SOFT-ERROR RESILIENCE FRAMEWORK FOR RELIABLE AND ENERGY-EFFICIENT COMPUTING ARCHITECTURES

by

FARIS SALIH ALGHAREB
M.Sc. University of Mosul, Mosul, 2009
B.Sc. University of Mosul, Mosul, 2007

A dissertation submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy
in the Department of Electrical and Computer Engineering
in the College of Engineering and Computer Science
at the University of Central Florida

Spring Term
2017

Major Professor: Ronald F. DeMara
ABSTRACT

Technology size scaling has induced emerging issues that impact reliability of nanoscale logic circuits. The first issue is there exists an increased impact of Process Variation (PV) on device performance, while the second reliability challenge that has been pronounced to impact robustness of logic circuits is an increased susceptibility of nanoscale device technology to soft error effects.

In particular, radiation induced-soft errors have become a key issue as they cause malfunctions in large-scale circuits and systems even for terrestrial applications. Similarly, operation at higher clock rate has also increased the probability of Single-Event Transient (SET) to be captured by a latch/flip-flop. In addition, the continuous shrinking of supply voltage reduces the required energy to induce transient and upset glitches at the susceptible nodes of a logic circuit. These reliability impacting factors have become major concern of recent device technology. Consequently, soft error mitigation techniques have become essential to improve systems’ reliability. In this dissertation, optimized soft error resilience solutions are proposed to improve robustness of sub-micron computing systems. The proposed approaches were developed to deliver energy-efficiency while incurring an acceptable speed performance degradation compared to the prior work. Moreover, our proposed approach for double error correction can tolerate double Single-Event Upset (SEU), simultaneously.

The necessity for energy-efficiency has given rise to run High Performance Computing (HPC) systems at Near-Threshold Voltage (NTV) region as opposed to the conventional approach of operating devices at nominal supply voltage. However, this can be achieved at the expense of increased reliability challenges since an increased Soft Error Rate (SER) has been identified as a significant concern for NTV operation of deeply-scaled CMOS logic paths. In this work, the impact of PV at the NTV on redundancy-based soft-error mitigation approaches was investigated to highlight the approach that can realize an increased performance in terms of energy-efficiency, critical datapath...
delay variation, speed degradation, and superior soft-error resilience.

Emerging spin-based devices are introduced as potential alternative and an intriguing candidate to alleviate scalability challenges of CMOS technology. These emerging spin-based devices offer many promising features, such as high speed for read/write access operations, near-zero static power, intrinsic immunity to radiation, non-volatility, and excellent scalability and great compatibility of MTJs with CMOS technology. However, they are still prone to the particle strikes due to the CMOS logic portion for read/write access operations, making the entire design susceptible to radiation-induced transient effects. Thus, there is a need to develop cost-effective and performance-efficient soft error handling techniques so that making this emerging nanoscale device technology is feasible. Herein, we design and evaluate for soft-error resilience Non-Volatile latching circuits that can achieve intriguing features, such as low energy consumption, high computing performance, superior soft errors tolerance, i.e., concurrently able to tolerate Multiple Node Upset (MNU), to potentially become as a mainstream solution for the aerospace and avionic nanoelectronics.
Dedicated to my family
ACKNOWLEDGMENTS

The acknowledgments page is optional. If you choose to use it, it should appear after the Abstract, but before the Table of Contents.
# TABLE OF CONTENTS

LIST OF FIGURES ........................................................................................................ xii

LIST OF TABLES ............................................................................................................ xv

CHAPTER 1: Introduction ............................................................................................... 1

1.1 Categories of Single-Event .................................................................................... 2

1.2 Soft Error Rate in Nanometer-scaled VLSI Devices ............................................. 3

1.3 Why Are Soft Errors Important? Needs for Soft-Error Resilience Logic Circuits . . . 3

1.4 FinFET Structure Devices

4

1.5 Contributions of the Dissertation .......................................................................... 5

CHAPTER 2: LITERATURE REVIEW ............................................................................ 8

2.1 SER-Induced System Dependability Issues ............................................................ 8

2.1.1 Reliability at Multiple Abstraction Levels ......................................................... 9

2.1.2 Radiation-Induced Soft-Errors ........................................................................... 11

2.1.3 Sources of Variations ....................................................................................... 12

2.2 Soft Errors in Logic Paths ..................................................................................... 14
2.2.1 Logic Path Masking Mechanisms ............................................................. 16
2.2.2 SET Characterizing and Modeling .......................................................... 19

2.3 SER Mitigation Techniques ........................................................................ 25
2.3.1 Device-Level Mitigation Techniques ...................................................... 27
2.3.2 Gate-Level Mitigation Techniques .......................................................... 28
2.3.3 Circuit-Level Mitigation Techniques ....................................................... 29
   2.3.3.1 State-of-the-Art Redundancy-based Techniques ............................. 30
   2.3.3.2 The Conventional Redundancy-based Techniques ......................... 33
2.3.4 System-Level Mitigation Techniques ..................................................... 44
   2.3.4.1 Soft Error Masking For Other Components ................................ 45
   2.3.4.2 Multi-Bit Upsets ............................................................................. 46

2.4 Soft Error Rate Trends and Technology Challenges ................................. 48
   2.4.1 Trends of Technology Scaling on SER .............................................. 50
   2.4.2 Trends of Voltage Scaling on SER ..................................................... 51
   2.4.3 Trends of Multi-Gate FinFET Devices on SER .................................. 52

2.5 Summary ..................................................................................................... 53

CHAPTER 3: TEMPORAL SELF-VOTING CHECKERS: ENERGY VS. RESILIENCE
TRADEOFFS .................................................................................................................. 55

3.1 Energy-Efficiency vs Fault Resilience ...................................................................... 55

3.2 The Proposed Approaches ...................................................................................... 56

   3.2.1 Temporal Self-Voting Logic (TSVL) Approach ...................................................... 56

   3.2.2 Hybrid Spatial and Temporal Redundancy Double-Error Correction Approach 60

3.3 Experimental Results and Analysis ........................................................................ 62

   3.3.1 Quantifying Fault Resilience for Proposed Techniques ........................................ 63

   3.3.2 Performance Evaluation for Proposed Techniques ................................................ 65

3.4 Summary .................................................................................................................. 69

CHAPTER 4: IMPACT OF PROCESS VARIATION IN THE NVT REGION ON SOFT
ERROR RATE .................................................................................................................. 71

4.1 Energy-Efficiency Through Near-Threshold Computing ............................................. 71

4.2 Trends of SER in NTV Region .................................................................................. 72

   4.3 Technology Scaling and PV Trends For SET ............................................................. 73

4.4 Experimental Setup .................................................................................................. 75

   4.4.1 Simulation Objectives, Tools and Workflow ........................................................ 75

4.5 Experimental Results and Analysis .......................................................................... 78
4.5.1 Quantifying Area Overhead for SE-Mitigation Techniques..........................78
4.5.2 Comparison of Delay and Energy Consumption ........................................79
4.5.3 Impact of Process Variation in NTV Region .............................................81
4.5.4 Impact of Supply Voltage in NTV Region ................................................84
4.5.5 Relationship of Area and FCER ..............................................................88
4.5.6 Benefit of Spatial Redundancy with Design Diversity .................................89
4.5.7 Constraints of Temporal Redundancy Approach ......................................90
4.8 Summary ......................................................................................................92

CHAPTER 5: Energy-Efficient and Soft-Error Resilient Non-Volatile Spintronic Flip-Flop Designs..................................................................................................................93

CHAPTER 6: High-Performance Double Node Upset-Tolerant Non-Volatile Flip-Flop Design.....................................................................................................................94

5.1 Hybrid CMOS/Spintronics: Alleviating CMOS Challenges ............................94
5.2 Fundamentals of Magnetic Tunnel Junctions ..................................................95
5.3 Technology Scaling Trends on Data Bit Upset ..............................................98
5.4 Radiation-Induced Transient Faults Emulation .............................................99
5.5 Proposed DNU-Tolerant NV Flip-Flop Circuit .............................................100
5.5.1 Functionality Analysis of the Proposed NV Flip-Flop ........................................101

5.5.2 Soft Error Rate Analysis of the Proposed NV Flip-Flop ........................................101

5.6 EXPERIMENTS AND RESULTS....................................................................................103

5.6.1 Evaluation of Area, Power, and Delay Overheads ..................................................104

5.6.2 Evaluation of SEU and DNU Immunity ....................................................................106

5.7 Summary ......................................................................................................................106

CHAPTER 7: CONCLUSIONS AND FUTUREWORK ..........................................................108

LIST OF REFERENCES .......................................................................................................109
# LIST OF FIGURES

2.1 Layered model of permanent and soft error sources and impacts..........................10

2.2 Random dopant fluctuation as a function of technology generations [1].............15

2.3 An example of logic datapath masking mechanisms............................................17

2.4 SET latching window of vulnerability [2]............................................................19

2.5 Estimated maximum SET widths based on modifying the n-well contact size of the pMOS device for 90nm technology process [3].....................................................22

2.6 Soft error rate of a logic and memory [1]............................................................24

2.7 Taxonomy of design levels for soft error mitigation schemes...............................26

2.8 Voting Circuits; (a) Majority voter circuit, (b) Self-voter circuit [4], (c) C-element in a temporal filter structure [5]. .................................................................34

2.9 Triple Module Redundancy (TMR) approach.......................................................35

2.10 Temporal redundancy approach [2]. ...................................................................38

2.11 Self-voting DMR logic circuit [4]. ........................................................................40

2.12 Structural DMR design for FIR filter [6]..............................................................44

3.1 Temporal Self-Voting Logic (TSVL) approach.....................................................58
Timing diagram of TSVL approach with SEU recovery occurred in the original, $Q_1$, and redundant, $Q_2$, flip-flop consecutively, $\text{masked} \_ \text{out}$ is an inverse of $\text{data} \_ \text{in}$. 

Hybrid Spatial and Temporal Redundancy Double-Error Correction approach.  

Area overhead evaluation of redundancy-based soft error masking techniques.  

Power consumption analysis of redundancy-based soft error masking techniques.  

Performance evaluation of redundancy-based soft error masking techniques.  

Local Clock Manager (LCM) to generate CLK2 and CLK3.  

Simulation framework developed to estimate the delay and energy for TMR. At least 1,000 samples are synthesized for each redundant system to conduct the statistical analysis.  

Speed degradation analysis at NTV for the 45-nm technology.  

Speed degradation analysis at NTV for the 16-nm technology.  

Mean energy consumption for redundant systems using 45-nm and 16-nm technologies.  

Delay variation for redundant systems using 45-nm and 16-nm technologies.  

Maintaining energy consumption at NTV using 45-nm technology process.  

Maintaining energy consumption at NTV using 16-nm technology process.
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.9</td>
<td>SETs pulse width rejection at NTV for 45-nm planar MOSFET and tri-gate 16-nm bulk FinFET</td>
</tr>
<tr>
<td>4.10</td>
<td>Mean fault coverage energy ratio analysis of redundancy-based approaches, $V_{DD}$ is scaled from nominal level (1.1V) to the NTV region (0.5V) with a decrement of 0.5V.</td>
</tr>
<tr>
<td>4.11</td>
<td>Relationship between area (number of cells in each benchmark circuit) and FCER.</td>
</tr>
<tr>
<td>6.1</td>
<td>Spin transfer torque switching mechanism.</td>
</tr>
<tr>
<td>6.2</td>
<td>The proposed DNU-tolerant soft error resilient nonvolatile flip-flop circuit.</td>
</tr>
<tr>
<td>6.3</td>
<td>Timing waveforms of the proposed DNU-tolerant NVFF.</td>
</tr>
</tbody>
</table>
## LIST OF TABLES

1.1 Trends of SER as impacted by technology scaling [1, 7, 8]. However, vertical channel devices, i.e., FinFETs, are introduced to reduce SER for technologies beyond 22nm. For instance, Intel recently released its processors within 14nm technology node. .................................................................4

2.1 Categories of exposures facing highly-scaled CMOS logic circuit design............13

2.2 Estimated maximum SET widths based on linear energy transfer (LET) [9]; taking into concentration the effect of Propagation-Induced Pulse Broadening (PIPB) and without the effect of PIPB phenomena........................................20

2.3 SBU and MCU ratio trends under technology scaling [10]. .........................47

2.4 Summary of SER mitigation techniques proposed at different abstraction design-level. .................................................................................................................49

2.5 Comparison between design-level abstraction of SE-mitigation approaches, where by each ($v'$, $v$) indicates relative (strength, limitation/weakness)........50

3.1 Possible error scenarios of HSTR-DEC approach; (NE = No Error; UE = Upset Error)...........................................................................................................61

3.2 Fault coverage analysis of TSVL approach. ................................................64

3.3 Mean area-power consumption and delay of redundancy-based approaches w.r.t. simplex design.............................................................................69
4.1 Mean energy reduction and speed degradation of temporal redundancy and SV-DMR approach versus TMR.................................................................81

4.2 Mean energy consumption and speed degradation of spatial, temporal, and hybrid redundancy approaches. .................................................................82

4.3 Normalized speed degradation of spatial, temporal, and hybrid redundancy (SV-DMR) w.r.t. simplex design. .................................................................84

4.4 Effect of reducing supply voltage under iso-energy constraints. .......................85

4.5 Fault coverage energy ratio analysis of redundancy-based approaches. ............90

4.6 Fault coverage energy ratio analysis of diversity-TMR arrangements............91

6.1 Parameters of STT-MTJ devices....................................................................97

6.2 Vulnerability and upset tolerance analysis for the proposed NVFF..................102

6.3 Performance penalties of the proposed NVFF circuit w.r.t. conventional CMOS-based flip-flop. ........................................................................105
CHAPTER 1: Introduction

This dissertation evaluates and presents the necessity for radiation-induced soft error mitigation techniques of nano-electronic computing system architectures. As technology trends of CMOS devices have improved the capability of contemporary processors, however, it has increased the susceptibility of VLSI circuits to transient faults. Thus, Soft Error Rate (SER) masking approaches are sought to increase reliability within area, speed, energy consumption, and fault masking resilience constraints. In this dissertation, energy-efficient and cost-effective redundancy-based soft error mitigation techniques to achieve resilience in VLSI circuits are proposed. In particular, resilient approaches are introduced to deal with radiation-induced transient effects, i.e., Single-Event Transient (SET) and Single-Event Upset (SEU), by carrying out tradeoffs between energy consumption and fault resilience. In addition, the effects of process variation at Near-Threshold Voltage (NTV) region on the redundancy-based soft error mitigation approaches were investigated to leverage the technique that achieves an increased performance, satisfies the user specified objective metrics, within an ISO-energy constraints. Whereas, the later part of this dissertation seeks to cover reliability of deeply-scaled hybrid CMOS/Spintronics (STT-MTJs) nonvolatile latching circuits, due to the radiation-induced soft errors in the CMOS-based logic portion that is utilized for read/write access operations. A study to completely and/or more efficiently address these transient effects to improve reliability will be conducted, so that making this emerging technology is feasible. This chapter discusses the necessity and motivation for techniques presented in this dissertation, and concludes with a list of contributions to the state-of-the-art as a result of this work.
1.1 Categories of Single-Event

The continued scaling of Complementary Metal Oxide Semiconductor (CMOS) technology process reduces the quantity of critical charge, $Q_{\text{crit}}$, transistor’s threshold voltage, a node capacitance, and the supply voltage ($V_{DD}$). This gives rise to increased susceptibility to radiation-induced soft errors caused by numerous energetic particles such as protons and heavy ions in space and alpha particles from packaging materials and neutrons from cosmic rays at sea level [11]. In particular, at the chip-level, various kinds of soft error effects can occur in CMOS logic circuits due to energetic radiation particles, including 1) particles that hit inside combinational logic and generate a transient pulse, the so-called single-event transient (SET) [12], which propagates through downstream logic, and eventually it might be captured by a storage element, i.e., latch/flip-flop, thereby causing an upset, 2) energetic particles that strike node(s) inside a storage element to cause a so-called single-event upset (SEU) that flips the bit state in a DRAM memory cell or a SRAM-based register, resulting in a soft error, and 3) energetic particle strikes on global signal lines, such as control signals or instruction lines, can result in a so-called single-event functional interrupt (SEFI) that causes a temporary malfunction, i.e., interruption of normal operation [13], as the wrong instruction might be executed [14]. The effect of a transient pulse depends on its duration, the state of the struck device (ON/OFF), and also on the type of struck node, i.e., drain side [15]. Since the corrupted data can be rewritten by new legitimate data, these temporary glitches are considered as soft errors. However, these can cause catastrophic failures in mission-critical applications. Thus, efficient protection schemes need to be evaluated with higher priority as we move towards NTC for energy-efficient operation.
1.2 Soft Error Rate in Nanometer-scaled VLSI Devices

Practically, soft errors occur when an energetic charged particle strikes a sensitive node of a circuit. This induces electron-hole pairs that cause charge/discharge at the struck transistor junction [16]. This depends on the amount of the collected charge, if the collected charge exceeds the critical amount of charge, $Q_{crit}$, it produces some momentary charge and flips the voltage state of some circuit node(s) thereby causing a temporary glitch. The effect of a transient pulse depends on its duration, the state of the struck device (ON/OFF), and also on the type of the struck node, i.e., drain side [15]. The rate of these errors defined as Soft Error Rate (SER), which is a measurement metric that is used to evaluate the sensitivity of a digital circuit to transient and upset faults [12].

1.3 Why Are Soft Errors Important? Needs for Soft-Error Resilience Logic Circuits

Earlier, soft errors were a major concern to the reliability of Integrated Circuits (ICs) for space applications. However, as technology has advanced and the transistor has become smaller and denser, soft errors have become a major contributor to reliability issues at ground-level [15]. The susceptibility of sub-micron electronics to radiation-induced transient effects has been of increased interest as this mechanism dominates the overall SER of large-scale logic circuits [1, 17]. In addition, the trend toward using smaller, denser, and faster CMOS devices has influenced logic circuits to experience larger leakage currents, thereby negatively impacting the reliability of deeply-scaled circuits by degrading their performance and robustness [18]. This is because the heat that results from large leakage currents can induce thermal noise, which can lead to bit upset in a memory element [19, 20]. Moreover, due to the increasing integration density and growing complexity of ICs, the number of susceptible nodes to transients faults have increased.

Taking all of these reliability impacting factors into consideration, Failure In Time (FIT)-based soft
Table 1.1: Trends of SER as impacted by technology scaling [1, 7, 8]. However, vertical channel devices, i.e., FinFETs, are introduced to reduce SER for technologies beyond 22nm. For instance, Intel recently released its processors within 14nm technology node.

<table>
<thead>
<tr>
<th>Design Rule (nm)</th>
<th>SER (Logic+Memory)</th>
<th>Total SER (Data Center)</th>
<th>Chip Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>180</td>
<td>1</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>130</td>
<td>15</td>
<td>10^5</td>
<td>10^4.4</td>
</tr>
<tr>
<td>90</td>
<td>25</td>
<td>10^5</td>
<td>10^4.8</td>
</tr>
<tr>
<td>65</td>
<td>50</td>
<td>10^5.5</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>60</td>
<td>10^6.25</td>
<td>10^5.3</td>
</tr>
<tr>
<td>32</td>
<td>75</td>
<td>10^7</td>
<td>10^5.7</td>
</tr>
<tr>
<td>22</td>
<td>90</td>
<td>10^7.5</td>
<td>10^6.15</td>
</tr>
<tr>
<td>16</td>
<td>10^2</td>
<td>10^8.8</td>
<td>10^6.75</td>
</tr>
</tbody>
</table>

errors are expected to dominate all failure mechanisms, and the SER might exceed the FIT [21–23]. As listed in Table 1.1, the SER of logic and memory per chip has increased 100-fold with technology scaling, especially for recent generations. This leads to a higher SER that recently has been considered a serious concern for reliable memory and logic circuits [24, 25]. Therefore, soft error suppression becomes essential to improve the reliability of resilient systems operating in harsh environments like space, marine, military, nuclear weapon, and even in commercial applications [15, 21, 22, 26].

1.4 FinFET Structure Devices

The High-κ/Metal Gate (HK-MG) devices with vertical channels, i.e., the fin-typed Field-Effect-Transistors (FinFETs) have been introduced to continue scaling and also show improved low-leakage and high performance over planar MOSFET devices [27, 28]. The tri-gate FinFET devices have been adopted by Intel at the 22nm technology node and beyond since they handle the problems of Short-Channel Effect (SCE) and provide more robust gate controllability than planar MOSFET
devices, while offering improved switching characteristics for potential scalability beyond 10nm half-pitch [27, 29, 30]. Additionally, multi-gate bulk FinFET devices improve variations and thus the system’s overall performance has significantly improved. Moreover, FinFET devices have exhibited higher tolerance to soft error effects, while being less prone to particle strikes compared to planar structure CMOS devices, thereby they have assisted in increasing the robustness of large-scale logic circuits [21,31]. In contrast, even though scaling the transistor volume helps in reducing the susceptible area, there exists an increased SER. This is due to two major factors including: the critical quantity of charge, $Q_{\text{crit}}$, of a node capacitance has aggressively diminished, while the other factor is that the integration capacity is doubling the number of transistors per unit area over time, leading to an increased SER as low energy particles potentially can deposit an adequate amount of charge, i.e., sufficient Linear Energy Transfer (LET), to cause soft errors.

1.5 Contributions of the Dissertation

The dissertation’s inspiration is partly to give an insight into the soft error mitigation techniques and their challenges related to the design of reliable systems, to assist in selecting the proper technique for satisfying specific soft error rate constraints. The main contributions of this dissertation are highlighted below.

**Energy-efficiency vs fault resilience:** In this work, a reliable and cost-effective, area and energy efficient, redundancy-based approaches have been developed to mitigate soft error effects at the circuit/module-level. The proposed soft error-tolerant techniques are able to tolerate both transient and upset faults and provide complete SEU protection, detection and masking. The presented schemes impose comparable speed performance compared to the previously presented self-voting DMR hybrid redundancy approach. In addition, the proposed approach incurs reduced area overhead as compared to the conventional TMR, aside from its ability to tolerate double SEUs, simul-
taneously. The efficacy of these approaches is quantified using Fault Coverage Power Ratio metric. This work is presented in Chapter 3.

