand, technology scaling aggravated ageing and process variations which may change the ranking of near-critical paths. A complex IC will have a huge number of potential critical paths that need to be monitored. Hence, the cost of the DVS system is becoming prohibitive. Two cost-efficient delay fault prediction sensors are presented in Chapter 3 and 4 which reduce the area overhead severely. It, therefore, brings the DVS system back to the scope of consideration as an ageing mitigation technique.

Existing ageing prediction techniques predict the lifetime by one-time ageing analysis or real-time ageing analysis. One time ageing analysis estimates the lifetime by using ageing models which assume a circuit running under a specific condition. However, ageing related factors especially workload, temperature and local supply voltage might be different in the case of each IC. One time ageing analysis gives a rough estimate of IC estimation without considering the actual operating condition, and this will estimate the lifetime degradation inaccurately. Existing real-time ageing analysis estimates the lifetime by using ageing models according to the actual operating condition. It requires acquiring various real-time information to assist the lifetime prediction. Extra circuitry, sensors and complex control unit are required motoring ageing related factors such as workload, temperature and voltage profiles. Compared with one-time ageing analysis, the real-time ageing analysis is able to estimate the lifetime accurately. However, it is complicated and expensive. The objective of this Chapter is to estimate the lifetime of a circuit accurately by using limited information from the DVS.

The premise of this chapter are: rather than adding extra circuitry and sensors specifically for real-time lifetime prediction system, our new real-time lifetime prediction system will be combined with the DVS system which is also an ageing mitigation and low power technique; The system will estimate the IC lifetime according to the limited information provided by the DVS system. It compares the operating voltage scaling time with a reference which can be estimated by ageing models or from other devices. The lifetime estimated by the new real-time lifetime prediction system was compared with the result of an ageing model. Our results indicate that lifetime is accurately estimated with less than 5% of error, which is a lot more precise compared with existing ageing prediction models.

5.2 Ageing Prediction System Design Principle

DVS is the adjustment of power and speed settings on a computing devices various processors, controller chips and peripheral devices to optimise resource allotment for tasks and maximise power saving when those resources are not needed. DVS allows devices to perform needed tasks with the minimum amount of required power. The technology is used in almost all modern computer hardware to maximise power savings, battery life and longevity of devices while still maintaining ready compute performance availability. Ageing models have been proposed for different levels of abstraction [19, 58] to estimate the lifetime of CMOS circuits.

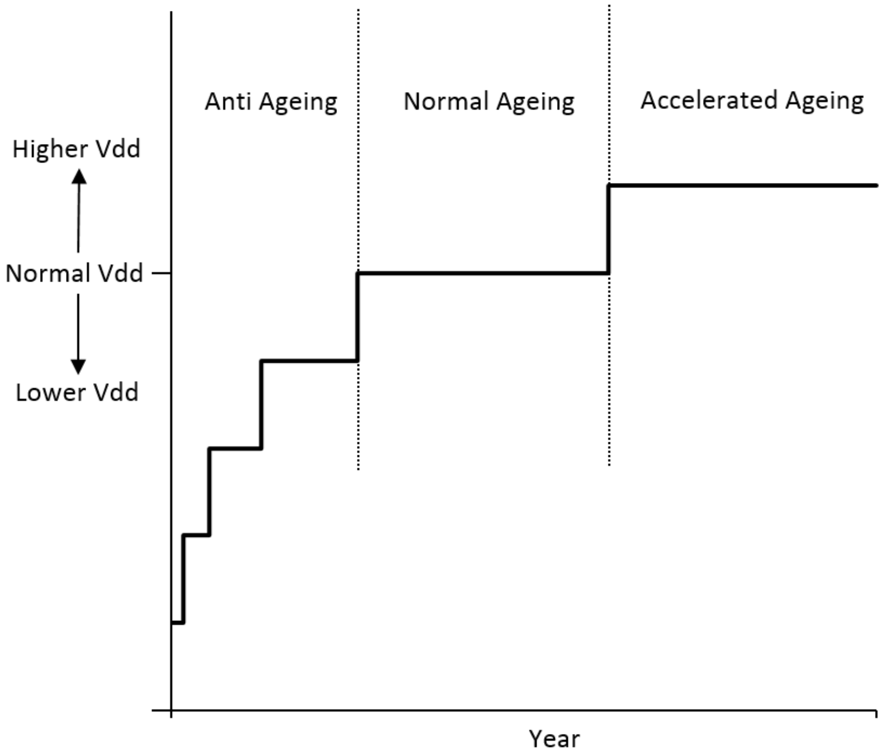


Figure 5.1: The Average Supply Voltage of a Circuit with DVS Over the Years

As Figure 5.1 shows, the average supply voltage of an IC scales up by the DVS over the years due to time-dependent ageing mechanisms. The time between two scaling point increases over time because the intrinsic delay of a circuit at low voltage states is longer than the intrinsic delay at high voltage states, and the increase of intrinsic delay degradation due to ageing mechanisms is faster at the beginning of ICs lifetime compared with the end of the ICs lifetime, stated in Chapter 2. There are three circuit ageing states during the circuits lifetime:

1. Anti-ageing: The circuit is operating at a relatively lower supply voltage compared with the circuit without DVS. Therefore, the degradation from ageing will be slower than the circuits without DVS at the anti ageing state, as ageing mechanisms is supply voltage dependent, such as NBTI.
2. Normal ageing: The circuit with DVS operates at the same supply voltage compared with the circuit without DVS. The speed of ageing will be the same as those circuits are running under the same condition. The circuit without DVS will die in this stage as the violation of timing constraint due to delay degradation from ageing.
3. Accelerated ageing: In this stage, the circuit will running at the supply voltage higher than it supposed to be, as the intrinsic delay of a circuit will violate the timing constraint without scaling up the supply voltage. Circuits will work under severe ageing as the high supply voltage. It, therefore, compensate the delay degradation from ageing phenomena. However, the circuit without DVS will fail before the end of the normal ageing state. It, therefore, extends the lifetime of an IC.

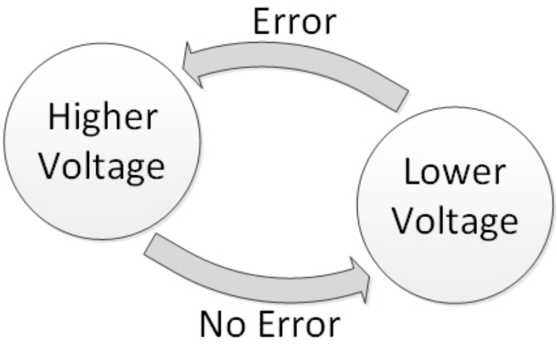


Figure 5.2: Voltage Scaling

The information of lifetime prediction from a DVS is very limited compared with other real-time ageing prediction approaches, which requires the information of operating time, temperature, input vector and supply voltages during the circuit lifetime. However, DVS is able to provide very useful information, which is the ageing stage and operating time.

Equation (5.1) shows the relationship between the delay degradation from BTI, delay at time zero, operating time and input vector duty cycle in certain conditions (See details in Section 2.3.2).

$$\Delta D_{bti} = K D_0 \alpha^n t^n \tag{5.1}$$

The lifetime of each IC changes according to its operating conditions, such as workload and temperature. Therefore, the duty cycle α, lifetime t and fitting parameter K might be different of each IC (See details in Section 2.3.2). The operating frequency (performance) of an IC is fixed for the same design at the same technology node which is defined by the intrinsic delay of the potential critical paths and the margin for the PVTA variations. Therefore, the margin might varies due to the process variation, but it will be fixed after the circuit's fabrication. The lifetime of a circuit ends when the margin is exhausted under it's worst operating condition, as a circuit might fail at high-temperature state even if it works at a relatively low temperature. Therefore the total intrinsic delay degradation during a circuit lifetime is fixed of each circuit after fabrication.

$$\frac{\Delta D_{btitotalc1}}{\Delta D_{btitotalc2}} = \frac{K_{c1} D_{0c1} \alpha_{c1}^n t_{c1}^n}{K_{c2} D_{0c2} \alpha_{c2}^n t_{c2}^n} \tag{5.2}$$

The comparison of the intrinsic delay degradation, $\Delta D_{btitotal}$, between two different circuits is shown in equation (5.2), where c1 and c2 represent two circuits which are under different PVTA variations: circuit one and circuit two.

$$\frac{\Delta D_{btitotalc1}}{\Delta D_{btitotalc2}} = Constant_{BTI} = \frac{K_{c1} D_{0c1} \alpha_{c1}^n t_{c1}^n}{K_{c2} D_{0c2} \alpha_{c2}^n t_{c2}^n} \tag{5.3}$$

The comparison of the intrinsic delay degradation between two different circuits will be a constant as the total ΔD_{bti} during the circuit's lifetime is certain, as shown in equation (5.3).

$$\frac{D_{0c1}}{D_{0c2}} = Constant_{D0} \tag{5.4}$$

The intrinsic delay at time zero will be of each individual chip will be settled after fabrication. Therefore, the ratio of the delay at time zero between two different circuits is a constant, as shown in equation (5.4).

$$\frac{\alpha_{c1}}{\alpha_{c2}} = Constant_{\alpha} \tag{5.5}$$

$$\frac{Kc1}{K_{c2}} = Constant_K \tag{5.6}$$

The same circuit will be running under the similar workload and environment[21], therefore, the ratio of the duty cycle (α) and the fitting parameter(K) between two different circuits are also constant, as shown in equation (5.5) and (5.6).

$$Constant_{BTI} = \sqrt[n]{Constant_K.Constant_{D0}.Constant_{\alpha}.\frac{t_{c1}}{t_{c2}}} \tag{5.7}$$

Therefore, the intrinsic delay degradation between two circuits can be derived as equation (5.7).

$$\frac{t_{c1}}{t_{c2}} = A \tag{5.8}$$

$$\frac{\Delta t_{c1}}{\Delta t_{c2}} = A \tag{5.9}$$

Hence, the comparison of the total lifetime between two circuits is a constant, where A is depending on the margin, K, D_0 and α , as shown in equation (5.8). Thus, the operating period between two circuits are proportional when the operating period corresponds to the same part of total delay degradation, as shown in equation (5.9).

$$T_{pd} = \frac{V_{DD}}{b(V_{DD} - V_{th})^{1.3}} \tag{5.10}$$

The DVS with ageing sensors can provide information on the supply voltage and the circuit's operating time between different voltage stage as Figure 5.1 shows. With the ageing prediction sensor, such as DMEDS and PCC, the maximum supply voltage will increase when the intrinsic delay at current voltage is going to violate the timing constraint under the worst operating conditions, as shown in Figure 5.2. In other words, the DVS reserves the margin by scaling down the supply voltage, and a margin will be released after the voltage scales up as the intrinsic delay is supply voltage dependent, as shown in equation (5.10) [68] (See the details in Section 2.1). Thus, the operating time that the maximum supply voltage reached to represents the ageing states of the whole lifetime. Therefore, the delay degradation between two maximum voltage scaling point represents a specific part of the total delay degradation. Hence, the operating time between two maximum voltage scaling point of two circuits is proportional. In other words, the lifetime of a circuit can be predicted by comparing remaining time between two maximum voltage scaling point with a circuit that the lifetime has already been determined, as shown in (5.11), where LT_{est} is the estimated lifetime of the target circuit, $t_{maxtarget}$ and t_{maxref} are the time between two maximum voltage scaling point of the target circuit and reference circuit respectively. The LT_{ref} is the lifetime of the reference circuit.

$$LT_{est} = \frac{t_{maxtarget}}{t_{maxref}} \times LT_{ref} \tag{5.11}$$

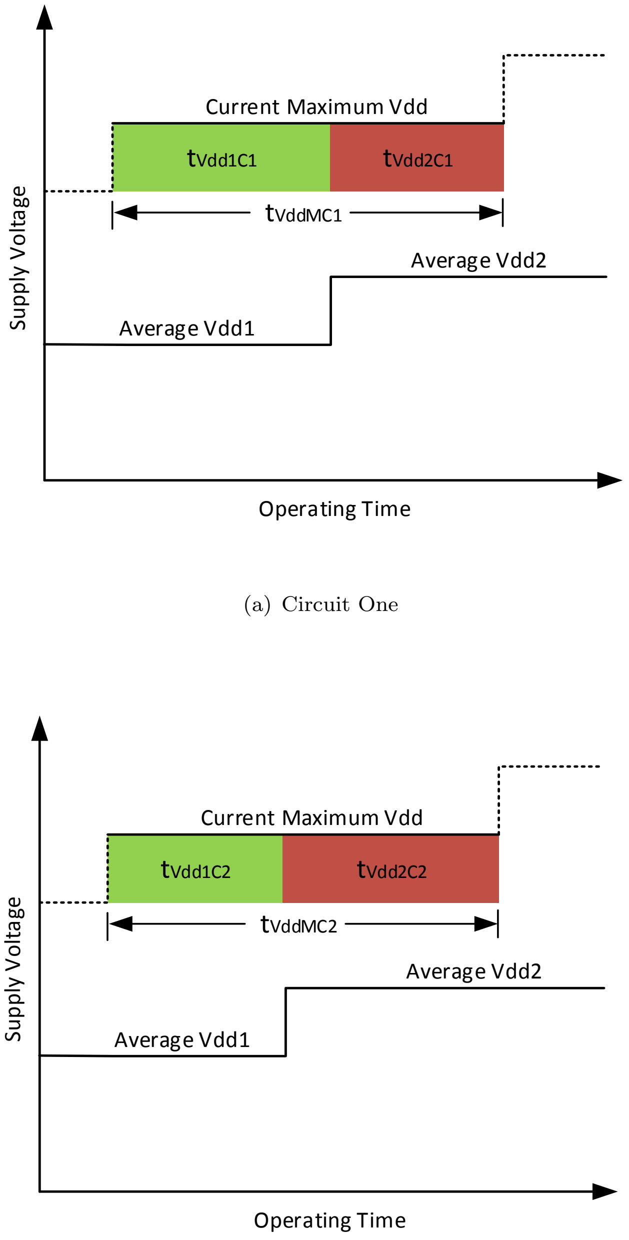
The Lifetime between two circuits will be close to each other. Therefore, the closest T_{maxref} should be the primary concern in terms of the choice of reference data. In practice, a circuit will not always run in the worst case. Therefore the average voltage level will be lower than the maximum voltage level. The average voltage level will define the fitting parameter K, as shown in equation (5.1), which represents the speed of ageing. Furthermore, the average voltage level might change between two maximum voltage scaling points, as shown in Figure 5.3, this constant A may change between two maximum voltage scaling points.

