Reliability Analysis and Improvement of Nanoscale CMOS Digital Circuits

By:
M.Sc. Andrés Felipe Gómez Chacón

A Dissertation submitted in partial fulfilment of the requirements for the degree of

Doctor in Electronics Sciences

At the
National Institute for Astrophysics, Optics and Electronics

Thesis Advisor:
Ph.D. Víctor Hugo Champac Vilela, INAOE

December 2017
Tonantzintla, Puebla, Mexico

@INAOE 2017
Copyright

The author hereby grants to INAOE permission to reproduce and to distribute copies of this thesis document in whole or in part
Abstract

Reliability degradation due to transistor aging is a major challenge in the design of digital integrated circuits for current and future technology nodes. Bias Temperature Instability (BTI) is the dominant aging mechanisms in digital circuits. BTI gradually increases the device’s threshold voltage ($V_{th}$) over the lifetime, which in turn degrades circuit speed, and ultimately, it may cause a faulty operation. Circuit performance degradation due to BTI is highly dependent on the operating conditions such as the executed workloads by the circuit and the operating temperature. Moreover, variability in devices parameter due to process variations aggravate the reliability issues. These effects need to be properly accounted during circuit design phase to obtain electronic products with improved lifetime.

Conventional worst-case one-time guardbans to deal with aging due to BTI are becoming too large with technology scaling, which leads to conservative designs with reduced performance. This thesis addresses BTI aging issues from two different perspectives:

1. Aging Monitoring:

   Aging sensors are introduced in the circuit critical paths (CP) in order to detect when they are close to violate timing specifications due to BTI. Then, failure-prevention actions (i.e. increase system clock period) can take place. This thesis proposes a methodology to efficiently identify, at an early design phase, the critical paths of the circuit that should be monitored for reliable on-line aging tracking. The major challenges addressed for critical path
selection are the uncertainty caused by process parameters variations, the uncertainty on spatial correlation between gates since gate placement in the layout is unknown at early design phase, and the uncertainty of the operating workload of the circuit, which is unknown in advance at the design phase.

2. Reliable Circuit Design:

Among the gates that belong to the critical paths of a circuit, some of them are more efficient to be re-sized to compensate aging degradation. This thesis proposes a circuit design methodology to reduce the guardband required for a circuit by smart selection and sizing of the circuit’s critical gates. Gate selection metrics are proposed to guide the sizing process of the circuits. One of the most important parameters taken into account by the metrics is the statistical path delay sensitivity to gate sizing, which depends on process variations, spatial correlation between gates, and the BTI degradation of the devices in the path. Classic worst-case aging assumptions are avoided by analyzing the circuit performance for various possible operating workloads and identifying a more realistic maximum degradation of the paths. In such way, area overhead due to worst-case aging assumptions is further mitigated. The proposed approach provides efficient trade-offs between gate sizing and guardband to compensate BTI aging under process variations effects.

The aforementioned proposals of this thesis have been implemented in an Aging-Aware Statistical Timing Analysis Tool written in C++ code. The benefit of the proposed methods were validated on different ISCAS circuits.

The contributions of this thesis provide a better understanding of how the aging mechanisms affect digital circuits. Moreover, new aging-aware design techniques are available to help designers to approach the complex problem of the design of reliable high-performance integrated circuits.
Acknowledgement

To the Consejo Nacional de Ciencia y Tecnología (CONACYT), for the economic support granted through the scholarship for doctorate studies No. 420129/ 264560

To the National Institute for Astrophysics, Optics and Electronics (INAOE), for giving me the opportunity to do the doctorate studies.

To my advisor, Dr. Víctor Hugo Champac Vilela, for his guidance and support in the course of my doctoral research.

To the professors:

Dr. Alejandro Díaz Sánchez
Dr. Gordana Jovanovic Dolecek
Dr. Luís Hernández Martínez
Dr. Héctor Luís Villacorta Mínaya
Dr. Roberto Gómez Fuentes

for their in-depth review and comments of this thesis.

To the staff of INAOE, for the support and help that they give to the students.
A mis padres, perseverancia es una virtud que aprendí de ustedes, y sin ella no habría culminado este trabajo.

A mi esposa, motivación para siempre seguir adelante.
Contents

Abstract iii

Acknowledgement v

List of figures xvii

List of tables xix

List of Acronyms xxi

1 Introduction 1

1.1 Semiconductor technology scaling 2

1.2 Process Parameters Variations 3

1.2.1 Systematic Variations 3

1.2.2 Non-Systematic Variations 4

1.3 Transistor aging 8

1.3.1 Hot Carriers Injection (HCI) 8

1.3.2 Time Dependent Dielectric Breakdown (TDDB) 9

1.3.3 Bias Temperature Instability 9

1.4 Reliability of digital integrated circuits 10

1.4.1 Reliability and Yield 11

1.4.2 Bathtub curve 12

1.4.3 Technology Scaling Impact on Reliability 13
# CONTENTS

1.5 Impact of aging and process variations on performance of digital integrated circuits .......................................................... 15

1.6 Justification of this thesis .......................................................... 18

1.7 Thesis organization .......................................................... 19

2 BTI Mechanism and its Impact on Digital Integrated Circuits 21

2.1 Physical Mechanisms for Bias Temperature Instability ............. 21

2.1.1 Reaction-Diffusion Model ................................................. 22

2.1.2 Trapping-Detrapping Model .............................................. 24

2.2 Long-Term BTI Predictive Model ........................................... 26

2.3 BTI Impact on Digital Circuits ............................................... 29

2.4 Aging-aware design techniques ............................................. 32

2.4.1 Design-time techniques .................................................. 32

2.4.2 Run-time techniques .................................................... 34

3 Aging-Aware Statistical Timing Analysis 37

3.1 Background .......................................................... 37

3.1.1 Basic definitions for timing analysis ................................... 38

3.1.2 Timing Analysis Methods .............................................. 39

3.2 Modeling of BTI Degradation ............................................... 44

3.2.1 Process Variations-Aware BTI Model ................................ 44

3.2.2 Workload Impact on BTI .............................................. 46

3.3 BTI-aware Statistical Timing Analysis Procedure .................... 49

3.3.1 Gate Delay Model ..................................................... 50

3.3.2 Cell Library Characterization .......................................... 53

3.3.3 Topological Paths Search ............................................. 55

3.3.4 Workload and Degradation Analysis ................................ 56

3.3.5 Path Statistical Delay Computation .................................. 59

3.3.6 Inter-Path Covariance ................................................ 61
4 Early Selection of Aging Critical Paths for Delay Degradation

4.1 Aging Monitoring Technique .................................. 64
4.2 Related work .................................................... 66
4.3 Challenges for Critical Path Selection ......................... 67
  4.3.1 Critical Path Definition .................................. 67
  4.3.2 Critical Paths dependence on process variations ......... 68
  4.3.3 Critical Paths dependence on operating workload ...... 69
4.4 Proposed Critical Paths Selection Condition .................. 71
  4.4.1 Handling Spatial Correlation Uncertainty ............... 72
  4.4.2 Handling Workload Uncertainty .......................... 80
4.5 Proposed methodology for critical path selection .......... 83
  4.5.1 Fast Path Pre-Filter ..................................... 84
  4.5.2 Potential Critical Paths Under Worst-Case Aging ...... 85
  4.5.3 Critical Path Selection Algorithm ......................... 86
4.6 Experimental Results on ISCAS Circuits ....................... 89
  4.6.1 Path Selection Results ................................... 90
  4.6.2 Validation of Spatial Correlation Approximation ....... 93
  4.6.3 Comparison with Other Works ............................. 100
4.7 Conclusions ..................................................... 101

5 Guardband Reduction by Efficient Selection and Sizing of Critical Gates

5.1 Related Work .................................................. 103
5.2 Considerations for Gate Sizing Metrics ....................... 106
  5.2.1 Sizing Impact on Gate Delay .............................. 106
  5.2.2 Sizing Impact on Path Delay .............................. 107
5.3 Proposed Gate Sizing Metrics ................................ 108
  5.3.1 Statistical Path Delay Sensitivity ......................... 108
<table>
<thead>
<tr>
<th>Section Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.3.2 Proposed Sizing Metrics</td>
<td>113</td>
</tr>
<tr>
<td>5.4 Methodology for Efficient Selection and Sizing of Critical Gates</td>
<td>118</td>
</tr>
<tr>
<td>5.4.1 PCP Identification Under Worst-Case Aging</td>
<td>119</td>
</tr>
<tr>
<td>5.4.2 Multiple Workloads Aging Analysis</td>
<td>119</td>
</tr>
<tr>
<td>5.4.3 Selection and Sizing of Critical Gates</td>
<td>122</td>
</tr>
<tr>
<td>5.5 Results from Application to ISCAS Benchmark Circuits</td>
<td>125</td>
</tr>
<tr>
<td>5.5.1 Sensitivity Approximation</td>
<td>125</td>
</tr>
<tr>
<td>5.5.2 Optimization Results</td>
<td>126</td>
</tr>
<tr>
<td>5.5.3 Benefit of the Multiple Workload-Aware Aging Analysis</td>
<td>128</td>
</tr>
<tr>
<td>5.5.4 Gate Sizing Optimization Comparison</td>
<td>132</td>
</tr>
<tr>
<td>5.6 Conclusions</td>
<td>134</td>
</tr>
<tr>
<td>6 Conclusions</td>
<td>137</td>
</tr>
</tbody>
</table>
List of Figures

1.1 Technology scaling following the behavior predicted by Moore’s law. The number of transistors per processor and the chip’s operating frequency increase. .................................................. 2
1.2 Optical Proximity Correction. Taken from [1]. ....................... 4
1.3 Classification of Spatially Correlated Process Variations. ........ 5
1.4 Spatial correlation behavior as function of distance [2]. .......... 6
1.5 Number of dopant for various technology nodes [3]. .............. 7
1.6 Hot Carriers phenomenon. ................................................. 9
1.7 Oxide breakdown due to conducting path through the gate oxide to substrate [4]. .......................................................... 10
1.8 Physical mechanisms for BTI. ............................................. 11
1.9 Consumer Electronics and Safety Critical Applications. Reliability is of great interest in Safety Critical Applications (Pictures from google images). ................................................. 12
1.10 Failure rate represented by the bathtub curve. ..................... 13
1.11 Gate Electric field increase vs. technology nodes [5]. ............ 14
1.12 Power consumption increase with technology scaling [6]. ........ 14
1.13 Total parameter variation vs. technology year [7]. ............... 15
1.14 Impact of device parameters shift at different levels of abstraction. 16
1.15 Impact of aging and process variations on the delay of the logic gate. Top: Transient behavior. Bottom: Histogram of the delay samples. .................................................. 16
1.16 Digital System and clock waveform ........................................... 17

2.1 Reaction-Diffusion BTI Model ......................................................... 23
2.2 Trapping-Detrapping BTI Model ................................................... 25
2.3 Long-term predictive model ......................................................... 26
2.4 Threshold voltage degradation dependence with different parameters. 28
2.5 Delay degradation of a path due to BTI ......................................... 30
2.6 Circuit delay degradation due to BTI for different ISCAS Circuits. 31
2.7 Critical path reordering due to different aging rates [8] .................. 32
2.8 Aging effects may cause a functional failure ............................... 32

3.1 Basic definitions for timing analysis ............................................. 38
3.2 Arrival times propagation in block-based timing analysis approach. 40
3.3 Propagation of gate delays through two paths ............................... 40
3.4 Deterministic Static Timing Analysis. Gates take a fixed delay ......... 41
3.5 Process Corners for STA .............................................................. 42
3.6 Computed delay of a 20 inverter chain path for each Monte Carlo trial. The obtained histogram is also shown. ....................... 43
3.7 Statistical Static Timing Analysis. Gates delays become a random variable ................................................................. 44
3.8 Threshold voltage shift due to BTI as function of operational time for different initial $V_{th}$ due to process variations ....................... 45
3.9 Histogram of the threshold voltage distribution for a fresh and aged device .................................................................................. 46
3.10 Workload applied to a circuit ...................................................... 47
3.11 Workload impact on delay ........................................................... 48
3.12 Workload as a Signal Probability ................................................... 49
3.13 Aging-aware path-based timing analysis procedure. 50

3.14 Behavior of the delay of an inverter gate as a function of important parameters. a) Gate size, b) Load capacitance, c)Input slew rate, d) Temperature 52

3.15 Gate delay as a function threshold voltage degradation. 53

3.16 Characterization process of an inverter gate. 54

3.17 Example of a digital circuit and its graph representation for searching the topological paths 55

3.18 Signal probability propagation rules for some basic gates. 57

3.19 Devices duty-cycle computation based on signal probabilities at gates inputs. 58

3.20 Electric model for temperature estimation [9]. 59

3.21 Two paths with structural correlation, which increases inter-path covariance. 62

4.1 Aging sensor operation for error prediction. 65

4.2 Aging Monitoring Framework 65

4.3 Critical Paths Definition Over time 68

4.4 Process variations impact on defining critical paths 69

4.5 Histogram of the delay of a PCP for various workload profiles (ISCAS c2670) 70

4.6 Statistical Delay Difference Between two Paths 71

4.7 Impact of $\sigma_{DD}$ on the selection probability of a path ($P_1 < P_2$). 75

4.8 Relationship between sum of path variances and inter-path covariances for some ISCAS circuits. 77

4.9 Analysis of the global correlation value ($\rho_g$) that maximizes the standard deviation of the delay difference between two paths. 78

4.10 Ratio of $\sigma_{DD}$ estimated with $\rho_g = 0$ to that estimated with $\rho_g = 1$ 79
4.11 Analysis of the degraded delay of two paths $i$ and $j$ in ISCAS c2670 for 1000 random-generated operating workloads. ........................................ 82
4.12 Behavior over the time of the delay distribution of two paths ($i$ and $j$). ................................................................. 83
4.13 Diagram of the proposed methodology ........................................ 84
4.14 Fast Path pre-filtering. ............................................................ 85
4.15 Selected critical paths vs. evaluated workload profiles ..................... 92
4.16 Impact of probability threshold on the CP set size and CPU time ($WL = 60$). ................................................................. 94
4.17 Approximated standard deviation of the delay difference between all the paths in the circuit and the LCP of the circuit Vs. the standard deviation of the delay difference obtained using spatial correlation extracted from circuit layout. ........................................ 95
4.18 Histograms of the ratio of of the estimated standard deviation of the delay difference using correlation approaches and the one obtained using spatial correlation information extracted from circuits layout. 97
4.19 Overestimation of critical paths selected using correlation bounds with respect to the critical paths selected using the proposed correlation approximation. ........................................ 98
4.20 Histogram of the probability to become critical of the missed paths of all ISCAS circuits. .................................................. 99
5.1 Nominal delay, Delay degradation, and Standard Deviation of the delay of an inverter gate as function of its size. ....................... 107
5.2 Impact of sizing a gate on the adjacent gates ................................. 108
5.3 Example of the magnitude of the components of the statistical path delay sensitivity to gate sizing (Equation 5.4). ...................... 110
5.4 A path example to illustrate the impact of sizing a gate on its neighboring gates in the path. ............................................... 111
5.5 Layout of an inverter gate. ............................. 114
5.6 Areas of some basic digital gates as function of gate size ...... 115
5.7 Circuit example to illustrate gate criticality .......................... 116
5.8 Flow of the proposed gate sizing optimization methodology .......... 118
5.9 Maximum delay degradation of some paths as function of the number of tested workload profiles. ............................ 121
5.10 Fast and Slow PCP sets ........................................... 123
5.11 Statistical delay sensitivity of a Path with respect to sizing each gate in the path. Longest path of C1908 circuit. .................. 126
5.12 Histograms of the $\mu + 3\sigma$ corner of the circuit aged delay obtained for an exhaustive number (20000) of multiple signal probability groups at main circuit inputs. .......................... 130
5.13 Number of iterations to achieve target guardband. ................. 134
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Aging-Aware Design Techniques</td>
<td>33</td>
</tr>
<tr>
<td>4.1</td>
<td>Main characteristics of the considered ISCAS circuits</td>
<td>90</td>
</tr>
<tr>
<td>4.2</td>
<td>Selected paths using the proposed methodology ($\varepsilon = 0.25%$)</td>
<td>92</td>
</tr>
<tr>
<td>4.3</td>
<td>Comparison of CPs selected using spatial correlation against those selected using spatial correlation extracted from circuits layout ($\varepsilon = 0.25%$)</td>
<td>98</td>
</tr>
<tr>
<td>4.4</td>
<td>Selected nodes to be monitored</td>
<td>100</td>
</tr>
<tr>
<td>4.5</td>
<td>Methodology Comparison</td>
<td>101</td>
</tr>
<tr>
<td>5.1</td>
<td>Results using multiple workloads for aging analysis</td>
<td>127</td>
</tr>
<tr>
<td>5.2</td>
<td>Results using a single workload for aging analysis</td>
<td>129</td>
</tr>
<tr>
<td>5.3</td>
<td>Results using worst-case aging analysis</td>
<td>130</td>
</tr>
<tr>
<td>5.4</td>
<td>Percentage of WLs for which 10% of guardband may be violated</td>
<td>131</td>
</tr>
<tr>
<td>5.5</td>
<td>Percentage of area overhead for target guardbands of 10% and 20% of using three selection and sizing methods</td>
<td>133</td>
</tr>
</tbody>
</table>
List of Acronyms

NBTI  Negative Bias Temperature Instability
PBTI  Positive Bias Temperature Instability
HCI   Hot Carriers Injection
TDDBD Time Dependent Dielectric Break-down
PDF   Probability Density Function
LTDM  Long-Term Degradation Model
RDF   Random Dopant Fluctuation
LCP   Longet Critical Path
PCP   Potential Critical Paths
SSTA  Statistical Static Timing Analysis
DSTA  Deterministic Static Timing Analysis
RD    Reaction-Diffusion
TD    Trapping/De-Trapping
DOE   Design Of Experiments
SP    Signal Probability
Chapter 1

Introduction

The relentless scaling of semiconductor technology has provided a steady increase in the performance of digital integrated circuits. However, reliability degradation caused by aging-induced and process-induced device’s parameters shifts have become a major concern in deeply scaled technologies, affecting the reliability of modern circuits. This chapter reviews the main reliability issues concerned to semiconductor technology scaling. The chapter is organized as follows: Section 1.1 presents the semiconductor technology scaling trends to achieve higher performance. Section 1.2 reviews the principal process variations issues presented in current and future technology nodes. Section 1.3 reviews the principal transistor aging mechanisms. Section 1.4 highlights the importance of chip reliability analysis and verification in safety-critical applications. Section 1.5 describes how these technology issues affect the reliability and performance of the digital integrated circuits. Section 1.6 provides a brief justification of the research work presented in this thesis. Finally, Section 1.7 gives a brief description of the organization of this thesis book.
1.1 Semiconductor technology scaling

During the past years, the development of complex electronics products, each time with better performance than the previous generation, has relied on device’s dimensions scaling. Smaller devices are capable to switch faster, which allow digital circuits to perform logic operations in a reduced amount of time. Moreover, higher integration density and lower power consumption are obtained [10]. As illustrated in Figure 1.1, both the number of transistors per processor and the chip’s operating frequency have increased over the years with technology scaling. Until now, the nanometric and sub-nanometric technology regime has been reached, as predicted by Moore’s Law [11]. An important implication of technology scaling is that the cost associated with the integrated circuit’s production is reduced, improving the semiconductor industry profit.

![Figure 1.1: Technology scaling following the behavior predicted by Moore’s law. The number of transistors per processor and the chip’s operating frequency increase.](image)

Unfortunately, as the technological and physical limits of the devices are reached, new design challenges have emerged. In deeply scaled technologies, device’s parameters may shift from the intended values at design due to fabrication process-induced variations and device aging [12,13]. Even these small variations on the intended device’s parameters can result in a significant change in circuit per-
1.2. PROCESS PARAMETERS VARIATIONS

Performance (i.e., delay) [14], which can make a circuit to violate specifications. Even more, circuits that meet design specifications right after manufacturing may eventually violate them due to aging effects. Assurance of lifetime reliability of modern digital integrated circuits under the combined effect of process variations and aging has become a major concern for designers. Therefore, an in-depth analysis of the impact of these mechanisms on circuit performance and reliability is mandatory in modern and future technology nodes. In such way, new design techniques can be developed to assure correct functionality of modern high-performance and complex circuits.