**Impact of PV at NTV on redundancy-based soft error mitigation techniques:** The energy and performance costs, which are crucial design considerations due to prevalent use of low-voltage operation in current computing systems, were deeper investigated. In particular, detailed energy and performance tradeoffs of redundancy-based approaches for soft error mitigation in 16-nm non-planar FinFET and 45-nm planar CMOS structures are provided. For process variation, a comparison between 45-nm planar CMOS and 16-nm FinFET CMOS structure was carried out regarding energy-efficiency. Furthermore, the effects of both delay variation and technology node scaling on redundancy-based soft error mitigation techniques at NTV for iso-energy constraints have been revealed to depict the quantitative results. Additional important insight of this work was to identify and quantify the increase in energy cost of spatial redundancy at contemporary and future technology nodes such as 16 nm. Monte-Carlo simulation results demonstrate that energy-efficiency benefits of scaled technology devices (16-nm) as compared to 45-nm node due to increase in performance variations.

The highlights of work presented in Chapter 4 are given below:

1. **Assessing Resilience vs Area of Soft-Error Masking schemes:** determining empirically the costs of spatial and temporal redundancy under the impact of technology scaling.

2. **Evaluating the Energy and Delay Cost of Redundancy at NTV:** determining the delay of redundant systems at NTV within a given energy budget.

3. **Soft-Error Resilience Sensitivity to Process Variation:** identifying the increased impact of $\sigma_{V_{th}}$ on alternative SER reduction strategies for 45-nm planar and tri-gate 16-nm bulk FinFET CMOS devices.
**Soft-Error Resilient Hybrid STT-MTJ/CMOS Latching Circuits:** Emerging spin-based devices are under intensive developments and investigations to improve their reliability. In this work, we design and evaluate of soft-error resilience hybrid STT-MTJ/CMOS latching circuits for non-volatile applications. The proposed approaches can achieve intriguing features, such as low energy consumption, high computing performance, superior soft error effects resilience, i.e., concurrently can tolerate MNU, to potentially become as a mainstream solution for the aerospace and avionic nanoelectronics.
CHAPTER 2: LITERATURE REVIEW

This chapter aims at providing a survey of SER mitigation techniques and their tradeoffs, which have been proposed at various levels of design abstraction to address these Soft Error Effects (SEE). To highlight the characteristics of each technique and demonstrate the accuracy of soft error mitigation schemes, we categorize them based on the design level of the protection scheme. The taxonomy of SER masking techniques relies on the Single Event Transient (SET) pulse generation at the device-level, propagation at the circuit-level, and capturing at the circuit/module-level. We also develop a concise taxonomy for the sources of induced soft-errors and variation. A comparison among the predominant mitigation schemes in terms of area, energy consumption, fault coverage, and design complexity is carried out to leverage the robustness of schemes that achieve the highest performance over the stated issues. Likewise, a review of the impact of technology and voltage scaling trends on the SER and its masking techniques is thoroughly discussed.

2.1 SER-Induced System Dependability Issues

In the nanometric technologies, integrated circuits are associated with the diminishing of device feature size, massive growth in integration density, and lower supply voltages. Thus, these desirable attributes have raised the vulnerability of the integrated circuits to soft errors as they impose risks for terrestrial and space applications. Next, we discuss system reliability at different abstraction layers and the sources of transient and permanent faults.
2.1.1 Reliability at Multiple Abstraction Levels

A system’s reliability can be impacted by different effects. Herein, we concentrate on the effects of soft errors as a major reliability issue in highly-scaled devices and systems. Suppression of soft errors can be achieved at different abstraction levels. As can be seen in Figure 2.1, a fault, error, and failure can occur at three distinguished layers. As depicted by the cumulative arc in Figure 2.1, during the mission each component may transition from viable status to faulty status for highly-scaled devices. This transition may occur due to cumulative effects of deep submicron devices such as Time Dependent Dielectric Breakdown (TDDB), which is caused by an electric field weakening of the gate oxide layer, the Total Ionizing Dose (TID) of cosmic radiation, Electromigration within interconnections, and other progressive degradations over the mission lifetime. Meanwhile, transient effects such as incident alpha particles, which ionize critical amounts of charge, ground bounce, and dynamic temperature variations may cause either long lasting or intermittent reversible transitions between viable and faulty status. A fault may affect the data that is being processed and then impact the final result and produce a failure. However, transient faults can be remedied at the Resource Layer, allowing the next layer’s input to be error-free, or faults may lie dormant whereby the physical resource is defective, yet currently unused. Later in the computations, dormant faults become active when such components are utilized, affecting the result of the resource layer.

The Behavioral Layer shown in Figure 2.1 depicts the outcome of utilizing viable and faulty physical components. Viable components result in correct behavior during the interval of observation. An error that occurs but does not incur any impact to the result of the computation is termed a silent error. Silent errors, such as a flipped bit due to a faulty memory cell at an address that is not referenced by the application or an error in a Triple Module Redundancy (TMR) system that occurs in a single instance but the final output is error-free as long as the other two inputs are correct, remain isolated at the Behavioral Layer without propagating to the application. On the other hand,
Figure 2.1: Layered model of permanent and soft error sources and impacts.

errors that are articulated propagate up to the Application Layer. The Application Layer depicts that correct behaviors contribute to sustenance of Compliant operation. Systems that are compliant throughout the mission at the application layer are deemed to be reliable. To remain completely Compliant, all articulated errors must be Concealed from the application and remain within the Behavioral Layer. Articulated errors that reach the application cause the system to have Degraded performance if the impact of the error can be tolerated. On the other hand, articulated errors that result in unacceptable conditions to the application incur a Failure condition, which might be recovered by replacing the effected cell(s) in the configurable devices [32], i.e., Field-Programmable Gate Arrays (FPGAs), or failures may be catastrophic in Application-Specific Integrated Circuits (ASICs) devices [33].
2.1.2 Radiation-Induced Soft-Errors

Sources of most temporary errors that impact system reliability are listed in Table 2.1. These sources are categorized based on radiation, variability, and charge sharing, as each one can result either in soft error or hard error/failure. Ionizing radiation can have different effects on CMOS chips, and this depends on the radiation environment in which the chips are utilized. Concerning CMOS circuits, there are two broad categories of radiation effects including Single Event Effects (SEE), which are due to a single strike of a particle with high ionizing power (thereby causing instantaneous failure); and Total Ionizing Dose (TID) effects, which are due to the progressive accumulation of defects caused by the movement of many particles (causing long-term failure), such as electrons or protons. Radiation-induced soft errors at the sea level consist of neutrons, photons, muons (generated from the interaction of secondary cosmic effects with the atmosphere), and pions. Additionally, alpha particles, that either originate from some metals used in integrated circuit fabrication or emitted from some elements that are used in the doping of semiconductors (radioactive impurities), is also an important source of error in sensitive circuits. Among them, roughly about 95% of these particles are neutrons [21, 24]. In SRAM cells, secondary particles from neutron collisions have the largest effect on $Q_{crit}$. Besides that, alpha particles from radioactive impurities also play a role in soft error sources in SRAMs [24]. Likewise, neutrons dominate SER in latches and combinational logic since they constitute a larger charge capacitance. However, as the device feature size is shrunk, $Q_{crit}$ becomes smaller, alpha particles contribute a comparable percentage to neutrons over the total SER [34].

Practically, enhancing the purity of the materials reduces the SER induced from alpha particles, whereas the effects of cosmic rays can be reduced by shielding the die area [24]. Shielding provides partial masking of electrons and low energy protons. In addition, shielding a circuit by packaging to prevent SEUs can increase the SER due to packaging as packaging materials themselves
can release alpha particles that induce soft errors. Likewise, long-term TID exposure can result in variations in threshold voltage, which might increase device leakage currents, and thereby increase power consumption. These are in addition to increasing the degradation of materials, leading to timing changes and thus decreased functionality. TID effects can be mitigated using radiation-hardened devices and shielding. However, these strategies are inadequate for robust systems with high reliability. Thus, utilizing redundancy techniques or changing the geometry of the device are still in use to improve the reliability. For example, the experimental results in [35] projected that SER for SOI is 5-fold reduced compared to substrate bulk technology. In addition, the contribution of neutron and alpha particle induced soft errors have found to be reduced roughly the same difference. Hence, the difficulty is not how to achieve a high reliability. However, achieving it with minimum area and power overheads and speed degradation is the aim. Thus, it is desired for SER mitigation schemes to allow tradeoffs between reliability and overhead of protection techniques. We identify these techniques in Section 2.3.

2.1.3 Sources of Variations

The lower portion of Table 2.1 identifies the sources of variability that impact reliability. Nanoscale devices are susceptible to process variations created by precision limitations of the manufacturing process. Sources of variability can be categorized into two groups: intrinsic (random) and extrinsic (systematic). Process variations on the gate oxide thickness ($t_{ox}$) and the gate length ($L_g$) are the most dominant extrinsic variations, whereas variation that is intrinsic in nature results from stochastic behaviors related to device fabrication including Random Dopant Fluctuation (RDF) (due to random number and position of impure dopants), Linear Edge Roughness (LER), Random Work Function fluctuation (WKF) (due to metal gate), and Interface Traps (ITs) [36, 37]. Practically, to alleviate the intrinsic-parameter fluctuations, the high-$\kappa$/metal-gate technology was introduced as an effective alternative.
Table 2.1: Categories of exposures facing highly-scaled CMOS logic circuit design.

<table>
<thead>
<tr>
<th>Category</th>
<th>Source</th>
<th>Failure Mechanism</th>
<th>Impact</th>
<th>Error Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background Radiation</td>
<td>Alpha particles(α)</td>
<td>Direct ionization</td>
<td>Single or Multi-Bit Upset (MBU)</td>
<td>Soft error</td>
</tr>
<tr>
<td></td>
<td>Proton(β)</td>
<td>Direct ionization</td>
<td>Single Event Upset (SEU)</td>
<td>Soft error</td>
</tr>
<tr>
<td></td>
<td>Beta(β) or Electron(θ)</td>
<td>Direct ionization</td>
<td>SEU</td>
<td>Soft error</td>
</tr>
<tr>
<td>Cosmic Rays</td>
<td>Photon: Gamma(γ), X-ray</td>
<td>Indirect ionization</td>
<td>SEU at sea level or Single Event latchup (SEL) in space level</td>
<td>Soft or hard error</td>
</tr>
<tr>
<td></td>
<td>Neutron(τ), heavy ions</td>
<td>Indirect ionization</td>
<td>SBU/MBUs at sea level or SEL in space level</td>
<td>Soft or hard error</td>
</tr>
<tr>
<td>Secondary Radiation</td>
<td>Sea level muon</td>
<td>Direct ionization</td>
<td>SEU</td>
<td>Soft error</td>
</tr>
<tr>
<td>Charge Sharing and SET Issues</td>
<td>Propagation-Induced</td>
<td>Increase width of</td>
<td>Increase SER</td>
<td>Soft error</td>
</tr>
<tr>
<td></td>
<td>Pulse Broadening (PIPB)</td>
<td>transient pulse</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>or Pulse Stretching</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pulse Quenching</td>
<td>Reduces width</td>
<td>Reduce SER</td>
<td>Soft error</td>
</tr>
<tr>
<td></td>
<td></td>
<td>and amplitude of</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Electromagnetic crosstalk</td>
<td></td>
<td>a SET</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supply Voltage Variation</td>
<td>Static Variation</td>
<td>Manufacturing</td>
<td>Time-dependent (increase/decrease output delay)</td>
<td>Bit error</td>
</tr>
<tr>
<td></td>
<td></td>
<td>imperfection</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dynamic Variation</td>
<td>Supply voltage variations, such as voltage drops by the effect of higher heat flux</td>
<td>Operating margin violation</td>
<td>Time-dependent</td>
<td>Bit error</td>
</tr>
<tr>
<td>Oxide Thickness Variation</td>
<td>Occurs during fabrication process</td>
<td>Manufacturing imperfection</td>
<td>Inherent unreliability effects in a logic circuit</td>
<td>Timing/Bit error</td>
</tr>
<tr>
<td>Gate Length Variation</td>
<td>Occurs during fabrication process</td>
<td>Manufacturing imperfection</td>
<td>Inherent unreliability effects in a logic circuit</td>
<td>Timing/Bit error</td>
</tr>
</tbody>
</table>

Considering all sources of variations, one of the most significant is RDF. Additionally, RDF is exacerbated as the device area decreases for bulk substrate devices. It was reported in [1] that the reduction in technology size accompanied by each technology generation decreases area roughly by half. Therefore, the channel dopant concentration decreases exponentially over time. Consequently, due to a few tens of dopant atoms in the channel for generations beyond 32nm, compared to thousands of dopant atoms in earlier technologies, the randomness will increase causing more variability [1]. However, in FinFET devices that have undoped channels, i.e., HK-MG Silicon-On-Insulator (SOI) developed by IBM, RDF is considered negligible and WKF, ITs, and LER become
the major sources of fluctuation [38,39]. The increased occurrence of process variation results in a distribution of threshold voltage ($V_{th}$). As the $V_{th}$ increases, the increase in switching time affects the delay performance of the circuit. Such variability is observed to be magnified by the continued scaling of process technology [39]. For example, the effect of RDF is magnified as the number of dopant atoms is fewer in scaled devices. For instance, the addition or deletion of just a few dopant atoms significantly alters transistor properties.

Nanoscale devices in CMOS technology require not only eliminating the fabrication challenges (extrinsic variation) but also mitigating intrinsic variation effects, which are essential for the characterization of deeply-scaled CMOS planar and FinFET structures [28, 36]. In the design levels, the Process Variation Effect (PVE) and RDF predominate at the circuit-level and device-level fluctuations [40]. Although all of these factors influencing variation participate a major role in impacting system reliability, recently, soft errors have been pronounced in the literature as predominant sources of reliability degradation, especially under the scaling impact at recent CMOS device technology nodes [21–23]. Since the characteristics of soft-errors impacting logic and memory elements differ, we review them separately in Section 2.2 and 2.3.

2.2 Soft Errors in Logic Paths

The contribution of soft errors in logic datapaths as opposed to memory elements is becoming significant as the technology node and supply voltage are scaled down, specifically at 65nm process size and beyond [12,23]. In the past, designers focused on soft error induced by SEUs to protect memories rather than protecting logic paths, i.e., latches/flip-flops and combinational logic gates, where transient logic errors were considered negligible as compared to SEUs of memories [15]. Since device shrinkage leads to a reduction in the amount of collected charge, the SER trend has been higher because of the reduced quantity of collected charge required to induce voltage and
current swings in the CMOS circuits. As research studies and electric design companies have tended to progressively advance technology processes including shrinking the size of transistors, reducing power supply, growing integration capacity, and increasing the switching speed of charging/discharging, it has reduced the circuit’s reliability and increased the complexity of submicron ICs [15, 16]. For instance, Dixit and Wood [10] examined SER in a microprocessor, and their experimental results revealed that SER is expected to exaggerate under technology node scaling. This causes the unprotected part of SER per microprocessor to be increased even though the SEU per bit has decreased. For instance, SER microprocessor logic (latch-origin) is considered negligible in earlier technology nodes, i.e., SER latch-origin is left unprotected. However, as device feature size has been decreased, SER in microprocessor logic becomes the major contribution of system failure [10, 25]. Thus, as with memory elements, it has become essential to protect logic paths against soft errors [12].

Figure 2.2: Random dopant fluctuation as a function of technology generations [1].
2.2.1 Logic Path Masking Mechanisms

In logic paths, the propagation of a transient pulse through downstream logic might be masked by three inherent masking mechanisms. These can prevent the propagation of a spurious transient pulse along a path towards the input of a flip-flop/latch, where it may be registered to cause an SEU [41]:

1. **Logical Masking:** corresponds to when a transient pulse does not affect the computation in other gates along the path towards the output for a given input vector,

2. **Electrical Masking:** due to the attenuation of the glitch while passing through subsequent logic gates, and

3. **Latching-window/temporal Masking:** occurs when a generated glitch does not occur within the setup/hold time window of a flip-flop.

To further elucidate these masking mechanisms, Figure 2.3 depicts a brief example. As can be seen, the generated SET pulse at gate G3 is masked by the logical masking since the output of a AND gate is logic 0 whenever one of its inputs is 0. On the other hand, the SET pulse at gate G5 is attenuated by the electrical masking due to the assumption that its width is smaller than the transition delay time of circuit G5, however, the SET pulse maintains its amplitude as passing through gate G4 since its width exceeds the transition delay of G4, as assumed in this example. While the SET at DFF1 and DFF2 is not captured to flip the bit state and cause an upset because it is either masked by the temporal masking, i.e., does not arrive during the setup/hold time of DFF2 or DFF3, or it arrives at the capturing time, but it was attenuated by the electrical masking. Thus its amplitude is inadequate to cause an upset. Meanwhile DFF3 undergoes an upset (bit state flip) since the transient pulse \((SET_2)\) arrives at the setup/hold time of DFF3.
Figure 2.3: An example of logic datapath masking mechanisms.

These masking mechanisms have been considered to significantly reduce SER in the logic datapaths as compared to that of a memory cell for the same technology process [41]. In addition, they are exploited in the SER mitigation approaches at the circuit/module-level as the means to eliminate the propagation of SET pulses. It is evident that logical masking is not impacted by operation at lower voltages (technology generation) since it is circuit design dependent. For example, a 0 logic on one input of a NAND gate or a 1 logic on one input of a NOR gate prevents the illegitimate pulse from being propagated to the next logic gate, or controlled by the state of the combinational logic.

On the other hand, if a SET pulse duration is roughly lower than the transition delay of a logic gate, it cannot propagate through a logic path. Whereas if its width exceeds twice the transition delay of a logic gate, it propagates through the logic without attenuation. While it might undergo some attenuation during its propagation if its duration is between 1-fold to 2-fold of the logic transition time [15]. At low supply voltages, electrical attenuation is lowered through logic datapaths because SET pulse widths become larger than the transition delay of a logic gate [10]. For instance, in the earlier technologies’ gate capacitances utilized large amounts of charge, and therefore, they were
less prone to experience upsets, and even if an upset occurs, it was most likely to be attenuated by electrical masking since the gate transition delay was larger than the SET pulse width [12]. Likewise, Mahatme et al. [23] projected that electrical masking is likely to diminish as device feature size is decreased due to a reduction in the gates’ capacitance and faster switching speed of transistors. Furthermore, as operating frequencies at lower supply voltage ($V_{DD}$) are expected to be low, it has been suggested that pipeline stages consist of fewer gates to regain lost throughput. This will consequently lower the benefit of both logical and electrical masking.

Similarly, temporal masking mechanism provides reduced benefits under technology scaling. A SET pulse might be registered in a flip-flop or latch if the logic transient time becomes shorter than the duration of the pulse width. Thus, it will no longer be masked by the temporal masking due to the overlapping with the setup/hold time. Furthermore, increasing the clock frequency raises the probability in which the SET pulses overlap with the clock event, a so-called window of vulnerability [15] as shown in Figure 2.4. For example, in [10] it is reported that the masking of the window for latching is decreased by shrinking the transistor geometry size and/or increasing the frequency rate. However, the experimental results of Munteanu and Autran [42] state that SET pulse rate is far below the expected failure rate where the logical, electrical, and temporal masking mechanisms help to suppress SETs and lower the predictable value of SER for 40nm and larger. Overall, incrementing the number of stages in a pipeline processor [41], technology node scaling, and voltage reduction can be anticipated to have detrimental impact on logic SER. Thus, it is sought to design effective soft error mitigation techniques for reliable low power applications. For instance, developing a method to select a hardening technique to protect a module of complex circuitry, so that evaluating and hardening according to the SER of combinational logic-origin or latch-origin is crucial for optimal performance [43]. In practice, designers seek the flexibility of achieving an optimal balance between the soft error coverage and the system requirements, with minimal overheads of power, area, and speed degradation as listed in Table 2.4.
2.2.2 SET Characterizing and Modeling

Recently, new methods have been carried out to model, estimate, and mitigate SET in digital circuits at different abstraction levels, and thus, various SER suppression strategies have been proposed. In earlier technologies processes, the largest part of soft errors causing fault in logic designs was resulting from memory-based SEUs. However, SET pulses from logic, i.e., latches and logic gates, have significantly increased when moving to advanced technology process or when increasing the integration densities [15, 16]. In particular, whether logic path originated SER or unprotected memory-originated SER will dominate the overall SER on a chip has been investigated in [10, 23]. The authors conclude that the logic SER should be emphasized for effective SER mitigation. Also, Mavis and Eaton [2] emphasize that transient glitches, i.e., SET pulses, in logic circuits have become a serious reliability challenge under the shrinking of the MOSFET feature size.

Modeling SETs pulse width under technology node scaling has been investigated widely. In the lit-
Table 2.2: Estimated maximum SET widths based on linear energy transfer (LET) [9]; taking into concentration the effect of Propagation-Induced Pulse Broadening (PIPB) and without the effect of PIPB phenomena.