Figure 5.3 shows the relationship between the circuit's operating time, the average supply voltage and the maximum supply voltage of two circuits during a part of the circuit's lifetime. Where t_{VddM} is the total time during the maximum supply voltage remains at current maximum Vdd, t_{Vdd1} and t_{Vdd2} is the time during the average supply voltage remains at Vdd1 and Vdd2 respectively while the maximum supply voltage is at current maximum Vdd, C1 and C2 represent circuit one and circuit two respectively. C1 and C2 are running under different conditions. Therefore C1 and C2 have a different lifetime. The supply voltage changes by the DVS according to the circuit's operating conditions. Therefore the proportion of the average Vdd will be different if two circuits age differently. As the Figure shows, the average supply voltage changes between two voltage scaling points in both cases. The proportion of t_{Vdd1} and t_{Vdd2} between those two cases are different. The comparison between t_{Vdd1} and t_{Vdd2} will be the constant A if and only if the proportion of t_{Vdd1} and t_{Vdd2} are identical. In the case shown in Figure 5.3, C2 is ageing faster than C1 as the proportion of VDD1 of C1 is higher than the proportion of VDD1 of C2. The shift of the ratio will cause the error of the estimated lifetime. In reality, it is very difficult to enumerate the reference data which covers each individual case. Therefore the system should select the nearest reference data for the lifetime estimation.

Therefore, the priority of selecting the reference data are:

1. Firstly, find the closest T_{maxref}.
2. Secondly, find the the closest rate between t_{Vdd1} and t_{Vdd2} for the comparison.

Figure 5.4 shows the lifetime estimation flowchart. The average supply voltage is continuously calculated and recorded when the system is running. The system then records



(a) Circuit One

(b) Circuit Two

FIGURE 5.3: Relationship Between Circuit's Operating Time and Supply Voltage

the time that the new highest VDD when it appears. The time of each value will only be recorded one time for the lifetime estimation. The lifetime estimation will be terminated when the highest VDD has reached the maximum value.

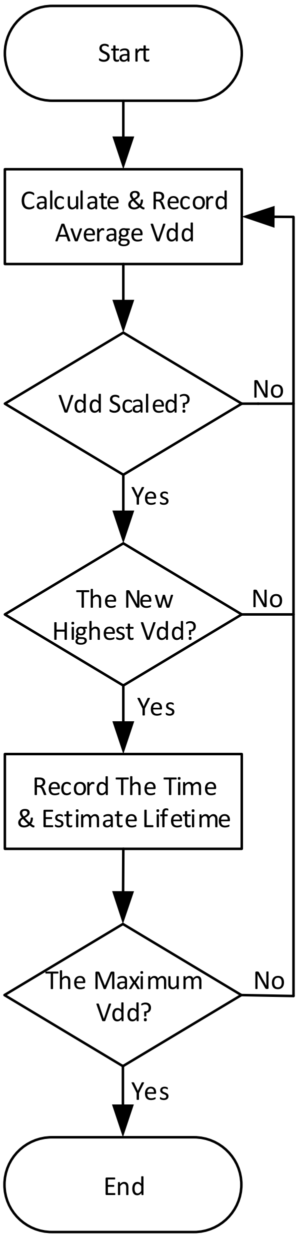


FIGURE 5.4: Lifetime Estimation Flow Chart

5.3 Reference Circuit Estimation

The work in this Chapter uses a long inverter chain to represent the long delay path of a complex IC circuit. It can be extended to a more complex circuit for the following reasons: The delay intrinsic delay degradation due to ageing will scale proportionally under the same operating condition (the same K), as shown in equation (5.1); The ratio of duty cycle $Constant_{\alpha}$ remains as a constant cause the signal probability remains the same when the circuit net-list is consistent.

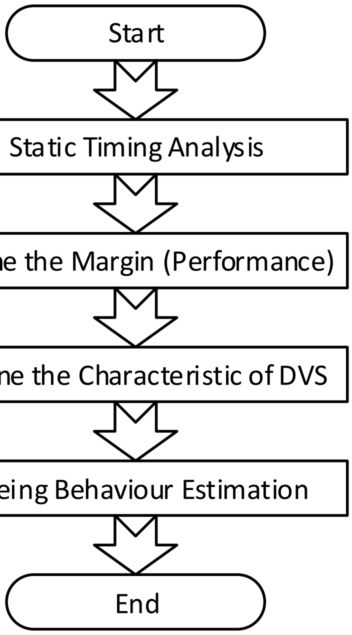

Start
Static Timing Analysis
Define the Margin (Performance)
Define the Characteristic of DVS
Ageing Behaviour Estimation
End

Figure 5.5: Reference Circuit Estimation Flow Chart

Figure 5.5 shows the flowchart of the reference circuit behaviour estimation. The first two steps are slandered procedure of a digital circuit design. Static timing analysis gives circuit net-list after considering the behaviour, timing, area constraints. It then provides the information of intrinsic delay of the circuit. The second step is to define the margin left for the circuit without DVS, which is usually inserted into the timing constraint for the PVTA variations after fabrication and during the circuit's lifetime. The margin of an IC is usually defined as intrinsic delay degradation in the worst PVTA conditions. The third step is to set the character of the DVS, such as the minimum unit of voltage scaling and the minimum and maximum supply voltage. Finally, the circuit behaviour will estimate considering the technology node of the design and the typical circumstances that the circuit will be running.

5.3.1 Define the Margin

Process Variations

The process variation will affect the propagation delay of the CMOS circuit. Figure 5.6 shows the worst and best case process variations of an inverter chain of 90 nm technology by using the process variation model in Spice. Where TT means typical charge and typical discharge, FF means fast charge and fast discharge and SS means slow charge and slow discharge. As the figure shows, the delay varies ±4% due to the process variations. However, the worst case will unlikely happen in reality as the process variation occurs randomly and the probability of worst and best case are very low in reality.

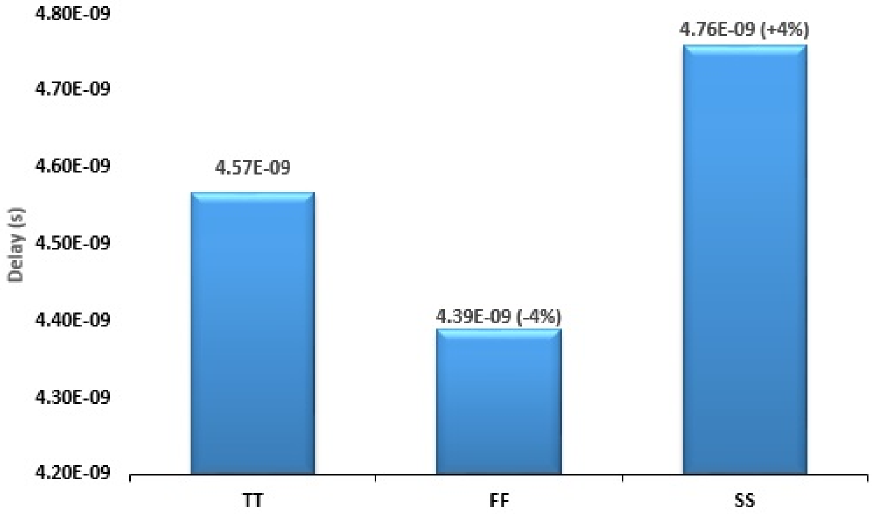


FIGURE 5.6: The worst and best case process variations of an inverter chain of 90 nm technology

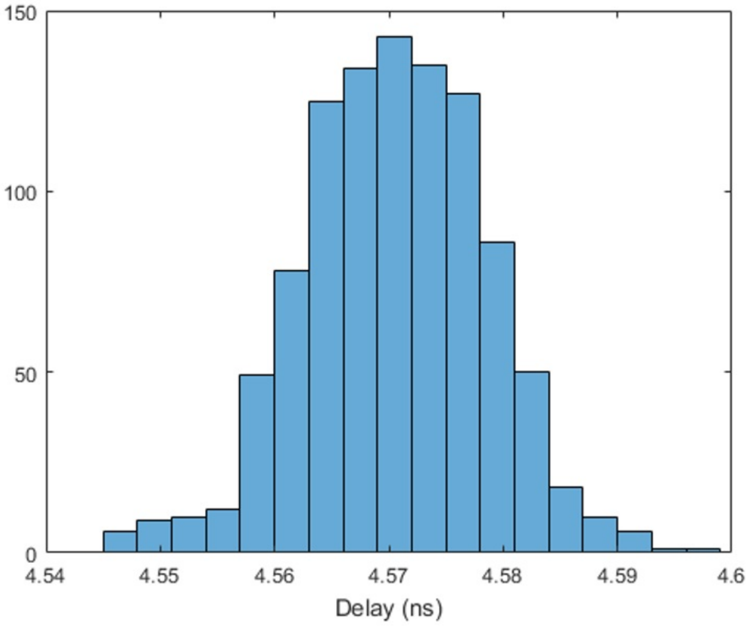


FIGURE 5.7: The random process variations distribution for an inverter chain of 90 nm technology sample size (1000)

Figure 5.7 shows the random process variation distribution for an inverter chain at 90 nm technology by using the spice process variation model with 1000 sample size. As the figure shows, the mean value of the delay is 4.57 ns which are the same with the typical case, and the worst and best case are not happening in those 1000 cases. Therefore, compared with the effect of process variations smaller circuit, in a complex or long circuitry, the process variation will not be significant compared with voltage and temperature variations in this technology node.

Voltage Variations

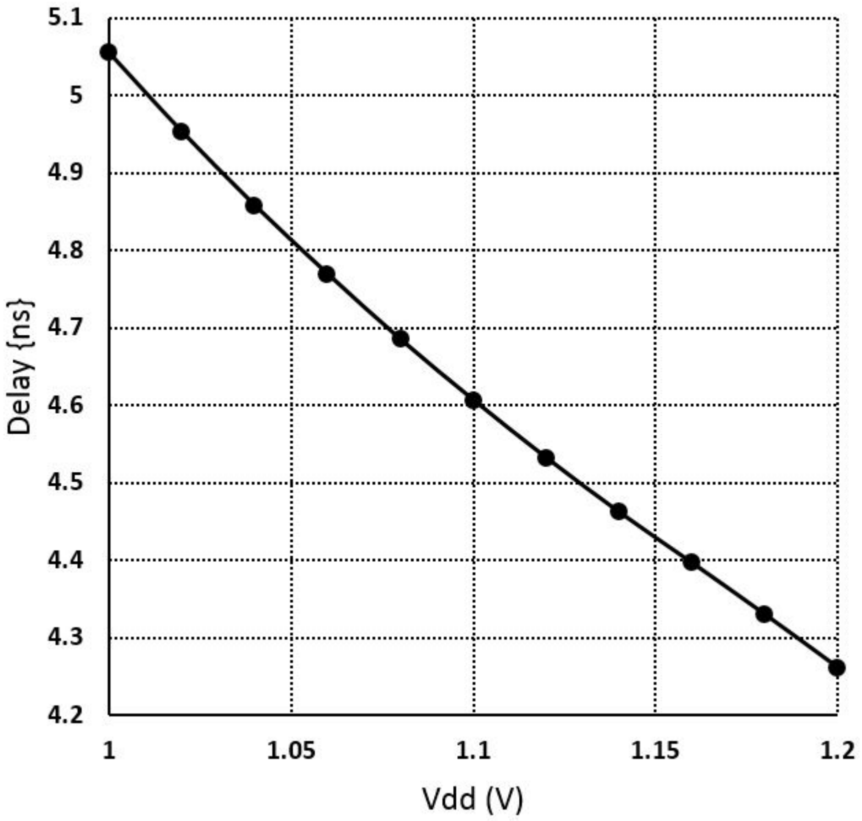


FIGURE 5.8: The Relation Between Supply Voltage and Delay

Figure 5.8 shows the relation between the supply voltage and logic delay of the long path in 90nm Technology. As the figure shows, the correlation between temperature and propagation delay is negative. Hence, means the performance of a digital circuit will decrease when the voltage goes down.

Temperature Variation

Figure 5.9 shows the delay relationship between temperature variation and propagation delay of an inverter chain using 90 nm technology in Spice. As the figure shows, the correlation between temperature and propagation delay is positive. The propagation delay increases rapidly when the temperature goes up, and this means the performance degradation of a CMOS circuit will become very significant while the circuit is running at the high performing stage.