1.2 Process Parameters Variations

Process parameters variations are caused by the randomness of manufacturing process due to imprecise process control, as well as the intrinsic variability of some fabrication steps (i.e., channel doping) [3,15]. Process variations are of static nature, which means that induced device parameters shifts remain constant over the lifetime. Process variations are classified into systematic and non-systematic variations.

1.2.1 Systematic Variations

Systematic variations are design-dependent variations (i.e., depend on layout characteristics), which are mainly presented in interconnections [16]. Systematic variations make the width, length, and shape of fabricated interconnections to differ from those intended at the design phase. A principal source of systematic variation is the inability to scale the wavelength of the light source for lithography, which has led to optical proximity effects [17]. Figure 1.2 (top) shows an example of the design target for a given interconnection, its corresponding mask used for lithography and the printed shape on the wafer. Detailed analysis of the layout
can be performed to account systematic variations at pre-manufacturing [18]. A resolution-enhancement technique called Optical Proximity Correction (OPC) is based on pre-distorting mask patterns to make printed shapes on wafer appear closer to the targeted shapes, as shown in Figure 1.2 (bottom) [1].

![Figure 1.2: Optical Proximity Correction. Taken from [1].](image)

### 1.2.2 Non-Systematic Variations

Non-systematic variations, also known as random variations, correspond to the truly uncertain component of physical parameter variations. Non-systematic variations are modeled as individual random variables for each device in a circuit. These random variables are usually assumed to be normal-distributed [18, 19]. Non-systematic variations themselves can be classified into spatially correlated variations (Die-to-Die and Within-Die) and independent (Pure Random) variations.

**Spatially Correlated Variations**

Spatially correlated variations are those parameter variations that tend to shift similarly between devices. These variations mainly affect the channel length (L), the channel width (W) and oxide thickness ($T_{ox}$) of the fabricated devices [19,20].

Spatially correlated variations arise at different fabrications steps, which define the spatial scale of the variation [21], as shown in Figure 1.3. Discrepancies
in the alignment of the wafers when processing one lot of wafers to the next may lead to Lot-To-Lot variations, which affect all the devices within the same wafer lot similarly. When processing a given wafer, discrepancies between the different masks used during photo-lithography may lead to Wafer-to-Wafer variations. Furthermore, small variations in the laser intensity and exposure time while a particular mask is scanned may lead to Reticle-to-Reticle Variations. At a smaller scale, variations across-mask may lead to within-die variations, which means that devices in the same die may have different critical dimensions [18]. As technology scales down, within-die variations have become a major contributor to the total parameters variations [7,19].

Spatially correlated variations cause that the parameter’s shifts of a device in a certain location will change with some degree of correlation with the parameter’s shifts of another device located in a different position. The amount of correlation between two devices has been shown to be a strong decaying function of the distance between them [2,22]. Figure 1.4 illustrates this phenomenon, where two devices closely located will be highly correlated, and two devices that are far away from each other will be poorly correlated. It is important to highlight that even two devices placed far away within a die will exhibit some degree of correlation due to variations from Reticle-to-Reticle to Fab-to-Fab scale, which impact nearly equal to all the devices within a die.

A model for robust extraction of spatial correlation between two devices $i$ and $j$ from the circuit layout was proposed in [2],

![Figure 1.3: Classification of Spatially Correlated Process Variations.](image-url)
1.2. PROCESS PARAMETERS VARIATIONS

\[ \rho_{ij} = \exp(-d_{ij}/CD) \]  

where \( \rho_{ij} \) is the degree of spatial correlation between the gates, \( d_{ij} \) is the distance in the layout between gates \( i \) and \( j \), and \( CD \) is the correlation distance, which is a constant that modulates the decaying rate of the correlation, and it has been shown to be roughly half of the chip width [22].

**Independent Variations**

Independent variations are those non-exhibiting any degree of correlation between devices. These variations lead to devices with different performance even when placed near to each other in the layout.

A major source of independent variations is the Random Dopant Fluctuation (RDF) of the atom impurities doping the channel region of a transistor. Small variations in the implant dose, energy, or angle of the ion beam lead to fluctuations in the discrete location of dopant atoms in the channel and source/drain regions. These variations mainly impact on the threshold voltage (\( V_{th} \)) of the devices [7]. Since the number of dopant atoms reduces as technology scales down (See Figure

Figure 1.4: Spatial correlation behavior as function of distance [2].
1.2. PROCESS PARAMETERS VARIATIONS

1.5), RDF has become of great concern in modern technology nodes [3].

![Figure 1.5: Number of dopant for various technology nodes [3].](image)

Threshold voltage variation due to RDF ($\sigma_{V_{th,RDF}}$) has been shown to be inversely proportional to the transistor area [23],

$$\sigma_{V_{th,RDF}} = \sigma_{V_{th0}} \cdot \sqrt{\frac{A_{VT}}{W_{eff}L_{eff}}} \quad (1.2)$$

where $A_{VT}$ is a technology constant dependent on the critical dimensions, $W_{eff}$ and $L_{eff}$ are the effective channel width and length, respectively, and $\sigma_{V_{th0}}$ is the threshold voltage variation of the technology.

**Modeling of Process Parameter Variations**

Due to the process variations, the physical and electrical parameters of each device have to be modeled as a random variable. Variations in channel length, channel width, oxide thickness and threshold voltage, have been shown to be the dominant contributors to device performance variations [19,20]. Due to the different sources of variations at Die-to-Die, Within-Die, and Independent scale, each parameter $P$ can be represented as the sum of each contribution to total parameter variation:
where $P_{\text{nom}}$ is the nominal parameter value at design, and $\Delta P_{D2D}$, $\Delta P_{WID}$ and $\Delta P_R$ are random variables representing the parameter variations due to Die-to-Die, Within-Die, and Independent process variations sources. These random variables are assumed Gaussian-distributed [18].

1.3 Transistor aging

Together with process variations, transistor’s aging has become a major concern in deeply scaled semiconductor technologies [24]. Unlike process variations that are static in nature, transistors aging mechanisms cause a gradual deterioration of the electrical parameters of the devices over the lifetime, which in turn degrades the performance of the circuit [25–27]. Aging mechanisms reduce the profit of semiconductor industry as they strongly impact on the lifetime reliability of the integrated circuits. The most important transistor aging mechanisms are described below.

1.3.1 Hot Carriers Injection (HCI)

HCI mechanism affects both NMOS and PMOS transistors. HCI occurs when the electric field at the drain-to-channel depletion region of a transistor is too high given kinetic energy to free electrons. Carriers are accelerated until they have enough energy to overcome the potential barrier of the $Si/SiO_2$ interface and leave the channel. As is shown on Figure 1.6 some of those hot carriers damage the gate oxide and the interface or get trapped in the oxide and form space charges. Both mechanisms lead to a degradation of the transistor characteristics [28]. The rest of the carriers contributes to the gate or bulk current. The end result of HCI is a degradation of transistor’s parameters such as effective mobility and
threshold voltage \[16\]. It is important to note that device’s damage due to HCI is not recoverable, hence it results in monotone transistor aging \[29\].

![Figure 1.6: Hot Carriers phenomenon.](image)

\[1.3.2 \ \text{Time Dependent Dielectric Breakdown (TDDB)}\]

TDDB, also known as gate oxide breakdown, is a failure mechanism caused by the long stress time of the \(Si - O_2\) at high electric fields. As illustrated in Figure 1.7, the oxide breakdown occurs when a conducting path through the gate oxide to substrate if formed due to electron tunneling current \[4\]. In addition to thin gate oxides, the saturating trend in supply voltage scaling results in a large electric field applied to the gate oxide, which increases the probability of dielectric breakdown.

\[1.3.3 \ \text{Bias Temperature Instability}\]

Bias Temperature Instability (BTI) is the dominant aging mechanism in digital circuits, affecting both NMOS and PMOS transistors biased at specific stress conditions: positive gate bias for NMOS (PBTI) and negative gate bias for PMOS (NBTI) transistors \[30\]. In previous technology nodes, the PBTI effect on NMOS transistors was negligible in comparison to the effect of NBTI. However, since the
introduction of high-k metal-gate technologies, PBTI effect has become a critical aging concern for NMOS devices [31]. At stress conditions, devices threshold voltage ($V_{th}$) is increased due to BTI. Two physical mechanisms have been identified as contributors to BTI degradation (See Figure 1.8):

- **Interface Traps**: Traps are generated at the $Si - SiO_2$ interface due to the breaking of weak $Si - H$ bonds under the presence of a high vertical electric field. The generated traps at the silicon-oxide interface increase the device’s threshold voltage with a time power law relation ($t^n$) [32] [33] [34].

- **Bulk oxide Traps**: There exist a number of oxide defect states. Due to tunneling through the pre-existing defects within the oxide, a defect (trap) can capture a charge carrier from the channel, which increases the threshold voltage [30,34].

### 1.4 Reliability of digital integrated circuits

Reliability is defined as the ability of a circuit to perform its required function under stated constraints (i.e., operating temperature, voltage, etc.) for a specified
1.4. RELIABILITY OF DIGITAL INTEGRATED CIRCUITS

Figure 1.8: Physical mechanisms for BTI.

lifetime [35]. The required degree of product reliability is application dependent. Figure 1.9 shows two different kinds of electronic applications. In the top: Consumer Electronics, which are electronic equipment intended for everyday use, most often in entertainment, communications and office productivity. Reliability of those electronic products is not very critical because their use is determined by a short time range and they can fail without causing major losses. In the bottom: Safety-Critical Applications, which are those applications whose failure could result in loss of life, significant property damage, or damage to the environment. Reliability is a great concern for safety-critical applications. There are many well-known examples in application areas such as medical devices, aircraft flight control, weapons, and nuclear systems. Furthermore, many modern information systems are becoming safety-critical in a general sense because their failure may result in a large financial loss or even loss of life [36].

1.4.1 Reliability and Yield

Reliability and Yield are two important parameters for the development of a new technology node. Yield can be defined as the probability of failure of a recently manufactured device \((t = 0)\), while reliability can be seen as the probability of a functional failure over an operating lifetime \((t > 0)\) of the device. A fabrication
1.4. RELIABILITY OF DIGITAL INTEGRATED CIRCUITS

process with low yield is unacceptable to start large-scale production, but even a process with high yield and low reliability is not economically sustainable at the long-term because many chips will fail to intended specification before the expected lifetime ends. Moreover, for safety-critical applications, circuits with a low degree of reliability lead to catastrophic results.

1.4.2 **Bathtub curve**

The *Bathtub Curve* (See Figure 1.10) illustrates a measurement of the number of failures per unit of time (failure rate) vs the operating time [37]. The failures that can occur in a digital circuit are classified according to the point in the lifetime that these failures occur. *Early Failures*, also known as Infant Failures correspond to those devices that quickly fails during operational lifetime. These failures are presented due to manufacturing defects, consequently, their occurrence reduces as the manufacturing process reaches maturity. In addition, most of these defects can be detected during the post-manufacture test. Therefore, most of the circuits with early failure defects do not reach the customer. *Random Failures* correspond to those failures presented during normal lifetime and are attributed to random external events or early aging. The failure rate during a normal lifetime is usually very small, leading to correct circuit functionality. *Aging Failures* are failures caused by the physical deterioration, or aging, of the devices. The time at
which these failures become dominant depends on technology parameters and the circuit’s operating conditions such as supply voltage, temperature, and workload.

![Bathtub Curve Diagram](image)

Figure 1.10: Failure rate represented by the bathtub curve.

### 1.4.3 Technology Scaling Impact on Reliability

The lifetime reliability of electronic products reduces with technology scaling, because transistor aging becomes more significant. One reason behind the increased impact of aging mechanisms is that constant field scaling [38] has not been achieved because it requires that threshold voltage scales proportionally to the feature size reduction. However, threshold voltage scaling is limited by the off-state current of the devices. Therefore, in order to achieve high-performance specification, supply voltage has not been scaled as aggressively as the devices feature size [39]. Another reason why supply voltage is not further reduced is to keep noise immunity under conservative levels. Hence, area scaling without proper supply voltage scaling results in higher electric fields across the transistor’s thin gate oxide, as illustrated in Figure 1.11.

Higher power density is also a major contributor to accelerated circuit aging in scaled technologies. As illustrated in Figure 1.12, the power consumption has been increasing with technology scaling in spite of the reduced supply voltage. This is due to the larger device density on a chip. However, high power densities
manifest in elevated operating temperatures along the chip, which further triggers the aging mechanisms [14].

A paralleling trend with technology scaling is that process parameters variations increase due imperfections in the manufacturing process of the ICs [40]. Figure 1.13 illustrates the total variations of some transistor parameters for different technology years. As can be seen, the relative percent of parameters unpre-
dictability is becoming significant with technology scaling. Furthermore, process variations impact with aging mechanisms and aggravate the lifetime reliability issue [41–43].

Figure 1.13: Total parameter variation vs. technology year [7].

1.5 Impact of aging and process variations on performance of digital integrated circuits

Figure 1.14 illustrates the impact of device’s parameters shifts on circuit performance for different levels of abstraction. Variations on device level parameters, which are caused by process variations and aging effects, result in a significant impact on electrical properties of devices (i.e., On-state current), which in turn affect circuit level parameters such as delay.

Due to the joint effect of aging and process variations, the delay of logic gates become highly uncertain. Figure 1.15 illustrates the transient behavior for a high to low output transition of a gate cell under both process variations and aging effects. As can be seen, the time at which the output signal crosses $0.5V$ ($\approx V_{DD}/2$) changes significantly due to process variations. Moreover, as the devices
1.5. IMPACT OF AGING AND PROCESS VARIATIONS ON PERFORMANCE OF DIGITAL INTEGRATED CIRCUITS

Figure 1.14: Impact of device parameters shift at different levels of abstraction.

deteriorate due to aging mechanisms (i.e., BTI), the time at which the transition occurs increases. As can be seen in the histograms, both the mean value and the shape (variance) of the delay distribution change due to aging [43,44].

Figure 1.15: Impact of aging and process variations on the delay of the logic gate. Top: Transient behavior. Bottom: Histogram of the delay samples.

A digital system is based on the structure shown in Figure 1.16(a). A combinational logic block receives various input signals, which are processed and prop-
agated through the logic gates to the outputs, where the resulting data is stored in memory elements (Flip-Flops). This task is performed in a synchronous way based on the clock signal. The period of the clock signal defines the available time to process the input logic data and propagate logic signals to the memory elements. The long-term effect of aging mechanisms and with process variations may result in the incapability to process the logic signals within the clock period. This occurs when the summation of all gate delays along a given path is larger than the clock period. Therefore, an incorrect logic data is stored in the Flip-Flops.

![Digital System and clock waveform](image)

Figure 1.16: Digital System and clock waveform

**Guardband for reliable operation**

As illustrated in Figure 1.16(b), a simple approach to assure safe circuit’s operation is to add a time guardband to the clock period. This guardband provides extra time for signal propagation through the critical signal paths having large delays [45,46]. The guardband is used not only to cover variations caused by aging and process, but also variations in the supply voltage and operating temperature.
of the system, among others. The size of the guardband depends on the impact of each variation source on circuit timing [47], and it is usually estimated based on worst-case assumptions (i.e., worst process, voltage, temperature, and aging) to guarantee circuit’s correct functionality [48]. However, this means that maximum circuit’s performance is reduced. Unfortunately, since aging and process variations are aggravated with technology scaling, the required guardbands are becoming too large, limiting the benefits of technology scaling.

1.6 Justification of this thesis

Operational failures that may arise because of aging and process variations effects are unacceptable for safety-critical applications. Hence, the development of reliability-aware circuit design techniques is mandatory for current and future technology nodes. This thesis addresses aging issues from two different perspectives:

1. Aging Monitoring:

Aging sensors are introduced in the circuit critical paths (CP) in order to detect when they are close to violate timing specifications due to BTI. Then, failure-prevention actions (i.e. increase system clock period) can take place. This thesis proposes a methodology to efficiently identify, at an early design phase, the critical paths of the circuit that should be monitored for reliable on-line aging tracking. The major challenges addressed for critical path selection are the uncertainty caused by process parameters variations, the uncertainty on spatial correlation between gates since gate placement in the layout is unknown at early design phase, and the uncertainty of the operating workload of the circuit, which is unknown in advance at the design phase.

2. Reliable Circuit Design:
Among the gates that belong to the critical paths of a circuit, some of them are more efficient to be re-sized to compensate aging degradation.

This thesis proposes a circuit design methodology to reduce the guardband required for a circuit by smart selection an sizing of the circuit’s critical gates. Gate selection metrics are proposed to guide the sizing process of the circuits. One of the most important parameter taken into account by the metrics is the statistical path delay sensitivity to gate sizing, which depends on process variations, spatial correlation between gates and the BTI degradation of the devices in the path. Classic worst-case aging assumptions are avoided by analyzing the circuit performance for various possible operating workloads and identifying a more realistic maximum degradation of the paths. In such way, area overhead due to worst-case aging assumptions is further mitigated. The proposed approach provides efficient trade-offs between gate sizing and guardband to compensate BTI aging under process variations effects.

The contributions of this thesis will provide a better understanding of how the aging mechanisms affect digital circuits. Moreover, new aging-aware design techniques will be available to help designers to approach the complex problem of the design of reliable circuits with high quality and yield.

1.7 Thesis organization

This thesis document is organized as follows: Chapter 1 has presented a brief introduction of the reliability issues in nanoscale technologies and the justification of this thesis. Chapter 2 describes the Bias Temperature Instability mechanism and its modeling for long-term reliability analysis of integrated circuits. The impact of BTI at circuit level and state-of-art aging-aware techniques are also analyzed. Chapter 3 presents the framework for BTI-Aware Statistical Static
Timing Analysis, which is used to simulate and analyze digital circuits under the joint effect of aging and process variations. Chapter 4 presents the proposed methodology to identify, at an early design stage, the critical paths due to BTI and process variations of a digital circuit. The results from ISCAS circuits are also given. Chapter 5 presents the proposed methodology for reliable circuit design using gate selection and sizing metrics. Finally, Chapter 6 summarizes the general conclusions of this work.
Chapter 2

BTI Mechanism and its Impact on Digital Integrated Circuits

Among aging mechanisms, Bias Temperature Instability (BTI) is well known to be a major lifetime reliability limiting factor. This Chapter presents a detailed description and analysis of the BTI physical mechanism and its impact on digital circuits performance. Furthermore, a review of the state-of-art circuit design techniques to deal with transistor reliability issues is given.

This chapter is organized as follows: Section 2.1 explains the physical mechanisms behind BTI degradation. Section 2.2 presents a widely used long-term model to predict the BTI-induced $V_{th}$ shift of a device within a circuit. Section 2.3 describes the impact that BTI aging has on circuit level performance of digital circuits. Section 2.4 describes the principal BTI-aware circuit design techniques.

2.1 Physical Mechanisms for Bias Temperature Instability

Bias Temperature Instability is an aging mechanism that affects both PMOS and NMOS transistors. The BTI effect on PMOS, called Negative-BTI (NBTI), occurs
when a Negative gate to source voltage is applied to the device. NBTI is considered the major reliability issue before the 45nm technology node. However, the BTI effect on NMOS device, called Positive-BTI (PBTI), has become important since the introduction of the high-k metal gate dielectric in sub-45nm technologies [49]. PBTI occurs when a positive gate to source voltage is applied to the device. At stress condition, trap generation at silicon-oxide interface and carriers trapping at oxide defects are two widely accepted BTI mechanisms responsible for the observed device’s threshold voltage ($V_{th}$) increase over the lifetime, resulting in a reduction of current drive capability [50]. An accurate estimation of the $V_{th}$ degradation that devices experience is critical in the design of reliable digital circuits. The BTI mechanism has been explained using two physics-based models: The Reaction-Diffusion Model and The Trapping-Detrapping Model [30].

2.1.1 Reaction-Diffusion Model

The reaction-diffusion (RD) model is one of the earliest approaches to explain BTI mechanism. It explains that traps are generated at the silicon-oxide interface due to the dissociation of weak $S_i - H$ bonds under the high vertical electric fields and elevated temperatures [33] [51].