<table>
<thead>
<tr>
<th>LET (MeV·cm²/mg)</th>
<th>Technology [90nm]</th>
<th>Technology [65nm]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SET Width with PIPB (ps)</td>
<td>SET Width w/o PIPB (ps)</td>
</tr>
<tr>
<td>10</td>
<td>900</td>
<td>125</td>
</tr>
<tr>
<td>20</td>
<td>1200</td>
<td>240</td>
</tr>
<tr>
<td>30</td>
<td>1220</td>
<td>270</td>
</tr>
<tr>
<td>40</td>
<td>1220</td>
<td>360</td>
</tr>
<tr>
<td>50</td>
<td>1300</td>
<td>440</td>
</tr>
<tr>
<td>60</td>
<td>1400</td>
<td>500</td>
</tr>
</tbody>
</table>

In the literature, some predictions of the transient pulse width trends under technology process scaling have been conflicting [9]. In fact, the impact of technology scaling on SETs is found to be complicated by several divergent trends contributing to SET characteristics, such as shrinking the device feature size (reduced $Q_{crit}$), reducing the supply voltage, higher clock rate, trends of datapath masking mechanisms (electrical and temporal masking), Propagation-Induced Pulse Broadening (PIPB) or pulse stretching, pulse quenching, n-well contact and spacing (to reduce parasitic bipolar effects), charge sharing (electromagnetic crosstalk), voltage/oxide thickness/gate length variation, altitude (for space applications), source of radiation/particles, and others [3, 12, 21, 22]. However, the results presented in [9] assist in elucidating some of the discrepancies in SET pulse width measurements presented by various researchers in the last decade, especially the research work that was done by Narasimham et al. [44]. In [9] the variation of transient pulse widths as a function of technology scaling was investigated. A consensus was reached through experimental verification whereby transient pulse widths in bulk 65nm CMOS technology overall was found to be reduced with technology scaling under the conditions that they evaluated, illustrated in Table 2.2.

The most significant outcome from Table 2.2 is that the maximum SET widths for a Linear Energy
Transfer (LET) of 60 MeV-cm²/mg are about 250 ps for 65nm, whereas they are about 1.4 ns for both 130nm and 90nm. Another crucial conclusion is that the Propagation-Induced Pulse Broadening (PIPB) also contributes a role in determining the SET pulse width. However, the broadening length that was considered in [44] was not realistic since the authors determined the longest SET pulse width by taking into account a linear chain of 1,000-inverters, whereas in real logic circuits the logic chains are relatively very short [9]. Thus, the results that are listed in the third and last column are more accurate and realistic. They demonstrate the trends of shrinking the width of SET pulse under technology generation by measuring SET pulse widths for different technologies, 130, 90, and 65nm. Clearly, the trends of transient pulse width are found to be decreased under technology scaling. A related work for determining the broadening SET pulse width in combinational logic was done in [45] using 65nm technology process from TSMC. Hamad et al. [45] projected that if the SET pulse width is sufficiently larger than the threshold combinational logic, it will propagate through logic paths without attenuation. For instance, their experimental results show that in a chain of six 4-input NAND gate the width of the injected SET broadens from 100 to 178 ps as passing through combinational logic, whereas it expands to 184 ps as passing through 4-input NOR gates chain. However, this variation, i.e., PIB, can be controlled by designing gates with balanced rising and falling times [46]. In addition, [47] report that SER of logic paths is not susceptible to the effect of PIB phenomena except for long datapaths circuits. Meanwhile, a SET that hits the control lines have been specified as a considerable portion of the overall SER per chip.

On the other hand, parasitic bipolar amplification is found to be a major contributor that causes long widths of transient pulse in deeply-scaled CMOS technologies [3]. Amusan et al. [3] measured the widths of transient pulses in a 90nm technology and they report that the parasitic bipolar effect and the widths of transient pulses greatly rely on the n-well contact area and spacing. The experimental results in [3] illustrate that the widths of SET pulses can be reduced significantly by modifying the area of the n-well contact while maintaining the well size constant, shown in Figure 2.5. Similarly,
Figure 2.5: Estimated maximum SET widths based on modifying the n-well contact size of the pMOS device for 90nm technology process [3].

Gadlage et al. [9] investigated SET pulse widths based on modifying the n-well contact size of the pMOS device, since parasitic bipolar amplification is more noticeable in a bulk pMOS device with an n-well and p-substrate [9]. Their experimental results show the same conclusion as that of results presented in [3]. Consequently, both results in [3] and [9] confirm that layout design techniques, i.e., n-well contact area, contribute a major role in determining SET pulse width. Thus, Gadlage et al. [9] concluded that if the results of one work is done using a robust scheme for n-well contacts and spacing, the SET pulse widths will be decreased with technology scaling. On the other hand, if a less robust n-well contact and spacing layout scheme is employed to experimentally measure SET pulse widths, the results will come to the conclusion that layout n-well contact and spacing will dominate the effects of transient pulse widths.

Overall, the variation in a SET pulse width can be either broadening its duration or attenuating its amplitude. For instance, it was predicted in [48] that, the SET pulse width is decreased as technology feature size is scaled down, where the critical transient pulse width is estimated to be
85, 70, and 55 pces for 180, 100, and 45nm, respectively. This trend is found to be true for both bulk and SOI devices. Similarly, estimating the width of the transient pulse, induced from neutron effects, under technology node scaling was done in [49] based on changing different design factors, including: drive strength, area of the diffusion drain, fan-out, temperature, and supply voltage. Results indicated that the SER of SET in logic circuits deceases roughly by 50% as technology is scaled from 90nm to 45nm for the same circuit and clock rate. Also, the authors concluded that the main factors that should be focused on are cell type, fan-out, and supply voltage. When reducing the drive strength or increasing fan-out, the impact of SET will decrease for the small SET pulse width (63 to 86 ps), but it will increase for the long pulse width (> 109 ps). Alternatively, supply voltage reduction will increase the rate of SET since less amount of collected charge can change the state in an output node at a logic cell.

Furthermore, numerous research studies emphasize that soft errors will increase and negatively impact the reliability of CMOS circuits as technology is aggressively scaled down [21, 23]. Thus, accurate and efficient methods for estimating SER in current technologies are sought. Diverse analytical techniques were recently presented to assess SER in semiconductor logic circuits. As following Moore’s law, it is predicted that SER per logic bit increases roughly by an 8 percent per technology generation. This is due to the number of transistors per unit area doubling. Thus, SER per unit area continually increases even though SEU per bit has decreased [1, 12, 21, 22], shown in Figure 2.6 [1]. For instance, Intel reported the same percentage for the generations from 0.25µ m to 90nm [7]. On the other hand, Ashraf et al. [50] proposed a novel approach to analyze and estimate the propagation of a transient fault through downstream computing logic. The proposed algorithm utilizes machine learning techniques to precisely track the impact and propagation speed in distributed MPI applications and provide more accurate prediction, thereby preventing the erroneous outcomes/conclusions that the statistical fault injection analysis might lead to, which is based on output variation. It was stated that the introduced fault propagation
Figure 2.6: Soft error rate of a logic and memory [1].

The framework can assist in realizing high resilience by improving vulnerability of HPC applications as it not only accurately predicts the number of corrupted memory locations into parallel MPI applications but also indicates how quickly a transient error propagates into the application’s state.

Recently, researchers have switched to using probabilistic methods for modeling SET since it provides accurate and fast measuring of SET pulse width in logic circuits under the effects of process variation. For example, Yao et al. [51] proposed a statistical SER analysis technique. Their experimental results emphasize that PV effects have made the analysis of SET faults more complicated as the technology process shrinks beyond 40nm. Also reported that process variations of effective channel length ($L_{\text{eff}}$), threshold voltage ($V_{th}$), and oxide thickness ($T_{ox}$) have significantly impacted SER by more than 50% of the static analysis result of SER, thus using only a static SER analysis methodology results in inaccurate estimation of SER in logic. Therefore, utilizing a probability distribution method to characterize SER is desirable since it takes into consideration the effects of process variations and reduces the computational time as compared to the Monte-Carlo
analysis. Similarly, a system’s overall performance can be improved by shifting from deterministic methods of design towards statistical and probabilistic methods due to a reduction in transistor’s variations.

2.3 SER Mitigation Techniques

In order to achieve high-reliability for contemporary computing systems, memory protection techniques (ECC), have been utilized. However, such techniques have recently become inadequate to completely and efficiently protect the reliability of VLSI circuits and systems. In particular, soft error effects (SEE) in logic circuits, i.e., latches/flip-flops and combinational logic, have become significant contributors to increased SER at the system-level [52, 53]. Therefore, it has become essential to ensure the integrity of logic paths by utilizing efficient soft-error mitigation techniques.

Figure 2.7 shows a contemporary taxonomy of soft error mitigation spanning multiple layers of abstraction and the general resolution characteristics. As memories are protected with *Error-Correcting Code (ECC)* and parity via information redundancy, the residual SER comes from logic circuits. In the literature, a wide range of strategies have been proposed to protect CMOS memory and logic circuits versus soft error effects, including radiation-hardening by process and material (RHBP) techniques, i.e., Silicon-on-Insulator (SOI), at the device-level, radiation-hardening by design (RHBD), i.e., spatial and temporal redundancy, at the circuit/module-level, and resilience by coding techniques, i.e., information redundancy, at the system-level [15, 21, 22, 54]. Several modifications to SER mitigation schemes at these levels have been adopted. Each approach exploits a design property to achieve system constraints while accommodating some inherent challenges.

Furthermore, SET hardening techniques can be categorized into three groups, including the point of SET origin (SET generation), along the datapath (SET propagation), and within the latch (SET
Figure 2.7: Taxonomy of design levels for soft error mitigation schemes.

capturing) [21, 22]. Schemes that tackle soft error at the SET’s origin concentrate on limiting the collected charge at the sensitive node of an OFF state transistor, i.e., the drain side. Limiting the collected charge diminishes SET pulse width and reduces SET voltage amplitude [22]. Reducing SET’s voltage amplitude boosts the electrical masking when a SET pulse propagates through downstream logic. Consequently, device resizing [55] for current drive is an obvious SET mitigation technique. Also, specifying the most critical logic data paths, which have nodes more susceptible to cause an error, is considered an effective concept. In particular, one of the most effective techniques for hardening SETs is inside the latch, which leverages the property that not all SET pulses will arrive at the setup/hold time of a flip-flop. Thereby, the transient pulses, which do not overlap with the window of vulnerability of a storage element, will not cause an upset [22]. Techn-
niques that are designed based on neglecting the irrelevant SET, i.e., not captured, are profitable concepts and they have been adopted to reduce the overheads of spatial redundancy approaches.

2.3.1 Device-Level Mitigation Techniques

At this level, designers concentrate on reducing the amount of collected charge by optimizing the fabrication process, i.e., using more advanced and sophisticated manufacturing processes to improve layout designs. Reducing the collected charge provides a reduction in SET pulse width and amplitude; as aforementioned, the former improves the temporal masking, whereas the latter assists in realizing better electrical masking. For example, switching to use Silicon-on-Insulator (SOI) device structure, a semiconductor fabrication technique that IBM pioneered using silicon oxide with pure silicon to realize CMOS circuits and microchip fabrication, improves both temporal and electrical masking due to its buried oxide that reduces the amount of collected charge. Likewise, CMOS SOI-based device structure is more resilient to multiple-cell upset than the bulk structure due to its inherent resistance to the effects of charge sharing, i.e., improve pulse quenching [21]. However, it is more susceptible to PIPB phenomenon, i.e., the change of pulse characteristics during propagation, as precipitated by body potential fluctuations, which results from having a floating body.

Other techniques that also exploit the inherent masking mechanisms are guard drain and guard ring for P-hit and N-hit mitigation, respectively [56]. Additionally, the channel doping concentration of well, or adding more well contacts eliminates or reduces soft error effects. Moreover, layout topology participates in determining the pulse quenching phenomenon that leads to reducing the width of the SET pulse and SER at the device-level. For instance, the experimental results in [57] conclude that benefits from pulse quenching are increasingly pronounced in the common n-well design as compared to the separate n-well design of an inverter chain. This confirms that lay-
out design techniques, i.e., n-well contact area, contribute a major role in determining SET pulse width, and thus reducing SER in CMOS circuits [57]. Furthermore, Lee et al. [58] introduced a new layout topology called Layout Design through Error-Aware Transistor Positioning (LEAP) that can be used to improve soft error resilience in terms of single event multiple upsets for sequential elements, such as latches and flip-flops. The principle of the proposed design is based on physically combining or placing the layout of susceptible contact nodes of any two OFF state transistors with an ON state transistor so that reducing the collected charge, and thus, compacts MBU malfunctions. The design principle of LEAP was tested on DICE-based and conventional flip-flop circuits, and the results showed that the LEAP-DICE achieves 5 and 2,000 orders of magnitude soft error resilience improvement, respectively. In addition, the LEAP-DICE flip-flop incurs negligible power and delay overheads compared to a conventional DICE flip-flop since the schematic structure and the operating behavior remain without any modification. However, since the principle of LEAP design utilizes more complicated intra-cell routing, it incurs roughly a 40% area overhead.

Similarly, Tang et al. [21] suggested that it is urgent to take the crosstalk into concentration for more robust SET suppression. In addition, adding capacitance in the feedback loop, between upset and recovery, has been successfully used to improve the speed penalty of SEU suppression, especially for the write operation of SRAM, and to improve the effectiveness of feedback resistance [59]. However, as we move towards advanced technology scaling processes, utilizing more robust fabrication processes adds more complexity and increases the production cost [25, 26].

2.3.2 Gate-Level Mitigation Techniques

Herein, soft error masking techniques rely on protecting the vulnerable nodes in logic paths. These schemes prevent SETs from being propagated through the datapath, including Cascade Voltage Switch Logic gate (CVSL) [60], critical node ranking, or reliable-aware logic synthesis, i.e., re-
placing cells with alternatives [61]. The drawbacks of these approaches are they impact the system’s speed performance higher than redundancy approaches at the circuit-level, thus they might not be attractive for high speed reliable systems designers. In addition, the use of gate resizing and gate cloning have been introduced as alternative to reduce the overheads of spatial redundancy, i.e., TMR, and improve SET masking of logic paths in CMOS technology. These techniques search to evaluate the most susceptible nodes of a logic circuit through a logic path. Once the susceptible nodes are determined, the logic gates are resized by increasing their capacitance, so that $Q_{crit}$ is increased and the prospect of SET occurrence is reduced. However, resizing the sensitive gate nodes does not guarantee a reduced SET pulse width [46]. Furthermore, it is costly and quite difficult to resize each sensitive individual node in designs consisting of over 1 billion transistors, thus considering some redundancy at a higher level of abstraction is required for complete and cost-effective SEU handling [26].

### 2.3.3 Circuit-Level Mitigation Techniques

Circuit-level SER mitigation techniques have been introduced as an intriguing alternative solution to reduce the design complexity and fabrication cost of SER suppression schemes designed at the device-level [62]. In the literature, numerous design techniques have been introduced at the circuit-level to mitigate soft errors in logic computing circuits and systems, such as spatial, temporal, and hybrid redundancy approaches. The positive aspect of these SER handling techniques is the ability to protect both sequential and combinational logic circuits, aside from mitigating both SET and SEU errors by utilizing a single scheme. However, they involve high area overhead, power consumption, and/or speed degradation. Thus, tradeoffs between these performance penalties and the fault masking coverage are desired, by hardening only the sequential logic portion, i.e., latches and flip-flops. This significantly saves area overhead, improves circuit and system reliability, and lowers production cost. In particular, protecting latches against SEU is a popular research field,
where replacing a regular flip-flop (unprotected) by a SEU-immune flip-flop improves soft error roughly by 10-fold [1, 63]. Next, we discuss the state of the art of radiation-hardening latching circuits and also the conventional redundancy-based mitigation approaches that are introduced to address the effects of soft error.

2.3.3.1 State-of-the-Art Redundancy-based Techniques

Here we discuss some state-of-the-art of radiation-hardening latching circuit designs. Notice that the condition of a transient pulse (SET) to upset a latch or flip-flop circuit is that its duration has to be larger than the feedback loop delay of the latching circuit [64]. There are several schemes that have been developed to increase the robustness of the latching circuits, so that they become resistant to radiation-induced transient errors. For instance, Dual Interlocked-storage Cell (DICE) [65] is one of prevalent designs for hardening against SEUs inside latches. The advantage of DICE is that it utilizes only 12 transistors as compared to 12-16 transistors in a redundant memory cell (hardened cell against SEU) [17]. However, recent investigations of radiation-based soft errors in hardened latches have shown that even though the DICE cell can be an efficient concept to suppress SEU inside latches, it can be very susceptible to upsets when both nodes controlled by the feedback process get struck and change the state of the deposited charge, especially under technology node scaling as the cell is packed very tightly [17]. For instance, at 130nm technology node the DICE cell is able to deliver about 10 times improvement [65] of circuit robustness over the conventional flip-flop, however, it exhibits reduced benefits under technology scaling, i.e., 1.4 times improvement for 40nm technology [66].

Jagannathan et al. [64] presented a radiation-hardening approach to protect latches/flip-flops. The architecture and the principle of the proposed latch is similar to that of a DICE latch where it has four storage nodes to store a single data bit and its complementary, however, the interconnection
topology is different from that of the DICE latch. More importantly, the Quatro latch has larger feedback loop delay, and thus, it realizes a higher immunity to SETs. Moreover, since the Quatro latch has less sensitive nodes and higher critical charge, $Q_{\text{crit}}$, it exhibits higher robustness to charge sharing multiple node upsets than the DICE latch [64]. For instance, the experimental results of [64] demonstrate that the Quatro latch achieves higher SER tolerance than both the conventional and DICE latches, more than 10 and 3 times improvement, respectively. In addition, it outperforms the DICE latch by consuming reduced area and power consumption, roughly about 30% and 40%, respectively. However, it still does not provide a complete immunity to multiple node charge sharing, as this issue is a serious reliability concern for large-scale logic designs that are packing tightly with enormous integration density.

The Built-In Soft-Error Resilience (BISER) technique for protecting latches and flip-flops against radiation-induced soft errors is also an efficient SEU-immune design [67]. This proposed latch utilizes double conventional latches operating in parallel, and a C-element is used to compare the latches’ outputs and determine the final output. In addition, a weak keeper buffer circuit is connected to the output node in order to provide the correct output in an event of upset when the latches’ outputs mismatch, as the output of C-element is floating (high impedance state) during the mismatch of the internal latches’ outputs. Likewise, the other benefit from using the weak keeper circuit is to fight the leakage current when the output of C-element floats due to both the pull-up PMOS device and pull-down NMOS device turn OFF. By considering soft errors in flips-flops only, the presented BISER technique improves chip-level SER by 10-fold at a power consumption penalty of 10.3%. However, the BISER design is only able to tolerate SEUs, and thus, it is not considered as an efficient solution to be embedded in large-scale logic paths of high-reliability systems.

Similarly, Katsarou and Tsiatoutas [14] presented a DICE-base charge sharing-double node upset (CS-DNU) latch that is able to tolerate double errors (upsets) simultaneously. The architecture of
this design is quite similar to the BISER design except that it uses double DICE cells operating in parallel instead of double conventional latches. The proposed DICE-based latch achieves an effective soft error masking coverage for both single and double node upsets, however, this is at the expense of high area overhead and power consumption.

Furthermore, numerous modifications have been carried out to increase the immunity of DICE cells against soft error effects [65], so that it can concurrently tolerate CS-DNU. The primary concept of the previously presented approaches is either increasing the feedback loop delay to introduce longer delay than a potential transient glitch that might hit a node of a DICE cell, or isolating the adjacent susceptible nodes using advanced layout design techniques. D’Alessio et al. [68] presented a Transistor DICE (TDICE) design by adding four NMOS transistors that are functioning only when the cell is written, whereas they are in OFF state during the hold mode, thereby enabling the proposed latch to achieve high speed performance. It was reported that the TDICE design assists in improving SEU resilience by three orders of magnitude of the critical charge for vulnerable nodes. However, the drawback of the presented TDICE circuit is that it is still susceptible to CS-DNU, simultaneously.

To address vulnerability of the TDICE design, Wang et al. [69] introduced a DICE-based latch with feedback transistors. The proposed latch effectively tackles the impediment of TDICE designs and reduces the probability of CS-DNU, simultaneously. Thus, its robustness is extremely improved. Similar to [68], four extra devices (two PMOS and two NMOS transistors), which are turned ON during the write access operations and turned OFF during the hold mode, are added. The disadvantage of the introduced DICE-based latch is that it has a single sensitive pair to cause DNU, simultaneously. However, this vulnerability can be alleviated by isolating the sensitive drain area occupied by this node pair on the cell’s layout. The area and power consumption penalties for the DICE latch with feedback transistors are estimated at 33% and 26% compared to conventional DICE cell, which incurs a roughly 5% speed degradation.
A novel approach, namely Reinforcing Charge Collection (RCC), that is based on adding some redundancy to increase the critical amount of charge at a vulnerable struck node was proposed in [70]. The RCC hardening technique is considered an efficient mitigation scheme that only requires less than 20% of area overhead and 28% of power consumption, while impacting speed performance about 6.4%. However, it is inadequate to be implemented in space applications, aside from its high-cost and reduced fault masking resilience under technology generations and voltage scaling. Next, we discuss the conventional redundancy-based mitigation approaches.

2.3.3.2 The Conventional Redundancy-based Techniques

The soft error resilience of conventional techniques can be obtained by applying four redundant strategies: spatial redundancy approaches such as Triple Module Redundancy (TMR), temporal redundancy approaches such as clock shadow latches, self-voting redundant system strategies (combining the spatial and temporal redundancy approaches), and diverse dual module redundancy, such as spatial redundancy with diversity. Redundancy techniques are effective to mitigate soft errors inside the latch and achieve a high level of protection by suppressing SEUs. They are utilized for designing mission-critical dependable systems so that achieving high reliability and availability while directly countering beneficial characteristics such as, low power, high speed, and minimal layout area. Thus, the challenge is to implement them with less penalties since they incur significant power and area overheads. Note that information redundancy schemes, such as Cyclic Redundancy Check (CRC) codes, are also used for fault handling, but at the system-level and they will be discussed briefly in Section 2.