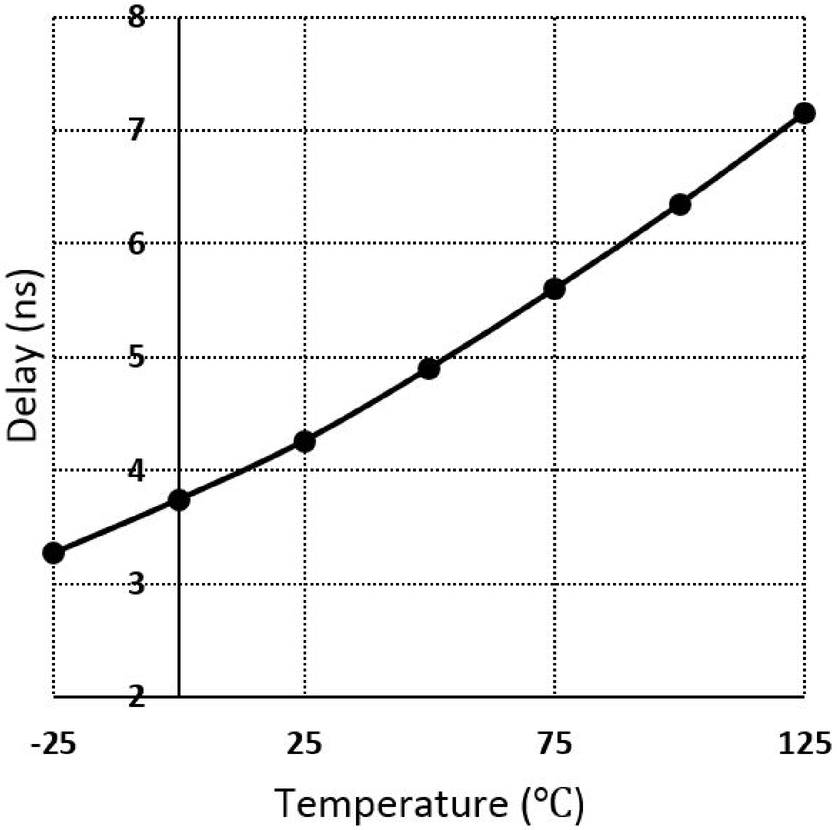


Figure 5.9: The Relation Between Temperature and Delay

Temperature((°C)	Delay (ns)	Delay degradation (%)
-25	384351	-24.22
0	4.41128	-13.03
25	5.07213	0.00
50	5.82282	14.80
75	6.64354	30.98
100	7.53145	48.49
125	8.46648	66.92

Table 5.1: Delay Degradation Comparison in Different Temperatures

Table. 5.1 shows the delay degradation of the inverter chain comparison in different temperatures. As the table shows, compared with the delay at room temperature,25℃, in the best case of temperature variation,-25℃, the propagation delay is decreased 24%, in the worst case of temperature variation,125℃, the propagation delay is increased about 67%.

The increase of propagation delay will lead to significant performance degradation, especially in the synchronous circuit system. Figure 5.10 shows the relationship between temperature variation, voltage variation and delay of an inverter chain at time zero with ±10% of voltage variation and -25% to 125%of temperature variation by using 90nm technology in Spice. Compared the delay degradation from process variation, as shown

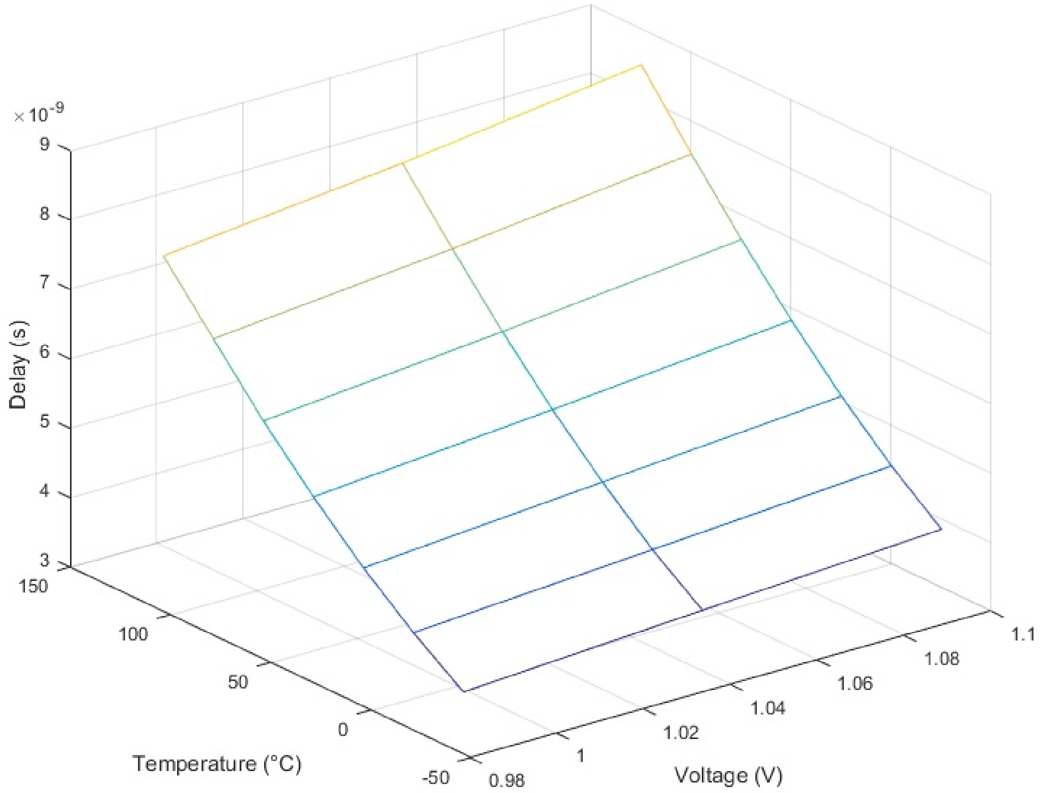


FIGURE 5.10: The Relationship between Temperature Variation, Voltage Variation and Delay Degradation at time zero

in Figure 5.6, the temperature variation has the most severe impact on propagation delay degradation.

Table 5.2 shows the intrinsic delay and the worst degradation of the long delay path. Where D_0 is the intrinsic delay at time zero and 25°C. The intrinsic delay degradation from PVT variations is estimated by SPICE with the worst process variation, Slow charge and Slow discharge (SS), voltage variation (-5%) and Temperature variation 125°C. 10% and 1% margin upon the D_0 are inserted for the delay degradation from NBTI and PBTI respectively, as the delay degradation from PBTI is about 10% of NBTI, [33]. Hence, the margin for the long delay path is defined as 4.377ns. The margin will be set as 4.345 for the circuit with DVS, as the ageing prediction sensors, like DMEDS, reserves a small safety margin for delay fault prediction, which is 32ps in this case.

TABLE 5.2: The Best and Worst Intrinsic Delay Degradation of a Long Delay Path

D_0	PVT Variations	NBTI Delay Degradation	PBTI Delay Degradation
4.26ns	3.44ns	0.852ns	0.085ns

Figure 5.11 shows the proportion of the timing constraint. As the Figure shows, the proportion of delay at time zero is less than 50% of the total timing constraint, which means the design without DVS will waste a lot of power during the ICs lifetime.

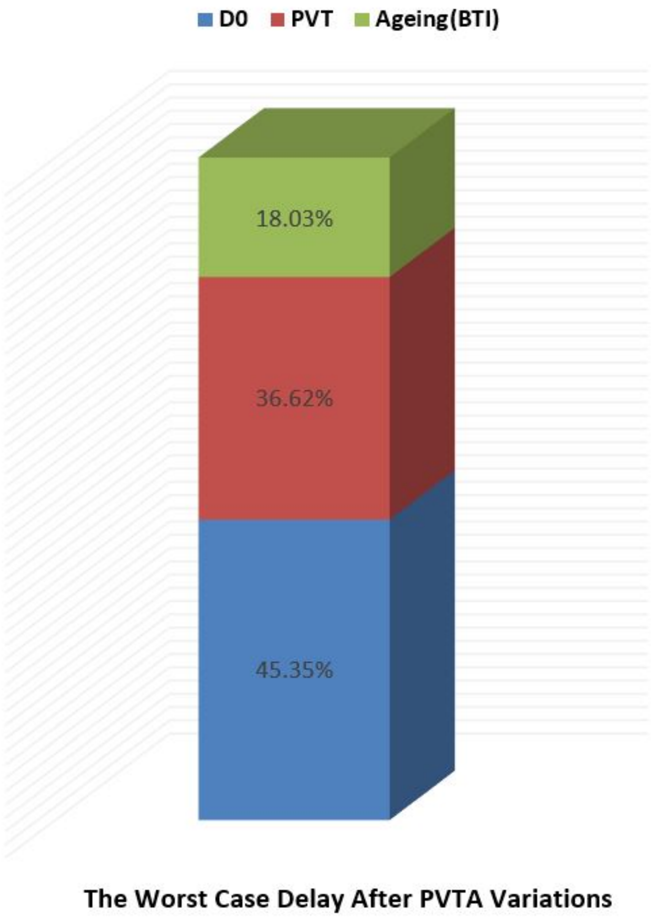


Figure 5.11: The margin for inverter Chain

5.3.2 Define the Characteristic of DVS

The ageing prediction algorithm compares the time between two highest scaling points as the lifetime is defined as the failure to meet the timing constraint while it is running at the worst condition during its lifetime. Therefore, the minimum reference voltage should be defined under the worst PVT condition, in case of the system running at a bad condition misses the first reference voltage.

Table 5.3: Intrinsic Delay with Different Vdd Under the Worst PVT variations

Vdd (V)	1.02	1.03	1.04	1.05	1.06
Delay (ns)	8.59	8.50	8.42	8.34	8.26

Table 5.3 shows the intrinsic delay with different supply voltage under the worst process and temperature variations of the long delay path. Where the worst temperature is defined at 125°C and the worst process are SS and voltage variation (-5%). As the table shows, the intrinsic delay decreases with the increase of the supply voltage. As the table shows, the delay intrinsic starts to violate the timing constraint when Vdd is under 1.03V. However, the delay at 1.03V is too close to the timing constraint, the

degradation from BTI grows in saturation curve during a circuit lifetime, as shown in equation (5.1). Therefore, the first reference will be a flash in the pan if 1.03V was set as the first reference. Hence, the minimum reference voltage is set at 1.04V. Existing DVS is able to scales the supply up until it exceeded the normal supply voltage of the technology node, [17, 40]. Assuming that the minimum scaling of a DVS is 0.04V and the highest Vdd is 1.24V. Then the reference voltage point will be 1.04V, 1.08V, 1.12V 1.16V, 1.20V and 1.24V.

5.3.3 Circuit Ageing Delay Estimation

To prove the derivation from Equitation (5.1) [33] to (5.8) is trustworthy, a different ageing model [10] for 90nm technology node NBTI threshold voltage shift estimation is applied to estimate the threshold voltage degradation after NBTI effect.

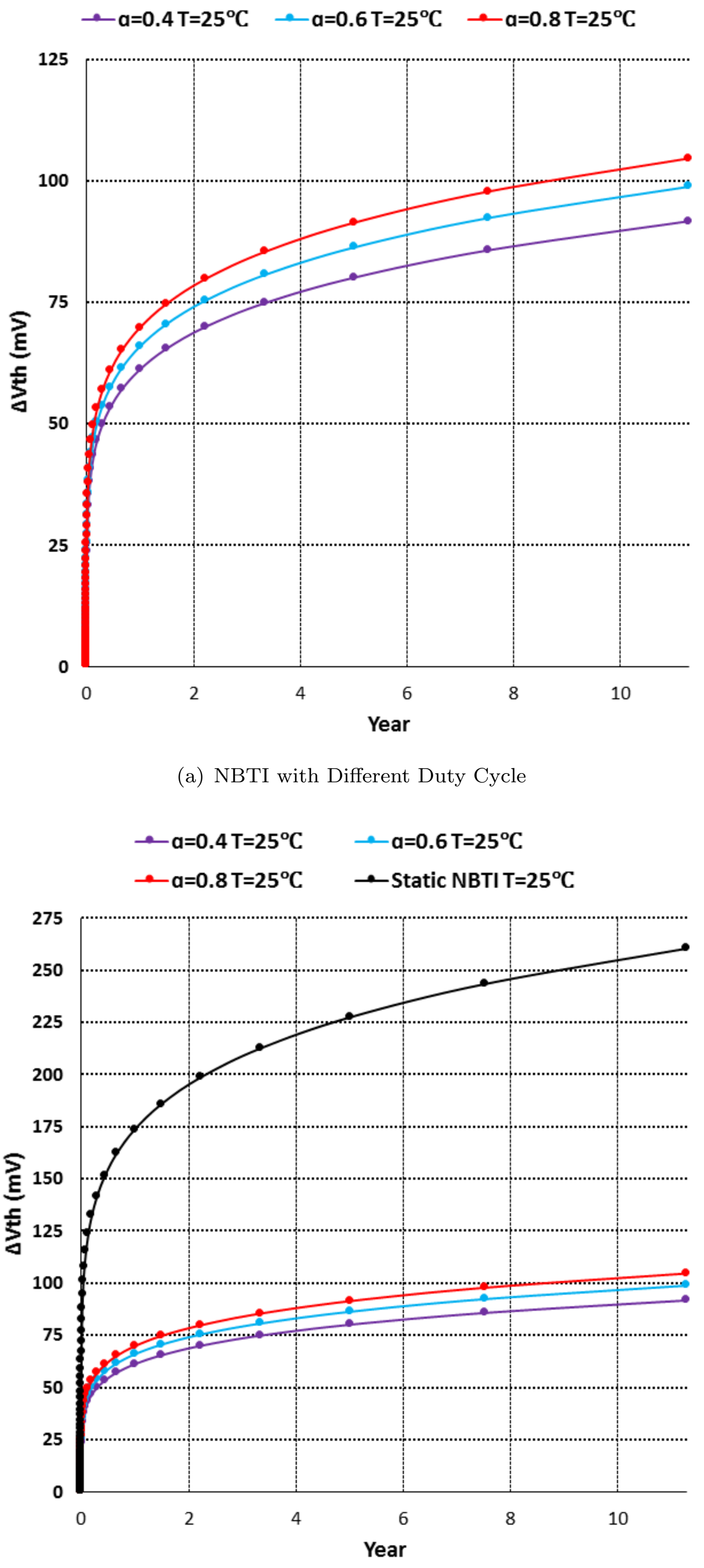
Duty cycle Dependence

The threshold voltage degradation from dynamic NBTI of a PMOS transistor depends on the duty cycle of the input signal. Figure 5.12 shows the threshold voltage degradation of static NBTI and dynamic NBTI for a minimum sized PMOS transistor of 90nm technology for different stress duty cycles at 25 °C with 100MHz switching activity and 1.2V Vgs over the time. As Figure 5.12 (a) shows, the Vth degradation will be stronger with higher duty cycle. As the results show in Figure 5.12 (b), ΔV_{th} grows over time, and the Vth degradation of static NBTI is much stronger than dynamic NBTI. Because of the recovery of ΔV_{th} is rapid at the beginning of recovery time. Compare to static NBTI, ΔV_{th} of dynamic NBTI is significantly decreased even though the duty cycle is close to one. Because of the correlation between duty cycle and ΔV_{th} is positive, the input vector control can mitigate the performance degradation from NBTI, as shown in, [37, 83].