Figure 2.1 illustrates the RD mechanism for BTI. In the silicon wafer, each $S_i$ atom is bonded with four other $S_i$ atoms, forming a tetrahedral structure, where all the states of the valence band of each $S_i$ atom are filled. At the silicon-oxide interface, most of the $S_i$ atoms are bonded with oxygen. However, due to structural imperfections, some $S_i$ atoms may remain unbounded. Those unsatisfied (dangling) bonds of the $S_i$ atom are called interface traps [52]. The interface traps become charged when occupied by holes or electrons, hence contributing to threshold voltage shift. In order to mitigate the number of interface traps, the $S_i - S_iO_2$ interface is passivated with hydrogen atoms during fabrication hence some weak $S_i - H$ bonds are formed at the interface, as shown in Figure 2.1.
However, when a gate voltage is applied (BTI stress), some weak $S_i - H$ bonds may break (Reaction) due to the high vertical electric field. The dissociation of $S_i - H$ bonds increases at elevated temperatures [33]. The released $H$ atoms then diffuse into the gate oxide (Diffusion) in the form of different $H - species$ ($H$ or $H_2$). Then, the unsatisfied dangling bond at the interface can capture a charge carrier from the channel via field-assisted tunneling (See Figure 2.1). This reduces the total carriers that flow through the channel, which can be seen as an increase in the threshold voltage of the device. When the stress is removed ($V_{gs} = 0$), the device experience a partial recovery, when hydrogen species diffuse back to the interface and passivize some of the dangling bonds. The recovery phase has an important impact on the estimation of BTI degradation in the AC operation of digital circuits, where devices are continuously switching between stress and recovery BTI phases. However, this recovery is partial, and the overall BTI effect tends to gradually $V_{th}$.

The overall rate of interface traps generation in the RD model can be described by the following equation [33]:

![Figure 2.1: Reaction-Diffusion BTI Model.](image-url)
2.1. PHYSICAL MECHANISMS FOR BIAS TEMPERATURE INSTABILITY

\[
\frac{\partial N_{it}}{\partial t} = k_F \cdot (N_o - N_{it}) - k_R N_H(0) N_{it}
\]  

(2.1)

where \(N_o\) is the initial number of \(Si - H\) bonds, \(N_{it}\) is the number of interface traps, \(k_F\) is the breaking rate constant of \(Si - H\) bonds, and it depends on the operating temperature, \(k_R\) is the recovery rate constant when H species diffuse back to the interface and it also depends on the operating temperature, \(N_H(0)\) is the number of hydrogen atoms at the \(Si - SiO_2\) interface.

2.1.2 Trapping-Detrapping Model

The Trapping-Detrapping (TD) model explains BTI issue due to the capturing of individual charge carriers at the gate oxide defects, which in turn changes the effective threshold voltage of the devices [53], [54]. This mechanism is also used to explain random-telegraph noise (RTN) effects.

Figure 2.2 illustrates the TD mechanism for BTI. It is assumed that some pre-existing traps (called oxide vacancies) are located in the dielectric of the device. These defects are presented due to the amorphous nature of the \(SiO_2\) [52]. When BTI stress is presented, trap energy (relative to the Fermi energy level) is modulated. If the trap gains sufficient energy, it may capture a charge carrier from the channel (via tunneling). A charge trapping event changes the number of carriers available to conduct current and also affects the mobility of channel charge carriers [34,55]. This leads to discrete steps in the threshold voltage of the device and the current flowing through the channel. When stress is removed, the trapped charge may be emitted after sufficient recovery time.

Trapping and detrapping events have been shown to be stochastic in nature [56]. The probability of trapping and de-trapping a charge depends on capture and emission time constants, which are strongly dependent on stress voltage and temperature. The probabilities of defect occupancy in case of capture (\(P_C\)) and
2.1. PHYSICAL MECHANISMS FOR BIAS TEMPERATURE INSTABILITY

![Figure 2.2: Trapping-Detrapping BTI Model.](image)

The traps that contribute to transistor deterioration due to BTI are those whose capture time constant is much larger than the emission time constant. Thus, the rate at which charge carriers are captured abruptly becomes larger than the rate at which carriers are emitted, and the number of trapped charge increases over the lifetime \[34\].

The effectiveness of the reaction-diffusion model and the trapping-detrapping model for prediction of the performance degradation of logic gates were investigated in \[57\]. It was concluded that both models match well to obtain the BTI
2.2 Long-Term BTI Predictive Model

The previous section presented two physical mechanisms to explain the BTI issue in semiconductor devices. However, detailed physical simulation of BTI is not practical for long-term reliability analysis of very large scale integrated circuits, where millions of devices are presented. Furthermore, each device may experience a different operating temperature and stress-recovery patterns (stress duty cycle), depending on the workload applied to the circuit, which further complicates physical simulations. Long-term predictive models aim to address the aforementioned issue by estimating the degradation of a device given the expected operating conditions. As illustrated in Figure 2.3, instead of performing detailed cycle-to-cycle stress and recovery phase simulation, a long-term model provides a tight upper bound for the degradation on the transistor.

The firsts long-term predictive models [58, 59] were based on the Reaction-Diffusion physics. These models express BTI degradation as a power-law function
of time ($\Delta V_{th} \sim t^n$). Recently, long-term BTI models based on the Trapping-Detrapping physics have been investigated [34,60], where BTI degradation is expressed as a logarithmic function of time ($\Delta V_{th} \sim \log(1 + C \cdot t)$).

Power-law models, based on the reaction-diffusion mechanism are still widely accepted in the literature [61–64]. The following closed form equation to calculate BTI-induced $V_{th}$ degradation was used [64],

$$\Delta V_{th} \approx K \cdot t_{ox} \cdot \sqrt{C_{ox} \cdot (V_{GS} - V_{TH0}) \cdot \exp(E_{ox}/E_{0}) \cdot \exp(-E_a/kT) \cdot \alpha^n \cdot t^n} \quad (2.3)$$

where $n$ is the time exponent, $t_{ox}$ is the gate oxide thickness, $E_{ox}$ is the vertical electric field, $T$ is the temperature, $k$ is the Boltzmann constant, $C_{ox}$ is the oxide capacitance per unit of area, $V_{TH0}$ is the initial (fresh) threshold voltage value, $E_a$ and $E_0$ are constants. The parameter $\alpha$ is the duty cycle, which refers to the percentage of time the device is in stress condition. The duty cycle of a device depends on the executed workload by the circuit. The duty-cycle solution of the long-term model avoids detailed cycle-to-cycle simulations and allows a fast estimation of device’s degradation. The constant parameter $K$ is a technology-dependent fitted constant, which may take a different value for for NBTI and PBTI.

The closed form of Equation 2.3 determines the threshold voltage degradation $\Delta V_{th}$ for a given time while taking into account the impact of important parameters. Figure 2.4a-d shows the behavior of both NMOS and PMOS threshold voltage degradation with different parameters obtained with the long-term predictive model (Equation 2.3) for a 32nm CMOS technology [65]. Figure 2.4(a) shows the dependence with stress time. As can be seen, at the beginning of the lifetime operation, BTI causes a sharp increase in the threshold voltage. This occurs because there is a high density of weak $Si-H$ bonds that can be broken. As more $H$ atoms are removed from the interface over the lifetime, the increase
2.2. LONG-TERM BTI PREDICTIVE MODEL

Figure 2.4: Threshold voltage degradation dependence with different parameters.

rate of Vth reduces. Note that this power-law behavior ($\Delta V_{th} \propto t^n$) is reported in the literature for the reaction-diffusion model [32]. Figure 2.4(b) shows the impact of the stress probability (duty cycle). Stress probability is defined as the percentage of time the device is under stress condition [66, 67]. A large stress probability means that most of the time, the transistor is at BTI stress, with little recovery time, which results in a larger degradation. On the other hand, a low stress probability indicates that the device experience large relaxation periods, which mitigates the total Vth deterioration. It is important to note that the specific stress probability experienced by the devices in a circuit depends on the input pattern applied to the circuit (i.e., the workload), which defines the signal probabilities (SP) at main circuit inputs. Depending on the circuit usage, these signal probabilities are translated into a stress probability for the devices at internal gates. Figure 2.4(c) shows that temperature has a significant impact
2.3 BTI Impact on Digital Circuits

The increment on $V_{th}$ reduces the logic gates current capability to charge or discharge capacitive loads. A reduction in the current capability increases the delay of the gates to propagate the logic functions in a digital circuit. The aged delay of a gate ($D_{age}$) can be computed once the $V_{th}$ shift of each of its transistors is predicted. The aged delay of a gate ($D_{age}$) is composed of the sum of the nominal gate delay ($D_n$) and the gate delay degradation caused by BTI ($\Delta D_{BTI}$) aging, as shown in Equation 2.4.

$$D_{age} = D_n + \Delta D_{BTI}(t)$$ (2.4)

As shown in Figure 2.5, each gate in a path may have a different degradation depending on its operating conditions such as temperature and the duty cycle of the devices in it. The total path delay degradation is given by the delay
2.3. BTI IMPACT ON DIGITAL CIRCUITS

degradation contribution of each gate. As circuit ages, the gate delays increase, and consequently, it is possible that a path fails timing specifications, although the specifications were fulfilled right after manufacturing. Therefore, BTI aging limits the lifetime of digital circuits. The paths that may cause such faulty behavior are usually referred as Critical Paths.

Figure 2.5: Delay degradation of a path due to BTI.

Figure 2.6 shows the delay degradation over the lifetime of two digital circuits. The degradation of each circuit may be quite different depending on the set of stress probabilities taken by the devices (represented $\alpha_1$, $\alpha_2$, and $\alpha_3$). If the input pattern applied to the circuit causes a large stress probability on critical devices, the delay degradation rate of the circuit increases. This behavior is of great concern in hardware security due to potential aging attacks [70]. Also, delay degradation of the circuits is different because of circuit’s topology characteristics, such as logic depth, and node arrangement of the gates, which impact on the stress probability of the devices [71, 72]. Several works in the literature show that the percentage of delay degradation caused by BTI can be around 10% to 30%, and its impact is aggravated in further scaled technologies [30].

The BTI impact on delay can be more pronounced in some gates than others because the specific stress probability that the devices experience during their operational time is non-uniform. Moreover, some gates may be more sensitive to the Vth shift due to BTI. Consequently, the delay degradation rate of two signal paths can also be quite different. This effect can lead to eventually reordering
Figure 2.6: Circuit delay degradation due to BTI for different ISCAS Circuits.

of the circuit critical paths as time progresses [8]. Figure 2.7(b) shows the path reordering behavior for the small circuit shown in Figure 2.7(a). Although the delay at output 9 is initially the slowest of the circuit, the delay at output 10 eventually become the slowest one after some aging time. This is different to classic timing analysis with no aging, where the paths of a circuit have a fixed timing order over time. However, under BTI effects, the original critical path may become non-critical and vice versa. This makes complex the critical path definition under aging effects. Furthermore, this effect may become more pronounced under process variations.

As shown in Figure 2.8, if BTI-induced delay degradation is significant, a circuit may fail to time specification before the end of its expected lifetime, even if the specification was fulfilled right after manufacturing [73]. Due to BTI, severely aged paths may perform signal transitions after the clock edged (delay specification), where the results of logic functions are stored in flip-flops. Consequently, an incorrect logic value is stored as the result of the logic functions. Such failures are unacceptable for safety-critical applications. Therefore, lifetime reliability analysis and improvement is required for the design of reliable, high-performance electronic systems.
2.4 Aging-aware design techniques

Design techniques have been explored in the past to deal with circuit delay degradation due to BTI aging [74, 75]. Table 2.1 shows an overview of the main aging-aware design techniques. They can be categorized in Design-Time techniques and Run-Time Techniques.

2.4.1 Design-time techniques

Design-time techniques are those that take place only once during circuit design phase. A simple design-time technique to account aging and process variations
Table 2.1: Aging-Aware Design Techniques

<table>
<thead>
<tr>
<th>Category</th>
<th>Technique</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design-Time</td>
<td>Guardbanding</td>
<td>[45] [76] [46] [77]</td>
</tr>
<tr>
<td></td>
<td>Synthesis/Library Optimization</td>
<td>[78] [79] [80] [81] [82] [75]</td>
</tr>
<tr>
<td></td>
<td>Logic Restructuring/Pin Re-Ordering</td>
<td>[71] [72] [83]</td>
</tr>
<tr>
<td></td>
<td>Dual $V_{DD}/V_{th}$ Assignment</td>
<td>[84] [85] [86]</td>
</tr>
<tr>
<td></td>
<td>Gate Sizing</td>
<td>[87] [88] [89] [90] [91] [92] [93] [42]</td>
</tr>
<tr>
<td>Run-Time</td>
<td>Input Vector Control</td>
<td>[94] [95] [96] [97]</td>
</tr>
<tr>
<td></td>
<td>Internal Node Control</td>
<td>[98] [99] [100] [101] [102]</td>
</tr>
<tr>
<td></td>
<td>Power Gating</td>
<td>[103] [104] [105]</td>
</tr>
<tr>
<td></td>
<td>Task Distribution</td>
<td>[106]</td>
</tr>
<tr>
<td></td>
<td>Dynamic Voltage/Frequency/$V_{th}$ Scaling</td>
<td>[107] [108] [109] [110] [111] [112] [113]</td>
</tr>
<tr>
<td></td>
<td>Error Prediction/Detection</td>
<td>[114] [115] [116] [117]</td>
</tr>
</tbody>
</table>

Guardbanding. As explained in Chapter 1 (See Figure 1.16(b)), a large extra time is usually added to the clock period to provide enough time for correct propagation of digital signals through the paths experiencing severe aging and process variations. However, this approach significantly penalizes performance as the impact of both BTI and process are increasing with technology scaling. An alternative guardbanding approach is to use greater than nominal supply voltage [77]. However, this increases both the power consumption and the degradation rate of the circuit.

High-level aging-aware synthesis perform circuit optimization to balance the post-aged delay of the paths instead of balancing the paths delays at design time. This is shown to be beneficial to extend the circuit lifetime under a given guarband [81]. Aging affected gates are replaced by a robust counterpart obtained from an aging-aware cell library. However, the design of that gate library can be very complex since gates must be designed to address different workloads.

Logic restructuring and node reordering techniques manipulate the stress probabilities of some influential devices to the deterioration of the critical paths of the circuit. To reduce the stress probability of critical devices, logic restructuring modifies the arrangement of super gates. A super gate is a group of connected
gates that is logically equivalent to a big AND/OR gate. By swapping some of
the gates, the delay deterioration of the super gate is mitigated due to reduced
stress probability of the critical devices [71]. Node reordering technique reduces
the stress probability of the devices within a gate by using the stacking effect of
the transistors: The stress probability of the transistor in the top of the stack
depends on the stress probability of the transistors in the bottom of the stack.
Therefore, the stress probability of the overall stack can be reduced by properly
chosen the input signal that drives each device in the stack. Although these tech-
niques are of low cost in terms of area and power, their effectiveness may not be
sufficient to compensate aging.

Dual $V_{DD}/V_{th}$ assignment aims the co-optimization of power consumption and
aging. A high (low) $V_{DD}$ ($V_{th}$) is assigned to the gates (transistors) in the critical
paths. On the other hand, for devices (gates) in non-critical paths, a low $V_{DD}$
($V_{th}$) is assigned to save power. A drawback of using multiple supply voltages is
the need of voltage level converters to connect gates in different $V_{DD}$ regime.

Gate sizing is one of the most effective approaches to compensate aging and
process variations. The current drive capability of the gates increases for wider
transistors at the cost of area and power overhead. Most of the gate sizing ap-
proaches are based either on formulating an optimization problem and solving it
using an exact optimization method, such as Lagrange Method. Other approaches
use heuristics to guide the optimization problem, leading to close to optimum re-
results with lower computational time than exact optimization methods.

2.4.2 Run-time techniques

Run-time techniques are those techniques that are dynamically applied to the
circuit over the lifetime operation.

Input Vector Control (IVC) and Internal Node Control (INC) are techniques
that take place only during circuit standby mode. When the circuit is under
normal operation conditions, these techniques are not applied. At standby, the input vector that minimizes the aging and static current consumption is applied to the circuit. The input vector that minimizes aging is the one that puts out of BTI stress those devices in the critical paths. Since the control-ability of internal nodes by main inputs is limited, the internal node control technique adds some logic at internal nodes to be able to manipulate the stress state of some critical transistors. Recently, a rejuvenation approach that applies the best stimuli that sets the critical devices at recovery state has been proposed in [97]. These rejuvenation stimuli are applied during small time intervals (not standby). One of the advantages of IVC and INC techniques is that the circuit can change between standby and normal operation modes quickly. However, when the duration of the standby mode is large enough, Power Gating technique, which disconnects the logic gates from the $V_{DD}/GND$ rails, may be more effective to minimize power and aging.

In multiprocessors systems, it is possible to perform task distribution to every single core dynamically. The work in [106] proposed an algorithm that assigns the task requiring more resources, hence more critical for performance, to those cores that remains with low aging while easy tasks are assigned to the most aging affected cores. The effect is to balance the degradation rate of the entire system.

Dynamic Voltage Scaling (DVS), Dynamic Frequency Scaling (DFS) and Adaptive Body Bias (ABB), are three widely known run-time techniques to address aging issues. DVS uses a low supply voltage level when the circuit is fresh, so that power consumption is reduced. As the circuit ages, the supply voltage level is increased to keep reliable operation. Similarly, DFS uses a higher frequency when the circuit is fresh, and it gradually reduces the operating frequency as circuit ages. In such way, close to best performance is achieved along the lifetime. ABB technique uses the body effect to control the threshold voltage of the device. As the devices age due to BTI, the body bias voltage is increased to compensate the performance deterioration. However, ABB technique is no longer efficient in
2.4. AGING-AWARE DESIGN TECHNIQUES

deeply scaled technologies due to the increased impact of variation in the threshold voltage. Moreover, in emerging technologies using FinFET devices, the body terminal is removed. Hence ABB is no longer applicable in those technologies.

Error prediction and error detection are interesting techniques for monitoring the aging condition of the circuit. Aging sensors are introduced at the end-points of the circuit critical paths to keep under surveillance their delay degradation. When the circuit delay degradation is larger than a given threshold, the sensors detect this condition and generate a warning signal. In the error prediction technique, the warning signal is generated before a faulty operation really occurs. Thus the delay of the critical paths is still lower than the clock period, and correct data is stored in the memory elements (i.e., Flip-Flops). In the Error detection scheme, the warning signal is generated when a faulty operation has occurred (the logic signal was not propagated to the memory elements under the given clock period). Therefore, error recovery strategies are needed. Error prediction/detection are usually combined with other run-time approaches such as task distribution and dynamic frequency-voltage scaling [110], so that each time an error prediction/detection occurs, self-healing mechanisms such as increasing supply voltage or reducing operating frequency take place. In order to reduce the complexity and overhead of delay monitoring, an accurate selection of those paths that really require being monitored is needed.
Chapter 3

Aging-Aware Statistical Timing Analysis

As discussed in previous chapters, temporal transistor degradation caused by reliability mechanisms such as NBTI/PBTI has a great impact on circuit timing performance. This chapter provides a detailed explanation of the process of computing circuit delay considering aging effects.

This chapter is organized as follows: Section 3.1 presents a background of the principal concepts for timing analysis. Section 3.2 presents the modeling of BTI degradation considering process variations effects and the impact of circuit workload. Finally, Section 3.3 provides a detailed explanation of the procedure for computing circuit delay using Statistical Static Timing Analysis and having into account the effects of aging due to BTI.

3.1 Background

Timing analysis is a very important step during circuit design. It is used mainly for timing verification to make sure that the circuit design meets timing requirements. Timing analysis is also important for circuit optimization, since it allows to identify circuit segments requiring timing improvement.
3.1. BACKGROUND

3.1.1 Basic definitions for timing analysis

The timing performance of a digital gate can be measured in terms of the delay that takes to propagate a digital signal from its inputs to its output. Figure 3.1(a) shows an example of a simple NAND gate experiencing a signal transition at one of its input nodes (node X). Note that signal propagation through the gate occurs only if the other input (node Y) is at the logic value that sensitizes the gate (logic 1 for a NAND gate). Figure 3.1(b) illustrates the waveforms at gate input X and gate output Z. Some basic definitions are as follows:

- $D$ is the delay of the gate.
- $SRI$ is the slew-rate of the input signal.
- $SRO$ is the slew-rate of the output signal.