Most of the redundancy techniques operate based on voting mechanisms. Figure 2.8(a) depicts a conventional 3-input majority voter circuit. The output changes state when two out of the three inputs change [4]. A transient glitch on any one input will be rejected by the voter, considering
the other two inputs are correct. However, a transient pulse occurring within the voter can incur a momentary output glitch that might be corrected in the next clock cycle. Thus no permanent SEU occurs. Figure 2.8(b) shows a circuit implementation of the self-voter, which is a 3-input majority voter configured to compare two inputs and the feedback of the output. The output of a self-voter changes state to a logical 1 when both its inputs are high, and becomes a logical 0 when the external inputs are low. In case of the two external inputs mismatch, output remains unchanged and thus prevents a SET from being propagated to the next stage [4]. Thus, the self-voter circuit is a state-holding element that behaves as a redundant copy of a latch or a flip-flop. Similar to the majority voter, a transient pulse occurring inside of the self-voter circuit results in a momentary glitch on the voter output. This is due to its input driven by a redundant copy of the same value. The self-voting circuit is also commonly referred to in the literature as a C-element gate.

The major benefit of a C-element design is its simplicity. A C-element maintains its output unchanged if the inputs differ, holds the previous state, or behaves as an inverter when the inputs are the same [5]. The C-element might be utilized as an evaluator between two components, as shown in Figure 2.8(c). A transient pulse that occurs at the first combinational logic, results in a discrepancy between nodes A and B, and thus, it will be masked by the C-element as long as its duration is less than the the delay of the buffer circuit. On the other hand, if the width of the SET
pulse is adequately large, it will pass through the C-element to the next combinational logic. This might cause an upset if the SET pulse is captured by a storage element. The delay of the buffer circuit depends on the width of the SET being rejected. As reported in [9], there has been some conflict in the expected SET pulse widths for recent technology nodes as discussed in Section 2.2. The C-element is used to mitigate any SET with a pulse width shorter than the considered delay, by using a small area overhead. In contrast, its drawbacks are that it impacts system performance and requires some modifications for the standard library cell or design flow, besides the circuit fails if the filter or flip-flop is hit. Next, we discuss the four categories of redundancy-based techniques that are employed to protect the mission-critical applications.

**Spatial Redundancy Strategies:** In the literature, numerous spatial redundancy techniques have been proposed to protect CMOS circuits against soft error effects. Modular redundancy is one of the most commonly used techniques where the design is protected by replicating the circuit $N$ times, $N$-Module Redundancy ($NMR$), and extra logic is included for error detection and correction. **Dual Modular Redundancy (DMR)** is configured by duplicating the module, i.e., $N$ equals two. In DMR, an error is detected if there exists a discrepancy between the outputs of identical modules. Conventional DMR is used to detect when a transient or upset error occurs, without error
correction, since the voting circuit is unable to identify which module is error-free [71]. *Triple Module redundancy (TMR)* is constructed by using three identical modules functioning in parallel, i.e., \( N \) equals three. Both DMR and comparison or TMR and majority voter are effective techniques to mitigate SER induced by a single SET or SEU in a logic circuit. The interest of spatial redundancy techniques are their resilience and the ability of employing them in most designs [15]. TMR is a typical example of a spatial redundancy approach to mask soft errors in logic paths, whereby the circuit or the logic datapath is triplicated and a majority voter decides the final output of the circuit [22]; Figure 2.9 depicts a simple example of TMR. Triple modular redundancy circuits are effective only in mitigating a single SET that strikes one of the modules, considering the other two modules are error free. This indicates that TMR circuits are unable to tolerate multiple upsets (SEUs) that influence multiple redundant modules simultaneously. Spatial redundancy is often employed in mission-critical applications to ensure system operation even in unforseen circumstances, such as autonomous vehicles, satellites, and deep space systems [72–74]. It has also been employed in commercial systems such as High-Performance Computing applications [75] where a significant increase in compute node availability is sought. Utilization of compute-node level redundancy at the processor, memory module, and network interface can improve reliability by a factor of 100-fold to 100,000-fold. This is because TMR provides 100% fault masking coverage for faults in single modules simultaneously, compared to the simplex arrangement (unprotected design). However, its major drawback is that it incurs a roughly 2-fold area and energy overhead [15]. Additionally, recent studies have reported that spatial redundancy techniques might increase the susceptibility of SER due to adding more area, and thus, larger sensitive nodes [76]. In short, the conventional TMR approach is suitable for high performance applications that can accommodate its inherent area and energy overheads.

**Temporal Redundancy Strategies:** In the earlier temporal redundancy approaches, input vectors are repeated for the same circuit instance to compare for discrepancy. If there is a mismatch
between both execution results, then a fault occurs and the system needs to rollback and repeat operation. This scheme is immune only to transient faults, cannot detect permanent faults, and in the best scenario, it degrades the system’s goodput, or correct throughput, by 50% as each circuit is checking half the time. An evolved temporal sampling methodology, associated with minimum overheads of area, power, and performance degradation, was proposed by Mavis and Eaton [2]; they describe how the new circuit design methodology can totally eliminate all soft errors, i.e., foreseeable SETs and SEUs, in any synchronous microcircuit. Adoptions and modifications on the basic temporal sampling approach had been done to construct practical circuit embodiments that can be utilized for space and terrestrial sub-micrometer electronics. The key contribution is the ability of implementing the proposed approach in combinational logic preceding a latch or utilizing it to protect the latch itself. In this approach, data for the same combinational logic can be sampled at three distinct instances to construct a voting arrangement while using a simplex instance of the datapath. As shown in the design of Figure 2.10 [2], the data is captured at three different time instances (T1, T2, and T3) by using three identical registers (flip-flops) triggered by three different clock signals (CLK1, CLK2, and CLK3). The relative latency between the clock signals is employed such that they are delayed by a phase shift (Φ1 and Φ2), which can be selected depending on the SET pulse width coverage and the propagation delay time of the circuit, so that the same data from the previous stage is latched in the registers. A majority voter is employed to determine the final output, so that when a soft error occurs in the datapath logic it will be stored in one of the flip-flops, and it should be rejected by the voter. The signal $Data_{in}$ represents the coming data from the previous datapath logic, whereas the final output signal represents the data sent to the next stage. To ensure that the same data is stored in all the registers, temporal redundancy should be constrained by timing constraints, where there is a phase shift $Φ_1$ between CLK1 and CLK2 and $Φ_2$ between CLK1 and CLK3. The timing constraints are presented in [77], so that the legitimate data is captured in the registers.
Figure 2.10: Temporal redundancy approach [2].

Avirneni and Somani [77] employed the temporal redundancy approach for mitigating soft error in a two stage pipeline processor. The authors proposed two approaches, called Soft Error Mitigation (SEM) and Soft and Timing Error Mitigation (STEM). The latter could be used to detect and correct the timing errors that occur when operating at very high frequencies, so-called over-clocking, in addition to detecting and correcting soft errors. Both approaches provide 100% fault coverage and since the proposed schemes are designed such that they do not cost any reduction in the system’s overall performance during the correct operations, i.e., error-free operations, they offer better performance. In terms of performance improvement, SEM approach attains 26.58% on average compared to a conventional TMR approach, while STEM outperforms SEM by 27.42%.

In [78], a heterogeneous spatial and temporal redundant FPGA-based system is designed and implemented by exploiting the linear transform of pipelineable applications. The proposed schemes are based on Concurrent Error Detection (CED). In the heterogeneous CED (hCED) based spatial redundancy, the redundant module is used as a checker for the original module and since it only operates as an error detection circuit, its size can be reduced, which is a major concept of the hCED scheme. Alternatively, hCED based temporal redundancy utilizes the original module as
the genuine module and also as the checker module by switching between them alternately. The major drawback of hCED is that it is considered an application dependent scheme since it needs a condition to identify a block as fault-free. Additionally, it can be used for error detection only (unable for correction).

**Self-Voting Dual Module Redundancy Strategies:** As shown in Figure 2.11, combining the DMR with the self-voting provides less area overhead compared to conventional TMR and less speed degradation (by the width of SET being mitigated) compared to temporal redundancy. The performance degradation is half that of temporal redundancy [4]. Self-voting DMR cannot tolerate multiple SETs similar to TMR. While achieving soft error masking equal to TMR, self-voting DMR must mask the effect of SET on data inputs, clock signals, and storage elements, i.e., flip-flops. In resilient systems for space or nuclear weapon applications, the voter circuit should be protected against soft errors. Figure 2.11 shows an example of protecting an edge-triggered flip-flop besides the protection of the voter circuit, where two majority voters are used at the register output to filter out any SEU. At the temporal sampling phase (see Figure 2.11), register R1 and R2 (DFF1 and DFF2) receive their inputs directly from the datapath logic, while register R3 receives its input from a self-voter that votes on two redundant combinational logic circuits to produce its output. Legitimate data is captured by the third flip-flop as long as the two external inputs match, in the absence of SET. On the other hand, if there exists a present mismatch between the external inputs, this might result in upsetting the third flip-flop. For instance, in case the previous state of the self-voter circuit is a 1 logic and a new data, which needs to be stored in the flip-flops, is a 0 logic, and if one of the redundant modules is hit, and a SET generates and reaches the self-voter circuit, this results in a mismatch, and the output of the self-voter circuit will be unchanged during the transient pulse duration. If the SET pulse overlaps with the setup/hold time of the third flip-flop, it will cause an upset/disrupt the third flip-flop’s state since a 1 logic is captured by the third flip-flop instead of a 0 logic. To address this problem, the third flip-flop should be triggered
Figure 2.11: Self-voting DMR logic circuit [4].

by a clock cycle delayed by the width of the SET pulse. On the other hand, if energetic particles hit the self-voter circuit during the latching cycle of the third flop-flop, this will result in an SEU in the third flop-flop. However, the data capturing phase will not fail as long as the other flip-flops registered the correct data. The main advantage for utilizing the self-voting DMR approach is that it requires to duplicate the datapath logic instead of being triplicated as in conventional TMR, while the sequential part remains triplicated. Based on the experimental results in [4], self-voting DMR provides a performance compromise between the temporal and spatial (TMR) redundancy approaches, where it consumes less area (by one redundant combinational logic) than TMR and provides half the speed degradation of the temporal redundancy scheme. However, for designs that operate with low speed, i.e., the critical delay is large (> 20ns), speed performance of self-voting DMR becomes comparable to the TMR [4]. Overall, the delay of TMR, temporal redundancy, and self-voting DMR are given by:

$$\delta_{\text{critical}} = \max (\delta_i)$$  \hspace{1cm} (2.1)

$$\delta_{\text{TMR}} = \delta_{\text{critical}} + \delta_{\text{voter}}$$  \hspace{1cm} (2.2)
\[ \delta_{Temp} = \delta_{datapath} + \delta_{voter} + 2 \cdot \delta_{SET} \quad (2.3) \]

\[ \delta_{SVDMR} = \delta_{critical} + 2 \cdot \delta_{voter} + \delta_{SET} \quad (2.4) \]

Where \( \delta_{critical} \), \( \delta_{voter} \), and \( \delta_{SET} \) represent the critical path delay of the datapath logic, delay of the voting logic, and delay of transient pulse width, respectively. \( 2 \cdot \delta_{voter} \) is required due to one majority voter and one self-voter are located in the longest critical datapath, whereas \( 2 \cdot \delta_{SET} \) is required to ensure that the legitimate data is captured at the registers. On the other hand, if \( A_{logic} \) represents the required logic area of a design and \( A_{comb} \), \( A_{seq} \), \( A_{voter} \), and \( A_{SVDMR} \) represents combinational logic, sequential logic, voter area, and self-voting DMR design area, respectively then the area overhead for the redundant systems are given by:

\[ A_{logic} = A_{cobm} + A_{seq} \quad (2.5) \]

\[ A_{TMR} = 2 \cdot A_{logic} + A_{voter} \quad (2.6) \]

\[ A_{Temp} = 2 \cdot A_{seq} + A_{voter} \quad (2.7) \]

\[ A_{SVDMR} = A_{comb} + 2 \cdot A_{seq} + 2 \cdot A_{voter} \quad (2.8) \]

Furthermore, area overhead depends on the combinational and register logic ratio in a design. Overall, self-voting DMR provides 10 – 24% area improvement over the TMR implementation, based on different combinational logic ratios in the design, low, medium, and high [4]. Gener-
ally, the higher the combinational ratio, the higher area improvement achieved. Practically, it is possible to combine self-voting DMR with TMR for optimizing large chip designs against SETs, where TMR is advantageous for critical datapath pipeline stages, whereas DMR is convenient for reducing area and energy overheads on pipeline stages with non-critical datapaths [4].

**Diverse Dual Module Redundancy:** The conventional DMR scheme is used to detect the occurrence of a soft error, without correction, since the voting circuit is unable to identify which module is error-free [79]. Recently, the DMR concept has been extended to provide low cost soft error rejection by monitoring the mismatch of the output, called *Diverse DMR (DDMR)*. Design diversity helps to prevent so-called *Common Mode Failures (CMFs)* [80]. By employing design diversity among the modules in spacial redundancy, different outputs are produced under CMF such that an error is detectable [81,82]. Also, design diversity is employed in temporal redundancy to detect permanent faults by performing the same operation redundantly by alternating physical circuits [83]. The main concept of DDMR is to use diversity with pattern mismatch, where two different structures with equivalent functionality utilized in parallel with an error location and masking circuit.

DDMR requires a discrimination circuit to be designed for each module type. When an error occurs, either structure produces a mismatch for one sample or pattern of samples and the error location and masking circuit monitors the output to determine which module structure is error-free [6]. Thus, the error pattern produced by module1 and module2 must be mutually exclusive. By recognizing the module that has a pattern of error when soft error hits, the output of the error-free module is used. The design drawback of this approach is to identify a pair of module implementations that can produce error patterns when a soft error occurs. Besides, a simple logic circuit must be used for error location and masking that only inspects the module outputs. Likewise, the redundant module should not be exceedingly larger than the unprotected module. However, *Digital Signal Processing (DSP)* is a promising field for DDMR due to the fact that DSP modules have
alternative implementations in addition to consuming significant area and energy [79]. Arithmetic circuits are another identified field for utilizing DDMR.

Figure 2.12 depicts a novel technique, namely structural DMR, of utilizing DDMR to protect the *Finite Impulse Response (FIR)* filter by using two different module structures, transpose of direct and mixed transpose-cascade structures [6]. Herein, the design diversity is used to identify and mask a soft error and thus FIR filters are constructed such that they provide different error patterns at the output. A single error in the transpose filter’s output, i.e., \( \hat{y} \), occurs in an upset event, whereas an SEU in any stage (except the final one) of the cascade filter produces consecutive multiple errors in output, i.e., \( \hat{y} \). A duplicated element is used to duplicate the cascade filter’s final stage, while an error correction (detecting and filtering out) circuit is used to check for consecutive mismatches between the filters outputs. When double or multiple consecutive mismatches are observed, the cascade filter must get upset, and thus, output of the transpose filter is selected as the final output. When only a single mismatch is observed, either the error occurs in the transpose filter or in the cascade filter’s final stage. Therefore, the final output is determined by comparing the duplicated element and the cascade filter’s output. When a discrepancy is detected, this indicates that the error must have occurred in the cascade filter. Otherwise, the transpose filter suffers a hit and an upset occurs. On the other hand, if the correction circuit becomes upset, this only changes which filter is selected as the error-free final output. Therefore, a single error cannot occur simultaneously in both cascade and transpose filters and in the error location and correction circuit, and thus, the output of both filters must be correct. Therefore, any one of them can be used as the final output. Experimental results show that the structural DMR FIR filter consumes 2.25x area compared to the non-protected filter while TMR FIR filter requires 3.003x [6].
2.3.4 System-Level Mitigation Techniques

At this level (information based), Error-Correcting Code (ECC), for memory protection, is the most attractive scheme immune to single-bit errors since it incurs less area, speed, and power overheads. While Multi-Cell Upset (MCU) can be suppressed by utilizing advanced coding approaches that are immune to multi-bit errors such as Cyclic Redundancy Check (CRC) codes, Reed-Solomon (RS) codes, Hamming code, et cetera [21, 24], these approaches require extra information along with the data to rollback the error-free data in the event of an upset. At the system-level, the design process is easier since it is achieved by software. However, under the highest level, the system synchronization becomes a major problem, thereby ceasing the running task or system clock is required for synchronization [84]. At this level, soft error suppression is achieved by concentrating on protecting memory cells. Next, we review how soft errors in memory cells are tackled.
2.3.4.1 Soft Error Masking For Other Components

SER has been aggravated under technology process scaling not only in logic circuits but also in memory cells. In the literature, numerous studies [10, 41, 85] have reported that SER in SRAM for a constant array area has increased, since the device feature size is shrunk, the sensitive area is increased. In modern SRAMs, as the density of devices has expanded beyond Moore’s law, recent experimental results showed that SEU susceptibility occurring from low-energy protons and alpha particles from radioactive impurities have become able to deposit adequate charges by direct ionization to cause transient and upset glitches due to an intrinsic reduction in the amount of deposit charge [21, 24]. Physically, mitigating SER in memory cells involves interleaving techniques such as placing the memory layout of a one bit of many different words adjacent on a chip. Likewise, Single Error Correct-Double Error Detect (SEC-DED) codes can be effective even with Multi-Cell Upsets (MCUs) by interleaving the SRAM cells, therefore, MCUs lead to a number of separate Single-Bit Upsets (SBUs) in multiple correction words, rather than Multi-Bit Upsets (MBUs) in a single word [86]. Unlike SRAM, Dynamic Read Access Memory (DRAM) soft error tends to decrease with subsequent technology generations. In fact, this is due to the $Q_{crit}$ amount is roughly constant in a DRAM cell, and thus it becomes more difficult to deposit the same amount of charge into a shrunk sensitive volume [86]. However, there exists a general agreement upon that MCUs should be addressed in DRAM since it is more susceptible for MCUs as density is increased.

On the other hand, recently, Ashraf et al. [87] showed that, even though compiler optimization has played a major contributor to realize high performance and efficiency for multi-core computing applications, their vulnerability towards transient errors has also increased, and generally the impact of vulnerability was found to be comparable to the level of code optimization. This indicates that highly-optimized code for HPC applications is more susceptible to soft errors than the unoptimized code [87].

45
2.3.4.2 Multi-Bit Upsets

Recently, as CMOS devices continue to shrink, advanced memory devices become impacted from multi-cell soft errors and thus MBU becomes a key reliability concern. MBU occurs if two or more bits in the same word are upset by a single event at one time. For instance, spatial MBUs occur when a single particle upsets multiple bits that reside within the same physical neighborhood, whereas temporal MBUs occur when two or more particle strikes independently upset distinct NMR instances. In contrast to SBU, MBU cannot be resolved by single-bit ECC, thus more powerful techniques are required to suppress SER. However, the term Multi-Cell Upset (MCU) refers to errors that occur in different rows of an array (multiple words), thus the serious concern is for MBU as it occurs in a single word.

As the technology generations move forward, MCU exhibits a negative trend as listed in Table 2.3 [10]. Clearly, the experimental results in [10] show that SRAM-MCU has increased from 2% to 38% as the technology node is scaled from 180nm to 40nm. The charge sharing phenomenon induced by SEU is considered one of the crucial mechanisms that cause MCUs. Another source, the parasitic bipolar effects resulting from the well structure strongly affects MCUs. However, Fuketa et al. [88] investigated alpha-particle-induced soft error in 10T subthreshold SRAM for 65nm and they conclude that the main cause of increasing MCU in the subthreshold region is the mechanism of charge sharing. Additionally, the variation of supply voltage and temperature also impact the rate of MBU [89,90]. Numerous fabrication concepts are proposed to reduce/eliminate MBU, such as adding more well contacts, increasing the distance between cell to the next well contact, guard ring, etc. Similarly, SOI technology exhibits less charge sharing, and thus, it is more robust in MCU suppression than CMOS bulk structures [21]. On the other hand, circuit-level mitigation techniques introduce reduced benefits under technology scaling. For instance, Mahatme et al. [91] projected that as the spacing between the sensitive nodes are reduced, the vulnerability
of hardened flip-flops increases. Furthermore, the combination of interleaving and ECC schemes are more efficient in reducing MBU rates. For example, it was stated in [92] that, MBU is reduced when the separations of data bits within each word are designed to be adequately large, whereas reducing spacing between bits in SRAMs increases the likelihood of MBU [91]. Clearly, with technology generation scaling, separating the bits of the same word becomes a complicated process, and it might not be feasible for deeply-scaled register sets, as it impacts area and power consumption. Thus, novel alternative techniques are sought.

Table 2.4 summarizes most of the discussed SER mitigation approaches. It can be seen that each approach has its inherent drawback(s), and there is no optimal solution that can provide efficient performance in terms of power consumption, area overhead, speed degradation, fault coverage, and design complexity. Thus, it can be inferred that a system’s performance is a tradeoff between the discussed issues. For instance, the C-element approach is a simple design that achieves an effective reduction for area overhead, but it incurs a speed degradation, limited by the width of the SET pulse, and also needs some modification for the standard library since it is implemented at the device-level. On the other hand, the self-voting DMR approach provides a comparable area overhead to DMR, making it advantageous for the applications that can tolerate some speed degradation while achieving energy saving and handling soft errors. Meanwhile, DDMR occupies a small area resulting in less speed degradation. These are powerful advantages, especially for DSP
applications and arithmetic circuits, but it requires one to identify a pair of module implementations that can produce a pattern mismatch. Thus, it is inconclusive to determine which approach can be utilized to achieve the design requirements without taking into account the application and its properties that the design is used for. Overall, selecting one of the discussed approaches to protect a design against soft errors depends on the application and the design constraints more than the mitigation scheme. Developing a new concept that can be used for a better optimization of the self-voting DMR approach or finding other promising fields that can be implemented using DDMR to protect a design against soft error effects are the current challenges for voting approaches based on DMR.

In summary, as can be seen in Table 2.5, circuit-level mitigation approaches are the most dominant techniques to protect logic paths against soft error effects since device-level techniques appear to become more difficult to implement and cost higher as device feature sizes are more shrunk. Meanwhile, effective alternative solutions that potentially offer intriguing features, such as being directly implementable on commercial technologies, able to tolerate SEU and MBU, while incurring acceptable area-energy and performance overheads under moderate cost are still sought to increase robustness of latching circuits at the circuit-level.