Temperature Dependence

The threshold voltage degradation from dynamic NBTI of a PMOS transistor depends on the operating temperature. Figure 5.13 shows the threshold voltage degradation due to dynamic NBTI for a minimum sized PMOS transistor of 90nm technology in temperature with 100MHz switching activity, Vgs=1.2V and 0.5 stress duty cycle over the years. As the estimate ΔV_{th} as shown in Figure 5.13, the correlation between temperature and ΔV_{th} are positive. As the temperature increases, the degradation of the Vth increases sharply.

In order to show the influence of temperature changes on the Vth degradation in detail, the Vth degradations are estimated and compared with relatively smaller temperature



(a) NBTI with Different Duty Cycle

(b) Compared with The Static NBTI

FIGURE 5.12: Long Term Dynamic NBTI with Different Duty Cycle

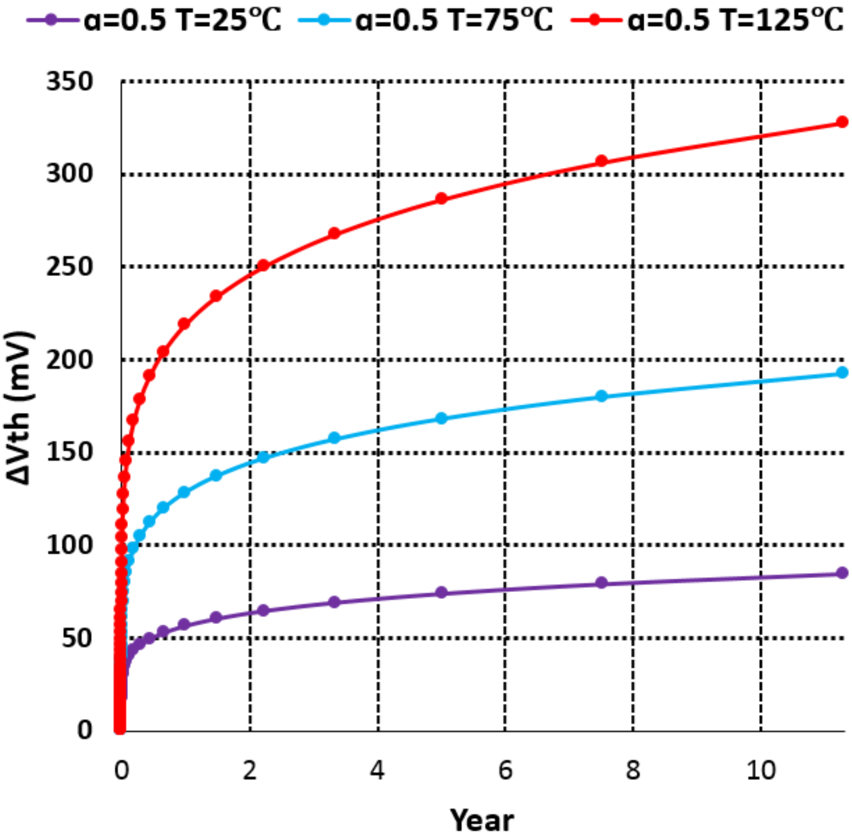


FIGURE 5.13: Long Term Dynamic NBTI with Different Temperature

Temperature(°C)	Delay (ns)	Delay degradation
25	82.99	30.00%
30	91.58	33.18%
35	100.65	36.47%
40	110.19	39.93%
45	120.19	43.55%
50	130.63	47.32%

TABLE 5.4: NBTI Vth Degradation Comparison in Different Operating Temperatures

changes shown in Table 5.4. According to the estimation results, shown in the Table, each temperature increase of 5 degrees, will lead to a 3% degrade of Vth.

Vgs Dependence

Figure 5.14 shows the threshold voltage degradation of dynamic NBTI for a minimum sized PMOS transistor of 90nm technology with different Vgs after 5 years. As the estimated results shown, the correlation between Vgs and ΔV_{th} is positive.

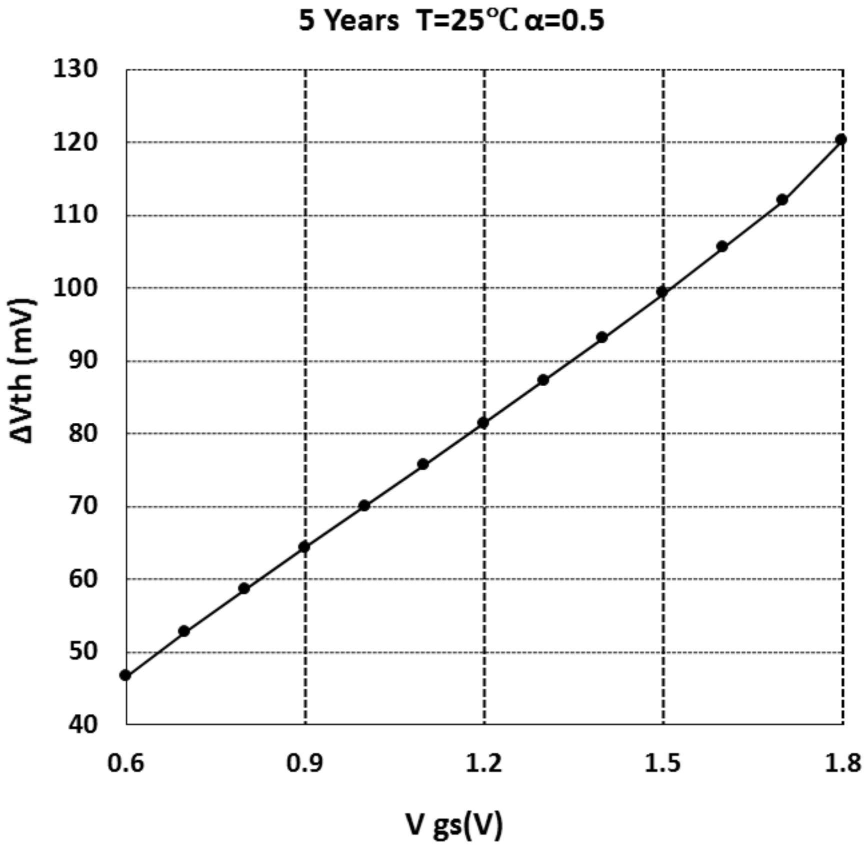


FIGURE 5.14: Long Term Dynamic NBTI with Different Vgs

5.3.4 Reference Circuit Ageing Behaviour Estimation

NBTI threshold voltage shift estimation is applied to estimate the behaviour after NBTI effect, which is introduced in the last Section, and the delay degradation from PBTI is considered as 10% of the NBTI. Figure 5.15 shows the circuit ageing behaviour estimation flow chart for the estimation of a circuit with DVS. The circuit with DVS will run at the lowest supply voltage at time zero under specific PVT variations. The supply voltage then increases slowly as time goes on. As the Figure shows, the first step of the estimations is scaled the Vdd to the lowest point without violating the timing constraint. It then increases the operating time to estimate the threshold voltage shift. A circuit will not always run under the worst voltage and temperature variations during its lifetime. The average temperature should be much lower than the worst case during a circuit's lifetime. Therefore, the voltage level should also be lower than the voltage level in the worst case. The worst case only applies to estimate the highest voltage scaling point in Spice. An average operating temperature and supply voltage should be applied during the estimation of delay degradation. As time goes on, the intrinsic delay will start to violate the timing constraint. There are two kinds of violations:

1. The violations under the worst operating conditions: In this case, the intrinsic

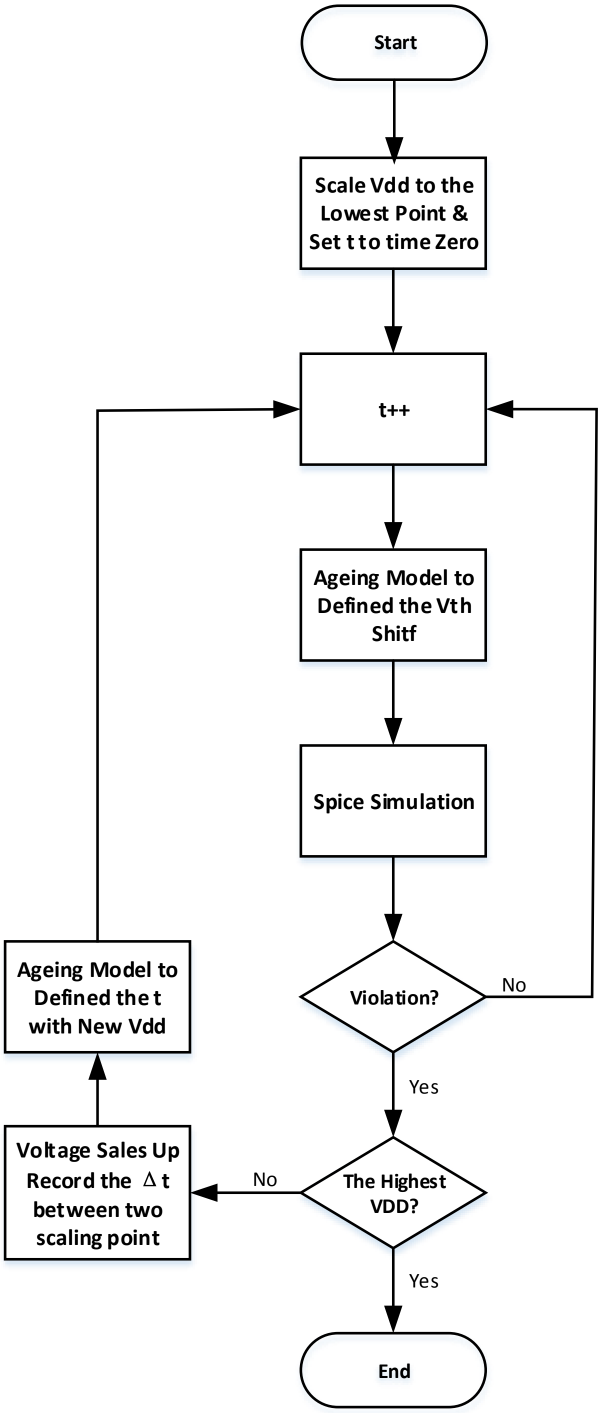

Start
Scale Vdd to the Lowest Point & Set t to time Zero
t++
Ageing Model to Defined the Vth Shitf
Spice Simulation
Violation?
No
Yes
The Highest VDD?
No
Yes
End
Voltage Sales Up Record the Δ t between two scaling point
Ageing Model to Defined the t with New Vdd

FIGURE 5.15: Circuit Ageing Behaviour Estimation Flow Chart

delay in the worst operating condition violates the timing constraint, start from 1.04V till 1.24V. The timing of the violation should the record as a reference point.

2. The violations under the average operating conditions: In this case, the intrinsic

delay will violate the timing constraint under the average operating condition. The time of the violation should be updated by import the ageing conditions back to the model. The fitting parameter K as shown in equation (5.1) should be bigger at the higher average voltage state than the K at relatively lower voltage state as the supply voltage dependence of NBTI, stated in Section 2.2.2. The operating time should be smaller if K goes bigger under the same ageing condition. The circuits ageing behaviour will then be estimated by an updated average supply voltage and operating time until the next violation. The total time is calculated by the time between two scaling point.

The Ageing behaviour estimation completes after the violation occurs at the highest voltage when the circuit is running under the worst operating condition, which is the end of the circuit's lifetime.

To provide the comparison of ageing mitigation the lifetime of the circuit without DVS is also estimated, as shown in Figure 5.16. The Figure shows the intrinsic delay estimation of the long path under the worst operating condition. The threshold voltage shift was estimated by using random process variation, random input vector, 1.20V supply voltage, 50°C average temperature and 125°C operating temperature in the worst case. As the figure shows, the intrinsic delay grows in saturation curve during its lifetime, which violates the timing constraint after 5.73 years.