The delay of the gate ($D$) is measured between the 50% transition point of the switching input and output waveforms. The slew-rate of the input ($SRI$) and output ($SRO$) signals measures how quickly the signal switches from one logic value to the other. Both $D$ and $SRO$ are important timing metrics of a gate and have been widely used for timing analysis of an entire circuit [118].

Figure 3.1: Basic definitions for timing analysis.
3.1.2 Timing Analysis Methods

Timing analysis methods estimate the delay of a circuit in a faster approach than performing circuit level electrical simulations based on SPICE engine. Timing Analysis can be classified into two different categories depending on the circuit traversal algorithm used: Block Based Timing Analysis and Path-Based Timing Analysis. Furthermore, depending on how the delay information is evaluated, timing analysis can be based on Deterministic Timing Analysis, Monte Carlo-Based or Statistical Timing Analysis.

Timing Analysis Categories

**Block-Based Timing Analysis** The goal of block-based timing analysis is to compute the maximum delay that takes to propagate a signal from circuit primary inputs to every gate output node. This delay is called the *Arrival Time* (AT) of a node. In this method, the gates of the circuit are divided into blocks depending on their logic level and the arrival times are propagated in a level-based order, from primary inputs to primary outputs [119]. Figure 3.2 illustrates the propagation of arrival times for a two-input NAND gate using the MAX operator. As shown, the arrival time at the output of a gate corresponds to the maximum of the inputs arrival times, plus the delay through the gate [120]. In order to compute the delay of the subsequent gates, the input transition slews are also propagated in a similar way.

One of the main advantages of block-based approaches is that computational time is linear with respect to the circuit size [121]. This is because arrival times are propagated only once for each gate. However, the use of MAX operator makes pessimistic the estimation of timing quantities, which may lead to over-design.

**Path-Based Timing Analysis** The goal of path-based timing analysis is to compute the specific delay of each signal path in the circuit. A signal path is any
Figure 3.2: Arrival times propagation in block-based timing analysis approach.

possible trajectory that a signal may traverse from a primary input to a primary output. Figure 3.3 illustrates the computation of the delay of two paths: Path 1 is composed of gates G0, G1, and G2, and path 2 is composed of gates G1 and G2. In this case, the path delay is computed using the operator SUM, which adds the delay contributions of each gate in the path trajectory. The delay of gates with multiple inputs is evaluated assuming that the non-switching inputs (also called path side-inputs) are tied at the logic value that sensitizes the gate.

Figure 3.3: Propagation of gate delays through two paths.

The main advantage of path-based timing analysis is its high accuracy. Another advantage is that it makes easier the identification of paths having large delays since the topology and delay information of every path is available. Path-based timing analysis is more computationally costly than Block-Based timing analysis. This is because the delay of a gate is re-computed for each path passing through the gate [7].

For reliability analysis and improvement, a high accuracy on computed delays is desired since an over-estimation of the circuit delay may lead to significant
over-design. Therefore, the path-based approach is used in this thesis for timing analysis.

**Timing Evaluation Methods**

The main methods to evaluate the delay of a circuit are: Deterministic Static Timing Analysis, Monte Carlo Analysis, and Statistical Static Timing Analysis.

**Deterministic Static Timing Analysis** Deterministic Static Timing Analysis (DSTA) computes circuit delay assuming that devices process parameters receive a fixed value [18], so that gate delays are deterministic (See Figure 3.4).

![Figure 3.4: Deterministic Static Timing Analysis. Gates take a fixed delay.](image-url)

In order to account for the impact of process variations on circuit timing using STA, a Corner-based analysis is usually adopted. The corner-based analysis relies on the assumption that if the circuit operates correctly under the most pessimistic conditions, then it will work correctly for all other process conditions [7]. Figure 3.5 illustrates the feasible corners of a fabrication process. The process corners are defined based on the current capability that devices take. Typically, Slow-Slow (SS), Typical-Typical (TT) and Fast-Fast (FF) process corners are of great concern. In the TT process corner, the parameters are assumed to take their nominal values. In the SS corner, the parameters shifts of all the devices in the circuit are chosen to minimize the device’s current capability while current capability is maximized in the FF corner. In order to compute delays, the gate
library needs to be characterized for the corresponding process conditions on each corner.

![Diagram of process corners]

**Figure 3.5: Process Corners for STA**

The corner-based method is computational inexpensive since only one timing computation run is performed for each corner. However, designing circuits based on process corners leads to significant over-design since the process corners are too pessimistic because the statistical behavior of correlated and independent variations is not considered.

**Monte Carlo Analysis** Monte Carlo analysis method is a brute-force approach to compute the statistical delay of a circuit. It consists on an exhaustive computation of the circuit delay for several circuit’s samples resulting from assigning a value to the process parameters of each device, which are modeled as a normal-distributed random-variable:

\[
W = N(\mu_W, \sigma_W^2) \quad L = N(\mu_L, \sigma_L^2) \quad T_{ox} = N(\mu_{T_{ox}}, \sigma_{T_{ox}}^2) \quad V_{th} = N(\mu_{V_{th}}, \sigma_{V_{th}}^2)
\]  

(3.1)

At each sample, each device process parameter receives a different value, which is generated according to the corresponding probability distribution function.
Then, DSTA is run to obtain the delay of the paths. This process is repeated several times to obtain a large enough number of samples for the desired accuracy of the statistical results.

Figure 3.6 shows an example of the computed delay for a 20 inverter gate chain at each Monte Carlo trial. The mean value and standard deviation of the path delay are obtained from the histogram distribution for all the computed delays. The larger the number of trials that are performed, the higher the accuracy of the results [7].

Monte Carlo method provides a very accurate prediction of circuit yield and reliability. However, the major drawback is its computational inefficiency since a large number DSTA runs are required for high accuracy.

**Statistical Static Timing Analysis**  Statistical Static Timing Analysis (SSTA) mathematically considers the probability distributions of the process parameters during delay computation, which in turn makes the gate delays to become a random variable, and gives a distribution of possible circuit’s delay outcomes rather than a single outcome. In such way, the statistical delay information of the paths (mean value and standard deviation) is obtained on a single delay computation run.
3.2. MODELING OF BTI DEGRADATION

As shown in Figure 3.7, in SSTA method, the statistical sum of the gate delay distributions is performed. The mathematical treatment of the statistical sum of the random variables representing each gate is usually based on simplified gate delay models (i.e., linear or second-order models) accounting for the impact of the process parameters.

\[
D_0 + D_1 + D_2 = D_{Path}
\]

Figure 3.7: Statistical Static Timing Analysis. Gates delays become a random variable.

Statistical Static Timing Analysis is less computationally costly than Monte Carlo method, while it keeps the statistical behavior of the process parameters into account. Therefore, SSTA is convenient for delay computation of large circuits.

3.2 Modeling of BTI Degradation

3.2.1 Process Variations-Aware BTI Model

Process variations have a strong influence on the degradation due to BTI that devices experience [43]. As can be observed in Equation 2.3, \(\Delta V_{th_{BTI}}\) depends on the initial value of \(V_{th} (V_{th0})\) of the device, which is impacted by process variations. Figure 3.8 illustrates the shift in threshold voltage due to BTI (\(V_{th_{BTI}}\)) as a function of lifetime for different process variations in the initial \(V_{th} (\Delta V_{th_{PV}})\). Data for this plot has been obtained using the long-term prediction model of Equation 2.3. As can be observed, positive (negative) values on \(\Delta V_{th_{PV}}\) produce a lower (higher) degradation rate. Thus, BTI-induced \(V_{th}\) degradation must be modeled as a random variable.
Figure 3.8: Threshold voltage shift due to BTI as function of operational time for different initial $V_{th}$ due to process variations

The impact of process variations in the long-term degradation of $V_{th}$ can be accounted by a first-order Taylor approximation of Equation 2.3 with respect to variations [122],

$$\Delta V_{th\_BTI} = (1 + S_v \cdot \Delta V_{th\_PV}) \cdot A \cdot \alpha^n \cdot t^n$$

(3.2)

where $\Delta V_{th\_PV}$ is the initial shift in $V_{th}$ due to process variations, and $A$, and $S_v$ are fitted constants that depend on technology node, temperature, and electric field, and $\alpha$ is the stress duty-cycle of the transistor.

The total variation in $V_{th}$ of a transistor $m$ is obtained by adding the shifts contributions due to BTI and Process variations ($\Delta V_{th\_BTI} + \Delta V_{th\_PV}$).

$$\Delta V_{th\_m} = \Delta V_{th\_BTI\_m} + \Delta V_{th\_PV\_m}$$

$$= (1 + S_{V,m} \cdot \Delta V_{th\_PV\_m}) \cdot A_m \cdot \alpha_m^n \cdot t^n + \Delta V_{th\_PV\_m}$$

(3.3)

$$= A_m \cdot \alpha_m^n \cdot t^n + (1 + S_{V,m} \cdot A_m \cdot \alpha_m^n \cdot t^n) \cdot \Delta V_{th\_PV\_m}$$

By analyzing Equation 3.3, it can be observed that the first term gives the BTI
impact on the mean value of the threshold voltage of the device, and the second term shows that BTI interacts with process variations through the Aging Factor (AF). The aging factor defines the BTI impact on threshold voltage variability. At $t = 0$ the total variation in $V_{th}$ is due to process variation only. However, as time increases aging effects cause changes in both the mean and standard deviation of $V_{th}$ [122]. This behavior can be appreciated in Figure 3.9, where a histogram of the threshold voltage distribution for a fresh and aged device is shown.

![Figure 3.9: Histogram of the threshold voltage distribution for a fresh and aged device.](image)

### 3.2.2 Workload Impact on BTI

A workload corresponds to the set of input vectors (0’s and 1’s) applied to the main circuit inputs while executing a given task or program, as illustrated in Figure 3.10. Each input vector (set of bits at main inputs) is usually evaluated in one clock period. When a logic vector is evaluated, the internal nodes of the circuit (n1 and n2) switch to set a logic value.

The operating workload of a circuit has a strong impact on the BTI-induced threshold voltage degradation of the devices since it influences the effective time
3.2. MODELING OF BTI DEGRADATION

the devices experience BTI stress and the operating temperature [123, 124], an consequently it impacts on the circuit delay degradation (See Figure 3.11). Depending on the input vector being evaluated, a set of devices enter in the stress condition of BTI for the current clock period. As the input vector alternates, the devices also alternate between the stress and the relaxation phase due to BTI. Therefore, the sequence of input logic vectors (workload) impacts on the stress time that each device experiences, and consequently, it defines the effective stress duty-cycle of each device. The operating temperature depends on the power dissipation of the circuit, which is mainly due to dynamic power consumption and leakage power consumption of the digital gates. A logic gate whose output node is constantly switching from one logic state to the other (due to the input vectors being applied) consumes a higher dynamic power. The leakage power of a gate depends on the logic state of its input nodes.

Logic Simulation is an approach for estimating the impact of the workload on the aging degradation that a circuit will experience. A large number of input vectors are applied to the circuit and evaluated at the logic level. Then, it is counted how many times each transistor was settled at stress condition, i.e., Positive Gate-to-Source bias for NMOS and Negative Gate-to-Source bias for PMOS. The duty-cycle of a device is approximated as the ratio of the number of times the device was at BTI stress condition to the total number of vectors. The operating temperature can also be estimated based on the average power consumption over

![Figure 3.10: Workload applied to a circuit.](image-url)
3.2. MODELING OF BTI DEGRADATION

Figure 3.11: Workload impact on delay.

the applied input vectors. However, performing such analysis for large-scale digital circuits is unpractical [66] since the exact sequence of input vectors applied to a circuit over the lifetime is very large.

A more efficient approach is to represent the workload as a set of Signal Probabilities applied at main circuit inputs:

\[ WL = \{SP_{n1}, SP_{n2}, ..., SP_{ni}\} \]  \hspace{1cm} (3.4)

where \(SP_{ni}\) corresponds to the signal probability at input node \(i\).

As shown in Figure 3.12, the signal probability is an estimation of the average time, in percentage, an input node is at logic 1 for a given workload (\(WL\)) being executed over the lifetime. One of the advantages of using signal probability representation of the workload is that the exact sequence pattern of the logic inputs is not needed.

In order to estimate the degradation of the devices, the signal probabilities are propagated to internal nodes using probability rules. Then, the duty-cycle and the operating temperature of the devices are obtained from the signal probabilities at internal nodes. Once duty-cycle and operating temperature are obtained, the BTI-
induced $V_{th}$ shift of each device can be computed using the long-term predictive model of Equation 3.3.

![Diagram of workload as a signal probability](image)

Figure 3.12: Workload as a Signal Probability.

Degradation analysis is usually performed assuming a representative workload of the circuit, which can be estimated if information of a typical bit pattern trace is available [89]. However, the specific workload that a circuit will experience over the lifetime is rarely known in advance at the design phase, which complicates reliability analysis and causes significant over-design due to worst-case assumptions. This issue will be addressed in the proposed design techniques in this thesis.

### 3.3 BTI-aware Statistical Timing Analysis Procedure

Traditional Timing Analysis does not consider the temporal degradation caused by aging mechanisms such as NBTI and PBTI. However, as technology continues to shrink, aging analysis has become a mandatory topic during IC design and verification to guarantee that timing constraints are met not only right after manufacturing but also over the circuit’s lifetime. This section provides a detailed explanation of the procedure of computing the delay of a circuit while accounting the effects of aging and process variations. As a result, the statistical timing behavior over the lifetime of the signal paths in a circuit is obtained.
An overview of the Aging-Aware timing analysis flow used in this work is presented in Figure 3.13. The process consists of four main steps. First, a pre-computing step, where the gate library is characterized via accurate SPICE simulations, takes place. Then, polynomials fitting the extracted data are obtained to allow a fast evaluation of gates delays using a gate delay model. Second, a signal path search step is performed to find all the topological paths that are in the circuit. Third, a workload and degradation analysis step is carried out. In this step, devices duty-cycles and operating temperature are estimated. With this information, $V_{th}$ degradation of devices is computed. Finally, the statistical delay of the paths is computed. As a result, the circuit timing performance over time is obtained.

![Diagram of Aging-Aware Timing Analysis Procedure](image)

Figure 3.13: Aging-aware path-based timing analysis procedure.

A more detailed explanation of each of the steps in the Aging-Aware SSTA flow is provided next.

### 3.3.1 Gate Delay Model

In order to estimate the delay of a circuit, timing analysis methods require a gate delay model capturing the dependence of gate delay with respect to important parameters such as the size of the gate ($K$), loading capacitance ($CL$), input slew-rate ($SRI$), and operating temperature ($T$). Furthermore, the model must account for the impact of parameters variations caused by process variations and
aging degradation ($\Delta P$). The delay of a gate can be represented as a function of those parameters:

$$D = f(K, CL, SRI, T, \Delta P)$$  \hspace{1cm} (3.5)

**Behavior of the delay of a gate**

The behavior of the gate delay with respect to important parameters can be observed in Figures 3.14(a)-3.14(d). The data for these Figures were obtained for an inverter gate using SPICE electrical simulations. Figure 3.14(a) shows that delay reduces when gate size increases. This is because the current capability of the gate increases as transistors are made bigger ($I_d \sim W$). It should be noted that the amount of delay reduction for larger gate sizes tend to reduce because the delay does not only depends on the gate current capability but also on its loading capacitance, which increases with gate size due to self-loading parasitics. Figure 3.14(b) shows that delay is closely a linear function of the gate loading capacitance. The behavior of gate delay with respect to the slew-rate at the switching input is shown in Figure 3.14(c). When the signal transition from a logic value to another takes a longer time, the delay of the gate also becomes longer. Figure 3.14(d) shows the impact of the operating temperature on gate delay. As can be observed, gate delay slightly increases when higher operating temperatures are presented.

**Statistical Linear Gate Delay Model**

For Aging-aware SSTA, the delay of each gate in the library is modeled as a linear combination of the random variables representing parameters shifts due to both process variations and aging effects [18, 122],

$$D = D_n + S_W^D \Delta W + S_L^D \Delta L + S_{t_{ox}}^D \Delta t_{ox} + \sum_{m}^{M} S_{Vth,m}^D \Delta Vth_m$$  \hspace{1cm} (3.6)
where $D_n$ is the nominal gate delay, $S_{W}^{D}$, $S_{L}^{D}$, $S_{tox}^{D}$ and $S_{Vth}^{D}$ are the gate delay sensitivities with respect to deviations in W, L, tox, and Vth, respectively. $M$ is the number of transistors in the gate. $\Delta Vth$ is composed of two deviation components, one due to aging effects and the other due to process variations (See Equation 3.3). As commented before, the delay of the gate is function of important parameters such as the size of the gate, loading capacitance, input slew-rate, and operating temperature (See Equation 3.5)

Equation 3.6 allows obtaining the statistical delay of a path by performing the statistical SUM operation using addition rules of linear combinations of random variables. The SUM operation will be explained later. The linear model is adequate for small variations as computational complexity remains low and the error due to discarded higher order terms can be neglected [18]. This is generally the case of intra-chip variations where the process parameter variations are relatively

Figure 3.14: Behavior of the delay of an inverter gate as a function of important parameters. a) Gate size, b) Load capacitance, c)Input slew rate, d) Temperature
small in comparison with the nominal values, and they can be approximated as normally distributed random variables [125].

The sensitivity coefficient of the gate delay to variations in a parameter $P$ is defined as the partial derivative of the gate delay with respect to changes in the parameter $P$ (See Equation 3.7). This derivative is evaluated at the nominal value of the parameter.

$$S_P^D = \frac{\partial D(K, CL, SRI, T)}{\partial P}$$

(3.7)

The behavior of the delay of an inverter gate as a function of shifts in $V_{th}$ is plotted in Figure 3.15, where it is evident a closely linear dependence with respect to $V_{th}$. Therefore, the linear model is well suited to account for such variations.

![Figure 3.15: Gate delay as a function threshold voltage degradation.](image)

3.3.2 Cell Library Characterization

In order to quickly and accurately estimate the statistical delay of a gate during timing analysis, every single gate in the technology library is pre-characterized with SPICE electrical simulations. Figure 3.16 illustrates the characterization process of an inverter gate having a low to high transition at its input. SPICE
simulations are run to measure the nominal delay and the delay sensitivities to device parameters (i.e. channel width, channel length, oxide thickness and threshold voltage) at various operating conditions given by combinations of input transition time ($SRIN$), the gate size ($K$), load capacitance ($CL$), and the operating Temperature ($Te$). The parameters combinations are obtained using full-factorial Design of Experiments (DOE), where a number of levels are assigned to each parameter. The larger the number of levels, the greater the resolution of the characterization process. In this work, 9, 16, 10, 5 levels were used for $SRIN$, $K$, $CL$, and $Te$, respectively. This process is repeated for each input transition type (low-to-high and high-to-low) and for each input node of each gate in the library. After SPICE simulations, tables with timing information for each gate are obtained, as shown in Figure 3.16. Then, polynomials that fit the extracted data are obtained as a function of $SRIN$, $K$, $CL$ and $Te$. Thus a mathematical expression if obtained to model the delay of each gate. The generated polynomials capture the timing behavior of the gates and are used for fast estimation of the statistical gate delays (using Equation 3.6) with close-to-spice accuracy.