2.4 Soft Error Rate Trends and Technology Challenges

Recent trends of combining technology and voltage scaling and the massive growth of device integration density have significantly increased the susceptibility of integrated circuits to SER. Herein, we review some trends and challenges of recent CMOS device/circuit design constraints on SER.
<table>
<thead>
<tr>
<th>Approach</th>
<th>Level</th>
<th>Scope</th>
<th>Beneficial Attributes</th>
<th>Performance Penalties and Costs</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOI structures [35]</td>
<td>Device</td>
<td>Optimizing the fabrication process, material and/or structure change.</td>
<td>SER is 5-fold lower than for bulk technology.</td>
<td>More susceptible to thermal effects and pulse broadening phenomenon.</td>
</tr>
<tr>
<td>Guard ring and guard drain [56]</td>
<td>Device</td>
<td>Limit the quantity of collected charge at the sensitive node of the transistor.</td>
<td>Diminishes SET pulse width (improves the temporal masking) and reduces SET voltage amplitude (boosts the electrical masking).</td>
<td>Increase fabrication cost under technology node scaling.</td>
</tr>
<tr>
<td>LEAP-DICE [58]</td>
<td>Device</td>
<td>Physically placing the layout of an invulnerable contact node between the susceptible ones.</td>
<td>Reduce the quantity of collected charge and thus reduce the probability of charge-sharing to cause MBUs.</td>
<td>Utilizes more complicated intra-cell routing, therefore; increases fabrication cost.</td>
</tr>
<tr>
<td>Device resizing [55], High-density well contacts [76]</td>
<td>Gate</td>
<td>Increase the size of the critical nodes of a gate.</td>
<td>Reduce SER.</td>
<td>Increase fabrication cost, besides they do not guarantee reduce SET pulse width.</td>
</tr>
<tr>
<td>Reliable-aware logic synthesis [61]</td>
<td>Gate</td>
<td>Creating an improved cell library to replace cells with alternatives.</td>
<td>Reduce the SET width and the diffusion drain area of a circuit by 30% and 40%, respectively.</td>
<td>Impact system’s speed performance higher than redundancy approaches.</td>
</tr>
<tr>
<td>DICE [65]</td>
<td>Circuit</td>
<td>Adding more transistors to protect the sensitive nodes in a latch or flip-flop.</td>
<td>Reduce the fabrication cost and design complexity, improves soft error roughly by 10-fold.</td>
<td>Vulnerable for charge sharing under technology node scaling.</td>
</tr>
<tr>
<td>BISER [67]</td>
<td>Circuit</td>
<td>Use double conventional latches operating in parallel, C-element and weak keeper to mask soft errors and fight against leakage currents.</td>
<td>Improve fault masking resilience by 10-fold at 10.3% of power consumption penalty.</td>
<td>Consume high area overhead and still susceptible to MBUs</td>
</tr>
<tr>
<td>SEM and STEM [77]</td>
<td>Circuit</td>
<td>Reduce the size of voting circuit by utilizing EDC circuit to detect errors and recover the system by rollbacking the correct data.</td>
<td>Average performance improvement of 26.58% compared to a conventional TMR and STEM outperforms SEM by 27.42%.</td>
<td>Incur the system’s performance and increase the complexity of the design.</td>
</tr>
<tr>
<td>Temporal redundancy [2, 15, 93]</td>
<td>Circuit</td>
<td>Triple the sequential paths with simplex combinational logic path or repeat the operation and check for discrepancy.</td>
<td>Reduce the inherent overheads of spatial redundancy.</td>
<td>Incur the system’s performance, depends on Equation (6.3).</td>
</tr>
<tr>
<td>Spatial redundancy [84, 94, 95]</td>
<td>Module</td>
<td>Double or triple the logic paths in a system.</td>
<td>Masking both SEU and STE.</td>
<td>Incur more than 200% of area and power overheads.</td>
</tr>
<tr>
<td>Self-voting DMR [4]</td>
<td>Module</td>
<td>Double the logic datapaths and triple the sequential logic.</td>
<td>10 – 24% area improvement over the TMR.</td>
<td>Compromised speed performance and area/power overheads as compared to spatial/temporal redundancy approaches.</td>
</tr>
<tr>
<td>Diverse DMR (DDMR) and structural DMR [6, 79]</td>
<td>Module</td>
<td>Design diversity with pattern mismatch, utilizes two distinguished structures with equivalent functionality.</td>
<td>Consume 2.25x area compared to 3.003x of TMR implementation.</td>
<td>Require a discrimination circuit and identifies a pair of module implementations that produce pattern mismatch.</td>
</tr>
<tr>
<td>hCED [78]</td>
<td>Module</td>
<td>Use CED scheme and exploit the linear transform of the pipelineable applications.</td>
<td>Reduce power consumption by reducing the size of the checker circuit.</td>
<td>Cannot correct errors and needs condition to identify a block as fault-free.</td>
</tr>
<tr>
<td>EEC and parity [10, 41, 85]</td>
<td>System</td>
<td>Require extra bits of information be stored with the data in memory cells.</td>
<td>incur less area, speed, and power overheads.</td>
<td>Vulnerable to MBUs, and ceasing the system clock for synchronization.</td>
</tr>
<tr>
<td>Combining ECC with interleaving scheme [92]</td>
<td>System</td>
<td>Separate data bits within each word adequately large.</td>
<td>Reduce the likelihood of MBU in SRAMs.</td>
<td>Complicated process and impacts area and power consumption.</td>
</tr>
</tbody>
</table>
Table 2.5: Comparison between design-level abstraction of SE-mitigation approaches, where by each ($\checkmark'$, –) indicates relative (strength, limitation/weakness).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>System-level</td>
<td>–</td>
<td>$\checkmark'$ high for both SEU and SET</td>
<td>– moderate</td>
<td>– -- not exploited</td>
</tr>
<tr>
<td>Circuit/Module -level</td>
<td>– medium / acceptable</td>
<td>$\checkmark'$ high for both SEU and SET</td>
<td>– moderate</td>
<td>$\checkmark'$ utilized considerably</td>
</tr>
<tr>
<td>Device-level</td>
<td>– medium / high</td>
<td>100% for SEU not guaranteed for SET</td>
<td>– high</td>
<td>– -- not exploited</td>
</tr>
</tbody>
</table>

2.4.1 Trends of Technology Scaling on SER

In the literature, several research efforts [22, 23, 96, 97] have been carried out to investigate the trend of soft error effects on the SER in CMOS logic designs under technology size scaling. Due to many divergent trends contributing to SET occurrences, the effect of technology scaling on the SET width is complicated [22]. Overall, an increase in SER is expected in current computing devices, despite decreasing the SEU rate per bit [10]. There exist several possible explanations to elucidate this issue. However, the most distinctly possible explanations are that under CMOS scaling, VLSI chips have become denser and smaller, therefore, its immunity towards particle strikes decreases significantly. With the massive growth in integration capacity in advanced CMOS technology, the susceptible nodes of the drain area have been increased, thereby leading to higher SER [10]. Additionally, in the earlier technologies, the main contributor of soft error was indirect ionization of neutrons. Even though the neutrons by themselves are unable to cause direct ionization, however, the high energy neutron collision that interacts with semiconductor materials induce sufficient ionizing energy depositions, which can cause SEUs. Therefore, the secondary particles of high energy neutrons have the largest effect on $Q_{crit}$ at sea level. Overall, as $Q_{crit}$ becomes smaller with
process scaling, alpha particles contribute as much as neutrons to the overall SER [24, 34]. This indicates that LET from alpha particles can generate greater than 1 MeV (equivalent to 44.5fc) by direct ionization, which is more than adequate to charge/discharge a node capacitance and flip the state of the struck node [24]. It is also reported that SEU susceptibility to low-energy protons and alpha particles from radioactive impurities is able to deposit adequate charges by direct ionization to cause SEUs due to an intrinsic reduction in the amount of deposited charge [21,24]. Meanwhile, the most highlighted explanation that elucidates why the current trend of soft error towards higher rates is that under technology scaling, the number of transistors per unit area doubles, thus SER per unit area continually increases even though SEU per bit has decreased [10, 12,22]. Thus, soft error effects have been pronounced in the literature as predominant sources of reliability degradation, especially under the scaling impact at advanced CMOS technologies.

Moreover, MCUs have become more pronounced due to reducing device dimensions; therefore, the designers of microprocessors and memory structures should take into account MCUs/MBUs in the SER protection techniques. However, the soft error historical studies in [22] and [21] report that the only fact is that SET rate has not yet exceeded the SEUs failure rate of a storage element.

2.4.2 Trends of Voltage Scaling on SER

Lowering the supply voltage is an effective approach for reducing energy consumption due to the reduction in both static and dynamic energy components, as there is expressive dynamic energy reduction due to a quadratic proportional to the $V_{DD}$, whereas the static energy has a linear dependence. This property can be exploited in a harsh environment, where some remote accessed systems might operate with serious constraints or power consumption budgets. For instance, low power designs can be protected more adequately by employing the redundancy approaches at lower $V_{DD}$ and lower performance. However, the literature has reported several studies regarding the im-
pact of voltage scaling of integrated circuits to soft errors where millions of transistors are used to construct the ICs. Primarily, the critical charge needed to cause a failure decreases linearly as $V_{DD}$ is scaled, and the SER exhibits an exponential effect based on critical charge [10, 44]. Such trends are consistent with decreasing feature sizes due to technology scaling [98]. The experimental results of Dixit and Wood [10] revealed that operation at lower $V_{DD}$ is expected to exaggerate SER. For instance, they state that SER increases by approximately 30% per each 0.1 V decade as $V_{DD}$ is decreased from 1.25V to 0.5V. In the Near Threshold Voltage (NTV) region, it is shown through both simulation and experiment at the 40nm and 28nm nodes, that SER originating in latches doubles when $V_{DD}$ is decreased from 0.7V to 0.5V. Thus, microprocessor energy reduction techniques, such as Dynamic Voltage Frequency Scaling (DVFS), might suffer higher SER, i.e., an increased rate of both SET and SEU, due to reducing $V_{DD}$ to minimize energy consumption, and therefore, negatively impact the microprocessor’s reliability [10].

2.4.3 Trends of Multi-Gate FinFET Devices on SER

To alleviate soft error effects under technology scaling, non-planar structure devices, i.e., FinFET, are introduced as alternatives that achieve increased tolerance to SEEs, i.e., SET and SEU [21]. The benefits of non-planar multi-gate devices on reducing SER in CMOS circuits have been emphasized in numerous studies [27–31, 36–38, 99, 100]. Li et al. [37] reported that FinFET devices have been introduced as alternatives with a 50% reduction of threshold voltage fluctuation of 16nm-gate HK-MG bulk FinFETs compared with the results of planar MOSFETs. Additionally, even though spatial MBU feasibility has increased due to technology scaling, the experimental results in [31] showed that non-planar devices offer a means to reduce SER in logic circuits and MCU in memory arrays. For example, 22nm tri-gate technology is shown to reduce neutron-induced SER at nominal voltage from 1.5-fold to 4-fold and alpha-particle SER in excess of 10-fold compared to a 32nm planar process. Moreover, Fang and Oates [99] projected that the neutron-induced SER
of bulk FinFET SRAM cells is roughly about 6% compared to the SRAM that is configured with non-planar devices. Indeed, this is due to how FinFET structures have shrunk sensitive areas, thereby, they are less prone to radiation effects. Consequently, non-planar devices can achieve reduced SER sensitivity relative to planar FETs [21, 99]. However, the WKF-induced threshold-voltage ($V_{th}$) fluctuation has been pronounced in the metal gate FinFET devices and reported in several studies. For instance, it was reported in [101] that minimal metal grain size plays a crucial role and can effectively reduce device variability resulting from the metal WKF fluctuation in FinFET devices. Likewise, Cheng and Li [102] estimated the fluctuation and suppression of the WKF-induced threshold voltage variation ($\sigma V_{th}$) and the results indicate that shrinking the metal grain size is an efficient scheme to alleviate or eliminate the WKF by the effect of the linear dependence between $\sigma V_{th}$ and grain size. While the preliminary results in [40] emphasize that in digital circuits most of the $\sigma V_{th}$ results from WKF and RDF, in addition to their impact of timing and power. Overall, an increase in SER is expected in deeply-scaled non-planar devices, i.e., FinFET structures that exhibit increased tolerance to particle strikes. This is due to lower $Q_{crit}$ and higher device density per unit area, i.e., roughly doubling, which leverages a higher strike probability.

2.5 Summary

CMOS technology is continually moving towards smaller structures, denser integration capacity, and lower supply voltage. Consequently, a small quantity of charge is stored on each internal node. Thus, the susceptibility of induced radiation will increase, leading to aggravated soft error reliability. In this chapter, a wide range of resilient SER mitigation techniques that exploit specific attributes of different design levels for energy-saving under a reliable protection level are discussed. These techniques have been adopted to mitigate soft error at different abstraction levels. Determining which scheme achieves better performance is a tradeoff between the design require-
ments and their challenges. However, SER mitigation techniques at the circuit/module-level can be convenient since not all SETs will cause an upset in logic paths. Overall, all the stated SER mitigation techniques involve some circuit performance degradation, and as technology trends for more shrinking and higher transistor density, soft error suppression schemes have become more complicated. Thus, more powerful SER mitigation schemes that consider layout modifications and/or circuit design topology with reduced area overhead and circuit performance degradation are sought. The inspiration of this survey has been to provide a compendium of design insights for soft error mitigation techniques to identify the most efficient hardening schemes, in order to serve as a guide for researchers and designers of reliable CMOS circuits and systems.
CHAPTER 3: TEMPORAL SELF-VOTING CHECKERS: ENERGY VS. RESILIENCE TRADEOFFS

In this chapter, we developed two circuit-level techniques, namely Temporal Self-Voting Logic (TSVL) and Hybrid Spatial and Temporal Redundancy Double-Error Correction (HSTR-DEC), to prevent the effects of soft errors in logic circuits, occurring due to Single-Event Upset (SEU) or Single-Event Transient (SET). TSVL and HSTR-DEC circuits can be utilized to improve the reliability of a logic path with minimal impact on circuit delay while achieving a high coverage and cost-effective SEU handling as compared to the previously presented redundancy-based soft-error masking approaches.

3.1 Energy-Efficiency vs Fault Resilience

SE-mitigation techniques are utilized for designing dependable systems so that high reliability and availability are realized while directly countering desirable attributes such as, high speed, low power, and minimal area overhead. However, adding extra logic to harden a design against soft error effects increases the number of susceptible nodes, leading towards the reduction of the system’s overall performance. Hence, the challenge is not how to realize a high reliability, however, achieving it with minimal area and energy overheads and speed degradation is the aim. Considering all these factors, tradeoffs between fault resilience and energy consumption were carried out in designing the proposed soft error handling approaches, so that slightly sacrificing fault making coverage to reduce the energy consumption of proposed schemes. The primary contribution of these SEU-tolerant approaches is that they eliminate error masking from the combinational data-path logic, thus, area and energy overheads are significantly reduced. It is worth noticing that the presented approaches achieve resilient fault masking coverage, higher than 96% for the worst case.
3.2 The Proposed Approaches

Achieving high reliability against transient faults poses significant challenges due to the trends of technology and voltage scaling. Thus, numerous soft error mitigation techniques [2, 4, 77, 78] have been proposed for masking Soft Error Rate (SER) in logic circuits. However, most soft error suppression approaches have significant overheads in terms of area, power consumption, and speed performance degradation. Herein, we introduce two novel efficient and cost-effective techniques to harden the logic paths against soft error effects. The proposed approaches concentrate on protecting the sequential logic elements, i.e., flip-flops or latches, from an upset while achieving significant area and power saving.

3.2.1 Temporal Self-Voting Logic (TSVL) Approach

In conventional Dual Module Redundancy (DMR) approach, the outputs of two identical modules are compared, and an error is detected where there is a discrepancy between them. Conventional DMR scheme is used to detect when a soft error occurs, without error correction, since voting cannot determine which of the two modules is error-free [71]. However, the proposed scheme, TSVL, can be utilized to detect and correct any upset in a flip-flop while incurring acceptable area, power, and performance penalties. As can be seen in Figure 3.1, the design is capable to detect and correct any SEU that hits one of the flip-flops or an SET that generates, propagates, and eventually is captured by one of the flip-flops. Herein, there are three possible scenarios in which an upset will occur. The first scenario is when an SET hits the combinational logic preceding the flip-flops and then propagates and arrives at the setup/hold time of the first/original flip-flop, overlapping with window of vulnerability. This will upset the original flip-flop, while the redundant flip-flop is error-free as long as it is triggered by a clock delayed by a phase shift greater than the width of the transient pulse. Thus, the comparator (XOR-gate) will assert the ErrorSEU signal to indicate the
occurrence of an upset, making the self-voter circuit determines the final output.

Hence, the third input of the self-voter is fed from a MUX that receives its first input directly from the datapath and the second input from the feedback of self-voter circuit; an SEU will be masked by the self-voter circuit regardless of whether it occurs in the original or in the redundant flip-flop. This is because the third input of the self-voter determines the output when the external two inputs mismatch. Since the third input of self-voter circuit relies on the output of the first MUX that drives its output based on the delayed clock rate (CLK2), the self-voter’s third input will be switched alternately between the direct combinational datapath, with the rising edge, and the feedback of self-voter circuit, with the falling edge. Thus, when an SET hits the combinational logic and is then captured by one of the flip-flops, its duration is significant. If it will have passed by the time that the first input’s rising edge arrives to the first MUX, then it will be ignored as long as the delay of CLK2 is larger than the generated SET. Therefore, the correct data will pass as the self-voter’s third input with the rising edge. The purpose for using a MUX circuit at the input of self-voter is to update the previous state of the self-voter circuit with the rising edge of the delayed clock (CLK2). Figure 3.2 depicts the validity of the proposed scheme, as two individual upsets occur in the original and redundant flip-flop, get corrected, and legitimate data is passed to the final output.

As a result, irrespective of whether the upset occurs in the original or redundant flip-flop, the final output is always correct as long as the third input of self-voter circuit alters based on CLK2. The second scenario is when an SET hits inside a flip-flop, an error (SEU) will be observed, and it will be masked in the same way as stated above. The last scenario is when the first MUX or the XOR-gate is hit by a transient pulse, in this case the final output is not effected as long as the output of both flip-flops are correct; therefore, any one of them can be selected as the final output. However, an SET that hits the second MUX circuit, that determines the final output, will cause a momentary glitch in the final output, as in the majority voter of a TMR. Note that signal CLK2 is
generated by buffering the main clock to add phase delay greater than the SET pulse width, which is a popular practice in the design of reliable circuits. The main difference between the proposed approach and the Self-Voting DMR approach presented in [4] is that TSVL functions are based on a simplex combinational datapath and double flip-flops, whereas Self-Voting DMR uses DMR for the combinational logic portion and TMR for the sequential logic portion, thus TSVL achieves significant reduction in area and power penalties as compared to the Self-Voting DMR approach; this will be discussed in Section 3.3.

As compared to spatial and temporal redundancy approach flip-flop based designs, TSVL provides a performance compromise between TMR and temporal redundancy approaches. The proposed scheme (TSVL) consumes less area (by two redundant logic modules) than TMR and provides higher performance in term of speed than temporal redundancy (roughly half the speed degradation

Figure 3.1: Temporal Self-Voting Logic (TSVL) approach.
Figure 3.2: Timing diagram of TSVL approach with SEU recovery occurred in the original, $Q_1$, and redundant, $Q_2$, flip-flop consecutively, masked_out is an inverse of data_in.

of the temporal redundancy scheme). Overall, the delay of TSVL approach is given by:

$$\delta_{TSVL} = \delta_{critical} + \delta_{voter} + 2 \cdot \delta_{Mux} + \delta_{SET}$$  \hspace{1cm} (3.1)$$

$\delta_{critical}$, $\delta_{voter}$, and $\delta_{SET}$ were discussed in chapter 2.

Based on Eq. 2.5, in chapter 2, the area overhead for TSVL approach is given by:

$$A_{TSVL} = A_{seq} + A_{voter} + 2 \cdot A_{Mux}$$  \hspace{1cm} (3.2)$$

Equation 3.1 considers the longest datapath delay. Thus, to meet the timing constraints of a radiation hardening technique, its main clock period must be greater than, or equal to, the longest logic datapath delay.
3.2.2 Hybrid Spatial and Temporal Redundancy Double-Error Correction Approach

Recently, CMOS technologies have been deeply-scaled, thereby a single particle strike that hits a sensitive node can affect more than a single storage element at a given instance of time [103]. This implies that the probability of double or even multiple errors occurring simultaneously has become high. Thus, soft error in flip-flops has become a concern due to the effect of charge sharing between the adjacent nodes [103]. Additionally, circuit-level mitigation techniques introduce reduced benefits under technology scaling. For instance, in [91] it is predicted that, as the spacing between the sensitive nodes is reduced, the vulnerability of hardened flip-flops increases. Therefore, optimized solutions which offer multiple error correction while minimizing performance degradation and energy overhead, imposed by the extra logic for protection, are sought.

Based on the above stated issue, we develop a new technique, namely Hybrid Spatial and Temporal Redundancy Double-Error Correction (HSTR-DEC), for soft error tolerance. As illustrated in Figure 3.3, the proposed approach utilizes DMR circuit for the combinational datapath with TMR for the sequential portion, i.e., flip-flops. The final output is determined by voting based on two masked outputs, nodes N1 and N2, and the masked output from the original and redundant datapaths. The proposed scheme significantly improves the error resilience as compared to previous traditional redundancy approaches due to its ability to tolerate double SEUs simultaneously, i.e., Multi-bit Upsets (MBUs) are tolerated. Moreover, HSTR-DEC approach can surpass not only the SEUs, but also can mask SETs that generate through datapath logic and are eventually captured by one of the flip-flops. Our approach has been motivated by the original technique proposed in [4], and the proposed changes have been designed by evaluating the limitations of error resilience, i.e., Multi-Bit Upsets (MBUs), regarding area and delay overheads. Table 3.1 shows the possible scenarios of soft error (SEU) that HSTR-DEC approach can tolerate. In case a single SEU occurs, the proposed approach is able to mask an upset regardless in which register the error occurs. On
Table 3.1: Possible error scenarios of HSTR-DEC approach; (NE = No Error; UE = Upset Error).