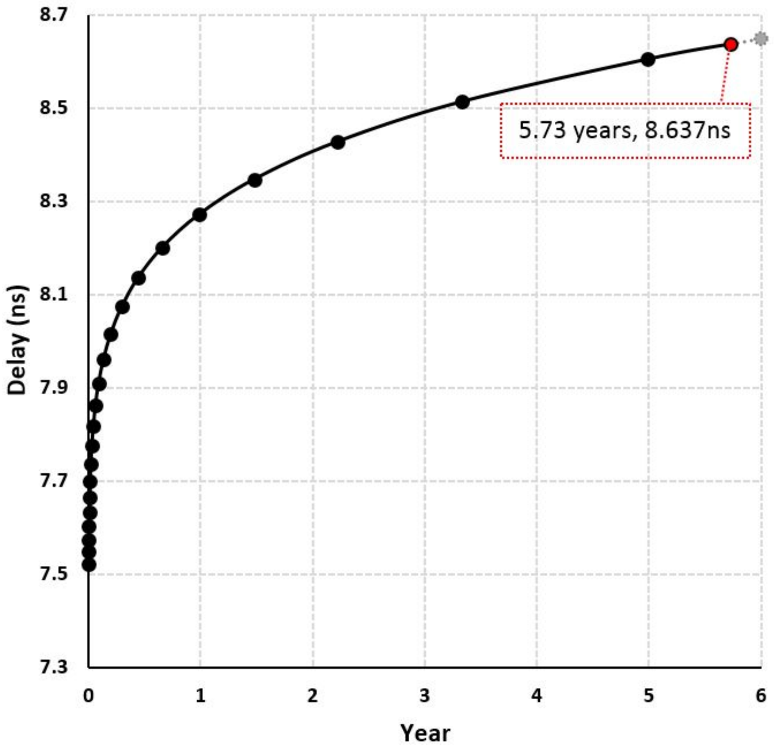


Figure 5.16: Circuit Lifetime Estimation When T=50°C

Figure 5.17 shows the lifetime estimation of a circuit with DVS. The intrinsic delay in the Figure is the simulation result from Spice under the worst operating condition. The delay degradation was estimated by using random process variation, random input vector, 50°C, average temperature, various supply voltage and 125°C operating temperature in the worst case. Compared with the result as shown in Figure 5.16 the delay unused timing margin is removed without violating the timing constraint. The timing constraint, in this case, is tighter compared with the circuit without DVS as the small safety margin left for the ageing prediction sensor. The time between two violation decreases over time as the margins are relatively small compared with the circuit without DVS and the degradation of the delay grows after at the beginning of the lifetime. The timing constraint is finally violated after more than 50 years when the supply voltage has reached the maximum value. However, the lifetime of the circuit after the violation, as the violation of the small safety margin will not cause the violation of the real timing constraint. The error signals will be triggered continuously by the ageing sensors until the end of circuit's lifetime. The lifetime of the circuit extends to 59.82 years. Furthermore, the lifetime increases more than 100% even though the supply voltage does not exceed the normal value of the supply voltage (1.20V) as the average operating voltage is smaller than the circuit without DVS, as shown in Figure 5.18.

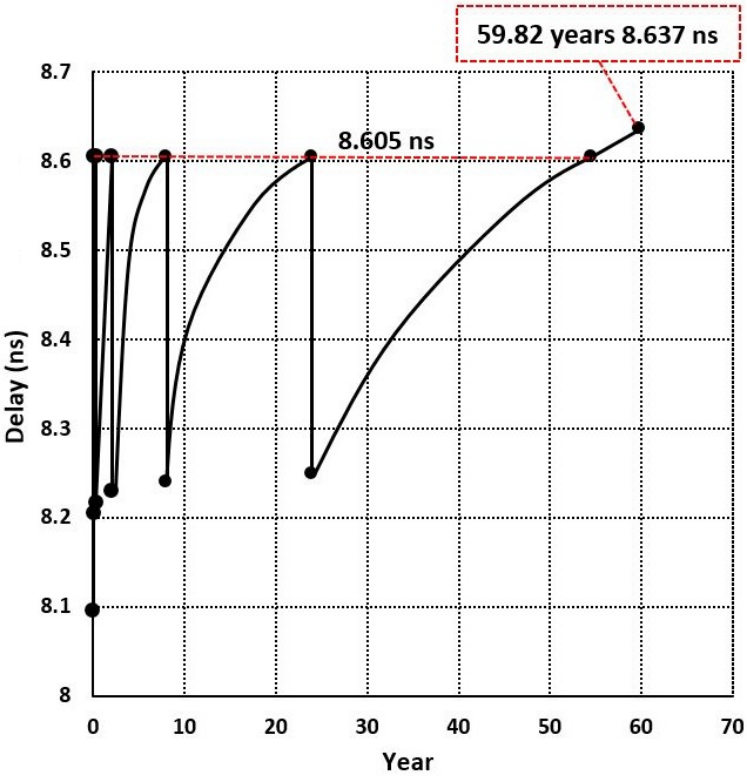


FIGURE 5.17: Lifetime Estimation of a Circuit With DVS When T=50°C

Figure 5.18 shows the average voltage level during the circuit's lifetime. The average supply of the circuit grows faster at the beginning of the lifetime compared with the end of the lifetime as the unbalanced ageing speed of BTI effect. The average supply voltage in each level is generally lower than the circuit without DVS, which will reduce power consumption significantly. On the other hand, the value of average voltage level affects the speed of ageing stated in Section 2.2.2. The time of average supply remaining is a valuable data which need to be counted and calculated by the DVS.

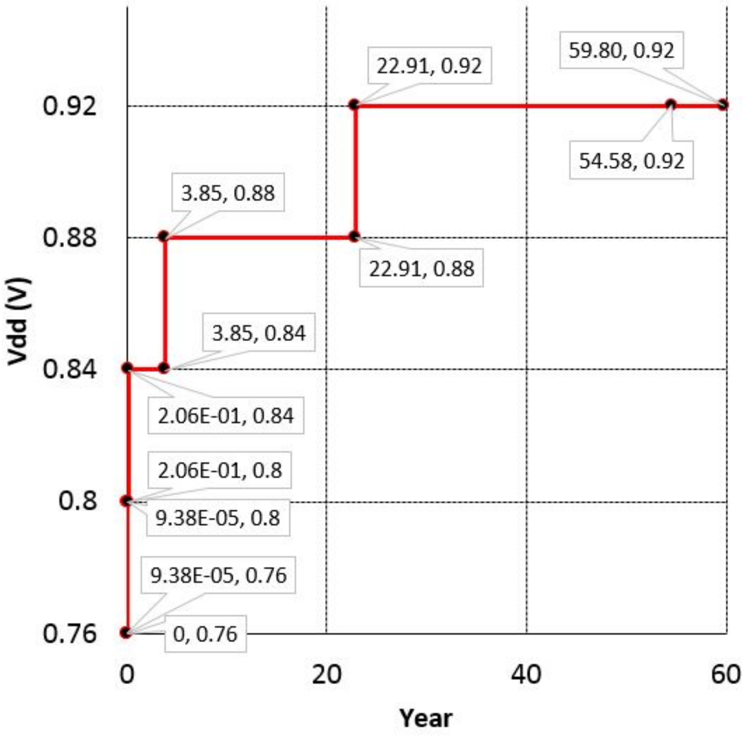


Figure 5.18: Average Vdd During the Circuit's Lifetime When T=50°C

$$Power = C.V^2.f \tag{5.12a}$$

$$PercentageofPowerSaving = 1 - \frac{V^2_{current}}{V^2_{Normal}} \times 100\% \tag{5.12b}$$

Equation 5.12a [12] shows the power consumption calculation of a digital IC. Where C is the total capacitance of the design, V is the supply voltage, and f is the operating frequency. The calculation in terms of power saving with and without voltage scaling can be derived as equation 5.12b because the total capacitance of the same circuit and operating frequency stays the same with or without DVS.

Table 5.5 shows the weighted average power saving compared with the circuit without DVS when the average operating temperature is at 50°C. As the Table shows, the circuit with DVS saves more power at the beginning of the lifetime. The power consumption then increases over time as the scaling up of the average voltage level. The weight of low voltage is very small. The average voltage level is mainly staying at 0.88V and 0.92V

TABLE 5.5: Weighted Average of Power Saving When Average T = 50°C

Voltage Level (V)	0.76	0.80	0.84	0.88	0.92
Time (Years)	9.50×10^{-5} (0.00%)	0.21 (0.34%)	3.65 (6.10%)	19.05 (31.86%)	36.9 (61.70%)
Power Saving	59.99%	55.56%	51.00%	46.22%	41.22%
Weighted Average of Power Saving : 43.46%					

which save 46.22% and 41.22% of power respectively. The weighted average of power saving is 43.46% compared with the original design. In practice, the DVS requires extra circuitry such as the insertion of sensors, which will also increase the power consumption of the circuit. However, the cost-efficient sensors such as DMEDS and PCC will not lead to an obvious growth in terms of power consumption. The power overhead is negligible compared with the power saving of the whole circuit.

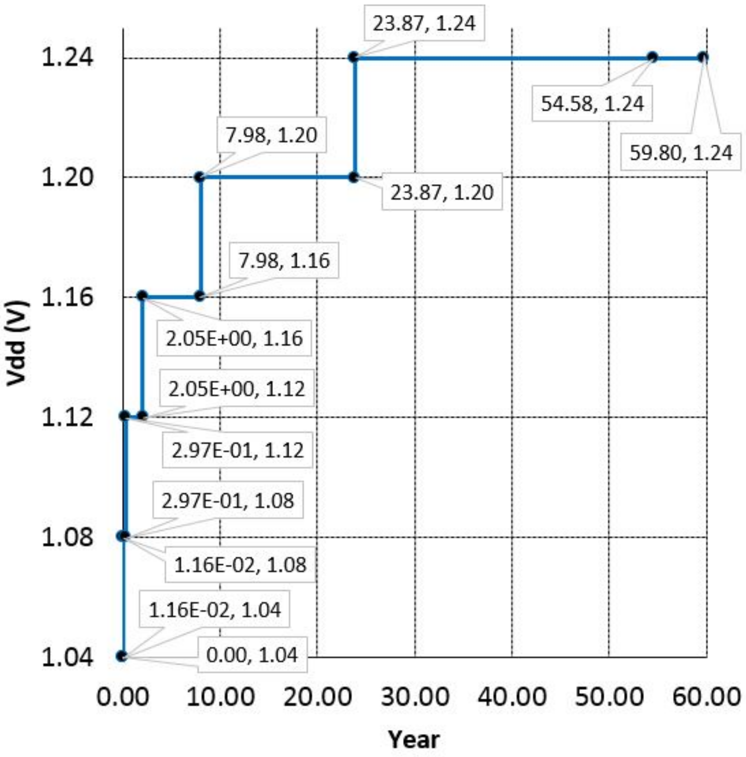


FIGURE 5.19: Maximum Vdd During the Circuit's Lifetime When T=50°C

Figure 5.19 shows the highest supply voltage during the circuit's lifetime. The DVS is not able to monitor the actual value of intrinsic delay, as shown in Figure 5.17. The remaining time of each highest voltage level during the circuit's lifetime symbolises the state of ageing as the time between two voltage scaling point is identical with the result from the relationship between maximum voltage level and the highest voltage scaling point, as shown in Figure 5.17. Therefore, the time between two voltage reference point is also a valuable data which need to be recorded by the DVS. The ratio of VDD is shown in Table 5.6 which is calculated by the data from Figure 5.19 and 5.18.

Table 5.6: The Ratio of Average VDD $T_{avg} = 50°C$

$T_{avg} = 50°C$	V_{max} (V)	1.08		1.12	1.16		1.20		1.24
	t_{max} (Year)	0.28		1.75	5.93		15.88		30.71
	V_{avg} (V)	0.8	0.84	0.84	0.84	0.88	0.88	0.92	0.92
	Ratio	0.32	0.68	1.00	0.30	0.70	0.94	0.06	1.00

5.3.5 Discussion

This Section introduces the method and the result of ageing behaviour estimation as a reference. A different model estimated the ageing degradation with the model used to deduce the theory which is stated in the last Section. The reference data comes from the co-simulation and analysis of an RD model and SPICE at the 90nm technology node. In practice, the DVS will collect the valuable data, as shown in Figure 5.18 and 5.19. It then analysed the data and compares the result with reference to estimate the lifetime of the circuit, as will be shown later.

5.4 Verification and Analysis

This section presents verification of the theory stated in Section 5.2. The data compared and analysed in this section was estimated under different operating conditions by the method shown in the last Section. Section 5.4.1 shows a detailed comparison between the same circuit running under two different operating conditions. In practice, one references data is not enough to estimate the lifetime under different circumstances. Therefore, more than one reference data should be provided in different circumstances. Section 5.4.2 shows a set of lifetime estimation of a circuit compared with more than one reference.

5.4.1 Lifetime Estimation and Comparison in Details

In order to provide comprehensive verification and analysis. The intrinsic delay degradation due to BTI is also estimated by the method shown in the last Section. Figure 5.20 shows the ageing behaviour estimation when the circuit is running with random process variation, random input vector, 1.20V supply voltage, 45°C average temperature and 125°C operating temperature in the worst case. As the figure shows, the circuit, in this case, will start to fail after 8.637 years. Compared with the result, as shown in Figure 5.16, the lifetime can be extended 50.73% with decrease 5°C reduction of the average temperature.