<table>
<thead>
<tr>
<th>$K$</th>
<th>$CL$</th>
<th>$SRIN$</th>
<th>$T$</th>
<th>$D$</th>
<th>$SRD$</th>
<th>$S^2_{RD}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.2FF</td>
<td>20ps</td>
<td>25°C</td>
<td>10ps</td>
<td>12ps</td>
<td>50</td>
</tr>
<tr>
<td>2</td>
<td>0.4FF</td>
<td>20ps</td>
<td>50°C</td>
<td>15ps</td>
<td>16ps</td>
<td>60</td>
</tr>
<tr>
<td>4</td>
<td>0.6FF</td>
<td>30ps</td>
<td>100°C</td>
<td>20ps</td>
<td>19ps</td>
<td>70</td>
</tr>
<tr>
<td>6</td>
<td>0.8FF</td>
<td>50ps</td>
<td>25°C</td>
<td>25ps</td>
<td>23ps</td>
<td>80</td>
</tr>
</tbody>
</table>

$D = f(K, CL, SRI, T)$

$SRD = f(K, CL, SRI, T)$

$S^2_{RD} = f(K, CL, SRI, T)$

$$S = S^2_{RD}$$

Figure 3.16: Characterization process of an inverter gate.

The pre-characterization process is done for all the basic static cells (Inverter, 2-4 input NAND and 2-4 input NOR). However, it can be easily extended to more complex cells. It must be noted that although the pre-characterization process may be time-consuming, it is performed only once for the entire gates library and the given technology node. Therefore, once the gate delay polynomials have been obtained, they can be used to evaluate the statistical delay of a gate in any general combinational circuit.
3.3.3 Topological Paths Search

A topological path represents a possible trajectory that a logic signal can take from a main circuit input to a main circuit output. All the possible paths that a logic signal may traverse need to be identified to verify their correct timing performance.

Given the gate-level netlist, a path search algorithm is used to find the topological paths of the circuit. The first step is to identify the preceding and subsequent gates of every gate in the circuit. In such way, a graph for the circuit structure is obtained. Figures 3.17(a) and 3.17(b) show an example of a digital circuit and its corresponding graph for searching the topological paths. Each vertex in the graph represents a node of the gate-level netlist, and the edges correspond to the logic gates that propagates a signal from one node to another. S and T nodes are the source and the sink terminals, respectively. All the primary inputs are connected to the source vertex while all the primary outputs are connected to the sink vertex.

![Digital Circuit](image1)

![Graph Representation](image2)

Figure 3.17: Example of a digital circuit and its graph representation for searching the topological paths

A *depth-first search* algorithm is used to traverse the graph from a principal output to a principal input. The path search is performed as follows:

1. Initialize and start the path structure with an output node.
2. Select and visit a non-marked adjacent vertex. Add the edge (gate) and vertex (node) to the path structure.

3. Repeat step 2 until a main input is reached. Mark the latest edge.

4. If all the edges arriving to a vertex are marked: a) Mark the vertex and its latest traversed output edge; b) Unmark the arriving edges to the vertex.

5. Repeat steps 1-5 until all the nodes connected to the output node (vertex) are marked.

6. Select other output node and repeat steps 1-6.

The result of the path search algorithm is a structure with information of the gates and nodes that compose each path in the circuit.

3.3.4 Workload and Degradation Analysis

The objective of this step is to estimate the stress duty-cycle and operating temperature that the devices experience given an operating workload profile for the circuit. In such way, an accurate prediction of the devices degradation can be made.

Signal Probability Propagation

In order to compute devices duty-cycle, the signal probabilities at main circuit inputs (representing the operating workload of the circuit) are propagated to the internal nodes first. Signal probability propagation is performed using basic probability rules as illustrated in Figure 3.18. For an inverter gate, the signal probability at its output node is equal to the probability of its input node to be at logic zero. The formula to propagate the signal probabilities for other gates can be easily derived based on its truth table (See Figure 3.18). For simplicity
3.3. BTI-AWARE STATISTICAL TIMING ANALYSIS PROCEDURE

purposes, the signal correlation that arises due to re-convergent fan-out is ne-
glected in this thesis. However, signal probability propagation methods taking
into account signal correlation are available in the literature [66].

\[
\begin{align*}
A & \quad Y \\
0 & \quad 1 \\
1 & \quad 0 \\
\end{align*}
\]

\[
\begin{align*}
A & \quad B & \quad Y \\
0 & \quad 0 & \quad 1 \\
0 & \quad 1 & \quad 1 \\
1 & \quad 0 & \quad 1 \\
1 & \quad 1 & \quad 0 \\
\end{align*}
\]

Figure 3.18: Signal probability propagation rules for some basic gates.

Devices Duty-cycle Computation

Duty-cycle (\(\alpha\)) of each transistor (See Equation 2.3) represents the effective per-
centage of the lifetime that the device is likely to be at BTI stress condition, i.e.,
when the absolute value of the gate-source voltage of the device is equal to the
bias voltage (\(V_{DD}\)) [8]:

\[
\alpha_m = P(|V_{gs,m}| = V_{DD}) \tag{3.8}
\]

Once signal probabilities at every internal node of the circuit are obtained,
the duty-cycle of each transistor can be computed. Figure 3.19 shows simple
equations to compute the duty-cycle (\(\alpha\)) of each transistor in some basic gates as
a function of the signal probabilities at gate inputs. It is important to note that
stacking effect due to transistors in series has been considered [59].
3.3. BTI-AWARE STATISTICAL TIMING ANALYSIS PROCEDURE

\[ \alpha = (1 - SP_A) \]
\[ \alpha = SP_A \cdot SP_B \]
\[ \alpha = SP_A \]
\[ \alpha = SP_B \]
\[ \alpha = (1 - SP_A)(1 - SP_B) \]

Figure 3.19: Devices duty-cycle computation based on signal probabilities at gates inputs.

**Temperature Estimation**

Temperature strongly influences BTI mechanism. The temperature profile of a circuit is obtained from the power consumption profile using the electric model show in Figure 3.20, which gives the following expression for the operating temperature of a gate:

\[ T_i = R_{J,i} \cdot P_i + R_{I-A} \cdot P_{total} + T_A \]  

(3.9)

where \( P_i \) is the power consumption (Static and Dynamic) of the gate \( i \), \( R_{J,i} \) is the junction to internal air heat resistance, \( P_{total} \) is the total circuit power consumption, \( R_{I-A} \) is the heat resistance from internal air to ambient, and \( T_A \) is the ambient temperature. This model assumes that devices affect one another through their influence on the internal chip temperature and allows a fast estimation of the chip thermal profile [9].

**Vth Degradation Prediction**

Once duty-cycle and operating temperature are obtained, the BTI-induced \( V_{th} \) shift of each device can be computed using the long-term predictive model of Equation 3.3.
3.3.5 Path Statistical Delay Computation

Once computed the degradation of each transistor in the circuit, the delay of each path can be obtained. The path delay is calculated for rising edge and falling edge separately. The process consists of applying a rise (or fall) edge at the path start point (i.e., main input of the circuit) and then statistically add the delay contribution of each gate in the path. For a path $p$ having $N$ gates, the statistical add operation of $N$ random variables ($D_i$) representing the delay of each gate in the path is performed:

$$D_p = D_1 + D_2 + ... + D_N = N(\mu_p, \sigma_p) \quad (3.10)$$

It should be noted that the statistical sum of random variables in SSTA is not as trivial as the simple sum of deterministic delays in DSTA. For SSTA, the add operation of gate delays requires computing the convolution of two probability distribution functions, which is computationally expensive [19]. In order to improve the speed of SSTA methods, the random variables representing each process parameter are usually assumed Gaussian-distributed [7]. Since gate delays are modeled as a linear combination of the transistors parameters variations, the path delay resulting from the addition of gates delays is also Gaussian distributed [126].

**Mean Value of Path Delay** The mean value of the path delay $\mu_p$ is the
3.3. BTI-AWARE STATISTICAL TIMING ANALYSIS PROCEDURE

simple sum of the mean value of the delay of each gate in the path.

\[ \mu_p = \sum_{i=1}^{N} \mu_{D,i} \]  

(3.11)

Using the linear gate delay model of Equation 3.6, and excluding the terms not related with process parameters variations, it can be easily obtained that the mean value of the path delay is:

\[ \mu_p = \sum_{i=1}^{N} D_{n,i} + \sum_{i=1}^{N} \sum_{m=1}^{M,i} S_{Vth,m}^{D,i} \cdot A_{m,i} \cdot \alpha_{m,i}^n \cdot t^n \]  

(3.12)

where the first summation adds the nominal delay of the gates in the path and the second summation adds the impact on the delay of the aging-induced \( Vth \) degradation of the transistors in the gates. When the circuit is young (\( t = 0 \)), the mean value of the path delay is defined by the nominal delay of the gates only. However, as operational time progresses, the mean value of the path delay increases due to the impact of devices threshold voltage degradation on the delay of each gate. The degradation rate of the path depends on the gates delay sensitivity coefficients to \( Vth \) shift.

**Variance of Path Delay** The computation of path delay variance requires to compute the covariance between each pair of gates in the path:

\[ \sigma_p^2 = \sum_{i=1}^{N} \sum_{j=1}^{N} COV(D_i, D_j) \]  

(3.13)

In order to compute the covariance between the delay of two gates, the following theorem that applies to linear combinations of random variables is used:

**Theorem 1** Given two linear combinations of random variables \( A \) and \( B \):

\[ A = a_0 \cdot X_0 + a_1 \cdot X_1 + \ldots + a_k \cdot X_k \]

\[ B = b_0 \cdot X_0 + b_1 \cdot X_1 + \ldots + b_k \cdot X_k \]  

(3.14)
where coefficients $a_i$ and $b_i$ represent the sensitivity of the function $A$ and $B$ to the $i^{th}$ random variables $X_i$ and $X_i$. The covariance between $A$ and $B$ is,

$$COV(A, B) = \sum_{i=1}^{k} \sum_{j=1}^{k} a_i \cdot b_j \cdot \rho_{ij} \cdot \sigma_{X,i} \cdot \sigma_{X,j} \quad (3.15)$$

where $\sigma_{X,i}$ and $\sigma_{X,i}$ are the standard deviation of the random variables $X_i$ and $X_j$, respectively, and $\rho_{ij}$ is the correlation between them.

By applying the the aforementioned theorem to compute the covariance between the delay of two gates modeled with the linear model of Equation 3.6 and assuming that there only exist correlation (due to spatial proximity) between variations on the same parameter type (i.e., $W$ of gate $i$ with $W$ of gate $j$), the covariance between the delay of two gates can be expressed as,

$$COV(D_i, D_j) = S_{W,i}^D \cdot S_{W,j}^D \cdot \rho_{ij} \cdot \sigma_{W,i} \cdot \sigma_{W,j} + S_{L,i}^D \cdot S_{L,j}^D \cdot \rho_{ij} \cdot \sigma_{L,i} \cdot \sigma_{L,j} + S_{tox,i}^D \cdot S_{tox,j}^D \cdot \rho_{ij} \cdot \sigma_{tox,i} \cdot \sigma_{tox,j} + \sum_{m} \sum_{k} S_{Vth,m}^D \cdot S_{Vth,k}^D \cdot \rho_{ij} \cdot AF_m \cdot \sigma_{Vth,i} \cdot AF_k \cdot \sigma_{Vth,j} \quad (3.16)$$

The spatial correlation between the two gates ($\rho_{ij}$) can be extracted from the circuit layout using the model of [2], which models spatial correlation as an exponentially decreasing function of the distance between two gates (See Equation 1.1).

### 3.3.6 Inter-Path Covariance

The inter-path covariance measures how the statistical delays of a pair of paths jointly change due to process variations. It is useful to determine whether a path can take delays larger than other.
Let’s consider two paths, a path A formed by K gates, and a path B formed by L gates. The inter-path covariance can be computed as:

\[
COV(D_{P,i}, D_{P,j}) = \sum_{i=1}^{K} \sum_{j=1}^{L} COV(D_{i,k}, D_{j,l})
\]  

(3.17)

where \(COV(D_{i,k}, D_{j,l})\) represents the covariance between the delay of the \(k^{th}\) gate in path A and the delay of the \(l^{th}\) gate in path B.

Inter-path covariance depends on how far are placed in the circuit layout the gates of one path from the gates of the other path. A special case to be considered is when the two paths share the same gate, as illustrated in Figure 3.21. In this case, the spatial correlation between the shared gates in the paths when computing \(COV(D_{i,k}, D_{j,l})\) becomes maximum (the unit). The case shown in Figure 3.21 is also known as structural (topological) correlation. Therefore, the structural correlation between paths increases the inter-path covariance.

![Figure 3.21: Two paths with structural correlation, which increases inter-path covariance.](image-url)
Chapter 4

Early Selection of Aging Critical Paths for Delay Degradation Monitoring

Delay degradation monitoring is a technique to deal with transistor aging by keeping under surveillance the delay degradation of the circuit’s critical paths. In such way, prevention actions can be performed when the circuit has aged. This chapter presents a methodology to identify the critical paths of a circuit for reliable aging monitoring. The proposed methodology uses a statistical selection criterion to decide whether a path should be monitored. The proposed methodology takes into account the fact that different paths may become critical for different operating workloads. In order to identify the critical paths early in the design stage, where no layout information is already available, a spatial correlation approximation has been proposed. This allows identifying the correct placement of the aging monitors before the final circuit layout has been done.
Aging monitoring is a smart way to deal with delay uncertainty caused by aging effects. Instead of adding worst-case margins to the clock period, the actual degradation of the circuit is monitored along the lifetime, and the system can react if degradation becomes significant. In order to measure the delay degradation of a circuit, in-situ circuit failure prediction sensors have been extensively investigated in the literature [115, 127–132]. As its name indicates, failure prediction sensors detect if a path is close to violating the clock period due to aging effects. Figure 4.1(a) illustrates the basic structure of an aging sensor and its behavior can be understood from the timing diagram shown in Figure 4.1(b). When the circuit is young (CP fresh), the CP transition occurs before the rising clock edge and also before a small timing window settled by the delay element (DE), where hazardous transitions are detected. Thus, the delayed signal (CP delayed) also makes its transition before clock edge and both the main flip-flop and the shadow flip-flop in the aging sensor store the same value. When CP wears due to BTI (CP aged), it performs a transition inside the detection window, so that each flip-flop stores a different value (note that the correct value is stored in the main flip-flop) and the sensor output signal is activated. The output of the prediction sensor can be used as control signal to perform countermeasures such as adaptive body biasing, supply voltage scaling, and frequency scaling, to guarantee reliable operation. Other structures for error prediction sensors have been suggested in [115, 127–132].

A basic framework for aging monitoring is shown in Figure 4.2. At design phase, the critical paths of the circuit are identified. The critical paths are those likely to become the longest path of the circuit and violate timing specifications. Then, aging sensors are placed at the selected paths in order to be able to keep under track their delay degradation. The clock period is initially settled to leave a small timing guardband to assure reliable operation until the time when the circuit aging is tested. During lifetime operation, delay-test is periodically performed. A
4.1. AGING MONITORING TECHNIQUE

![Diagram](a) Error Prediction Architecture  (b) Waveform

Figure 4.1: Aging sensor operation for error prediction.

test vector is applied to the circuit activating each of the critical paths, and if the sensor output is activated, which indicates that a path is close to violate timing specifications, preventive actions are taken, such as increasing the clock period. In such way, close-to-best performance is achieved over the lifetime.

![Diagram](Design Phase, Design Circuit for Desired Performance, Select Critical Paths to be Monitored, Aging Sensor Insertion on Critical Paths, Periodical On-line Delay Test, Error Predicted?, Preventive Actions (i.e. Decrease Frequency), Lifetime ≥ 0, Design Phase = 0)

Figure 4.2: Aging Monitoring Framework

A major concern for aging monitoring is the identification of the critical paths to be monitored at design phase. A cost-efficient methodology for critical paths selection is needed to guarantee that aging sensors are effectively introduced only at output nodes of those paths strictly requiring to be monitored.
4.2 Related work

Different approaches for aging analysis and critical path selection have been explored in the past. Aging-Aware Deterministic Static Timing Analysis (DSTA) is used in [133], [134], to identify the critical paths of a circuit. Those paths that under aging effects exceed a threshold delay value are considered as critical. However, circuit aging strongly depends on process variations (PV), and Aging-Aware DSTA cannot capture such combined effect of PV and BTI.

Aging-Aware Statistical Static Timing Analysis was used in [135] for circuit lifetime optimization. The critical paths to be optimized were those paths whose $\mu + 3\sigma$ of the aged delay was larger than the clock period (delay constraint). However, it should be noted that under a frequency scaling framework, the clock period of the circuit is not unique, as it is periodically adjusted to the actual longest path of the circuit. Therefore, although some paths may exceed the initial clock period, they would not trigger any logic fault if their delays are always shorter than the delays of the monitored paths.

In [115], [136] the joint effect of aging and process variations are considered. The selection of critical paths was improved by computing the probability of a degraded path under analysis ($t = 10$ years) to take a delay longer than the Longest Critical Path (LCP) of the circuit with no aging ($t = 0$ years). There are two main issues in these approaches: 1) Worst-case duty-cycle assumptions were made (i.e. the path under analysis is assumed under maximum aging whereas the LCP is assumed fresh), which leads to a large number of unnecessarily selected critical paths; 2) These approaches require spatial correlation information to perform Aging-Aware SSTA, but correlation information can be extracted only after final circuit layout has been made. However, critical paths need to be identified at early stages of the design flow in order to introduce test circuitry. Furthermore, it should be noted that some paths that are not covered by monitoring the LCP only may be covered by monitoring another path.
The main contributions of this work are:

1. Critical paths are selected statistically using a greedy algorithm and a spatial correlation approximation that allows performing SSTA early in the design flow.

2. Critical paths are selected for various realistic operating workload profiles. Thus, different aging conditions of the paths are taken into account during path selection and classic worst-case aging assumptions are avoided.

### 4.3 Challenges for Critical Path Selection

Selecting more paths than required leads to additional area overhead introduced by non-required aging sensors, and extra online test effort. On the other hand, monitoring fewer paths than required may lead to lower reliability since potential errors may not be predicted, and consequently, the clock frequency may not scale as demanded by the circuit.

The identification of the critical paths that should be monitored is a complex problem because of the following challenges: 1) The critical Paths depend on process variations and need to be selected statistically; 2) The critical paths depend on the operating workload of the circuit, which is usually unknown in advance at the design phase. These challenges are described in more detail later. Let’s first define the concept of critical path.

#### 4.3.1 Critical Path Definition

A path is said to be a Critical Path (CP) if it can become the slowest path of the circuit over the lifetime.

Due to the combined effect of aging and process variations, there is a set of paths satisfying the above condition. Periodical online-test should be done over the critical paths, as they are the ones likely to fail first due to aging. Figure
4.3 illustrates a possible temporal behavior of some paths in a circuit and how clock period should adapt accordingly. Path P1 is initially the longest path of the circuit hence the clock period is settled slightly larger than P1 delay. Each time the sensors detect when P1 is close to exceeding the clock period due to BTI degradation, the clock period is increased to keep correct operation. However, path P2, that is not a critical at \( t = 0 \) can also become the longest path of the circuit after a given operational. Therefore, path P2 also requires being monitored. Note that the delay of path P3 is always shorter than P1 and P2 over the entire lifetime. Thus monitoring P1 and P2 (the critical paths of the circuit) is sufficient to assure reliable operation.

Figure 4.3: Critical Paths Definition Over time

An efficient critical path selection should identify only those paths that may become the slowest path of the circuit over the lifetime as those paths “fail-first” than the others. However, this is not a trivial problem due to the following challenges.

### 4.3.2 Critical Paths dependence on process variations

Process variations play an important role in determining whether a path is likely to become critical and should be monitored. Figure 4.4 illustrates the delay distributions of various paths in a circuit. The path P1 is taken as reference and we want to know if the other paths can become slower than path P1. As can be
observed, identification of those paths that can become slower than P1 do not only depends on the mean value of the paths delay. Due to process variations, the probability of the delays of a path to be longer than the delays of other path depends on the variance of each path and also on the covariance between them. As explained in previous Chapters, spatial correlation further impact on these parameters and consequently on the probability of a path to become critical.

The critical paths should be selected statistically by means of Statistical Static Timing Analysis (SSTA). However, conventional SSTA approaches require detailed information on gate placement in the corresponding layout to account for spatial correlation between devices, but this information is not usually available in an early design phase [137–139], where the critical paths should be identified to introduce aging sensors and other test circuitry.