<table>
<thead>
<tr>
<th>of SEUs</th>
<th>Case</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>N1</th>
<th>N2</th>
<th>Masked Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single</td>
<td>I</td>
<td>UE</td>
<td>NE</td>
<td>NE</td>
<td>v'</td>
<td>v'</td>
<td>v'</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>NE</td>
<td>UE</td>
<td>NE</td>
<td>v'</td>
<td>v'</td>
<td>v'</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>NE</td>
<td>NE</td>
<td>UE</td>
<td>v'</td>
<td>v'</td>
<td>v'</td>
</tr>
<tr>
<td>Double</td>
<td>I</td>
<td>UE</td>
<td>UE</td>
<td>NE</td>
<td>UE</td>
<td>v'</td>
<td>v'</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>UE</td>
<td>NE</td>
<td>UE</td>
<td>v'</td>
<td>v'</td>
<td>v'</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>NE</td>
<td>UE</td>
<td>UE</td>
<td>v'</td>
<td>UE</td>
<td>v'</td>
</tr>
</tbody>
</table>

The other hand, in case there are two SETs generated and propagated through the datapath at the same time, they might be captured by R1 and R3 only, whereas R2 will capture the legitimate data since it is triggered by a clock which is delayed by a phase shift greater than the generated transient pulse. There exist three possible scenarios for two SEUs to occur simultaneously. In case I, both register R1 and R2 are upset, therefore, the output of the first self-voter circuit, node N1, is incorrect while node N2 is correct due to the second self-voter votes based on R2, R3, and the redundant direct datapath during the rising edge of the clock. Thus, N2 is error-free as long as both R3 and the redundant datapath are correct. The same scenario, i.e., case III, occurs when both R2 and R3 get upset. In case II, register R1 and R3 get upset, meanwhile nodes N1 and N2 are correct. This is because both self-voter circuits vote based on R2, which is error-free, and the original and redundant datapaths which are also error-free as long as the width of SET is less than 200 psec. However, the proposed scheme is unable to recover when all the flip-flops, i.e., R1, R2, and R3, become upset at the same time. This indicates that the proposed scheme fails to recover only when the energetic particles strike inside the flip-flops and are adequately large to flip the bit state at all registers during the same clock period.
3.3 Experimental Results and Analysis

To assess alternatives that can mask Soft Error Effects (SEE) under the impact of transistor scaling, Synopsys Design Compiler was used to synthesize the proposed approach using 15nm technology process based on NanGate open source library. Since TSVL does not require any modifications or adoptions at the device-level, the standard cell libraries were used to implement the circuits. Thus, no complexity or additional fabrication cost will be associated with the proposed scheme under technology generations. Additionally, area and energy costs are minimized in the generation of the delayed clock signal (CLK2) by using buffers that introduce maximum delay with minimum power consumption and area usage.
3.3.1 Quantifying Fault Resilience for Proposed Techniques

The ability of the proposed hardening approaches to suppress soft errors, was tested by randomly injecting transient and upset faults at the gate-level design. As a case study, a set of ISCAS89 benchmark circuits, with a different combinational and sequential logic ratio, have been synthesized using Synopsys Design Compiler and validated with fault injection technique, presented in [104], where 7000 faults, both SET and SEU, were injected into the gate-level Verilog code describing the benchmark circuit constructed with the proposed approach. SEUs were injected into the input node of a flip-flop with the rising edge, whereas SETs were injected into a node of a logic gate, and they were injected at arbitrarily chosen locations and times. The fault coverage of the proposed approach is listed in Table 3.2 for a set of ISCAS89 benchmark circuits constructed with TSVL using 15nm technology process.
Table 3.2: Fault coverage analysis of TSVL approach.

<table>
<thead>
<tr>
<th>Benchmark Circuit</th>
<th># of Cells</th>
<th># of Injected Faults</th>
<th># of Masked Faults</th>
<th># of Unmasked Faults</th>
<th>Fault Coverage %</th>
<th>SV-DMR [4] FCPR</th>
<th>TSVL FCPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>S27</td>
<td>13</td>
<td>500</td>
<td>487</td>
<td>13</td>
<td>97.4</td>
<td>20.40</td>
<td>23.75</td>
</tr>
<tr>
<td>S349</td>
<td>127</td>
<td>4000</td>
<td>3883</td>
<td>117</td>
<td>97.075</td>
<td>3.51</td>
<td>4.35</td>
</tr>
<tr>
<td>S444</td>
<td>163</td>
<td>7000</td>
<td>6814</td>
<td>186</td>
<td>97.34</td>
<td>2.13</td>
<td>2.5</td>
</tr>
<tr>
<td>S838</td>
<td>418</td>
<td>7000</td>
<td>6798</td>
<td>202</td>
<td>97.11</td>
<td>1.73</td>
<td>2.21</td>
</tr>
<tr>
<td>S1423</td>
<td>683</td>
<td>7000</td>
<td>6779</td>
<td>221</td>
<td>96.84</td>
<td>0.68</td>
<td>0.804</td>
</tr>
<tr>
<td>S9234</td>
<td>1369</td>
<td>7000</td>
<td>6737</td>
<td>263</td>
<td>96.24</td>
<td>0.365</td>
<td>0.455</td>
</tr>
</tbody>
</table>
The analysis of simulation results indicates that TSVL is able to recover from soft errors by 100% of SEUs and 97% of SETs, thus, on average, it can roughly suppress 99% of soft errors. In addition, since SER mitigation techniques offer a tradeoff between the error resilience and overheads of protection techniques, we calculate a new metric called Fault Coverage Power Ratio (FCPR) by dividing the fault coverage by the power consumption, as the goal is to maximize the fault coverage and minimize the power consumption. Thus, the higher the value of FCPR, the better the design. As can be seen in the last two columns of Table 3.2, TSVL realizes higher FCPR for all the selected benchmark circuits, regardless whether the combinational or the sequential logic ratio is higher. Also, TSVL achieves higher FCPR for the small circuits. Notice that the unmasked faults in the S1423 and S9234 are large because these benchmark circuits utilize large number of flip-flops, 74 and 228 respectively, thus the probability of an SET hitting the second Mux circuit is increased. Thus, the proposed approach can be advantageous for soft error tolerant designs constrained with a tight energy consumption budget with a minimal impact on the reliability by 1% for the soft error effects.

3.3.2 Performance Evaluation for Proposed Techniques

In this section, area overhead, power consumption, and speed degradation are quantified for the proposed techniques. The area overhead for the TSVL is estimated by measuring the implementation area in $um^2$, and it turned out that the proposed scheme requires 54.53 $um^2$, whereas 66.5, 64.9, and 43.4 $um^2$ is needed to configure S27 benchmark circuit with TMR, Self-Voting DMR, and temporal redundancy approach, respectively. Figure 3.4 shows the area overhead for a set of ISCAS89 benchmark circuits with a distinct combinational and sequential logic ratio. It can be seen that the proposed approach imposes an area overhead comparable to the temporal redundancy, while achieving significant reduction as compared to Self-Voting DMR or TMR approach. In addition, Self-Voting DMR approach realizes an area reduction when the sequential logic portion is
Figure 3.4: Area overhead evaluation of redundancy-based soft error masking techniques.

high, while the proposed approach saves significant area either if the sequential logic portion is high or if the combinational logic portion is high. On average, TSVL approach imposes less area overhead by 22.02% and 36.84% as compared to Self-Voting DMR approach, presented in [4], and TMR, respectively. In terms of power consumption, TSVL realizes a power consumption improvement of 20.1% and 35.55% over the Self-Voting DMR and conventional TMR approach, respectively, as depicted in Figure 3.5. Therefore, for applications which seek to protect a design against soft errors with constant energy budget, it is advantageous to utilize TSVL approach as it imposes high level of protection, besides achieving significant reduction in area-energy overhead as compared to other hardening redundancy-based techniques.

In terms of performance, TSVL provides a comparable speed performance to that of Self-Voting DMR approach, shown in Figure 3.6, as both incur a phase delay within a period that depends on
the SET pulse width. Overall, TSVL outperforms Self-Voting DMR by 2.15%. A transient pulse of 200 psec was considered in order to generate CLK2 that triggers the second and third flip-flop of TSVL and Self-Voting DMR approach, respectively. Thus, an SET pulse within a width larger than the considered phase delay might be captured by a flip-flop if it arrives at the setup/hold time thereby causing an upset in that storage element, unless it is masked via the electrical or logical masking mechanism.

Interestingly, HSTR-DEC approach is able to tolerate an SET within a width wider than 200 psec. However, in this case the transient pulse should occur in the original datapath only, $Data_{in}$, and the redundant datapath should be error-free at the same time, and vice versa. This indicates that HSTR-DEC is able to tolerate a single SET with a width larger than the considered phase delay as long as the other datapath is correct. On the other hand, HSTR-DEC scheme is able to tolerate
two simultaneous SETs within a width up to 200 psec. While TMR, temporal redundancy, and self-voting DMR approaches achieve complete fault masking coverage for SEU in single module simultaneously, HSTR-DEC approach realizes complete fault masking coverage for SEU in single or double module concurrently. In addition to its increased level of reliability, the proposed scheme consumes acceptable extra logic for error masking. The evaluation results indicate that HSTR-DRC approach reduces area and power overheads roughly by 18.2% and 16.83%, respectively, and imposes the performance by 20.17% as compared to conventional TMR using 15nm technology. On the other hand, it improves speed performance by 15.38% while incurring 19.23% of area overhead as compared to temporal redundancy approach. Likewise, it achieves comparable overheads in terms of power, area, and delay as compared to previous works of hybrid redundancy approach, presented in [4], as both utilize duplicated datapaths and triplicated flip-flops. However, our approach is able to tolerate double upsets, occurred at the same time, whereas conventional
Table 3.3: Mean area-power consumption and delay of redundancy-based approaches w.r.t. simplex design.

<table>
<thead>
<tr>
<th>Design Implementation</th>
<th>Normalized Area</th>
<th>Normalized Power</th>
<th>Normalized Delay</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial Redundancy</td>
<td>3.14x</td>
<td>3.11x</td>
<td>1.19x</td>
</tr>
<tr>
<td>Temporal Redundancy [2]</td>
<td>1.88x</td>
<td>1.83x</td>
<td>1.69x</td>
</tr>
<tr>
<td>Self-Voting DMR [4]</td>
<td>2.48x</td>
<td>2.46x</td>
<td>1.4x</td>
</tr>
<tr>
<td>Proposed Approach (TSVL)</td>
<td>1.96x</td>
<td>1.91x</td>
<td>1.37x</td>
</tr>
<tr>
<td>Proposed Approach (HSTR-DEC)</td>
<td>2.6x</td>
<td>2.54x</td>
<td>1.43x</td>
</tr>
</tbody>
</table>

TMR, temporal redundancy, and SV-DMR are unable to detect and correct double upsets simultaneously. Table 4.2 lists the area, power consumption, and speed degradation penalties of the redundancy-based soft error mitigation techniques, normalized to the simplex (non-redundant) design. In addition to increased levels of fault coverage, the proposed approaches are considered a compromise solution ranging between the spatial and temporal redundancy approaches, as they realize a high protection against soft error effects, MBUs tolerance for HSTR-DEC, and impose acceptable overheads in terms of area, power, and performance degradation.

3.4 Summary

As technology continues to scale, CMOS logic becomes denser, thus its immunity to wards particle strikes decreases significantly. In this work, efficient and cost-effective redundancy-based Soft Error (SE) handling approaches are proposed, namely TSVL and HSTR-DEC. They are utilized to prevent the effects of soft errors in logic paths, occurring due to SEU or SET. The proposed approaches are validated by injecting transient and upset glitches. Experimental results indicate that TSVL approach can cover soft errors, on average, roughly by 99% while realizing an improvement of 22.02% and 2.15% for area and speed degradation compared to the previous Self-Voting DMR approach. Meanwhile, HSTR-DEC approach realizes complete masking coverage for single
and double SEUs while incurring comparable area and delay overheads as compared to the prior hybrid redundancy approach, aside from its energy-efficiency as compared to conventional TMR approach.
CHAPTER 4: IMPACT OF PROCESS VARIATION IN THE NVT REGION ON SOFT ERROR RATE

In this chapter, we investigate the effect of both technology size scaling and delay variation at Near-Threshold Voltage (NTV) region on redundant systems, regarding energy versus performance tradeoffs within an iso-energy constraints. We develop new results for the evaluation of alternatives to mask Single Event Transients (SET) in combinational logic and Single Event Upsets (SEU) in storage elements for three commonly utilized redundancy approaches, namely, spatial, temporal and hybrid of both spatial and temporal. The performance, and energy impact of each approach is quantified at NTV operation. Additionally, the impact of increased effect of threshold voltage variation at NTV is assessed for all redundant systems. We also investigate the effect of technology scaling by comparing the energy and performance variation of 45-nm MOSFET planar and 16-nm High-κ/Metal Gate (HK-MG) bulk FinFETs structures as modeled by PTM NanGate open source library via simulations in HSPICE. Monte-Carlo simulations reveal the influence of the threshold voltage ($V_{th}$) variation in redundant systems at NTV to mask a transient pulse in combinational logic and/or an upset glitch in a storage element.

4.1 Energy-Efficiency Through Near-Threshold Computing

Supply voltage ($V_{DD}$) scaling is widely recognized as the most effective strategy to reduce power dissipation / energy consumption for VLSI logic devices. Total energy consumption is quantified by dynamic and static energy components which are both dependent on supply voltage. The dynamic energy has a quadratic relationship with $V_{DD}$, whereas the static energy has a linear dependence. Operation at Near Threshold Voltage (NTV) region indicates that supply voltage is between the nominal voltage and the threshold voltage of the transistor. In fact, NTV operation
slightly reduces supply voltage to 100-200mV above the threshold voltage of the transistors, allowing for significant improvement in energy-efficiency [105], up to six orders of magnitude of energy savings as compared to operating at nominal voltage [106]. Taking all of these factors into consideration, NTV operation provides a 10 times delay improvement compared to operation in the sub-threshold region with only a 50% reduction in energy savings [106]. Hence, operation in the NTV region is sought, as it provides an energy-efficient operating point. Herein, we assess the relative performance of temporal and spatial alternatives from 800mV to 500mV using a 16-nm Predictive Technology Model for multi-gate transistor (PTM-MG) library. We also investigate how a portion of the energy savings obtained can be utilized to tune resilience by means of redundant implementations.

4.2 Trends of SER in NTV Region

Operation with $V_{DD}$ in the NTV region is sought for highly-scaled CMOS logic circuits due to its balancing of minor performance degradation relative to its significant power savings [107]. While NTV offers an attractive approach to balance energy consumption versus delay for power-constrained applications such as high-performance computing, there is a need to evaluate its reliability implications through increase in soft errors [105] and performance variation due to higher impact of threshold voltage variation [108]. In particular, radiation-induced SEUs which cause soft errors can increase significantly in this operating region [10]. Furthermore, the SER is exacerbated by two complex interacting factors: technology scaling and reduced $V_{DD}$. As SER has an exponential dependence on the critical charge, $Q_{crit}$, reducing $V_{DD}$ will increase the impact of SETs since the amount of $Q_{crit}$, which needs to be collected in order to change the state of an output node at a logic cell, is reduced. Primarily, the critical charge needed to cause a failure decreases as $V_{DD}$ is scaled and SER has an exponential dependence on $Q_{crit}$ [10]. As operation at NTV is
expected to exaggerate the trends of SER, the increase in SER can have a significant effect on the reliability of logic elements inside of deeply-scaled VLSI systems operating at NTV.

4.3 Technology Scaling and PV Trends For SET

In the literature, some predictions of the transient pulse width trends under technology process scaling have been conflicting [9]. In fact, the impact of technology scaling on SETs is found to be complicated by several divergent trends contributing to SET characteristics, such as reducing $Q_{\text{crit}}$, Propagation-Induced Pulse Broadening (PIPB), pulse quenching, n-well contact area and spacing (to reduce parasitic bipolar effect), trends of datapath masking mechanisms, charge sharing, and source of radiation/particles [9, 22, 23]. However, the results presented in [9] assist in elucidating some of the discrepancies in SET pulse widths measurements presented by various researchers for the last decade. Gadlage et al. [9] experimentally measured the transient pulse widths in a bulk 130, 90, and 65-nm CMOS technology, and they report that SET pulse width overall was found to be reduced under technology scaling for the conditions which they evaluated. However, an overall increase in SER is expected in recent computing devices, even though the SEU per bit has decreased due to lower $Q_{\text{crit}}$ and higher device density per unit area, i.e., roughly doubling, which results in a higher strike probability, even in deeply-scaled non-planar devices that exhibit increased tolerance to particle strikes. Moreover, the analyses of the previous soft error studies have predicted that soft error in combinational logic and latches/flip-flops would dominate the overall CMOS chip errors, and it will exacerbate under technology process scaling [21]. Particularly, in [67] it is predicted that SER contribution of logic circuits to total chip SER, roughly estimated at 60%, exceeds SRAM which contributes 40% of SER during execution. Therefore, SER logic-based should be emphasized for effective SER mitigation [23].

On the other hand, to alleviate the Process Variation (PV) associated with technology scaling,
FinFET devices have been introduced as an alternative with 50% reduction of threshold voltage fluctuation for 16-nm-gate HK-MG bulk FinFETs compared with the results of the planar MOS-FETs [37]. In particular, performance improvement of nano-scale CMOS devices requires not only eliminating a variety of fabrication challenges (systematic or extrinsic variation) but also mitigating random or intrinsic variation effects including Random Dopant Fluctuation (RDF), Line-Edge Roughness (LER), Interface Traps (ITs), and Work-Function (WK) variations, which are crucial for device characterization of nanometer scale CMOS planar and FinFET structures [28, 36]. The non-planar devices offer a means to reduce SER. For example, 22-nm tri-gate technology is shown to reduce neutron induced SER at nominal voltage from 1.5-fold to 4-fold and alpha-particle SER in excess of 10-fold compared to a 32-nm planar process [31].

In this work, intra-die variations for both CMOS 45-nm planar and tri-gate 16-nm bulk FinFET technology nodes are simulated using the Monte-Carlo method in HSPICE. For 45-nm, the random effects are modeled through the variation in $V_{th}$ caused by RDF and LER effects [39]. Similarly, for 16-nm the variation is due to RDF, WK, and ITs effects [28, 37]. The standard deviation $\sigma V_{th}$ values of 25.9mV for 45-nm process and 28.7mV for 16-nm process are adopted from [39] and [37], respectively. Finally, we restrict our discussion to show how these PV effects combine in redundant systems of logic datapaths to exhibit a higher mean delay than a simplex system. While numerous studies in the literature discuss recent innovations for SE-mitigation in memory devices, we focus herein on logic critical datapaths due to its increasing importance. The redundant system performance is determined by the worst-case delay out of any of the constituent modules. For instance, the delay of a TMR arrangement can be predicted to exceed that of any single module since the module instance with maximum delay determines its final delay. Generally, if the worst-case delay of module instance $i$ of a redundant system is $\tau_i$, then the overall delay of the TMR
system $T_{TMR}$, temporal system $T_{temp}$, and hybrid system $T_{SVDMR}$ are given by:

$$
T_{TMR} = \max_{1 \leq i \leq 3} \left( \tau_i \right) + \delta_{voter} \tag{4.1}
$$

$$
T_{temp} = \tau_i + \delta_{voter} + 2 \times \delta_{SET} \tag{4.2}
$$

$$
T_{SVDMR} = \max_{1 \leq i \leq 2} \left( \tau_i \right) + 2 \times \delta_{voter} + \delta_{SET} \tag{4.3}
$$

Where $\delta_{voter}$ and $\delta_{SET}$ represent the delay of the voting logic and the delay of transient pulse width, respectively, which contribute directly to the critical delay. $2 \times \delta_{SET}$ is required as a phase shift between CLK1 and CLK3, see Figure 2.10, to ensure that the legitimate data is captured at the registers, whereas $2 \times \delta_{voter}$ is required in SV-DMR because one majority voter and one self-voter are located in the longest critical datapath [4].

4.4 Experimental Setup

4.4.1 Simulation Objectives, Tools and Workflow

Experiments are carried out to analyze the overheads for the above-mentioned soft error mitigation techniques in terms of energy consumption and the $\sigma V_{th}$ impact on the propagation delay. For this case study, an inverter chain composed of 26 Fanout-of-4 inverters with registered outputs was synthesized and simulated using 45-nm MOSFET planar and 16nm-gate High-$\kappa$/Metal Gate (HK-MG) bulk FinFETs structure based on PTM-based NanGate open source library [109], using an HSPICE simulation. Monte-Carlo simulations were carried out to implement the threshold variation in spatial, temporal, and hybrid (SV-DMR) redundancy schemes. These simulations vary the $V_{th}$ of the transistor in the netlist based on a Gaussian distribution having a mean equal to
the nominal model card for PTM and $V_{th}$ as provided in [37, 39]. The $\sigma V_{th}$ can be adapted to accommodate local and global variations, or their combined effects as considered in this work.

First, operation at NTV increases the effect of threshold variation on the datapath delay of the soft error mitigation techniques redundancy-based. Therefore, using Eq. 6.1 through Eq. 6.3, we quantify the delay performance impact of threshold voltage variation for each mitigation technique. For example, Eq. 6.1 is used to determine the delay of a TMR system. The overall delay in this case is determined by the slowest critical datapath among the three modules and the delay of the voter circuit. On the other hand, there exists only a single datapath in the temporal redundancy approach while intricate details about the working of the voter circuit determine the overall delay as quantified in Eq. 6.2. The design of the temporal redundancy scheme is depicted in Figure 2.10, in chapter 2. It can be observed that the data is captured at three different time instances (T1, T2, and T3) by using three identical registers (flip-flops) triggered by three different clock signals (CLK1, CLK2, and CLK3). The relative latency between the clock signals is employed such that they are delayed by a phase shift ($\Phi_1$ and $\Phi_2$), where there is a phase delay $\Phi_1$ between CLK1 and CLK2 and $\Phi_2$ between CLK1 and CLK3. To further clarify, Figure 4.1 illustrates the internal working of Local Clock Manager (LCM) which is used to generate the required clock signals. Finally, the delay of $\Delta T$ can be selected depending on the SET pulse width coverage, thereby the total datapath delay is determined by summing the datapath and voter circuit’s delay, while considering the slowest clock rate (CLK3) as delay of $2 \times \delta_{SET}$ should be added.