Figure 5.21 shows the lifetime estimation of a circuit with DVS when random process variation, random input, 45°C average temperature, various supply voltage and 125°C

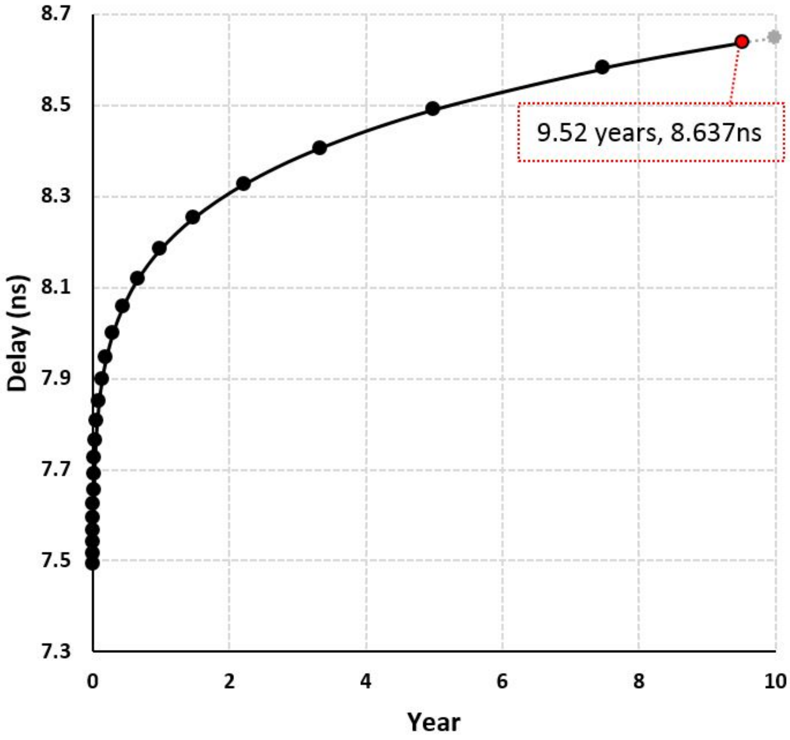


FIGURE 5.20: Circuit Lifetime Estimation when T=45°C

operating temperature in the worst case are applied. The lifetime of the circuit, in this case, is increased 10.24 times compared with circuit without DVS. Compared with the reference data, which the lifetime is in increased 9.44 times, the DVS is more effective in terms of ageing mitigation when the average temperature is lower.

Figure 5.22 shows the average voltage level during the circuit's lifetime when the average operating temperature is 45°C. Compared with the result from the reference circuit, the average supply voltage generally remains at a lower voltage in this case, which will result in better ageing mitigation and power saving.

TABLE 5.7: Weighted Average of Power Saving When Average T = 45°C

Voltage Level (V)	0.76	0.80	0.84	0.88	0.92
Time (Years)	6.88×10^{-3} (0.01%)	1.24 (1.16%)	13.40 (12.51%)	56.32 (52.61%)	36.10 (33.72%)
Power Saving	59.99%	55.56%	51.00%	46.22%	41.22%
Weighted Average of Power Saving : 45.24%					

Table shows 5.7 shows the weighted average power saving compared with the circuit without DVS when the average operating temperature is 45°C. As the Table shows, the weighted average of power saving is 45.24% compared with the circuit without DVS. Compared with the reference data, the DVS saves more power in this case as the increase of the weight at the lower voltage state during the circuit's lifetime.

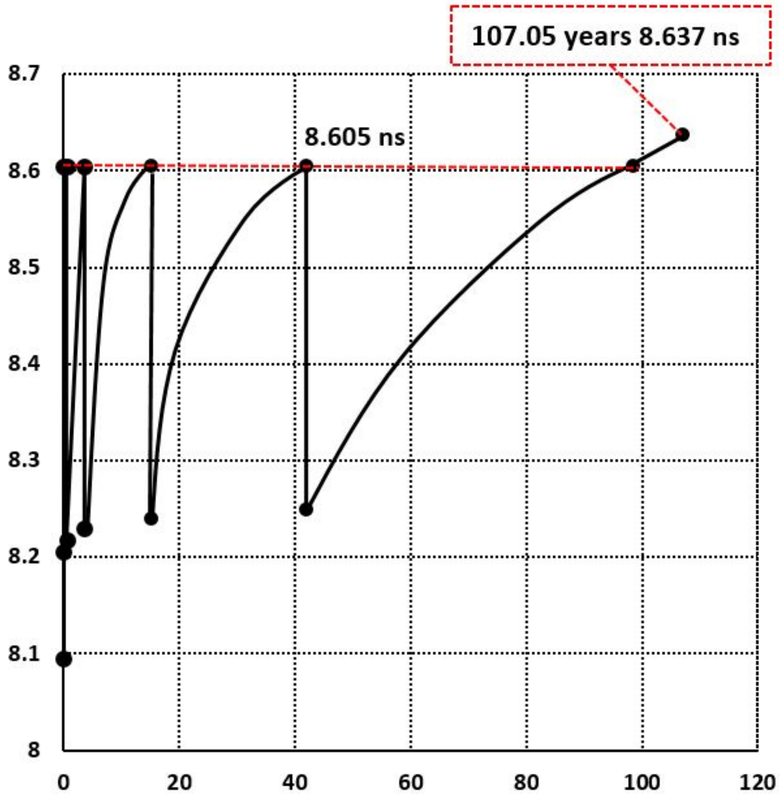


FIGURE 5.21: Lifetime Estimation of a Circuit with DVS when T=45°C

Figure 5.23 shows the highest supply voltage during the circuit's lifetime. The as the comparison in terms of average supply voltage, the phase of the curve, as shown in this Figure shifts right compared with the reference data. The lifetime will then be estimated merely by comparing the remaining time at the highest supply voltage stage as the phase shift direction of the data represents ageing speed (the remaining time at the average) and the data represents the ageing states are identical. The result of lifetime prediction is shown in Table 5.8.

TABLE 5.8: Lifetime Estimation When T=45°C

T_{avg}	V_{max} (V)	1.08		1.12		1.16		1.20		1.24	
50°C	t_{max} (Year)	0.28		1.75		5.93		15.88		30.71	
	V_{avg} (V)	0.8	0.84	0.84		0.84	0.88	0.88	0.92	0.92	
	Ratio	0.32	0.68	1.00		0.30	0.70	0.94	0.06	1.00	
45°C	V_{max} (V)	1.08		1.12		1.16		1.20		1.24	
	t_{max} (Year)	0.51		3.07		11.44		26.75		56.78	
	V_{avg} (V)	0.80		0.8	0.84	0.84	0.88	0.88		0.88	0.92
	Ratio	1.00		0.23	0.77	0.96	0.04	1.00		0.52	0.48
	LT_{est} (Year)	107.6		104.3		115.4		100.7		110.2	
	Error %	0.49		-2.57		7.80		-5.90		2.91	

Table 5.8 shows the lifetime estimation for the circuit when the average temperature is 45°C. Where T_{avg} is the average temperature that the circuit is operating, V_{max}

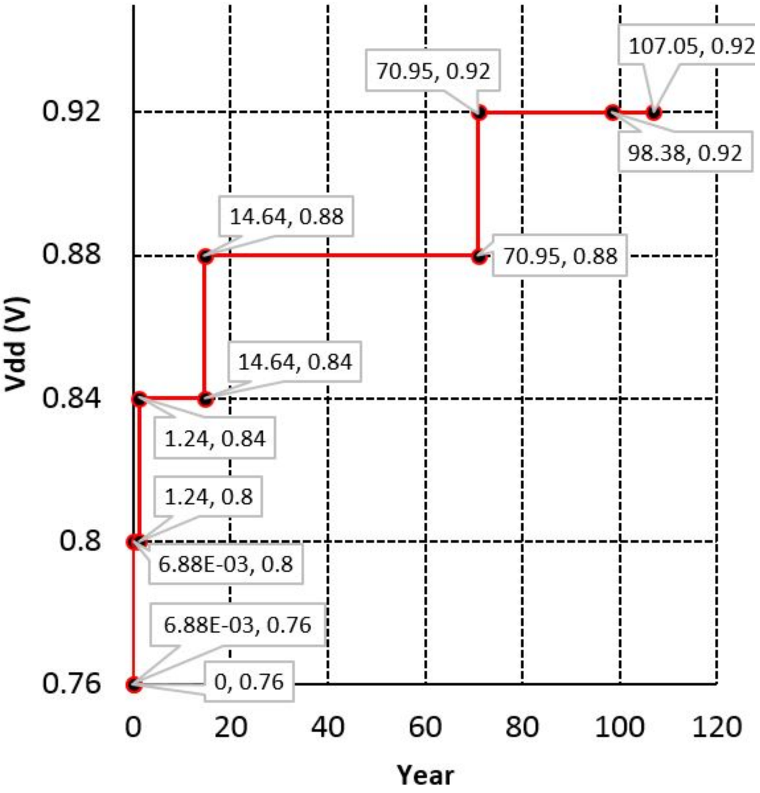


FIGURE 5.22: Average Vdd During the Circuit's Lifetime When T=45°C

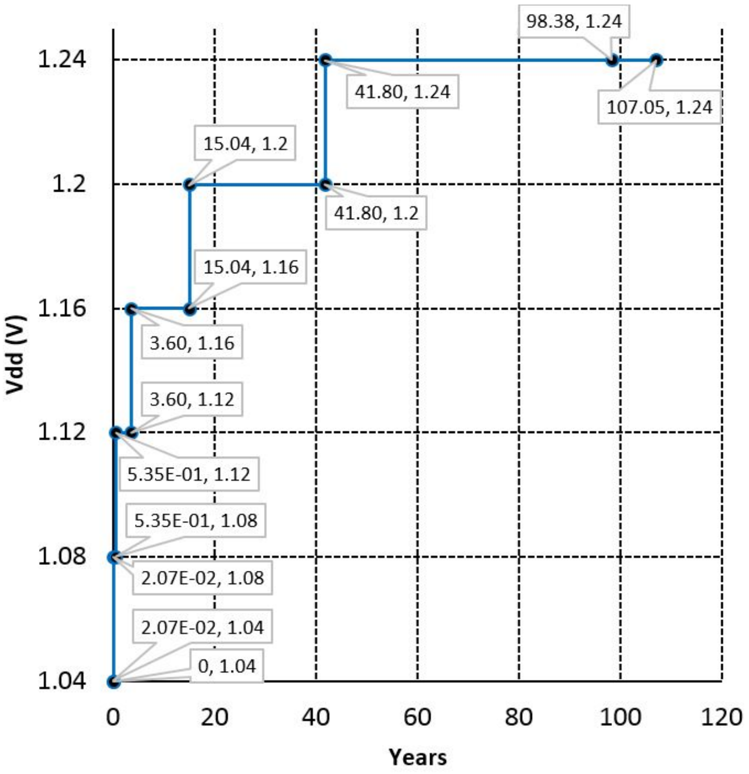


FIGURE 5.23: Maximum Vdd During the Circuit's Lifetime When T=45°C

is the maximum voltage level, t_{max} is the remaining time at each maximum voltage level, V_{avg} is the average voltage level, LT_{est} is the estimated lifetime and LT_{ref} is the lifetime of the reference circuit. The Lifetime is estimated by comparing t_{max} between the reference circuit and the target circuit, as shown in equitation 5.11. As the table shows, the lifetime is estimated with maximum 7.8% of error, which is caused by the unbalanced ratio of the average voltage and the mismatch between two ageing models.

5.4.2 Lifetime Estimation and Comparison with More than One Reference

Table 5.9 shows a set of lifetime estimation when the circuits are running under different average temperature with random process variation and input vectors.

Table 5.9: Reference Data When T= 55°C and 50°C

T_{avg}	V_{max} (V)	1.08		1.12		1.16		1.20		1.24	
55°C	t_{max} (Year)	0.15		1.02		3.42		8.49		18.23	
	V_{avg} (V)	0.8	0.84	0.84	0.88	0.88		0.88	0.92	0.92	0.96
	Ratio	0.11	0.89	0.26	0.74	1		0.29	0.71	0.84	0.16
T_{avg}	V_{max} (V)	1.08		1.12		1.16		1.20		1.24	
50°C	t_{max} (Year)	0.28		1.75		5.93		15.88		30.71	
	V_{avg} (V)	0.8	0.84	0.84		0.84	0.88	0.88	0.92	0.92	
	Ratio	0.32	0.68	1.00		0.30	0.70	0.94	0.06	1.00	

Table 5.10 shows the lifetime estimation for the circuit when the average temperature is from 51% to 54%. As the table shows, the lifetime is estimated with maximum 4.16% of error (absolute value). Existing ageing prediction models will produce about 3% of error in every 1°C temperature difference, as shown in Table 5.1. Therefore, the accuracy of the ageing prediction system presented in this Chapter is a lot more precise compared with existing ageing prediction models, [10, 77].

5.5 Concluding Remarks

In this chapter, we propose an idea of lifetime prediction system, and the ageing prediction controller receives the error signal from the ageing prediction sensors such as DMEDS and PCC, it then compares the data with reference to analysis the ageing of the device. The ageing data of the reference comes from the existing ageing model. A warning signal will be given after the analysis. The user of the device will receive a warning message with the estimated lifetime of the device. Our results indicate that the use of the proposed system for lifetime prediction system can accurately estimate the lifetime of the IC compared with the data from the ageing model, the error is less than 5% with limited reference data. Together with the DVS, the lifetime can be extended about 10 times under a reasonable operating condition.