### 4.3.3 Critical Paths dependence on operating workload

The aging rate of the circuit is highly sensitive to the executed workload since it defines the average time that devices are at stress condition (duty-cycle). Figure 4.5 shows a histogram of the aged delay of a path in ISCAS c880 circuit for 1000 different workload profiles. As shown, overall path delay degradation has a Gaussian-like distribution, as was also found in [140,141]. The delay degradation
that the path experiences significantly changes depending on the executed work-
load. For this example, up to $50\text{ps}$ of different delay degradation can occur for
different workload profiles.

Figure 4.5: Histogram of the delay of a PCP for various workload profiles (ISCAS
c2670)

A major issue to consider the workload impact on BTI and circuit delay degra-
dation is that the exact operating workload that a circuit will experience over the
lifetime is uncertain at the design phase. If the workload of a circuit is known
in advance, the circuit can be optimally designed. Otherwise, if the workload is
not (exactly) known, conservative assumptions have to be made to assure that
the circuit does not fail during its specified lifetime. State-of-art approaches solve
workload uncertainty issue by assuming a specific value for the SP at each input
of the circuit (usually $SP = 0.5$ is assumed) [71, 81, 89]. However, this assumption
may not be reliable enough if the actual workload applied to the circuit differs
from that assumed during design. Other approaches assume worst-case degrada-
tion conditions, settling the devices under near-static stress ($\alpha \approx 1$) and elevated
temperatures [142, 143]. As shown in Figure 4.5, the aged delay estimated for a
path using worst-case assumptions is much longer than the actual delay that the
path can take for various realistic operating workloads. Although circuit design
under this assumption guarantees reliable operation, significant over-design (area
and power overhead) may be obtained.
4.4 Proposed Critical Paths Selection Condition

A path should be monitored if it can become the longest path of the circuit over the lifetime. This can be expressed as follows:

**Selection Condition:** A path \( i \) should be monitored if for a given operating workload its probability over the lifetime to take delays longer than every other path \( j \) is greater than a probability threshold (\( \varepsilon \)).

The above selection condition can be expressed as follows:

\[
P(D_i > D_1) \geq \varepsilon \land P(D_i > D_2) \geq \varepsilon \land \ldots \land P(D_i > D_j) \geq \varepsilon \quad (4.1)
\]

Equation 4.1 is based on comparing the distributions between a pair of paths. If for all the feasible workload conditions a path \( i \) does not satisfy Equation 4.1 for at least one path \( j \), the path \( i \) does not require to be monitored as the delays of other paths are always longer, which means that the other path will fail first.

The probability threshold (\( \varepsilon \)) defines the maximum allowed probability of a non-monitored path to becoming the longest path of the circuit. If this happens, a faulty operation may occur. Therefore, the probability threshold (\( \varepsilon \)) establishes a trade-off between the desired degree of circuit reliability and the corresponding number of paths to be monitored.

![Statistical Delay Difference Between two Paths](image)

**Figure 4.6: Statistical Delay Difference Between two Paths**

For a given pair of paths \( i \) and \( j \), the probability of path \( i \) to be greater than
4.4. PROPOSED CRITICAL PATHS SELECTION CONDITION

path $j$ can be computed from the bivariate joint probability density function. However, the two-dimensional problem is transformed to a uni-dimensional problem using the delay difference ($DD_{ij} = D_i - D_j$) random variable, and evaluating its probability of being greater than zero, as shown in Figure 4.6,

$$P(D_i > D_j) \geq \varepsilon \Rightarrow P(DD_{ij} > 0) \geq \varepsilon$$

(4.2)

The delay difference between two paths is normally distributed with a mean value ($\mu_{DD}$) and a variance ($\sigma^2_{DD}$) given by:

$$\mu_{DD,ij} = \mu_{D_i} - \mu_{D_j}$$

(4.3a)

$$\sigma^2_{DD,ij} = \sigma^2_{D_i} + \sigma^2_{D_j} - 2COV(D_i, D_j)$$

(4.3b)

where $\mu_{D_i}$ and $\mu_{D_j}$ are the mean delays of the paths under analysis, $\sigma_{D_i}$ and $\sigma_{D_j}$ are the standard deviations of the delays of the paths under analysis, and $COV(D_i, D_j)$ is the inter-path covariance between $D_i$ and $D_j$. Once computed $\mu_{DD,ij}$ and $\sigma^2_{DD,ij}$ between two paths, the probability of Equation 4.2 is obtained by integrating the PDF of $DD_{ij}$ in the region where it takes values greater than zero. The integral is approximated by Gaussian quadrature method with 15 points [144].

Note that both, $\mu_{DD,ij}$ and $\sigma^2_{DD,ij}$ depend on the operational time and the operating workload, which is uncertain during design. Furthermore, $\sigma^2_{DD,ij}$ also depends on the spatial correlation of the gates, which depends on the placement coordinates of the gates in the final layout, which are unknown at early design phases. The following sections explain how to handle uncertainty due to unknown spatial correlation and operating workload.

4.4.1 Handling Spatial Correlation Uncertainty

The spatial correlation that arises due to devices proximity in the circuit layout has a significant impact on the selection probability of a path. As commented in
Chapter 3, spatial correlation can be extracted from the final circuit layout using the model of [2]. However, critical path selection and aging sensor insertion need to be performed at an early design stage, where the coordinates of the transistor locations in the layout are unknown. The following lines explain the spatial correlation approach used to perform SSTA at an early design stage, without spatial correlation information from the circuit layout.

The standard deviation of the delay difference between two paths $i$ and $j$ can be expressed as:

$$\sigma_{DD,ij}^2 = \sigma^2_{D_i} + \sigma^2_{D_j} - 2 \text{COV}(D_i, D_j)$$

$$= \sum_{x \in i} \sum_{y \in i} \rho_{xy} \cdot \sigma_x \cdot \sigma_y + \sum_{x \in j} \sum_{y \in j} \rho_{xy} \cdot \sigma_x \cdot \sigma_y - 2 \cdot \sum_{x \in i} \sum_{y \in j} \rho_{xy} \cdot \sigma_x \cdot \sigma_y$$

(4.4)

where $N$ and $M$ are the numbers of gates in path $i$ and $j$, $\sigma_x$ and $\sigma_y$ are the standard deviation of the delays of gates $x$ and $y$ in the corresponding path, and $\rho_{xy}$ is the spatial correlation between the gates. $\rho_{xy}$ in the first and the second summation relates two gates in the same path while $\rho_{ij}$ in the third summation relates one gate in path $i$ and one gate in path $j$ when computing inter-path covariance.

The values of spatial correlation required to compute the standard deviation of the delay difference can be represented by a correlation matrix as shown in Equation 4.5. The term $\rho_{m_i,n_j}$ represents the spatial correlation between gates $n$ and $m$ in paths $i$ and $j$, respectively. Matrix indexes highlighted in green correspond to the spatial correlation between gates in the same path. The diagonal of the matrix highlighted in red corresponds to the self-correlation of a gate, which is the unit, no matter where the gate is placed. Those values are used to compute each path variance (First two summations of Equation 4.4). Matrix
indexes highlighted in blue correspond to the spatial correlation between a gate in one path and a gate in the other path. Those values are used to compute the inter-path covariance (Third summation of Equation 4.4). Note that in case of structural correlation, some of the spatial correlation values relating gates in the different paths become the unit.

In order to select critical paths under conservative assumptions due to spatial correlation uncertainty, $\sigma_{DD}^2$ should be approximated in such way that its value is maximum, maximizing the probability of the Delay Difference variable to take values greater than zero, as illustrated in Figures 4.7(a) and 4.7(b).

The approximation of $\sigma_{DD}^2$ can be made assuming spatial correlation bounds, using a similar idea to that in [137–139]. Analyzing Equation 4.4, it can be observed that an upper bound for $\sigma_{DD}^2$ can be obtained as follows:

- Assume that the spatial correlation for computing the variances of the delay of paths $i$ and $j$ is the unit (gates are assumed fully correlated). Thus
4.4. PROPOSED CRITICAL PATHS SELECTION CONDITION

![Diagram showing impact of \( \sigma_{DD} \) on selection probability of a path (\( P_1 < P_2 \)).](image)

Figure 4.7: Impact of \( \sigma_{DD} \) on the selection probability of a path (\( P_1 < P_2 \)).

- Assume that the spatial correlation for computing the inter-path covariance between paths \( i \) and \( j \) is zero (except if a structural correlation exists). Thus estimated inter-path covariance is minimum.

In such way, a maximum value for the sum of paths variances (\( \sigma_{D_i}^2 + \sigma_{D_j}^2 \)) is obtained, and a minimum value is obtained for inter-path covariance (\( \text{COV}(D_i, D_j) \)). Thus overall \( \sigma_{DD} \) is maximized. However, this approximation would result in a significant overestimation of the \( \sigma_{DD} \) values that are obtained after gate placement in the layout, and consequently, in the number of paths selected for delay monitoring [114, 145].

In order to develop a more efficient method to estimate \( \sigma_{DD} \) without spatial correlation, an analysis of the behavior of the terms in Equation 4.4 was done using spatial correlation information extracted from the final layout of some circuits. For this experiment, the layouts of the circuits were done using Mentor Graphics EDA tool. The gates were placed in the layout using the auto-place option. Then, Statistical Static Timing Analysis was performed using spatial correlation information extracted with the model of [2]. Then, the standard deviation of the delay difference (Equation 4.4) between each path \( i \) in the circuit and a reference path \( j \) was computed. The path with the longest mean delay was used as reference.
4.4. PROPOSED CRITICAL PATHS SELECTION CONDITION

path \((j)\) for this experiment.

Figures 4.8(a) and 4.8(b) plot the terms of Equation 4.4 obtained for this experiment for two ISCAS circuits. Each dot in the figures represents a different path in the circuit. As can be observed, there exists a relationship between the sum of path variances and the corresponding inter-path covariance. For those paths with larger variances, which are usually the paths with the larger mean delay too, the inter-path covariance with the reference path tends to be greater. Those paths whose delay is larger than the 90% of the maximum delay of the circuit are highlighted in green in the Figure. Those paths with largest delays are the ones with larger probabilities to become critical. As can be seen, both the sum of path variances and the corresponding inter-path covariance tend to be larger for those paths. This trend was also observed for other ISCAS circuits. These results demonstrate that the correlation bounding approach does not capture such dependency. Based on this observation, a first approximation is made:

**Correlation Approximation I:** A global correlation \((\rho_g)\) value is assumed to estimate the standard deviation of the delay difference between two paths.

Correlation Approximation I means that all \(\rho_{m,n_j}\) in Equation 4.5 are replaced by a single \(\rho_g\) value, except when gates \(m\) in path \(i\) is equal to the gate \(n\) in path \(j\), where \(\rho_g\) becomes the unit. Some paths in a given circuit could be out of the trend observed in Figures 4.8(a) and 4.8(b), and consequently, the global correlation approximation may be inaccurate for them, leading either to an overestimation or an underestimation of the \(\sigma_{DD}\). The impact of this approximation will be analyzed later in the results part.

A second issue is to determine the best \(\rho_g\) value to be used to estimate \(\sigma_{DD}\) in such way that most of the critical paths are correctly selected. Therefore, selected \(\rho_g\) should maximize \(\sigma_{DD}\). The experiment illustrated in Figure 4.9(a) is used to analyze which global correlation value (between 0 and 1) maximizes \(\sigma_{DD}\) between two paths. The standard deviation of the delay difference between two paths \(i\) and \(j\) of different depths is computed using different global correlation values.
4.4. PROPOSED CRITICAL PATHS SELECTION CONDITION

Figure 4.8: Relationship between sum of path variances and inter-path covariances for some ISCAS circuits.

Path \( j \) is used as a reference path and has a fixed depth of 20 inverters, while path \( i \) depth is changed from 1 to 20 inverters. Figure 4.9(b) shows how \( \sigma_{DD} \) changes as function of the depth of path \( i \) for various \( \rho_g \). As can be observed, for \( \rho_g = 0 \), \( \sigma_{DD} \) is low for short paths, and it increases linearly as the depth of path \( i \) (\( N \)) gets closer to the depth of the reference path \( j \) (\( M = 20 \)). On the other hand, for \( \rho_g = 1 \), \( \sigma_{DD} \) is large for short paths, but it reduces as the depth of path \( i \) (\( N \)) gets closer to the depth of the reference path \( j \) (\( M = 20 \)). Interestingly, for long paths \( \sigma_{DD} \) approximated using \( \rho_g = 0 \) is larger than \( \sigma_{DD} \) approximated using \( \rho_g = 1 \). As shown in Figure 4.9(b), for path depths larger than \( N \approx 13 \), using \( \rho_g = 0 \) maximizes \( \sigma_{DD} \). This behavior can be deduced analytically with the following simple equations, assuming that the variance of each inverter gate is the same [19]:

\[ \sigma_{DD} \approx \sum_{k=1}^{M} \sigma_k^2 \rho_g(2k-1) \]
4.4. PROPOSED CRITICAL PATHS SELECTION CONDITION

(a) Example of Two paths
(b) $\sigma_{DD}$ as function of depth of path $i$ (N)
(c) Ratio of $\sigma_{DD}$ computing with $\rho_g = 0$ to that computed with $\rho_g = 1$

Figure 4.9: Analysis of the global correlation value ($\rho_g$) that maximizes the standard deviation of the delay difference between two paths.

$$\sigma^2_{D_i} = [N + \rho_g (N^2 - N)] \cdot \sigma^2_{inv}$$

$$\sigma^2_{D_j} = [M + \rho_g (M^2 - M)] \cdot \sigma^2_{inv}$$

$$COV(D_i, D_j) = N \cdot M \cdot \rho_g \cdot \sigma^2_{inv}$$

(4.6)

Using Equations 4.6, $\sigma_{DD}$ using $\rho_g = 0$ and $\rho_g = 1$ are given by:

$$\sigma^2_{DD, \rho_g=0} = (N + M) \cdot \sigma^2_{inv} - 0$$

(4.7a)

$$\sigma^2_{DD, \rho_g=1} = (N^2 + M^2) \cdot \sigma^2_{inv} - 2 \cdot N \cdot M \cdot \sigma^2_{inv}$$

(4.7b)

From Equation 4.7a, it is evident that the first term, corresponding to the sum
4.4. PROPOSED CRITICAL PATHS SELECTION CONDITION

of path variances always increases as function of $N$ while nothing is subtracted since covariance becomes zero. From Equation 4.7b, the first term (sum of variances) increases as function of $N^2$, while subtracted term (covariance) increases as function of $N \cdot M$, and since $M > N$, the covariance term increases faster than the sum of path variances as $N$ increases, and therefore, $\sigma_{DD,\rho_g=1}$ reduces, as shown in Figure 4.9(b). The ratio of $\sigma_{DD}$ computed using $\rho_g = 0$ to that computed using $\rho_g = 1$ as function of the depth of path $i$ ($N$) is plotted in Figure 4.9(c). As can be observed, for long paths (path with a large number of gates), the $\sigma_{DD}$ value computed using $\rho_g = 0$ is greater than the computed using $\rho_g = 1$. Long paths are those with larger delays in the circuit. Hence, the global correlation value should be set to zero ($\rho_g = 0$) to approximate $\sigma_{DD}$ at early design stages. The behavior of Figure 4.9(b) was observed on ISCAS circuits. For example, Figure 4.10, shows $\sigma_{DD,\rho_g=0}/\sigma_{DD,\rho_g=1}$ obtained for every path in ISCAS circuit C1908. As can be seen, for those paths with larger delays, $\rho_g = 0$ provides a larger $\sigma_{DD}$.

![Figure 4.10: Ratio of $\sigma_{DD}$ estimated with $\rho_g = 0$ to that estimated with $\rho_g = 1$](image)

Based on the previous analysis the second approximation is proposed:

**Correlation Approximation 2:** A global correlation value of $\rho_g = 0$ is assumed to approximate $\sigma_{DD}$ between two paths.

By using the proposed correlation approximations, the delay difference be-
4.4. PROPOSED CRITICAL PATHS SELECTION CONDITION

tween two paths can be estimated at an early design phase without spatial correlation information. Moreover, this approach is more realistic than classic pessimistic correlation bounds.

4.4.2 Handling Workload Uncertainty

Operating workload affects the selection probability of a path. As mentioned before, the specific workload that the circuit will experience over the lifetime is hardly known in advance at the design phase. State-of-art approaches usually assume pessimistic workload scenarios to assure that the critical paths that can be presented for any operational workload are identified. In [136, 146], the longest critical path at $t = 0$ is used as reference path ($j$) for critical path selection. The critical paths are selected evaluating Equation 4.1 assuming that the reference path ($j$) remains unaffected by aging, while the path under analysis ($i$) experiences worst-case aging as all the transistors in the path are assumed at near-static stress ($\alpha \approx 1$). In such way, the mean value of the delay difference is maximized. However, this approach is unrealistic since such unbalanced degradation may not occur during normal circuit operation. Furthermore, all the paths experience some degradation under realistic operating conditions. In [145], the degradation of the paths was estimated for various feasible workload scenarios. A maximum/minimum degradation of each path was identified. Then, Equation 4.1 was evaluated assuming that the reference path ($j$) experiences minimum aging while the path under evaluation was assumed under maximum aging. This improves the workload conditions assumed for path selection, but such unbalanced degradation is still pessimistic.

Figures 4.11(a) and 4.11(b) show scatter plots of the mean value of the aged delay (after 10 years of aging) of two paths in ISCAS c2670 for several different workload profiles. For these figures, the reference path ($j$) is the path with the largest nominal delay in the circuit (LCP), and two cases of paths under analysis
(i) are shown:

a) The delay degradation of path $i$ has a low correlation with the delay degradation of path $j$.

b) The delay degradation of path $i$ has a high correlation with the delay degradation of path $j$.

As can be observed in Figures 4.11(a) and 4.11(b), even different operating workloads may cause similar degradation in the paths. This aging correlation between paths may arise due to nodes signal probability correlation, paths topology (i.e., similar logic gates are presented in the paths) and the existence of shared gates between paths. This realistic behavior of delay degradation is not considered in classic approaches. Figure 4.11(c) shows a histogram of the mean value of the delay difference (See Equation 4.3a) for the cases in 4.11(a) and 4.11(b). As shown, the maximum realistic $\mu_{DD}$ that can exist between the two paths is much smaller than that computed using the max/min aging approach in [145]. Therefore, the probability of the path $i$ to take delays greater than path $j$ is much smaller than the estimated using such assumption.

Based on the above observations, it is proposed to perform path selection for various realistic random-generated workload profiles. By analyzing different workload profiles, a more efficient selection of critical paths can be made. It is important to trade-off the computational effort for analyzing various workloads and the degree of circuit reliability.

It must be noted that for a given operating workload, the probability of a path to take longer delays than other path changes over the lifetime depending on the aging rate of each path. This behavior is captured by the temporal dependence of the paths delay difference. Therefore, paths that are not critical at the beginning of the lifetime may become critical at the end of the lifetime [134]. Taking into account this reordering behavior is complex as exhaustive evaluation of crit-
4.4. PROPOSED CRITICAL PATHS SELECTION CONDITION

The proposed critical path selection condition (Equation 4.1) as a function of operational time is computationally costly.

The computational cost of computing the reordering probability over the time between a pair of paths can be lowered by evaluating Equation 4.2 only at the point in the time where the reordering probability is maximum. Figure 4.12 shows the behavior over the time of the delay distribution of two paths (i and j) for two degradation cases: a) The aging rate of path i is greater than the aging rate of path j (Figure 4.12(a)), and b) The aging rate of path i is lower than the aging rate of path j (Figure 4.12(b)). As can be seen, due to monotonically increase of paths delays as a function of aging time exponent (n) [147], the maximum reordering probability between the paths occurs either at the beginning or at the end of

Figure 4.11: Analysis of the degraded delay of two paths i and j in ISCAS c2670 for 1000 random-generated operating workloads.

(a) Low Delay Degradation Correlation (b) High Delay Degradation Correlation

(c) Histograms of the delay difference
4.5 Proposed methodology for critical path selection

The flow of the proposed methodology is shown in Figure 4.13. For timing analysis, the linear gate delay model of Equation 3.6 was used to capture the effect of process variations and BTI-induced Vth shift on gate delay. The linear model parameters are obtained during gate library pre-characterization, as explained in Chapter 3.