In addition, the tradeoff analysis for the hybrid (Self-Voting DMR [4]) mitigation approach, 2.11 discussed in chapter 2, was carried to for delay variation calculation. We calculate slowest datapath delay among two modules, the original and the redundant module. However, it uses two voting circuits, one majority voter and one self-voter, therefore, $2 \times \delta_{voter}$ is required for the longest delay datapath. In addition, it also incurs a phase delay within a period that depends on the SET pulse width to delay the clock rate of the third flip-flop, more details can be found in [4]. Thus, the total
Figure 4.1: Local Clock Manager (LCM) to generate CLK2 and CLK3.

The datapath delay is determined by summing the critical datapath and the delay of two voter circuits, while considering the slowest clock rate (CLK2) as quantified in Eq. 6.3.

The considered equations are leveraged in HSPICE circuit simulation tool to calculate the critical path delay for each scheme, so that investigating the effect of threshold variation on each SER mitigation technique. The simulation framework is illustrated in 4.2. The Monte-Carlo simulations were conducted to utilize at least 1,000 experimental runs for every design implementation, therefore, the datapath delay of TMR system is obtained by calculating the mean value of the datapath delay of each module for 1,000 runs. Then the module instance with maximum datapath delay determines the delay of the TMR system as indicated in Eq. 6.1. Mean values are reported for each case. This scenario is repeated 1,000 times to establish mean value for every supply voltage value, as the supply voltage was altered from the nominal $V_{DD}$ (1.1V) to 0.5V by a discernment of 0.5V for the 45-nm CMOS planar structure, whereas it is altered from the nominal $V_{DD}$ (0.8V) to 0.5V for the 16-nm bulk FinFET structure. Furthermore, the energy consumption is computed by accumulating the energy requirement of the mean value operating at a frequency of $1/T_{TMR}$. The same setup is altered for other redundancy-based SE mitigation techniques presented in this work.
4.5 Experimental Results and Analysis

4.5.1 Quantifying Area Overhead for SE-Mitigation Techniques

It is known that the area overhead for the TMR is more than 200% including voting logic, whereas for temporal redundancy only the latches in the design grow in number while all the combinational logic elements remain unchanged. Thus, area overhead incurred can be expected to be roughly twice the area of the latches in the overall design. Indeed, a design which consists of a large sequential ratio and/or a large word-width occupies larger area as more flip-flops are required. On the other hand, area overhead of SV-DMR scheme presents a compromise between spatial and temporal redundancy approaches. For the fault coverage, the spatial, temporal, and hybrid redundancy
approaches achieve complete fault masking coverage for SEU in single module simultaneously. However, temporal and hybrid redundancy are capable of detecting and correcting all upsets occurring on registers, but they are unable detect any transient pulse width exceeding 160 psec, as explained later on. Therefore, to maintain high coverage for all redundant systems in order to enable a fair comparison, a higher value of SET pulse width is considered herein than estimated in prior works [9].

4.5.2 Comparison of Delay and Energy Consumption

Results Figure 4.5 depict energy consumption for the simplex (non-redundant) and redundant arrangements. It is observed that temporal approach consumes a comparable amount of energy compared to the unprotected circuit. However, it incurs an average speed degradation of 28.19% compared to TMR. In addition, as listed in Table 4.1 the experimental results depict an increment in the speed degradation for both temporal and hybrid redundancy approaches with scaled supply voltage. This is due to buffering the clock signals causes more delay with scaling $V_{DD}$, buffering the clock signals will be discussed late in Section 4.4.5. Similarly, the speed degradation of both TMR and SV-DMR is seen to be produced from both considering the slowest critical delay path due the $V_{th}$ variation and the delay of voter(s) circuit, besides considering the slowest clock rate for SV-DMR as quantified in equation 6.3. Figure 4.3 shows speed degradation analysis for redundant systems at NTV region using 45-nm technology process.

To further analyze the impact of increased variability on energy overhead of soft error resilient designs, experiments were conducted using the 16-nm PTM-MG HP model. As shown in Table 4.1, which illustrates the energy reduction and speed degradation for different supply voltages, the temporal redundancy realizes higher average energy saving as compared to TMR implementation, 54.14% and 56.21% for 45-nm and 16-nm, respectively. This is partly due to the fact that
the multi-gate FinFET devices have less leakage than the planar MOSFET devices [27, 28]. In term of speed performance, the speed degradation, see Figure 4.4, of both temporal and hybrid redundancy is reduced while utilizing multi-gate bulk FinFET devices because the latter achieves more robust gate controllability and thus improves the system performance. Thus, both temporal or hybrid redundancy approaches are able to exploit the benefits of scaled technology devices. The overall comparison of energy and delay for all redundant system is presented in Table 4.2. On average, temporal scheme consumes 34.51% more energy than the simplex circuit, whereas the energy overhead is 207.19% and 123.25% for TMR and SV-DMR, respectively. However, temporal redundancy degrades system performance by 37.79% within a $\delta_{SET} = 150$ psec for transient pulse width using 16-nm technology.

Figure 4.3: Speed degradation analysis at NTV for the 45-nm technology.
Figure 4.4: Speed degradation analysis at NTV for the 16-nm technology.

Table 4.1: Mean energy reduction and speed degradation of temporal redundancy and SV-DMR approach versus TMR.

<table>
<thead>
<tr>
<th>V_{DD}</th>
<th>45nm</th>
<th>16nm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Temporal</td>
<td>SV-DMR</td>
</tr>
<tr>
<td></td>
<td>Energy</td>
<td>Speed</td>
</tr>
<tr>
<td>1.1 V</td>
<td>55.6%</td>
<td>21.22%</td>
</tr>
<tr>
<td>0.8 V</td>
<td>54.41%</td>
<td>26.73%</td>
</tr>
<tr>
<td>0.65 V</td>
<td>53.9%</td>
<td>31.9%</td>
</tr>
<tr>
<td>0.55 V</td>
<td>53.6%</td>
<td>33.16%</td>
</tr>
<tr>
<td>0.5 V</td>
<td>52.45%</td>
<td>35.76%</td>
</tr>
<tr>
<td>Ave.</td>
<td>54.14%</td>
<td>28.19%</td>
</tr>
</tbody>
</table>

4.5.3 Impact of Process Variation in NTV Region

Table 4.3 depicts the speed degradation in the NTV region of redundant systems normalized to the simplex design implementation. It can be seen that the speed degradation exacerbates with
lowering $V_{DD}$ for redundant implementations. This implies that the variation in the output delay was found to be increased under scaling down $V_{DD}$ from the nominal level to 0.5V for both the 16-nm bulk FinFET and 45-nm planar MOSFET. The critical path and SET pulse width for spatial and temporal redundancy respectively incur more speed degradation with down scaling of the supply voltage. As illustrated in Figure 4.6, the delay variation of temporal redundancy is lower than
the variation of both TMR and SV-DMR, even the variation of 16-nm is beneath that of 45-nm technology node for spatial and hybrid redundancy approach. This implies that delay variation increases for either higher redundancy or increased sophistication of the fault resolution circuit. However, the latter impacts the delay variation higher as can be seen in the curves of SV-DMR approach in Figure 4.6 for both 45-nm and 16-nm technology node. Thus, temporal redundancy can be utilized to alleviate the effect of delay variation at NTV. In addition, the 16-nm node-based temporal system improves the normalized speed degradation by 12.4\% compared to the implementation of a temporal system based on 45-nm planar, and it also enhances the average normalized energy saving by 5.74\% compared to the 45-nm planar structure. Consequently, soft error resilient low power designs can be protected more adequately by employing the temporal approach at lower $V_{DD}$. 

Figure 4.6: Delay variation for redundant systems using 45-nm and 16-nm technologies.
Table 4.3: Normalized speed degradation of spatial, temporal, and hybrid redundancy (SV-DMR) w.r.t. simplex design.

<table>
<thead>
<tr>
<th>(V_{DD})</th>
<th>45nm</th>
<th>16nm</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8V</td>
<td>1.105x</td>
<td>1.33x</td>
</tr>
<tr>
<td>0.75V</td>
<td>1.107x</td>
<td>1.34x</td>
</tr>
<tr>
<td>0.7V</td>
<td>1.109x</td>
<td>1.36x</td>
</tr>
<tr>
<td>0.65V</td>
<td>1.11x</td>
<td>1.37x</td>
</tr>
<tr>
<td>0.6V</td>
<td>1.11x</td>
<td>1.39x</td>
</tr>
<tr>
<td>0.55V</td>
<td>1.12x</td>
<td>1.42x</td>
</tr>
<tr>
<td>0.5V</td>
<td>1.13x</td>
<td>1.44x</td>
</tr>
</tbody>
</table>

4.5.4 Impact of Supply Voltage in NTV Region

NTV operation reduces supply voltage to 100-200mV above the threshold voltage of the transistors, allowing for significant improvement in energy-efficiency with a reasonable performance impact [105]. The energy savings can be utilized for either reduced power consumption or to increase resilience via redundant implementation. Thus, the near-threshold region allows for consideration of interesting tradeoffs. For example, a design with spatial or temporal redundancy can be utilized as a means to increase reliability within the same energy budget as a simplex (unprotected) system operating at nominal voltage. This is valid provided that the increase in delay, and thus corresponding drop in performance and area costs are acceptable. Note that pursuit of increased reliability is predicated upon the assumption that the source of variability in the near-threshold region is due to variation in \(V_{th}\) for which this work is restricted. Based on these assumptions, the tradeoffs in Figure 4.7 and 4.8 are highlighted. For instance, it shows the feasibility of the temporal approach at around 0.97V on average given an identical energy budget of a simplex system operating at nominal voltage of 1.1V (for 45-nm technology). While TMR and hybrid redundancy can be employed to protect the design at lower supply voltage, \(\sim 0.67V\) and \(\sim 0.8V\), respectively. On the other hand,
Table 4.4: Effect of reducing supply voltage under iso-energy constraints.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Simplex</td>
<td>1.1</td>
<td>1x</td>
<td>1x</td>
<td>0.8</td>
<td>1x</td>
<td>1x</td>
</tr>
<tr>
<td>TMR</td>
<td>0.67</td>
<td>1.059x</td>
<td>2.7x</td>
<td>0.5 ~ 0.55</td>
<td>1.057x</td>
<td>1.615x</td>
</tr>
<tr>
<td>Temporal</td>
<td>0.97</td>
<td>1.087x</td>
<td>1.68x</td>
<td>0.7 ~ 0.75</td>
<td>1.072x</td>
<td>1.37x</td>
</tr>
<tr>
<td>SV-DMR</td>
<td>0.8</td>
<td>1.037x</td>
<td>2.35x</td>
<td>0.6 ~ 0.65</td>
<td>1.014x</td>
<td>1.48x</td>
</tr>
</tbody>
</table>

For 16-nm technology process, results emphasize that utilizing temporal redundancy approach at NTV provides better energy reduction with scaling, shown in Figure 4.8. The temporal scheme can be employed to mitigate soft errors with the nominal energy budget at a supply voltage between 0.7V to 0.75V. Meanwhile, spatial and hybrid redundancy should operate between 0.5V to 0.55V, or 0.6V to 0.65V, respectively, to maintain the energy consumption at the same level as the simplex design implementation. In this case, using either TMR or SV-DMR in NTV region will degrade the performance more than utilizing temporal redundancy as shown in Table 4.4, where it depicts performance degradation with constant energy budget for redundant systems compared to simplex design implementation. Thus, while maintaining energy-efficiency temporal redundancy scheme provides better performance in terms of energy saving, speed degradation, and variation in output delay with technology scaling at NTV. Therefore, for applications which seek to protect a design against soft errors with constant energy dissipation budget as compared to nominal operation, it is advantageous to utilize temporal redundancy approach to achieve that at lower $V_{DD}$ since scaling down the supply voltage is the most effective method to reduce energy consumption.

To further invigorate tradeoffs between the resilience of fault masking coverage and performance penalties of protection techniques in terms of power consumption and delay degradation, the evaluation of the redundant arrangements was done by calculating a new metric called Fault Coverage Energy Ratio (FCER), which divides the fault masking coverage over the energy consumption.
Figure 4.7: Maintaining energy consumption at NTV using 45-nm technology process.

The aim is to maximize the resilience of fault coverage and minimize the consumed energy, and thus, the higher the value of FCER metric, the better the protection arrangement.

\[
FCER = \frac{\text{fault coverage}}{\text{time delay} \times \text{power consumption}} \tag{4.4}
\]

The summary of the FCER analysis for the redundancy-based soft error mitigation techniques is listed in Table 4.5. Note that, only the fault masking coverage for TSVL approach is listed in Table 4.5, while the masking coverage for other redundancy-based techniques is assumed to be 100%. Additionally, Figure 4.10 shows the average of FCER under scaling down \( V_{DD} \) from the nominal level (1.1V) to 0.5V for 45nm technology. As can be seen in Table 4.5, at the nominal supply voltage, the hybrid redundancy scheme (TSVL) achieves the best tradeoffs in terms of FCER.
Figure 4.8: Maintaining energy consumption at NTV using 16-nm technology process.

among the redundancy-based soft error mitigation approaches for most of the selected benchmark circuits, irrespective of whether the combinational or the sequential logic ratio is higher. While the temporal redundancy realizes high or acceptable FCER when the combinational logic ratio is high, whereas TMR achieves desirable/acceptable FCER either when the sequential logic ratio is high or when the benchmark circuit occupies a small area. On the other hand, at the NTV region, the temporal redundancy approach realizes an increased benefit in terms of FCER, meanwhile TSVL still achieves competitive FCER compared to the temporal approach at NTV operations. Consequently, the TSVL approach can be utilized to harden logic paths against radiation-induced soft errors that are constrained with an iso-energy consumption budget with a minimal impact on the masking coverage roughly by 1.5% of transient errors. Likewise, the temporal approach offers promising attributes at NTV operations while encountering an acceptable/reasonable performance delay variation and degradation.
Figure 4.9: SETs pulse width rejection at NTV for 45-nm planar MOSFET and tri-gate 16-nm bulk FinFET.

4.5.5 Relationship of Area and FCER

Herein, experiments are conducted to explore the relationship between the SER and area overhead of redundant systems by using a set of ISCAS89 benchmark circuits with various number of cells. It turns out that the ratio for all circuits is the same except that the power is positively correlated with area and the probability of soft error is positively correlated with area. Thus there is an overall impact of $FCER = O(\text{Area}^2)$ relationship as can be seen in Figure 4.11. Meanwhile, since the TSVL approach offers a reduced SER while incurring modest performance penalties in terms of occupied area, power consumption, and speed degradation, it is considered an intriguing approach that addresses or alleviates the drawback(s) of spatial and temporal redundancy approaches.
Figure 4.10: Mean fault coverage energy ratio analysis of redundancy-based approaches, $V_{DD}$ is scaled from nominal level (1.1V) to the NTV region (0.5V) with a decrement of 0.5V.

4.5.6 Benefit of Spatial Redundancy with Design Diversity

Here, diversity-enabled TMR arrangements are evaluated based on the FCER metric that is defined in Section 4.5.4 to highlights the advantages/disadvantages of design-diversity. Table 4.6 lists the summary for some design arrangements of TMR systems. Results indicate that the inverter-based chain achieves the highest FCER, but it incurs the highest delay variation to impact the speed performance. Additionally, these results are based on a single error (SE) injection scenario considered for each diversity-TMR inverter chain, since the conventional TMR approach is only able to tolerate a single error at a time. Thus, the diversity-TMR arrangements might be used to alleviate/address the issue of MBU, aside from its benefit to tolerate the CMFs. Meanwhile, the delay variation results in time-dependent bit error, when the delay variation exceeds the clock
### Table 4.5: Fault coverage energy ratio analysis of redundancy-based approaches.

<table>
<thead>
<tr>
<th>Benchmark Circuit</th>
<th># of Cells</th>
<th>Fault Coverage of TSVL (%)</th>
<th>Redundancy Approach</th>
<th>FCER ( (V_{DD} = 1.1,V) )</th>
<th>FCER ( (V_{DD} = 0.75,V) )</th>
<th>FCER ( (V_{DD} = 0.55,V) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>S27</td>
<td>13</td>
<td>97.4</td>
<td>TMR</td>
<td>12.82</td>
<td>12.25</td>
<td>10.28</td>
</tr>
<tr>
<td>S349</td>
<td>127</td>
<td>97.075</td>
<td>TMR</td>
<td>3.60</td>
<td>3.4467</td>
<td>2.8929</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Temporal</td>
<td>3.77</td>
<td>3.6617</td>
<td>3.9745</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SV-DMR</td>
<td>3.73</td>
<td>3.7022</td>
<td>2.8993</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TSVL</td>
<td>4.26</td>
<td>4.3210</td>
<td>3.4546</td>
</tr>
<tr>
<td>S444</td>
<td>163</td>
<td>97.34</td>
<td>TMR</td>
<td>2.25</td>
<td>2.1523</td>
<td>1.8065</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Temporal</td>
<td>2.19</td>
<td>2.1330</td>
<td>2.3153</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SV-DMR</td>
<td>2.13</td>
<td>2.1165</td>
<td>1.6574</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TSVL</td>
<td>2.71</td>
<td>2.7488</td>
<td>2.1976</td>
</tr>
<tr>
<td>S838</td>
<td>418</td>
<td>97.11</td>
<td>TMR</td>
<td>1.55</td>
<td>1.4937</td>
<td>1.2537</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Temporal</td>
<td>1.82</td>
<td>1.7662</td>
<td>1.9172</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SV-DMR</td>
<td>1.66</td>
<td>1.6468</td>
<td>1.2896</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TSVL</td>
<td>2.027</td>
<td>2.0595</td>
<td>1.6466</td>
</tr>
<tr>
<td>S1423</td>
<td>683</td>
<td>96.84</td>
<td>TMR</td>
<td>0.798</td>
<td>0.7630</td>
<td>0.6404</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Temporal</td>
<td>0.748</td>
<td>0.7247</td>
<td>0.7866</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SV-DMR</td>
<td>0.724</td>
<td>0.7179</td>
<td>0.5622</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TSVL</td>
<td>0.924</td>
<td>0.9392</td>
<td>0.7509</td>
</tr>
<tr>
<td>S9234</td>
<td>1369</td>
<td>96.24</td>
<td>TMR</td>
<td>0.223</td>
<td>0.2140</td>
<td>0.1797</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Temporal</td>
<td>0.278</td>
<td>0.2690</td>
<td>0.2920</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SV-DMR</td>
<td>0.224</td>
<td>0.2219</td>
<td>0.1738</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TSVL</td>
<td>0.282</td>
<td>0.2888</td>
<td>0.2309</td>
</tr>
</tbody>
</table>

Period. Therefore, considering the transient error (SET/SEU) and timing error when calculating the fault masking coverage will provide a more accurate evaluation for the FCER metric, which is a means to further highlight the pros and cons of design diversity when used with spatial redundancy.

#### 4.5.7 Constraints of Temporal Redundancy Approach

Herein, experiments are conducted to explore the tradeoffs of performance vs SET pulse width for the temporal approach. To generate the clock signals for temporal redundancy approach, shown in Figure 2.10, a Local Clock Manager (LCM) is employed, presented in [77], offering advantages
in terms of complexity and energy consumption. CLK1 is the same as CLK (main clock), while CLK2 and CLK3 are produced by utilizing clock buffers. The number of buffers depends on the phase shift between the clocks. The latter depends on SET pulse width. The transient pulse width is estimated in [9, 110] with nominal voltage operation and 65-nm technology node between 25 psec to 125 psec at the sea level. In this work, a relatively higher pulse width value of 160 psec and 150 psec was chosen to accommodate the worst case and incorporate the effect of using reduced
technology node size of 45-nm and 16-nm, respectively. As illustrated in Figure 4.9, the considered delay, $\delta_{\text{SET}}$, increases as lowering the supply voltage. Thus, operation at NTV increases the delay for the clock buffers, where the number of buffers were determined based on nominal value of $V_{DD}$, and thus this makes the temporal redundancy scheme more robust to reject larger SET pulse width than operation at nominal voltage. Overall, the temporal redundancy approach is able to reject SET pulse widths ranging between 380 psec to 850 psec and 170 psec to 230 psec when the supply voltage is scaled between 0.7V to 0.55V, for 45-nm and 16-nm technology process, respectively. This demonstrates the benefit of temporal approach for these parameters. However, this at the expense of performance degradation as stated in Section 4.4.2.

4.8 Summary

Mitigating soft errors at NTV can provide a range of alternatives across metrics of area, speed, and power. Thus, it would be advantageous to consider NTV operation within contemporary constraints of the design, such as minimum energy within a soft error mitigated design, maximum speed given an energy budget, or a tradeoff between these issues. In terms of delay variation, temporal redundancy incurs less variation (22%) at NTV ($V_{DD} = 550\text{mV}$) under technology node scaling as compared to both TMR and SV-DMR (31.6% and 35.2%, respectively), even the variation of 16-nm is beneath that of 45-nm technology node for both. Thus, the drawback of TMR and SV-DMR is the need to accommodate the slowest module delay. On the other hand, temporal redundancy provides higher energy saving, but it requires consideration of the SET pulse duration. Thus, for soft error resilient low power designs, designers can select temporal redundancy at lower supply voltage thereby allowing for significant improvement in energy-efficiency at NTV under technology generation, while facing an acceptable delay variation and a reasonable speed degradation.
CHAPTER 5: Energy-Efficient and Soft-Error Resilient Non-Volatile Spintronic Flip-Flop Designs

This chapter will discuss implications for my current research work.
CHAPTER 6: High-Performance Double Node Upset-Tolerant Non-Volatile Flip-Flop Design

This chapter discusses implications for future emerging spin-based latching circuits. Recently, emerging spin-based devices are introduced as an intriguing candidate to alleviate leakage currents and continue the scalability of CMOS technology. However, their immunity to radiation-induced Soft Error Effects (SEEs) needs to be adequately addressed. In this chapter, a radiation-immune hybrid Spin Transfer Torque Magnetic Tunnel Junction (STT-MTJ)/CMOS flip-flop is designed and evaluated for nonvolatile applications. The proposed nonvolatile flip-flop circuit achieves attractive features, such as low standby power dissipation (21% less than CMOS-based design), high computing performance, and superior SEEs resilience (concurrently can tolerate DNU) to potentially become as a mainstream solution for the aerospace and avionic nanoelectronics.