Table 5.10: Target Data When T= 51°C - 54°C

54°C	V_{max} (V)	1.08		1.12		1.16		1.20		1.24	
	t_{max} (Year)	0.17		1.13		3.77		9.56		20.7	
	V_{avg} (V)	0.80	0.84	0.84	0.88	0.88		0.88	0.92	0.92	0.96
	Ratio	0.18	0.82	0.91	0.09	1.00		0.41	0.59	0.98	0.02
	LT_{est} (Year)	38.45		37.59		37.40		38.21		38.53	
	Error %	0.53		-1.73		-2.22		-0.10		0.73	
53°C	V_{max} (V)	1.08		1.12		1.16		1.20		1.24	
	t_{max} (Year)	0.19		1.31		4.18		10.80		22.88	
	V_{avg} (V)	0.80	0.84	0.84		0.84	0.88	0.88	0.92	0.92	
	Ratio	0.27	0.73	1.00		0.03	0.97	0.53	0.47	1	
	LT_{est} (Year)	42.98		43.58		41.47		43.16		42.58	
	Error %	-0.68		0.71		-4.16		-0.25		-1.59	
52°C	V_{max} (V)	-2.31		2.31		-1.89		-4.24		2.1	
	t_{max} (Year)	0.22		1.44		4.68		12.23		25.22	
	V_{avg} (V)	0.8	0.84	0.84		0.84	0.88	0.88	0.92	0.92	
	Ratio	0.37	0.63	1		0.12	0.88	0.66	0.34	1	
	LT_{est} (Year)	46.99		49.21		47.19		46.06		49.11	
	Error %	-2.31		2.31		-1.89		-4.24		2.10	
51°C	V_{max} (V)	1.08		1.12		1.16		1.20		1.24	
	t_{max} (Year)	0.25		1.59		5.26		13.9		27.8	
	V_{avg} (V)	0.8	0.84	0.84		0.84	0.88	0.88	0.92	0.92	
	Ratio	0.51	0.49	1		0.20	0.80	0.79	0.21	1	
	LT_{est} (Year)	53.39		54.33		53.04		52.34		54.13	
	Error %	-0.37		1.38		-1.03		-2.33		1.01	

Chapter 6

Conclusions

6.1 Conclusions

The feature size of the technology node has been shrunk by more than 700 times, in the last five decades. The technology scaling results in a significant achievement regarding performance, power-efficiency and device density of the semiconductor devices. However, the side effect leads to the increasing of reliability issues. The process, voltage, temperature variations and ageing-induced device degradation are becoming major reliability concerns in modern semiconductor technologies.

All the shortage of excising state-of-the-art techniques have been clearly listed in Chapter 1 :

1. Technology scaling arises the reliability issues of IC by the PVTA variations.
2. The cost of existing sensors implementation is very high, and the implementation usually replaces the original Flip Flop of the design as a sensor, which brings more reliability issues of ICs. The sensors are suffering from the metastability, and usually not suitable for ageing prediction.
3. Existing lifetime prediction techniques are either unreliable or not cost-efficient, and the state-of-the-art ageing mitigation techniques are usually sacrifices the performance of an IC.

The background study for this research fulfils the first shortage. The limitation of the existing research has been addressed and improved in this research.

The second shortage objective is fulfilled by the two cost cost-efficient ageing sensors, stated in Chapter 3 and 4 respectively. The sensor proposed in those Chapters can monitor multiple paths at the same time. Both sensors are the external sensor which can monitor the delay fault without replacing the original Flip Flop. Moreover, the proposed, DMEDS, not only reduced the area overhead significantly but also resistance to the metastability, therefore increased the reliability for the delay fault prediction. Our results indicate that using the PCC sensor for delay fault monitoring in a 32-bit MIPS can lead to a significant saving in the area and power overheads, compared to the use of canary flip-flops [40]: by two-thirds and one-third, respectively. Moreover, the use of the proposed sensor for delay fault monitoring across 10 paths can lead to a significant saving in area overhead compared to Razor [29], and Canary [40]: 87.59%, 77.67%, respectively.

The third shortage has fulfilled the idea of lifetime prediction system, and the ageing prediction controller receives the error signal from the ageing prediction sensors such as DMEDS and PCC, it then compares the data with reference to analysis the ageing of the device. The ageing data of the reference comes from the existing ageing model. Our results indicate that the use of the proposed system for lifetime prediction system can accurately estimate the lifetime of the IC compared with the data from the ageing model. The error is controlled within 5% with limited reference data. More ever together with the DVS, the lifetime can be extended about 10 times without under a reasonable operating condition without sacrifice the performance.

6.2 Future work

The ageing prediction system outlined in Chapter 5 presents the theory and the method for the lifetime prediction by using the limited information provided by the DVS. The test circuit for the verification of the ageing prediction system is a simple inverter chain. More complex circuits are expected for the lifetime prediction, although the thesis has proved the ageing prediction system will be functional in a complex circuit. However, the workload of ageing analysis for a complex circuitry is equivalent to a new PhD project. The research of ageing analysis for complex circuitry work is carried out with the research presented in this thesis simultaneously in the Electronic Systems and Devices research group at the University of Southampton, [33, 34]. The incomplete work can be fulfilled by combining the method from those two PhD project.

Based on the finding presented in this thesis, two directions for future research are identified and described as follows:

1. A reliable Cross-layer design regarding ageing. Existing research groups focus on the ageing mitigation focus on their area of expertise. However, the ageing effect can be mitigated in many ways. A cross-layer reliable design can mitigate the ageing in many aspects.

2. Ageing Prediction by using the big data analysis. The work in Chapter 5 applies the minimal information estimated the lifetime accurately. Big data analysis together with the Internet of Things (IoT) will further improve the accuracy regarding lifetime prediction.

References

[1] Haider Muhi Abbas, Basel Halak, and Mark Zwolinski. Bti mitigation by anti-ageing software patterns. *Microelectronics Reliability*, 79:79–90, 2017.

[2] Shady Agwa, Eslam Yahya, and Yehea Ismail. Ersut: A self-healing architecture for mitigating pvt variations without pipeline flushing. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 63(11):1069–1073, 2016.

[3] Muhammad Ashraful Alam, Haldun Kufluoglu, Dhanoop Varghese, and Souvik Mahapatra. A comprehensive model for pmos nbti degradation: Recent progress. *Microelectronics Reliability*, 47(6):853–862, 2007.

[4] Muhammad Ashraful Alam and Souvik Mahapatra. A comprehensive model of pmos nbti degradation. *Microelectronics Reliability*, 45(1):71–81, 2005.

[5] Massimo Alioto, Elio Consoli, and Gaetano Palumbo. General strategies to design nanometer flip-flops in the energy-delay space. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 57(7):1583–1596, 2010.

[6] A Asenov, S Roy, RA Brown, G Roy, C Alexander, C Riddet, C Millar, B Cheng, A Martinez, N Seoane, et al. Advanced simulation of statistical variability and reliability in nano cmos transistors. In *Electron Devices Meeting, 2008. IEDM 2008. IEEE International*, pages 1–1. IEEE, 2008.

[7] Sanghyeon Baeg, Pierre Chia, ShiJie Wen, and Richard Wong. Dram failure cases under hot-carrier injection. In *Physical and Failure Analysis of Integrated Circuits (IPFA), 2011 18th IEEE International Symposium on the*, pages 1–3. IEEE, 2011.

[8] Sanghyeon Baeg, Hyeonwoo Nam, Pierre Chia, ShiJie Wen, and Richard Wong. Ac-dc factor sensitivity for dram components lifetime under hot-carrier injection. In *Reliability Physics Symposium (IRPS), 2011 IEEE International*, pages 2D–1. IEEE, 2011.

[9] R Jacob Baker. *CMOS: circuit design, layout, and simulation*, volume 1. John Wiley & Sons, 2008.

[10] Sarvesh Bhardwaj, Wenping Wang, Rakesh Vattikonda, Yu Cao, and Sarma Vrudhula. Predictive modeling of the nbti effect for reliable design. In *Custom Integrated Circuits Conference, 2006. CICC'06. IEEE*, pages 189–192. IEEE, 2006.

[11] Song Bian, Michihiro Shintani, Shumpei Morita, Hiromitsu Awano, Masayuki Hiromoto, and Takashi Sato. Workload-aware worst path analysis of processor-scale nbti degradation. In *Proceedings of the 26th edition on Great Lakes Symposium on VLSI*, pages 203–208. ACM, 2016.

[12] Shekhar Borkar. Design challenges of technology scaling. *IEEE micro*, (4):23–29, 1999.

[13] Shekhar Borkar. Designing reliable systems from unreliable components: the challenges of transistor variability and degradation. *Ieee Micro*, 25(6):10–16, 2005.

[14] Shekhar Borkar, Tanay Karnik, Siva Narendra, Jim Tschanz, Ali Keshavarzi, and Vivek De. Parameter variations and impact on circuits and microarchitecture. In *Proceedings of the 40th annual Design Automation Conference*, pages 338–342. ACM, 2003.

[15] William Bornstein, Robert Dunn, and Tony Spielberg. Field degradation of memory components from hot carriers. In *2006 IEEE International Reliability Physics Symposium Proceedings*, 2006.

[16] Keith A Bowman, James W Tschanz, Shih-Lien L Lu, Paolo A Aseron, Muhammad M Khellah, Arijit Raychowdhury, Bibiche M Geuskens, Carlos Tokunaga, Chris B Wilkerson, Tanay Karnik, et al. A 45 nm resilient microprocessor core for dynamic variation tolerance. *IEEE Journal of Solid-State Circuits*, 46(1):194–208, 2011.

[17] David Bull, Shidhartha Das, Karthik Shivashankar, Ganesh S Dasika, Krisztian Flautner, and David Blaauw. A power-efficient 32 bit arm processor using timing-error detection and correction for transient-error tolerance and adaptation to pvt variation. *Solid-State Circuits, IEEE Journal of*, 46(1):18–31, 2011.

[18] Andrea Calimera, Enrico Macii, and Massimo Poncino. Nbti-aware power gating for concurrent leakage and aging optimization. In *Proceedings of the 2009 ACM/IEEE international symposium on Low power electronics and design*, pages 127–132. ACM, 2009.

[19] Andrea Calimera, Enrico Macii, and Massimo Poncino. Design techniques for nbti-tolerant power-gating architectures. *IEEE Trans. Circuits Syst. II, Exp. Briefs*, 59 (4):249–253, 2012.

[20] Alejandro Campos-Cruz, Esteban Tlelo-Cuautle, and Guillermo Espinosa-Flores-Verdad. Advances in bti modeling for the design of reliable ics. In *Electrical Engineering, Computing Science and Automatic Control (CCE), 2016 13th International Conference on*, pages 1–4. IEEE, 2016.

[21] Vikas Chandra. Quantifying workload dependent reliability in embedded processors. In *Design Automation Conference (ASP-DAC), 2014 19th Asia and South Pacific*, pages 474–477. IEEE, 2014.

[22] Ankush Chaudhary and Souvik Mahapatra. A physical and spice mobility degradation analysis for nbti. *IEEE Transactions on Electron Devices*, 60(7):2096–2103, 2013.

[23] Cheng-I Chen and Gary W Chang. An efficient prony-based solution procedure for tracking of power system voltage variations. *IEEE Transactions on Industrial Electronics*, 60(7):2681–2688, 2013.

[24] Xiaoming Chen, Yu Wang, Huazhong Yang, Yuan Xie, and Yu Cao. Assessment of circuit optimization techniques under nbti. *IEEE Design & Test*, 30(6):40–49, 2013.

[25] Binjie Cheng, Andrew R Brown, and Asen Asenov. Impact of nbti/pbti on sram stability degradation. *IEEE Electron Device Letters*, 32(6):740–742, 2011.

[26] Lih-Yih Chiou, Chi-Ray Huang, and Ming-Hung Wu. A power-efficient pulse-based in-situ timing error predictor for pvt-variation sensitive circuits. In *Circuits and Systems (ISCAS), 2014 IEEE International Symposium on*, pages 1215–1218. IEEE, 2014.

[27] Kun-Wei Chiu, Yu-Guang Chen, and Chao Lin. An efficient nbti-aware wake-up strategy for power-gated designs. In *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2018*, pages 901–904. IEEE, 2018.

[28] H-L Chou, Chih-Fang Huang, and Jianya Gong. Dimension dependence of unusual hci-induced degradation on n-channel high-voltage demosfet. *Electron Devices, IEEE Transactions on*, 60(5):1723–1729, 2013.

[29] Shidhartha Das, David Roberts, Seokwoo Lee, Sanjay Pant, David Blaauw, Todd Austin, Krisztián Flautner, and Trevor Mudge. A self-tuning dvs processor using delay-error detection and correction. *IEEE Journal of Solid-State Circuits*, 41(4): 792–804, 2006.

[30] Shidhartha Das, Carlos Tokunaga, Sanjay Pant, Wei-Hsiang Ma, Sudherssen Kalaiselvan, Kevin Lai, David M Bull, and David T Blaauw. Razorii: In situ error detection and correction for pvt and ser tolerance. *IEEE Journal of Solid-State Circuits*, 44(1):32–48, 2009.

[31] Jie Ding, Dave Reid, Campbell Millar, and Asen Asenov. Investigation of sram using bti-aware statistical compact models. In *Solid-State Device Research Conference (ESSDERC), 2013 Proceedings of the European*, pages 186–189. IEEE, 2013.

[32] Shengyu Duan, Basel Halak, Rick Wong, and Mark Zwolinski. Nbti lifetime evaluation and extension in instruction caches. In *ERMAVSS@ DATE*, pages 9–12, 2016.

[33] Shengyu Duan, Basel Halak, and Mark Zwolinski. An ageing-aware digital synthesis approach. In *Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD), 2017 14th International Conference on*, pages 1–4. IEEE, 2017.

[34] Shengyu Duan, Mark Zwolinski, and Basel Halak. Lifetime reliability-aware digital synthesis. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, (99):1–12, 2018.

[35] Dan Ernst, Nam Sung Kim, Shidhartha Das, Sanjay Pant, Rajeev Rao, Toan Pham, Conrad Ziesler, David Blaauw, Todd Austin, Krisztian Flautner, et al. Razor: A low-power pipeline based on circuit-level timing speculation. In *Microarchitecture, 2003. MICRO-36. Proceedings. 36th Annual IEEE/ACM International Symposium on*, pages 7–18. IEEE, 2003.