The methodology starts from the gate-level circuit netlist. A path search algorithm (See section 3.3.3) is used to find all the topological paths of the circuit. Since the number of topological paths may be too large in modern digital integrated circuits, a fast path pre-filtering step based on Static Timing analysis with process corners is done. This reduces the number of paths that are analyzed statistically. The filtered paths are next analyzed statistically using the proposed
4.5. PROPOSED METHODOLOGY FOR CRITICAL PATH SELECTION

The following subsections explain in detail each of the steps of the proposed methodology.

4.5.1 Fast Path Pre-Filter

The fast path pre-filter step allows identifying those paths having some probability to become critical under excessive aging and process corner variations. Timing de-
lay information of all topological paths is obtained by conventional Static Timing Analysis based on process corners.

Figure 4.14: Fast Path pre-filtering.

Figure 4.14 illustrates the pre-filtering step. Those paths taking delays at the Slow-Slow process corner plus some pessimistic percentage ($p\%$) of aging larger than the delay of the fresh Longest Critical Path (LCP) at Fast-Fast corner are included in the pre-filtered path set. The value of $p\%$ represents the highest delay degradation due to BTI that a circuit may suffer, which has been reported to be around $15−25\%$ for 10 years \[25,148\]. In this work, we select a value of $p = 25\%$. By using this overestimated value, it is assured that no path with possibilities to become critical for any process and aging condition will be discarded.

### 4.5.2 Potential Critical Paths Under Worst-Case Aging

The set of pre-filtered paths is reduced even more by identifying the statistical Potential Critical Paths (PCP). A PCP is a path that under worst-case aging conditions may exceed the initial (fresh) delay of the Longest Critical Path (LCP) of the circuit, which is the path with the largest mean delay at $t = 0$. This means that if the clock period is adjusted according to the LCP only, the PCPs may trigger a faulty operation over the lifetime. This condition is expressed as follows:

**PCP Condition:** A path $i$ is said to be a PCP if the probability of its aged
delay to be longer than the fresh delay of the LCP is larger than a user-defined probability threshold \( \varepsilon \):

\[
P(D_{i,t=10} > D_{LCP,t=0}) \geq \varepsilon
\]  

(4.8)

where \( D_{i,t=10} \) is the degraded delay of path \( i \) after 10 years of aging and \( D_{LCP,t=0} \) is the delay of the LCP with no aging. Here, the delay degradation of path \( i \) is evaluated under worst case conditions given by a near-static duty-cycle (\( \alpha \approx 1 \)) and elevated operating temperature (\( T = 110^\circ \)).

Paths that do not meet this condition do not require to be monitored since they would never activate an error prediction sensor because the delay of the LCP is always longer for the entire lifetime (the LCP will always fail first), even when the paths suffer extreme aging and the LCP remains fresh.

### 4.5.3 Critical Path Selection Algorithm

The final selection of critical paths takes place over the reduced set of PCPs. The selection algorithm takes into account that:

- Different paths may become critical depending on the workload profile at main circuit inputs. Therefore, various workload profiles are applied to the circuit to identify the different critical paths that may exist.

- Some of the paths not covered by the LCP could be covered by another path. Therefore, a path is also compared against other previously selected paths to determine if it requires to be monitored. In such way, the final set of CPs is reduced.

Algorithm 1 describes the procedure for final critical path selection. The set of PCPs is sorted in a decreasing order according to the nominal delay (no-BTI and no-PV) of the paths. The set of critical paths is initialized with the LCP of the circuit since this path has the largest chance to cover most of the other
ALGORITHM 1: Fine Critical Path Selection

**Input:** Potential Critical Paths

**Output:** Critical Paths to be monitored

Sort PCPs paths from largest to shortest delay.

Initialize CP set: $\text{Path}[1].\text{Critical}=\text{YES}$, $\text{Num.CPs}=1$

for $WL = 1$ to $WL = \text{MaxWL}$ do

\[
\text{generate_propagate_SP}() \\
\text{compute_temperature}() \\
\text{compute_duty_cycle}() \\
\text{compute_A_and_S}() \\
\text{Run SSTA on PCPs}
\]

for $i = 2$ to $i = \text{Num.PCPs}$ do

if $\text{PCP}[i].\text{Critical}==\text{YES}$ then

The PCP $i$ is already selected.

Continue with next PCP.

end

Path Prob cum = 1

for $j = 1$ to $j < i$ do

if $\text{PCP}[j].\text{Critical}==\text{YES}$ then

$\text{Path Prob} = \max(P(D_{ij,t=0}>0), P(D_{ij,t=10yr}>0))$

$\text{Path Prob cum} = \min(\text{Path Prob}, \text{Path Prob cum})$

if $\text{Path Prob cum} < \varepsilon$ then

Monitored PCP $j$ covers PCP $i$

$\text{PCP}[i].\text{Critical}=\text{NO}$

break

end

if $\text{Path Prob cum} \geq \varepsilon$ then

Monitored PCP $j$ does not cover PCP $i$

Try with next monitored PCP $j$

end

end

if $\text{Path Prob cum} \geq \varepsilon$ then

PCP $i$ can be greater than all current monitored PCPs $j$

$\text{PCP}[i].\text{Critical}=\text{YES}$

$\text{Num.CPs}++$

end

end

paths. Critical path selection is performed for a set of $\text{MaxWL}$ random-generated workload profiles ($WL = \{WL_1, WL_2, ..., WL_N\}$), which are intended to represent realistic operating scenarios for the circuit. Each workload profile is represented by the Signal Probabilities (SP) at main circuit inputs,

\[ WL_i = \{SP_{1,i}, SP_{2,i}, ..., SP_{NI,i}\} \quad (4.9) \]

where $WL_i$ is the workload profile $i$, and $SP_{NI,i}$ is the signal probability as-
4.5. PROPOSED METHODOLOGY FOR CRITICAL PATH SELECTION

Signed to the input NI for the workload profile i. A uniform random generator between 0 and 1 is used to obtain the signal probability assigned to each input.

For each workload profile, the signal probabilities are propagated to internal nodes using basic probability rules. Then, the stress duty-cycle (\( \alpha \)) and operating temperature of each transistor in the circuit are computed. Once duty-cycle and operating temperature are obtained, the device’s parameters \( A_m \) and \( S_{V,m} \) of the process-variation aware BTI model (Equation 3.3) are updated. Then, aging-aware SSTA is run over the PCP set to update the paths timing information according to the analyzed workload.

The selection condition (Equation 4.1) of each non-monitored PCP \( i \) to take delays longer than each already selected CP \( j \) (initially the LCP only) is computed. Since the exhaustive computation of Equation 4.2 between each pair of PCPs is computationally expensive, Algorithm 1 uses a greedy heuristic based on the PCPs ranking to reduce the number of paths comparisons. The heuristic takes into account the following aspects to speed-up computational time:

1. If a path is identified as critical (i.e. none of the actual CP covers the path delays), the path is removed from the PCP set and it is included in the CP set. Then, its selection condition is not evaluated for the rest of the workloads.

2. Each PCP is only compared against the CPs with a larger mean delay.

3. Usually, the first slowest CPs are capable to cover most of the paths. Therefore, each PCP is usually compared against few CPs.

In order to determine if a path needs to be monitored, the following procedure is carried out (See Algorithm 1): For a PCP \( i \), its selection probability (Path.Prob) is initialized at the unit. This value is going to be reduced by evaluating the probability of PCP \( i \) to take delays larger than other CP \( j \) being monitored. Equation 4.2 is computed only with respect to those paths \( j \) already
in the CP set (being monitored) and having a larger nominal delay \((j < i)\). Note that the maximum selection probability of the one obtained at \(t = 0\) or \(t = 10\) years of aging is stored in \(\text{Path}_\text{Prob}\) variable and compared against the probability threshold \((\varepsilon)\). If the computed probability for PCP \(i\) is lower than \(\varepsilon\), then PCP \(i\) does not require to be monitored, and its selection probability is no longer computed against the other remaining critical paths. This approach (See highlighted line in blue in the Algorithm) further reduces computational time. In case the computed probability for a PCP \(i\) is larger than \(\varepsilon\), which means that monitoring the CP \(j\) does not cover the delays of the PCP \(i\), the selection probability process continues with the next CP \(j\) being monitored. Note that the variable \(\text{Path}_\text{Prob}\) stores the minimum selection probability that PCP \(i\) took with respect to the CPs with it was compared. If none of the current selected CPs \(j\) can cover the delays of the PCP \(i\), the PCP \(i\) is included in the critical path set which need to be monitored.

After applying Algorithm 1, those paths that are likely to take delays greater than all the other paths in the circuit due to BTI-induced delay degradation and process variations are effectively selected to be under constant surveillance by aging sensors. Hence, reliable circuit tuning can be performed.

### 4.6 Experimental Results on ISCAS Circuits

The proposed methodology has been applied to some ISCAS 85/89 benchmark circuits implemented in a 32nm technology library [65]. The algorithm has been written in C++ code. In addition, the layout of the ISCAS circuits has been obtained to compare the results obtained using the global correlation approximation with those obtained using layout information. The experiments were performed on an Intel(R) Xeon(R) CPU E5-2609 V2 @ 2.50GHz.

Table 4.1 shows the main characteristics of the analyzed circuits. Second and third columns give the number of gates and topological paths of the circuits. IS-
CAS circuits of different sizes and complexity have been considered to evaluate the scalability of the proposed methodology. Columns fourth and fifth show the nominal delay (no-BTI and no-PV) and the Design-Time guardband required to cover BTI aging (10 years) and process variations under a conventional design (non-adaptive). This guardband was computed under worst-case aging assumptions. On average, a 51.20% of guardband over the nominal delay is required. This imposes a significant performance loss.

Table 4.1: Main characteristics of the considered ISCAS circuits

<table>
<thead>
<tr>
<th>Circuit</th>
<th>Num. Gates</th>
<th>Num. Paths</th>
<th>$D_{\text{nom}}$ (ps)</th>
<th>Design-Time Guardband (ps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C432</td>
<td>168</td>
<td>82364</td>
<td>600.25</td>
<td>312.32</td>
</tr>
<tr>
<td>C880</td>
<td>226</td>
<td>4335</td>
<td>351.50</td>
<td>194.19</td>
</tr>
<tr>
<td>C1908</td>
<td>244</td>
<td>15638</td>
<td>329.21</td>
<td>164.84</td>
</tr>
<tr>
<td>C2670</td>
<td>393</td>
<td>3490</td>
<td>492.12</td>
<td>235.56</td>
</tr>
<tr>
<td>C3540</td>
<td>748</td>
<td>787401</td>
<td>731.75</td>
<td>352.53</td>
</tr>
<tr>
<td>C5315</td>
<td>1139</td>
<td>24666</td>
<td>433.09</td>
<td>213.25</td>
</tr>
<tr>
<td>C7552</td>
<td>1365</td>
<td>43614</td>
<td>426.43</td>
<td>218.3</td>
</tr>
<tr>
<td>S838</td>
<td>279</td>
<td>1714</td>
<td>350.47</td>
<td>210.53</td>
</tr>
<tr>
<td>S1423</td>
<td>657</td>
<td>44726</td>
<td>1427.29</td>
<td>628.13</td>
</tr>
<tr>
<td>S5378</td>
<td>1174</td>
<td>11728</td>
<td>339.77</td>
<td>176.68</td>
</tr>
<tr>
<td>S9234</td>
<td>1817</td>
<td>29071</td>
<td>542.51</td>
<td>273.91</td>
</tr>
<tr>
<td>S13207</td>
<td>2403</td>
<td>585308</td>
<td>748.13</td>
<td>371.39</td>
</tr>
<tr>
<td>S35932</td>
<td>10902</td>
<td>26146</td>
<td>221.73</td>
<td>123.282</td>
</tr>
</tbody>
</table>

4.6.1 Path Selection Results

Table 4.2 shows detailed results of the selected paths and CPU time at the PCP selection and final path selection steps. For critical path selection, a probability threshold of $\varepsilon = 0.25\%$ has been used to provide a 99.75% degree of circuit reliability. The second column of Table 4.2 gives the number of identified PCPs. The percentage of PCPs out of the total topological paths is also given in the parentheses. As can be seen, the percentage of PCPs may change significantly, depending on the circuit topology, how balanced are the delays of the paths and how sensitive are the paths to BTI-aging. These PCPs were selected using worst-
4.6. EXPERIMENTAL RESULTS ON ISCAS CIRCUITS

case aging conditions as in some state-of-art works [115, 136]. However, monitoring such number of paths could be expensive due to online test effort and area overhead of the inserted sensors. The computational time of the PCPs selection is proportional to the number of paths in the circuit since each path is compared against the LCP of the circuit. Final selection of critical paths has been applied for different cases of workloads. The number of selected critical paths (CPs) and the computational time (CPU) are given for four cases (1, 20, 40, and 60) of the number of random-generated workload profiles. For workload = 1, a reduced set of paths is identified as critical. This large reduction occurs because the proposed approach takes into account that the monitored paths degrade over the lifetime, which means that the clock period is gradually adjusted according to these paths, and consequently, various paths with lower aging cannot trigger any fault. For most of the circuits, the number of selected critical paths increases when more workloads profiles are analyzed during the fine selection step. This is because the new workloads induce more aging in some of the paths that were not selected first. For some circuits, a limited number of new critical paths (CPs) are identified as the number of workloads increases (i.e., circuits s1423, s838 and c2670). However, for other circuits, the number of new critical paths increases more significantly as the number of workloads increases (i.e., s13207). The CPU time increases as more workload are used. The total CPU time of the proposed methodology is the addition of both the CPU time for the coarse selection and the CPU time for the fine selection. An interesting observation is that only a few new critical paths are found when the number of tested workload profiles increases from 40 to 60. Figure 4.15 support this observation by showing the behavior of the number of critical paths (normalized with respect to the number of paths selected with one single workload) with respect to the number of tested workload profiles for some circuits. These results suggest that for some circuits, selecting critical paths for a moderate number of workload profiles may be enough to identify most of the paths that may become critical for any given workload operating condition. However, some
circuits may be more sensitive to the used workloads than others, and they may require a greater number of workload to assure high reliability. Designers should trade-off the degree of circuit reliability, which increases by detecting paths that may become critical under certain workload conditions, with the CPU cost to be paid.

Table 4.2: Selected paths using the proposed methodology ($\varepsilon = 0.25\%$)

<table>
<thead>
<tr>
<th>ISCAS</th>
<th>Coarse Selection</th>
<th>Fine Selection</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PCPs</td>
<td>CPU (sec)</td>
</tr>
<tr>
<td></td>
<td>Workloads=1</td>
<td>Workloads=20</td>
</tr>
<tr>
<td></td>
<td>Workloads=40</td>
<td>Workloads=60</td>
</tr>
<tr>
<td>c432</td>
<td>14319 (53.8%)</td>
<td>66.32</td>
</tr>
<tr>
<td>c880</td>
<td>2516 (59.8%)</td>
<td>2.40</td>
</tr>
<tr>
<td>c1908</td>
<td>11154 (71.32%)</td>
<td>8.67</td>
</tr>
<tr>
<td>c2670</td>
<td>752 (21.54%)</td>
<td>1.80</td>
</tr>
<tr>
<td>c3540</td>
<td>100167 (12.72%)</td>
<td>629.33</td>
</tr>
<tr>
<td>c5315</td>
<td>9903 (40.14%)</td>
<td>13.10</td>
</tr>
<tr>
<td>c7552</td>
<td>17650 (40.46%)</td>
<td>21.42</td>
</tr>
<tr>
<td>s838</td>
<td>372 (21.7%)</td>
<td>0.48</td>
</tr>
<tr>
<td>s1423</td>
<td>4896 (10.94%)</td>
<td>151.27</td>
</tr>
<tr>
<td>s5378</td>
<td>2016 (17.18%)</td>
<td>3.11</td>
</tr>
<tr>
<td>s9234</td>
<td>2840 (9.75%)</td>
<td>13.45</td>
</tr>
<tr>
<td>s13207</td>
<td>170913 (29.19%)</td>
<td>5.66</td>
</tr>
<tr>
<td>s35932</td>
<td>13907 (53.18%)</td>
<td>224</td>
</tr>
</tbody>
</table>

Figure 4.15: Selected critical paths vs. evaluated workload profiles

Figure 4.16 shows the impact of increasing the probability threshold ($\varepsilon$) for critical path selection. Figure 4.16(a) shows the reduction in the percentage of the number of critical paths that need to be monitored for $\varepsilon$ values of 1% and 2%.
4.6. EXPERIMENTAL RESULTS ON ISCAS CIRCUITS

The percentage is computed with respect to the results using $\varepsilon = 0.25\%$ (Table 4.2). When larger $\varepsilon$ is allowed, fewer paths need to be monitored. The reduction in the CP set is because those paths with a lower probability to become critical are discarded. On average, the critical path set is reduced in 28% and 41% for $\varepsilon = 1\%$ and $\varepsilon = 2\%$, respectively. When $\varepsilon$ is increased, CPU time is reduced due to two reasons: 1) The set of PCP selected under worst-case aging conditions is reduced, which means that the number of paths analyzed in the fine selection process is lower; and 2) In the fine selection algorithm, a PCP is more likely to be covered by the first selected Critical Paths. Therefore, the number of pair of paths comparisons reduces. Figure 4.16(b) shows the reduction in CPU time obtained for the analyzed circuits. On average, CPU time is reduced by 13.7% and 19.49% for $\varepsilon = 1\%$ and $\varepsilon = 2\%$, respectively. Small values for the probability threshold are recommended for higher degrees of reliability.

4.6.2 Validation of Spatial Correlation Approximation

Due to the spatial correlation approach used to perform SSTA at an early design state, the estimated standard deviation of the delay difference between paths may differ from the exact value that is obtained using spatial correlation information extracted from final circuit layout with the model of [2].

Figure 4.17 shows a comparison between the standard deviation of the delay difference computed with the proposed correlation approximation and the one computed using spatial correlation extracted from circuit layout. To obtain the figures, $\sigma_{DD,ij}$ was computed between the Longest Critical Path (LCP) and every other path in the circuits for a given workload profile. Each dot represents a path, and the dots highlighted in orange correspond to those paths whose probability to become slower than the LCP was greater than the probability threshold ($\varepsilon$). As can be observed in Figures 4.17(a) and 4.17(b), for the critical paths, the estimated $\sigma_{DD,ij}$ with the proposed correlation approach is slightly larger than
4.6. EXPERIMENTAL RESULTS ON ISCAS CIRCUITS

Results

Impact of increasing the probability threshold

CPs reduce in 28% and 41% for $\varepsilon = 1\%$ and $\varepsilon = 2\%$ (on average)

(a) Reduction on selected critical paths.

(b) Reduction on CPU time.

Figure 4.16: Impact of probability threshold on the CP set size and CPU time ($WL = 60$).

the one computed using spatial correlation information. This imply that the approximated selection probability of the paths is larger than the one obtained using layout information, and therefore, all the critical paths are selected with the proposed approach. However, some paths may be out of this trend, as shown in Figure 4.17(c) obtained from circuit C7552. As can be seen, the estimated $\sigma_{DD,ij}$ is lower than the one obtained using layout information for some paths. This may occur if the gates within a path are close to each other but the two paths are far from each other. However, it should be highlighted that those paths out of this trend may still be selected/covered due to the following reasons:
4.6. EXPERIMENTAL RESULTS ON ISCAS CIRCUITS

- The mean value of the delay difference also influences the estimated probability for a path.

- If the estimated probability using the correlation approximation is still larger than the threshold probability, the path is selected.

- The path may be covered by a path other than the LCP.

Figure 4.17: Approximated standard deviation of the delay difference between all the paths in the circuit and the LCP of the circuit Vs. the standard deviation of the delay difference obtained using spatial correlation extracted from circuit layout.