5.1 Hybrid CMOS/Spintronics: Alleviating CMOS Challenges

Currently, CMOS logic designs are moving toward their physical limits, and therefore, their continued scalability to sustain Moore’s law becomes challenging in near future [111]. In addition, static power in CMOS logic and memory circuits of technologies beyond 45nm has increased due to increasing leakage currents [112]. To address this issue, a hybrid STT-MTJ/CMOS circuit design is developed as a potential alternative and an intriguing candidate to address scalability challenges [113]. Moreover, the property of Non-Volatility (NV) facilitates the capability of instantly turning on devices that are normally OFF. Thus, the energy consumption demand can be reduced. Due its efficient recall operation and compatibility with CMOS processing, the STT-MTJ NV flip-flop has been targeted for nonvolatile processors. In particular, it is an intriguing concept to address power failures from energy harvesting [114, 115].
On the other hand, the issue of radiation-induced transient effects, i.e., Single-Event Transient (SET) and Single-Event Upset (SEU), in the CMOS-based circuits for read/write access operations remains unaddressed [116]. In addition, as the spacing between the sensitive nodes is reduced with technology size scaling, the vulnerability of SEU-immune flip-flops increases. Thus, soft errors in flip-flops have become a concern due to the effect of charge sharing between the adjacent nodes to cause Double Node Upset (DNU), simultaneously [103, 117]. Consequently, efficient solutions to completely and efficiently address this issue are sought to improve the reliability of hybrid STT-MTJs/CMOS nonvolatile sub-micron electronics [111], thereby making this technology feasible. In response to these aforementioned challenges, we have developed a power-efficient with superior soft error resilience Non-Volatile Flip-Flop (NVFF) design that significantly reduces the potential Single-Event Double Node Upset (SEDNU) rate due to charge sharing at susceptible nodes. The experimental results validate its resilience, in terms of DNU-tolerance, and its efficiency, in terms of occupied area, speed performance, and power dissipation, compared to the prior work in [118] and [119]. The proposed design is the first NVFF circuit that has been demonstrated to tolerate charge sharing DNU, simultaneously.

5.2 Fundamentals of Magnetic Tunnel Junctions

An MTJ consists of two ferromagnetic (FM) layers, called fixed layer and free layer, which are separated by a thin oxide barrier, i.e., MgO [120]. There are two different magnetization configurations for FM layers, parallel (P) and antiparallel (AP), according to which MTJ resistance changes between $R_P$ and $R_{AP}$, respectively. MTJ resistance is determined by the angle ($\theta$) between the magnetization orientations of fixed layer and free layer due to the tunnel magnetoresistance (TMR) effect [121]. The MTJ resistance in P ($\theta=0^\circ$), and AP ($\theta=180^\circ$) states is expressed by
the following equations [112,122,123]:

\[
R(\theta) = 2 R_{MTJ} \times \frac{1 + TMR}{2 + TMR + TMR \times \cos \theta} \tag{6.1}
\]

\[
R_{MTJ} = \frac{t_{ox}}{Factor \times Area \times \sqrt{\phi} \exp(1.025 \times t_{ox} \times \sqrt{\phi})} \tag{6.2}
\]

\[
TMR = TMR(0)/1 + \left(\frac{V_b}{V_h}\right)^2 \tag{6.3}
\]

Where \(V_b\) is the bias voltage, and \(V_h = 0.5V\) is the bias voltage when TMR is half of the \(TMR_0\), \(t_{ox}\) is the oxide thickness of MTJ, \(Factor\) is obtained from the resistance-area product value of the MTJ that relies on the material composition of its layers, \(Area\) is the surface area of the MTJ, and \(\phi\) is the oxide layer energy barrier height.

A promising approach, which is known as spin transfer torque (STT), is proposed by Slonczewski [124] to switch the MTJ states. In the STT approach, a bidirectional spin-polarized current (\(I_{MTJ}\)) is required for switching the free layer nanomagnet configuration of the MTJ, as shown in Figure 6.1. The P or AP configuration of the MTJ is determined by the direction of the current that flows through it. The required bidirectional current could be produced by means of simple MOS-based circuits. According to the relative amplitude of the \(I_{MTJ}\) and the switching critical current (\(I_C\)), STT switching behavior can be categorized into precessional region (\(I_{MTJ} > I_C\)), and thermal activation region (\(I_{MTJ} < I_C\)). To have high switching speed, MTJ is required to work in precessional region. The below equation describes the switching duration of the MTJ in this region [125]:

\[
\frac{1}{(\tau_{STT})} = \frac{2}{C + \ln(\pi^2 \Delta)} \frac{\mu_B P}{\eta \mu_B P (1 + P^2)} (I_{MTJ} - I_C) \tag{6.4}
\]
Table 6.1: Parameters of STT-MTJ devices.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTJ Area</td>
<td>MTJ Length ( \times ) MTJ Width ( \times ) ( \frac{\pi}{4} )</td>
<td>( 40nm \times 40nm \times \frac{\pi}{4} )</td>
</tr>
<tr>
<td>( t_{ox} )</td>
<td>Thickness of oxide barrier</td>
<td>0.85 nm</td>
</tr>
<tr>
<td>( P )</td>
<td>Spin Polarization</td>
<td>0.52</td>
</tr>
<tr>
<td>( R_P, R_{AP} )</td>
<td>MTJ Resistances</td>
<td>( 3.22 , \text{KΩ}, , 6.44 , \text{KΩ} )</td>
</tr>
<tr>
<td>TMR( R_0 )</td>
<td>TMR ratio</td>
<td>100%</td>
</tr>
</tbody>
</table>

Figure 6.1: Spin transfer torque switching mechanism.

Where \( \tau_{\text{STT}} \) is the mean duration for precessional switching region, \( C = 0.577 \) is the Euler’s constant, \( \Delta \) is the thermal stability factor, and \( m \) is the free layer magnetic moment. In this paper, Verilog-A is utilized to model the behavior of STT-MTJs based on the aforementioned physical equations [113]. Then, the model is leveraged in SPICE circuit simulator to validate the functionality of the designed circuits using experimental parameters listed in Table 6.1.
5.3 Technology Scaling Trends on Data Bit Upset

In recent sub-micron CMOS electronics, the latching circuits have become more susceptible to multiple node upset (MNU), simultaneously, due to the effect of charge sharing phenomenon. This is because the charge/discharge node capacitance of a circuit and the spacing between the vulnerable nodes are reduced to ten(s) of nanometers under the aggressive scaling of technology generations [103]. As a result, a higher Soft Error Rate (SER) is anticipated due to single particle strikes upsetting double nodes concurrently. Moreover, separating the bits of the same word becomes a complicated process, and it might not be feasible for deeply-scaled register sets, as it impacts area and power consumption [69]. Therefore, optimized solutions which offer multiple error correction capability while minimizing performance degradation, area overhead, and power dissipation, imposed by the extra logic for protection, are sought. Next we discuss the proposed soft error resilient NVFF design.
5.4 Radiation-Induced Transient Faults Emulation

In the earlier technologies, the main contributor of soft error was indirect ionization of neutrons. Even though the neutrons by themselves are unable to cause direct ionization, however, the high energy neutron collision that interacts with semiconductor materials induce sufficient ionizing energy depositions, which can cause SEUs. Therefore, the secondary particles of high energy neutrons have the largest effect on $Q_{\text{crit}}$ at sea level. Overall, as $Q_{\text{crit}}$ becomes smaller with the process scaling, alpha particles contribute as much as neutrons to the overall SER [24, 34]. This indicates that Leaner Energy Transfer (LET) from alpha particles can generate greater than 1 MeV (equivalent to 44.5fc) by direct ionization, which is more than adequate for recent technologies to charge/discharge a node capacitance and flip state of a struck node [24]. It was also reported that SEU susceptibility to low-energy alpha particles from radioactive impurities is able to deposit adequate charges by direct ionization to cause SEUs due to intrinsic reduction in the amount of deposit charge [21, 24].

As discussed in Section 5.1, the MTJs are inherently immune to radiation-induced transient faults. However, the COMS-based portion for read/write access operations needs to be hardened. To imitate the effect of alpha particles hit, a double exponential current is used. This current source was injected/connected into vulnerable nodes during the sense mode mode for the master-latch and the latching mode for the slave latch. The injected current is given by:

$$I_{\text{inj}}(i) = -\frac{Q_{\text{inj}}}{\tau_1 - \tau_2} (e^{-\frac{i}{\tau_1}} - e^{-\frac{i}{\tau_2}})$$

(6.5)

Where the parameter $Q_{\text{inj}}$ represents the amount of injected charge and it depends on the technology node size (50 fC was considered for 45nm), whereas $\tau_1$ and $\tau_2$ represent the time constant with typical values of 150 and 50 ps, respectively [126, 127]. Thus, herein, $\alpha$-particles effect is
considered the major source of radiation-induced transient faults.

5.5 Proposed DNU-Tolerant NV Flip-Flop Circuit

Herein, we have developed a power-efficient with superior soft error resilience NVFF design. Our approach has been motivated by the techniques presented in [111] and [69], based on the combination of reliable Pre-Charge Sense Amplifier (PCSA) circuit [111] and a double upset tolerant DICE latch with feedback transistors [69]. The proposed modifications have been designed by evaluating the limitations of power consumption, soft error resilience, and volatility of CMOS-based large-scale logic circuits, regarding area overhead and speed performance degradation. Figure 6.2 depicts the design architecture of the presented NVFF circuit. As can be seen, the structure of the proposed NVFF circuit is composed of three components including: logic control for write access operations, nonvolatile master latch, and volatile slave latch. The write circuit employs STT switching mechanism to switch the configurations of the MTJs since the STT technique is considered the most mature and advanced writing mechanism due to its low current usage, e.i., 50µAmps@40nm [128], and high-speed performance. In addition, the transmission gates (TG1 and TG2) were added to isolate the write circuit from the sensing path during the sensing mode, so that synchronizing the read/write access operations and preventing the read disturbance. The magnetic master latch utilizes 17 transistors in sense amplifier circuit configuration that is highly robust against particle strike induced charge sharing DNU, whereas the slave latch uses 18 transistors in DICE latch with feedback transistors configuration that is substantially immune to single-event induced double upset.
5.5.1 Functionality Analysis of the Proposed NV Flip-Flop

Default operation mode: We consider the SEU-free case. When the clock signal (CLK) is low and the complement clock signal (CLK) is high, i.e., falling edge of the clock, the coming data bit at the input is written into the NV master latch by reconfiguring the state of the MTJs based on the bit value. This is achieved by the STT switching mechanism that generates bidirectional current to switch the free layer of the storage MTJ cell to represent a 0 or 1 logic value as compared to the reference MTJ. During the rising edge of the CLK signal, when CLK is high and CLK is low, the magnetic master latch is functioning in the transparent mode of operation, and the sensed data bit from the NV MTJ cells is sent to the output through TG3. Whereas when the CLK signal is low and CLK is high, the slave is functioning in the latching mode of operation, and the output is taken from n2 of DICE latch through TG4 with the falling edge of the CLK. Meanwhile, since the feedback path delay of the proposed NVFF is isolated from the input-to-output delay path during the rising edge of the clock signal, the proposed DNU-tolerant NVFF achieves excellent sensing and \( \text{CLK} \rightarrow \text{to} \rightarrow \text{Q} \) delay times as the overall delay is acceptable in comparison to that of an unprotected flip-flop. The proposed NVFF was evaluated by simulations using HSPICE, and Figure 6.3 shows the timing waveforms. As can be seen in Figure 6.3, both nodes \( S_0 \) and \( S_1 \) are precharged to \( V_{DD} \) during the falling edge of the CLK, while the data bit and its complementary are sensed at these nodes during the rising edge of the CLK. The output, \( Q \), is updated from node \( S_1 \) with the rising edge of the CLK, while the logic value of node \( n_2 \) is sent to \( Q \) during the falling edge.

5.5.2 Soft Error Rate Analysis of the Proposed NV Flip-Flop

In this section, the robustness of the introduced NVFF against transient errors is evaluated. In order to estimate the Soft Error Rate (SER) for the proposed NVFF, we have used the model presented
Figure 6.3: Timing waveforms of the proposed DNU-tolerant NVFF.

Table 6.2: Vulnerability and upset tolerance analysis for the proposed NVFF.

<table>
<thead>
<tr>
<th>Latching Circuit</th>
<th>Total $\Phi$ of Nodes</th>
<th>$\Phi$ of Vulnerable Nodes</th>
<th>Possible $\Phi$ Node Pairs</th>
<th>Unprotected Nodes Pairs</th>
<th>Fault Masking Coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Master-Latch</td>
<td>6</td>
<td>S0, S1, S2, S3</td>
<td>15</td>
<td>(S0,S3) and (S1,S3)</td>
<td>100 for SEU 86.67 for DNU</td>
</tr>
<tr>
<td>Slave-Latch</td>
<td>8</td>
<td>n0, n1, n2, n3, v0, v1</td>
<td>28</td>
<td>(n0,v0)</td>
<td>100 for SEU 96.43 for DNU</td>
</tr>
<tr>
<td>The Proposed NVFF</td>
<td>14</td>
<td></td>
<td>91</td>
<td>3 pairs</td>
<td>100 for SEU 96.7 for DNU</td>
</tr>
</tbody>
</table>

in [129], which determines the SER of a CMOS-based logic circuit by adding the total number of
vulnerable nodes \( n \) in the circuit [20]. Thus, the total SER is given by [20]:

\[
SER = \prod_{i=1}^{n} \frac{WOV_i}{T_{CLK}} \cdot k_i \cdot e^{-\beta \cdot Q_{on}(i)}
\]  

(6.6)

Where \( n \) represents the number of nodes, \( k_i \) is a constant proportional to the area of the node, and \( \alpha \) and \( \beta \) are fitting parameters [129]. While window of vulnerability of node \( i (WOV_i) \) refers to the time interval during which an energetic particle striking a susceptible node \( i \) might result in a soft error [20].

As stated in [129], a transient fault strikes at an input node of a latch circuit is not a major concern since it has a marginal impact. Thus, in this work, the resilience of fault masking of the hardened latches has been evaluated by considering only transient faults hitting the internal (feedback loop nodes) and output nodes when the CLK single is asserted. Table 6.2 lists the overall number of nodes that cause major reliability impact. In addition, the most vulnerable nodes of each latch (master/slave) are identified based on their susceptibility. These nodes, i.e., (S0, S1, S2, and S3) for master-latch and (n0, n1, n2, n3, v0, and v1) for slave-latch, result in a latch upset when hitting by an energetic particle with adequate LET to flip the latch state. These nodes are considered susceptible area because they are connected to the drain side of OFF state transistors. Further details on how the most susceptible node pairs are identified can be found in [69, 111].

5.6 EXPERIMENTS AND RESULTS

Experiments are carried out to analyze the overheads for the proposed soft error resilient NVFF. The proposed DNU-tolerant NV flip-flop was synthesized and simulated using HSPICE simulations based on 45nm HK-MG bulk CMOS PTM-based NanGate open source library. The con-
ducted simulations for the proposed NVFF circuit is performed with supply voltage of 1.1 V and room temperature. In order to make the design work using a single clock rate for both the TGs and the MTJs’ sensing circuit, a relatively higher switching current of roughly 100 µA was used to quickly switch the configuration of the MTJs at a delay of ~2 ns during the falling edge of the clock. Thus, the frequency is set at 250 MHz for the entire circuit. However, this is at the expense of high write energy as will be discussed in Section 6. For a fair comparison purpose, all designs in [69, 118, 119] have been re-simulated using the same CAD tools, i.e., the same technology node (45nm) and the same STT-MTJ model based on Verilog-A.

5.6.1 Evaluation of Area, Power, and Delay Overheads

Performance penalties of the proposed radiation-hardening NVFF are analyzed and evaluated in terms of area overhead, efficiency of power consumption, and propagation delay to validate the superiority of the presented NVFF over the selected prior work in [119] and [118]. Table 6.3 summarizes the performance penalties and soft error robustness of the presented NVFF, normalized to the CMOS-based flip-flop design presented in [69]. The occupied area of the proposed NVFF is estimated in equivalent Unit Size Transistors (USTs) required to construct the NVFF design, adopted from [130]. Simulation results reveal that, the proposed NVFF incurs 17% and 13% of area overhead and power consumption, respectively, while the design recently presented in [118] incurs 37% of area overhead and 24% of power consumption penalty. Additionally, the proposed NVFF realizes an acceptable performance degradation (25%) in term of overall time delay ($T_{D-Q} = T_{su} + T_{CLK-Q} = T_{sense}$) and relatively associated with low leakage power (21% less than CMOS-based design). The design presented in [119] incurs the lowest area and power overheads, however, it does not provide protection for both SEU and DNU.

On the other hand, the drawback of the proposed NVFF is that it consumes high energy for write
Table 6.3: Performance penalties of the proposed NVFF circuit w.r.t. conventional CMOS-based flip-flop.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>CMOS-Based FF [69]</th>
<th>NVFF [119]</th>
<th>NVFF [118]</th>
<th>Proposed NVFF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area (USTs)</td>
<td>1</td>
<td>0.57x</td>
<td>1.37x</td>
<td>1.17x</td>
</tr>
<tr>
<td>Power Consump. (µW)</td>
<td>1</td>
<td>0.48x</td>
<td>1.24x</td>
<td>1.13x</td>
</tr>
<tr>
<td>Overall Delay (ns)</td>
<td>1</td>
<td>1.72x</td>
<td>1.78x</td>
<td>1.25x</td>
</tr>
<tr>
<td>Leakage Power (µW)</td>
<td>1</td>
<td>0.43x</td>
<td>0.91x</td>
<td>0.79x</td>
</tr>
<tr>
<td>Write Energy (fJ)</td>
<td>1</td>
<td>10 ~ 11x</td>
<td>9x</td>
<td>5 ~ 6x</td>
</tr>
<tr>
<td>Nonvolatility</td>
<td>x</td>
<td>v'</td>
<td>v'</td>
<td>v'</td>
</tr>
<tr>
<td>SEU/DNU Tolerant</td>
<td>v'/v'</td>
<td>x/x</td>
<td>v'/x</td>
<td>v'/v'</td>
</tr>
</tbody>
</table>

access operations since a high switching current ($I_{MTJ} = 100µA$) with a shot duration ($\sim 2$ ns) is utilized to change the states of the MTJs, but it is less than those in [118] and [119]. Thus, improved developments to lower the write energy and improve the reliability, scholastic behaviors, of STT switching mechanism is sought so that hybrid STT-MTJ/CMOS circuits can be directly implemented on commercial technologies [131]. The presented NVFF incurs moderate area overhead, low leakage power, and high computing performance. Thus, it can be utilized as a base component to design more complicated hybrid Spintronics/CMOS nonvolatile logic and memory circuits.
5.6.2 Evaluation of SEU and DNU Immunity

In order to quantify the robustness of the presented design against radiation-induced soft errors, Monte-Carlo simulations were carried out. A transient current source was connected to each vulnerable node, identified in Table 6.2, with a maximum pulse amplitude of 0.28 mA and 800 ps of duration. This current source was injected into a single node to emulate a SEU, whereas double current sources were connected to a double adjacent susceptible nodes to imitate the effect of charge sharing induced DNU. All these simulations were carried out via HSPICE simulations. The results of the transient simulation demonstrate that the presented NVFF is robust to achieve 100% of SEU masking, while the probability of unmasked DNU is found to be roughly less than 3.3%, as listed in Table 6.2. This vulnerability of DNU results from the fact that the hardened PCSA circuit has two vulnerable node pairs (S0 and S3; S1 and S3), and the DICE latch has a single susceptible node pair (n0 and v0), as shown in Figure 6.2. However, this vulnerability can be alleviated by isolating the device sensitive drain area occupied by these node pairs on the cell’s layout, which is beyond the scope of this work.

Meanwhile, among the designs listed in Table 6.3, the presented NVFF is the only nonvolatile latching circuit that is able to highly tolerate charge sharing induced DNU, simultaneously. It is worth noticing that each latch circuit of the proposed NVFF is able to tolerate DNU, simultaneously, which makes the presented NVFF circuit an attractive soft error resilience storage element for nanoscale device technology.

5.7 Summary

In this chapter, a soft error resilient NVFF circuit that can achieve intriguing attributes is presented. Our proposed NVFF achieves reduced area and power penalties (14.6% and 10.3%, respectively),
while offering speed performance improvement of 29.5% as compared to the prior work in [118]. In addition, it significantly reduces leakage power by 21% compared to the CMOS-based design. Thus, the proposed soft error resilient NVFF can be utilized to harden the critical or vulnerable registers against radiation-induced transient effects, including both SEU and DNU.
CHAPTER 7: CONCLUSIONS AND FUTURE WORK

This final chapter presents a summary of achievements of this dissertation and integration into state-of-the-art. In addition, the drawback(s) and inherent limitations of proposed techniques are discussed. Likewise, recommendations to improve circuit performance in terms of energy-efficiency and resilience of fault masking coverage are elaborated here. The final section presents a list of possible future work.
LIST OF REFERENCES