[36] Bahar Farahani and Saeed Safari. Cross-layer custom instruction selection to address pvta variations and soft error. *Microelectronics Reliability*, 55(11):2423–2438, 2015.

[37] Farshad Firouzi, Saman Kiamehr, and Mehdi B Tahoori. Power-aware minimum nbti vector selection using a linear programming approach. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 32(1):100–110, 2013.

[38] Matthew Fojtik, David Fick, Yejoong Kim, Nathaniel Pinckney, David Harris, David Blaauw, and Dennis Sylvester. Bubble razor: An architecture-independent approach to timing-error detection and correction. In *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*, pages 488–490. IEEE, 2012.

[39] Michael Frank et al. Modifying operating parameters of a device based on aging information, March 11 2014. US Patent 8,671,170.

[40] Hiroshi Fuketa, Masanori Hashimoto, Yukio Mitsuyama, and Takao Onoye. Adaptive performance compensation with in-situ timing error predictive sensors for subthreshold circuits. *IEEE Transactions on very large scale integration (VLSI) systems*, 20(2):333–343, 2012.

[41] Andres F Gomez and Victor Champac. Early selection of critical paths for reliable nbti aging-delay monitoring. *IEEE Transactions on very large scale integration (VLSI) systems*, 24(7), 2016.

[42] Paul Heremans, Rudi Bellens, Guido Groeseneken, and Herman E Maes. Consistent model for the hot-carrier degradation in n-channel and p-channel mosfets. *Electron Devices, IEEE Transactions on*, 35(12):2194–2209, 1988.

[43] Akio Hirata, Hidetoshi Onodera, and Keikichi Tamaru. Estimation of short-circuit power dissipation for static cmos gates. *IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences*, 79(3):304–311, 1996.

[44] Sebastian Höppner, Yexin Yan, Bernhard Vogginger, Andreas Dixius, Johannes Partzsch, Felix Neumärker, Stephan Hartmann, Stefan Schiefer, Stefan Scholze, Georg Ellguth, et al. Dynamic voltage and frequency scaling for neuromorphic many-core systems. In *Circuits and Systems (ISCAS), 2017 IEEE International Symposium on*, pages 1–4. IEEE, 2017.

[45] Brett Howse and Ryan Smith. Tick tock on the rocks: Intel delays 10nm, adds 3rd gen 14nm core product kaby lake. *Google Scholar*, 2015.

[46] R HSPICE. Reference manual: Commands and control options. Technical report, Version D-2010.03-SP1, June 2010. http://www. synopsys. com/Tools/Verification/AMS Verification/CircuitSimulation/HSPICE/Pages/default. aspx (accessed on October 14, 2011), 2012.

[47] Yanxiang Huang, Meng Li, Chunshu Li, Peter Debacker, and Liesbet Van der Perre. Computation-skip error mitigation scheme for power supply voltage scaling in recursive applications. *Journal of Signal Processing Systems*, pages 1–12, 2016.

[48] Inductiveload. *The stage-by-stage architecture of a MIPS*, volume 1. Inductiveload, 2009.

[49] Hai Jiang, SangHoon Shin, Xiaoyan Liu, Xing Zhang, and Muhammad Ashraful Alam. The impact of self-heating on hci reliability in high-performance digital circuits. *IEEE Electron Device Letters*, 38(4):430–433, 2017.

[50] Ushio Jimbo, Junji Yamada, Ryota Shioya, and Masahiro Goshima. Applying razor flip-flops to sram read circuits. *IEICE Transactions on Electronics*, 100(3):245–258, 2017.

[51] H-F Jyu, Sharad Malik, Srinivas Devadas, and Kurt W Keutzer. Statistical timing analysis of combinational logic circuits. *IEEE Transactions on Very large Scale integration (VLSI) systems*, 1(2):126–137, 1993.

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

[53] Kelin J Kuhn, Martin D Giles, David Becher, Pramod Kolar, Avner Kornfeld, Roza Kotlyar, Sean T Ma, Atul Maheshwari, and Sivakumar Mudanai. Process technology variation. *IEEE Transactions on Electron Devices*, 58(8):2197–2208, 2011.

[54] Yang Lin and Mark Zwolinski. A cost-efficient self-checking register architecture for radiation hardened designs. In *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 149–152. IEEE, 2014.

[55] Jani Mäkipää, Matthew J Turnquist, Erkka Laulainen, and Lauri Koskinen. Timing-error detection design considerations in subthreshold: An 8-bit microprocessor in 65 nm cmos. *Journal of Low Power Electronics and Applications*, 2(2): 180–196, 2012.

[56] C Martins, J Pachito, J Semião, IC Teixeira, and JP Teixeira. Adaptive error-prediction aging sensor for on-line monitoring of performance errors. In *Proceedings of the 26th Conference on Design of Circuits and Integrated Systems-DCIS*, 2011.

[57] S Mhira, V Huard, A Benhassain, F Cacho, S Naudet, A Jain, C Parthasarathy, and A Bravaix. Dynamic adaptive voltage scaling in automotive environment. In *Reliability Physics Symposium (IRPS), 2017 IEEE International*, pages 3A–4. IEEE, 2017.

[58] Evelyn Mintarno, Joëlle Skaf, Rui Zheng, Jyothi Bhaskar Velamala, Yu Cao, Stephen Boyd, Robert W Dutton, and Subhasish Mitra. Self-tuning for maximized lifetime energy-efficiency in the presence of circuit aging. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 30(5):760–773, 2011.

[59] Mohd Syafiq Mispan, Basel Halak, Zufu Chen, and Mark Zwolinski. Tco-puf: A subthreshold physical unclonable function. In *Ph. D. Research in Microelectronics and Electronics (PRIME), 2015 11th Conference on*, pages 105–108. IEEE, 2015.

[60] Mohd Syafiq Mispan, Basel Halak, and Mark Zwolinski. Nbti aging evaluation of puf-based differential architectures. In *On-Line Testing and Robust System Design (IOLTS), 2016 IEEE 22nd International Symposium on*, pages 103–108. IEEE, 2016.

[61] Jan M Rabaey, Anantha P Chandrakasan, and Borivoje Nikolic. *Digital integrated circuits*, volume 2. Prentice hall Englewood Cliffs, 2002.

[62] Hans Reisinger, O Blank, Wolfgang Heinrigs, A Muhlhoff, Wolfgang Gustin, and C Schlunder. Analysis of nbti degradation-and recovery-behavior based on ultra fast vt-measurements. In *Reliability Physics Symposium Proceedings, 2006. 44th Annual., IEEE International*, pages 448–453. IEEE, 2006.

[63] Gaole Sai, Basel Halak, and Mark Zwolinski. Multi-path aging sensor for cost-efficient delay fault prediction. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 65(4):491–495, 2018.

[64] Christian Schlünder, Jörg Berthold, Fabian Proebster, Andreas Martin, Wolfgang Gustin, and Hans Reisinger. On the influence of bti and hci on parameter variability. In *Reliability Physics Symposium (IRPS), 2017 IEEE International*, pages 2E–4. IEEE, 2017.

[65] Dieter K Schroder. Negative bias temperature instability: What do we understand? *Microelectronics Reliability*, 47(6):841–852, 2007.

[66] J Semiao, D Saraiva, C Leong, A Romao, MB Santos, IC Teixeira, and JP Teixeira. Performance sensor for tolerance and predictive detection of delay-faults. In *Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT), 2014 IEEE International Symposium on*, pages 110–115. IEEE, 2014.

[67] Mark Zwolinski Shengyu Duan and Basel Halak. Lifetime reliability-aware digital synthesis. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, To be published.

[68] Toshinori Takeshita, Tohru Ishihara, and Hidetoshi Onodera. Guidelines for effective and simplified dynamic supply and threshold voltage scaling. In *VLSI Design, Automation and Test (VLSI-DAT), 2016 International Symposium on*, pages 1–4. IEEE, 2016.

[69] Bernard M Tenbroek, Michael SL Lee, William Redman-White, R John T Bunyan, and Michael J Uren. Self-heating effects in soi mosfets and their measurement by small signal conductance techniques. *Electron Devices, IEEE Transactions on*, 43 (12):2240–2248, 1996.

[70] Martine Turgeon, Alan M Wing, and Lawrence W Taylor. Timing and aging: Slowing of fastest regular tapping rate with preserved timing error detection and correction. *Psychology and Aging*, 26(1):150, 2011.

[71] Stefanos Valadimas, Andreas Floros, Yiorgos Tsiatouhas, Angela Arapoyanni, and Xrysovalantis Kavousianos. The time dilation technique for timing error tolerance. *Computers, IEEE Transactions on*, 63(5):1277–1286, 2014.

[72] Stefanos Valadimas, Yiorgos Tsiatouhas, and Angela Arapoyanni. Timing error tolerance in nanometer ics. In *On-Line Testing Symposium (IOLTS), 2010 IEEE 16th International*, pages 283–288. IEEE, 2010.

[73] Stefanos Valadimas, Yiorgos Tsiatouhas, and Angela Arapoyanni. Cost and power efficient timing error tolerance in flip-flop based microprocessor cores. In *Test Symposium (ETS), 2012 17th IEEE European*, pages 1–6. IEEE, 2012.

[74] Luca Vandelli, Luca Larcher, Dekel Veksler, Andrea Padovani, Gennadi Bersuker, and Kenneth Matthews. A charge-trapping model for the fast component of positive bias temperature instability (pbti) in high-gate-stacks. *Electron Devices, IEEE Transactions on*, 61(7):2287–2293, 2014.

[75] Harry JM Veendrick. Short-circuit dissipation of static cmos circuitry and its impact on the design of buffer circuits. *Solid-State Circuits, IEEE Journal of*, 19(4):468–473, 1984.

[76] Jinn-Shyan Wang and Shih-Nung Wei. Process/voltage/temperature-variation-aware design and comparative study of transition-detector-based error-detecting latches for timing-error-resilient pipelined systems. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 2017.

[77] Wenping Wang, Vijay Reddy, Anand T Krishnan, Rakesh Vattikonda, Srikanth Krishnan, and Yu Cao. Compact modeling and simulation of circuit reliability for 65-nm cmos technology. *Device and Materials Reliability, IEEE Transactions on*, 7(4):509–517, 2007.

[78] Wenping Wang, Zile Wei, Shengqi Yang, and Yu Cao. An efficient method to identify critical gates under circuit aging. In *Computer-Aided Design, 2007. ICCAD 2007. IEEE/ACM International Conference on*, pages 735–740. IEEE, 2007.

[79] Wenping Wang, Shengqi Yang, Sarvesh Bhardwaj, Rakesh Vattikonda, Sarma Vrudhula, Frank Liu, and Yu Cao. The impact of nbti on the performance of combinational and sequential circuits. In *Proceedings of the 44th annual Design Automation Conference*, pages 364–369. ACM, 2007.

[80] Yangang Wang and M Zwolinski. Impact of nbti on the performance of 35nm cmos digital circuits. In *Solid-State and Integrated-Circuit Technology, 2008. ICSICT 2008. 9th International Conference on*, pages 440–443. IEEE, 2008.

[81] Yao Wang, Sorin Cotofana, and Liang Fang. A unified aging model of nbti and hci degradation towards lifetime reliability management for nanoscale mosfet circuits. In *Proceedings of the 2011 IEEE/ACM International Symposium on Nanoscale Architectures*, pages 175–180. IEEE Computer Society, 2011.

[82] Yu Wang, Xiaoming Chen, Wenping Wang, Varsha Balakrishnan, Yu Cao, Yuan Xie, and Huazhong Yang. On the efficacy of input vector control to mitigate nbti effects and leakage power. In *Quality of Electronic Design, 2009. ISQED 2009. Quality Electronic Design*, pages 19–26. IEEE, 2009.

[83] Yu Wang, Hong Luo, Ku He, Rong Luo, Huazhong Yang, and Yuan Xie. Temperature-aware nbti modeling and the impact of standby leakage reduction techniques on circuit performance degradation. *IEEE Transactions on Dependable and Secure Computing*, 8(5):756–769, 2011.

[84] Shi-Jie Wen et al. New dram hci qualification method emphasizing on repeated memory access. In *2010 IEEE International Integrated Reliability Workshop Final Report*, pages 142–144, 2010.

[85] Martin Wirnshofe. *Variation-aware adaptive voltage scaling for digital CMOS circuits.* Springer, 2013.

[86] Martin Wirnshofer, Nasim Pour Aryan, Leonhard Heiss, Doris Schmitt-Landsiedel, and Georg Georgakos. On-line supply voltage scaling based on in situ delay monitoring to adapt for pvta variations. *Journal of Circuits, Systems and Computers*, 21(08):1240027, 2012.

[87] Zhang XiaoWen and En YunFei. The hci effect reliability evaluation of cmos process. In *Electron Devices and Solid-State Circuits (EDSSC), 2014 IEEE International Conference on*, pages 1–2. IEEE, 2014.

[88] Bo Yang, Emanuel Popovici, Michael Alan Quille, Andreas Amann, and Sorin Cotofana. A supply voltage-dependent variation aware reliability evaluation model. In *Nanoscale Architectures (NANOARCH), 2016 IEEE/ACM International Symposium on*, pages 79–84. IEEE, 2016.

[89] Yiqun Zhang, Mahmood Khayatzadeh, Kaiyuan Yang, Mehdi Saligane, Nathaniel Pinckney, Massimo Alioto, David Blaauw, and Dennis Sylvester. 8.8 irazor: 3-transistor current-based error detection and correction in an arm cortex-r4 processor. In *2016 IEEE International Solid-State Circuits Conference (ISSCC)*, pages 160–162. IEEE, 2016.

[90] John Zimmerman, Jodi Forlizzi, and Shelley Evenson. Research through design as a method for interaction design research in hci. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 493–502. ACM, 2007.