The proposed spatial correlation approximation is more realistic than approaches using correlation bounds for computing the standard deviation of the delay difference between paths. As commented before, using correlation bounds
results in a significant overestimation of the standard deviation of the delay difference, and consequently, on the selection probability of a path. Figure 4.18 shows histograms of the ratio of $\sigma_{DD,ij}$ computed using correlation estimation approaches at early design phase to the one computing using spatial correlation from circuit layout. Histograms for both the ratio corresponding to the $\sigma_{DD,ij}$ obtained with the proposed correlation approximation and with correlation bounds are shown. As can be observed, when the proposed correlation approximation is used, the ratio of $\sigma_{DD,Early}/\sigma_{DD,Layout}$ is closer to the unit, which means that the proposed correlation approximation provides a good estimation of the $\sigma_{DD,ij}$ between two paths. It is also evident that using correlation bounds is over-pessimistic, which may result in a significant number of paths being unnecessarily monitored. Figure 4.19 shows, for some ISCAS, the overestimation of the number of paths selected using correlation bounds to estimate $\sigma_{DD}$ early in the design flow with respect to the case when the proposed correlation approximation is used. The selected paths are overestimated in up to 70X times, which may represent a larger area overhead due to extra aging sensors and online delay test effort overhead.

Table 4.3 shows a detailed comparison of the critical path set obtained with the proposed methodology against the critical path set obtained using circuit layout information. For both cases, path were selected under 60 different randomly generated workload profiles. Spatial correlation has been extracted from the final circuit layout using the model of [2].

Table 4.3 shows that there are some wrong and missed selected paths. A path is said to be a wrong path (missed path) if it was selected (unselected) using the spatial correlation approach and unselected (selected) using spatial correlation extracted from circuit layout. As mentioned before, the spatial correlation approximation seeks to maximize $\sigma_{DD}$ to increase critical paths coverage. Although the proposed approach may select more paths than those needed, it should be noted that it is much more realistic than over-pessimistic approaches using correlation bounds and/or worst-case aging assumptions. As shown in Table 4.3,
there were selected some wrong paths, which may increase periodical test effort and circuitry area. However, few missed paths that remain without surveillance are obtained for some circuits. Missed paths are obtained because the selection probability of a path computed using the spatial correlation approach underestimates the selection probability of the path computed using spatial correlation extracted from the Layout. Figure 4.20 shows a histogram of the selection probability computed with spatial correlation of layout for each of the missed paths of all the ISCAS circuits. Most of the missed paths have low probabilities to become critical (less than 2%), and they may not represent a major reliability concern for the circuit. Only few missed paths got larger probabilities than 2%. Column 6 of Table 4.3 shows the maximum selection probability that a path took.
4.6. EXPERIMENTAL RESULTS ON ISCAS CIRCUITS

Figure 4.19: Overestimation of critical paths selected using correlation bounds with respect to the critical paths selected using the proposed correlation approximation.

Table 4.3: Comparison of CPs selected using spatial correlation against those selected using spatial correlation extracted from circuits layout ($\varepsilon = 0.25\%$)

<table>
<thead>
<tr>
<th>ISCAS</th>
<th>CPs Correlation Approach</th>
<th>CPs Correlation from Layout</th>
<th>Wrong Paths</th>
<th>Missed Paths</th>
<th>Missed Probab. (%)</th>
<th>Penalty GB (ps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>c432</td>
<td>1653</td>
<td>1100</td>
<td>553</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>c880</td>
<td>360</td>
<td>216</td>
<td>144</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>c1908</td>
<td>2599</td>
<td>1520</td>
<td>1079</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>c2670</td>
<td>109</td>
<td>83</td>
<td>26</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>c3540</td>
<td>616</td>
<td>190</td>
<td>428</td>
<td>2</td>
<td>0.32</td>
<td>3.70 (1.05%)</td>
</tr>
<tr>
<td>c5315</td>
<td>806</td>
<td>370</td>
<td>436</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>c7552</td>
<td>763</td>
<td>464</td>
<td>329</td>
<td>30</td>
<td>0.94</td>
<td>7.60 (3.48%)</td>
</tr>
<tr>
<td>s838</td>
<td>36</td>
<td>27</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s1423</td>
<td>16</td>
<td>8</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s5378</td>
<td>268</td>
<td>97</td>
<td>171</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s9234</td>
<td>69</td>
<td>59</td>
<td>34</td>
<td>24</td>
<td>6.81</td>
<td>58.04 (21.2%)</td>
</tr>
<tr>
<td>s13207</td>
<td>1029</td>
<td>198</td>
<td>831</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s35932</td>
<td>354</td>
<td>187</td>
<td>191</td>
<td>24</td>
<td>0.72</td>
<td>10.93 (8.85%)</td>
</tr>
</tbody>
</table>

among the missed paths for each circuit. For circuits C3540, C7552 and s35932, this probability is below 1%. Only for circuit s9234, the selection probability of a missed path was up to 6.26%. However, this corresponded to only 1 specific path in the circuit. In order to guarantee reliable operation of the missed paths, minor modifications could be done to the final circuit design. For example, gate sizing could be applied to the small set of missed paths. Another approach is to
add a small guardband penalty to the clock period. Column 7 of Table 4.3 shows the guardband penalty that needs to be added to the clock period for the circuits with missed paths. This guardband penalty is significantly smaller than the nominal guardband to cover aging and process variations of the conventional designs. The percentage of required guardband penalty compared to nominal guardband in conventional designs are shown in parenthesis. In the worst case, the guardband penalty is 21.2% of the nominal guardband for circuit s9234. For the other circuits having missed paths, the guardband penalty is smaller. These results demonstrate that the proposed methodology provides efficient critical path selection at early stages of the design flow, which allows reliable online delay monitoring and clock frequency scaling.

Figure 4.20: Histogram of the probability to become critical of the missed paths of all ISCAS circuits.

Once the critical paths to be monitored are identified, aging sensors (See Figure 4.1(a)) are inserted at the output nodes to keep under surveillance circuit delay degradation. Only one single aging sensor per group of those selected paths converging at the same output node is required. This minimizes the total number of aging sensors required per number of critical paths. Table 4.4 shows a comparison of the number of nodes where aging sensors should be placed according to the
4.6. EXPERIMENTAL RESULTS ON ISCAS CIRCUITS

selected paths using the proposed correlation approximation and using correlation obtained from circuit layout. The total number of outputs of each ISCAS circuit is also given. Only few conventional FF require to be modified to introduce the aging sensor. However, the number of nodes to be monitored may increase for large industrial circuits, where several paths having balanced delays may become critical. As can be observed, the number of aging sensors match well for most of the circuits. There are only three circuits where a different number of sensors was obtained. In circuit C7552, 18 and 17 nodes were selected using the proposed correlation and the correlation information extracted from circuit layout, respectively. There were two wrong nodes (due to the wrong paths) and only one missed node (due to the missed paths). In circuit s5378 there was only one wrong node. The largest difference on the selected nodes is in circuit s35932, where 21 missed nodes were obtained. This large difference in the number of selected nodes is because each missed path is isolated with respect to the other paths in the circuit. It should be noted that circuit s35932 is the one with the largest number of output nodes. As can be seen, the missed nodes in circuit s35932 closely correspond to the number of missed paths. As commented before, the missed paths have in fact low probability to become critical and can be covered with little modifications to the final circuit.

<table>
<thead>
<tr>
<th>ISCAS</th>
<th>c432</th>
<th>c880</th>
<th>c1908</th>
<th>c2670</th>
<th>c3540</th>
<th>c5315</th>
<th>c7552</th>
<th>s838</th>
<th>s1423</th>
<th>s5378</th>
<th>s9234</th>
<th>s13207</th>
<th>s35932</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output Nodes</td>
<td>7</td>
<td>26</td>
<td>25</td>
<td>50</td>
<td>22</td>
<td>123</td>
<td>108</td>
<td>33</td>
<td>79</td>
<td>203</td>
<td>234</td>
<td>714</td>
<td>1760</td>
</tr>
<tr>
<td>Correlation Approach</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>18</td>
<td>4</td>
<td>1</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>32</td>
</tr>
<tr>
<td>Correlation from Layout</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>17</td>
<td>4</td>
<td>1</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>53</td>
</tr>
</tbody>
</table>

4.6.3 Comparison with Other Works

A comparison of the main features of the proposed methodology with other state-of-art methods is shown in Table 4.5. The approach in [134] does not consider
4.7. CONCLUSIONS

A methodology to select the critical paths of a circuit for reliable Dynamic Frequency Scaling under BTI-aging and process variations effects have been proposed.
4.7. CONCLUSIONS

The statistical selection condition assures that those paths likely to fail first over the lifetime are included in the critical path set for reliable application of DFS. The proposed methodology addresses three major challenges for path selection: 1) Critical Path statistical behavior due to process variations, 2) Critical Path dependence on workload profile, which is rarely known in advance at the design phase, and 3) Critical Path reordering probability over the time due to aging. The proposed methodology can be applied at an early design phase as it uses a spatial correlation approach to performing SSTA. Hence no detailed layout information is required to extract spatial correlation between gates. It has been shown that a trade-off can be established between the computational time and the critical paths set identified for various operating workloads scenarios. Furthermore, it was shown that most of the critical paths are identified by analyzing a moderate number of workload profiles. A spatial correlation approach was used to perform aging-aware SSTA early in the design flow, where the critical paths need to be identified. Few missed paths are obtained compared to an approach using spatial correlation information extracted from circuit layout. However, those few missed paths, which are unlikely to become critical, could be covered with little effort using design gate sizing or adding a small guardband penalty to the clock period. Additionally, it has been shown that the number of critical paths selected with the proposed approach is significantly lower than that selected using worst-case aging scenarios. Hence, the proposed approach further mitigates online test effort.
Chapter 5

Guardband Reduction by Efficient Selection and Sizing of Critical Gates

Guardbands added to the clock period to assure reliable circuit operation are becoming unacceptably large as devices continue to shrink, due to the increasing impact of aging and process variations effects. This chapter presents a methodology to reduce the guardband of a circuit towards a desired target guardband by efficient selection and sizing of critical gates considering BTI aging and PV effects. The proposed approach uses metrics to identify those gates providing efficient guardband reduction with as small as possible area overhead.

5.1 Related Work

Various aging-aware design techniques already exist in the literature. Gate-level size optimization is a widely used approach to address aging and process variations issues. This method tries to size the gates to achieve optimal trade-offs between delay, area, and reliability. This requires a method to find and re-size
the most beneficial gates. In [91], the design optimization of a full-adder circuit based on extensive SPICE simulations was presented. However, SPICE-based reliability analysis is computationally unfeasible for large-scale integrated circuits. The works [78, 82], built gate libraries robust to aging by sizing the transistors in a gate according to the duty cycle that the devices experience. However, these approaches require more detailed guidance to determine where to place the robust gates within a circuit as the gates having more delay degradation may not be the most influential to overall circuit timing. In [149], it is proposed to increase the size of the gates in the critical paths of the circuit and having a delay degradation larger than a given threshold (i.e., 5%). The proposed approach takes into account the maximal load capacitance that a gate of a given size can drive. However, not all the gates in the critical paths have the same impact on circuit delay degradation. Therefore, they should be treated differently. In [89, 92] an optimization problem minimizing circuit area for a given delay constraint under aging is formulated and solved using Lagrangian relaxation. However, these methods may become complex for large circuits, especially if process parameters variations are considered.

The concept of gate criticality metrics under aging effects was introduced in [142]. Gate criticality metrics provide a fast estimation of how efficiently the delay degradation of the circuit can be improved at a given area cost. Thus, designers can optimize the circuit by increasing (decreasing) the size of the gates with higher (lower) metric score. Different gate metrics have been proposed in [44, 133, 135, 142]. These metrics combine parameters representing the nominal delay (with no-PV and no-BTI) sensitivity with respect to gate sizing, the number of critical paths impacted by the gate, and the delay degradation of a gate. Gates are selected based on the metric score ranking, and design actions take place on the selected gates. In [133, 142], the selected gates are replaced by their aging-robust counterparts from an aging-aware gate library (such as that in [82]). In [44], [135], the size of the gates with the highest metric score is iteratively increased until the
desired timing constraint is met. However, area saving by size-decreasing in gates with little impact on delay is not considered. Also, the used metrics do not consider the impact of sizing a gate on both the degradation and the standard deviation of the paths delay (under PV), which may limit the efficiency of the gate selection and sizing process.

BTI-induced delay degradation strongly depends on the workload executed by the circuit as it defines the duty cycle of each device in the circuit. Unfortunately, the exact operating workload executed by a circuit over the lifetime is unpredictable and hardly known in advance at the design phase. Therefore, a major limitation of the aforementioned aging-aware optimization approaches is that they either assume worst-case duty cycle or a specific duty cycle profile at main circuit inputs for aging estimation. While the first approach leads to conservative designs with excessive area overhead, the second approach may not be reliable enough if the actual signal probabilities that the circuit experiences during normal operation differ from those used at the circuit design phase. Recently, a sizing approach considering the distribution of paths delay degradation for various different duty cycle profiles has been proposed in [141]. However, degraded path delays are optimized for the mean value of the expected degradation over the set of feasible operating workloads, which do not guarantee reliable operation. Furthermore, the effect of process variations is not considered.

The contributions of this thesis for reliable circuit design are:

1. New statistical gate sizing metrics are proposed. The metrics include the impact of gate sizing on BTI delay degradation and standard deviation of delay reduction. A fast approximation of the statistical delay sensitivity of a path with respect to sizing of a gate is proposed. The sizing process considers gate sizing-up to improve delay and gate sizing-down to mitigate area overhead.

2. A multiple workloads aware sizing algorithm is proposed. The paths delays
are estimated for various different signal probabilities scenarios at main inputs. Then, the paths are optimized for the signal probability scenario that causes largest delay degradation. In such way, a more accurate estimation of the maximal paths delay degradation is made.

5.2 Considerations for Gate Sizing Metrics

5.2.1 Sizing Impact on Gate Delay

When the size of a gate is changed, its timing behavior is modified. The statistical delay of a gate has three components: the nominal mean delay, the mean delay degradation due to BTI and the standard deviation of the delay due to joint effects of BTI and process variations (See Equation 5.1). These delay components are obtained from the linear gate delay model 3.6.

\[
D_i = N(\mu, \sigma) = N(\mu_{Dn} + \mu_{\Delta D}, \sigma_D)
\]  \hspace{1cm} (5.1)

Figure 5.1 shows the initial (nominal) delay, the gate delay degradation due to BTI after 10 years of aging, and the standard deviation of the gate delay of an inverter gate as a function of the gate size. Nominal delay reduces because transistors current capability to charge or discharge capacitive loads increases, while gate delay degradation reduces because gate delay becomes less sensitive to $V_{th}$ instability. Similarly, standard deviation reduces as the gate delay is less sensitive to process parameters variations.

An important parameter measuring the impact of sizing a gate on its timing behavior is the statistical delay sensitivity, which measures the change in the gate delay per unit of gate size increase. From Figure 5.1, it is concluded that statistical delay sensitivity to gate sizing ($S_D^K$) is composed of (See Equation 5.2):

1. A nominal delay sensitivity component ($S_{Dn}^K$).
5.2. CONSIDERATIONS FOR GATE SIZING METRICS

2. A delay degradation sensitivity component ($S_{K}^{\Delta D}$).

3. A standard deviation sensitivity component ($S_{K}^{\sigma D}$).

\[ S_{K}^{D} = S_{K}^{Dn} + S_{K}^{\Delta D} + S_{K}^{\sigma D} \]  

(5.2)

Figure 5.1: Nominal delay, Delay degradation, and Standard Deviation of the delay of an inverter gate as function of its size.

It should be highlighted that conventional sizing approaches [44,133,135,142] do not consider the two last sensitivity components regarding the impact of sizing a gate on the delay degradation and on the standard deviation of the delay. Including these effects in the statistical delay sensitivity allows to select more efficiently the best gates to be sized.

5.2.2 Sizing Impact on Path Delay

When the size of a gate is increased, it does not only impact on the gate timing itself, but also on the timing of the adjacent gates [92].

As illustrated in Figure 5.2, when the size of a gate within a path is increased, its input capacitance increases, thus impacting of the delay of its preceding gate in the path. Similarly, the output transition time of the gate reduces due to its
5.3 Proposed Gate Sizing Metrics

Gate selection metrics are proposed to guide the optimization process. The gate selection metrics are intended to identify the best critical gates to be sized in the critical paths to improve circuit guardband efficiently.

5.3.1 Statistical Path Delay Sensitivity

The statistical path delay sensitivity to gate sizing measures how much impact has sizing a gate on the path delay. Therefore, it provides information of the most efficient gate to be sized to improve the statistical delay of a path.

Knowing the mean and standard deviation for a given aging time for all the gates in a path using Equation 3.6, the path delay PDF \( D_p = N(\mu_{D_p}, \sigma_{D_p}) \) is

Figure 5.2: Impact of sizing a gate on the adjacent gates
5.3. PROPOSED GATE SIZING METRICS

obtained by:

\[
\mu_{D,p} = \sum_{i=1}^{N} \mu_{D_i} = \mu_{D_{n,p}} + \mu_{\Delta D,p} \tag{5.3a}
\]

\[
\sigma_{D,p} = \sqrt{\sum_{i=1}^{N} \sum_{j=1}^{N} \rho_{ij} \cdot \sigma_{D_i} \cdot \sigma_{D_j}} \tag{5.3b}
\]

where \(\mu_{D_i}\) is the mean delay of the gate \(i\) for the given aging time, \(\sigma_{D_i}\) and \(\sigma_{D_j}\) are the standard deviation of the gate delay \(i\) and \(j\), respectively. The parameter \(\rho_{ij}\) corresponds to the correlation between gate delays. Note that the mean value of path delay has a nominal component and a component due to aging effects. Also, the standard deviation of the paths depends on aging effects, as the threshold voltage variability changes due to aging, as was shown in Equation 3.3.

The statistical path delay sensitivity is defined as the derivative of the \(\mu + 3\sigma\) of the path delay distribution to sizing a gate \(i\) in the path:

\[
S_{K_i}^{D_p} = \frac{\partial \mu_{D_p}}{\partial K_i} + 3 \frac{\partial \sigma_{D_p}}{\partial K_i} \tag{5.4}
\]

\[
= \left[ \frac{\partial \mu_{D_{n,p}}}{\partial K_i} + \frac{\partial \mu_{\Delta D,p}}{\partial K_i} \right] + 3 \cdot \frac{\partial \sigma_{D_p}}{\partial K_i}
\]

where \(K_i\) is the size of the gate \(i\) in the path, \(\mu_{D_p}\) and \(\sigma_{D_p}\) are the mean value and the standard deviation of the aged path delay obtained from Equations 5.3a and 5.3b, respectively. \(\mu_{D_{n,p}}\) and \(\mu_{\Delta D,p}\) correspond to the mean value of the nominal (fresh) path delay and the mean value of the delay degradation of the path.

As can be seen from Equation 5.4, three components influence the statistical path delay sensitivity to sizing a gate. The component related to the nominal delay (no aging and no PV), the component related due to aging effects and the component related to process variations. Figure 5.3 shows the magnitude of the components of the statistical path delay sensitivity to gate sizing for the path.
5.3. PROPOSED GATE SIZING METRICS

Figure 5.3: Example of the magnitude of the components of the statistical path delay sensitivity to gate sizing (Equation 5.4).

example shown in the inset Figure. As can be observed, the component related to the nominal path delay is the largest. However, the components due to the impact of aging on the mean delay and the impact of process variations are also significant. It is worth to mention that the aging component depends on the degradation of the gate. A gate whose devices has a larger aging also exhibit a larger $\frac{\partial \mu_{D_{\text{nd}}}}{\partial K_i}$.

It is also important to note that spatial correlation plays an important role in the magnitude of the derivative of the standard deviation of the path delay with respect to the size of a gate in the path. Figure 5.3 shows two cases: when all the gates in the path are placed far away, and their spatial correlation is almost zero ($\rho = 0$), and the case when all the gates are placed very close to each other, having a full spatial correlation ($\rho = 1$). Therefore, it can be concluded that those gates that have a higher correlation with the other gates in the path may be preferable to optimize.

Approximation of the Statistical Delay Sensitivity of a Path

The computation of Equation 5.4 may be computational costly as it implies to evaluate the statistical distribution of the aged path delay for both the current size of the gate and when the size of the gate is changed by a small perturbation.
5.3. PROPOSED GATE SIZING METRICS

(this is done to compute the derivatives). Therefore, for a path with N gates, the statistical path delay would have to be computed $N + 1$ times to compute the statistical delay sensitivity with respect to the size of each gate, which is unpractical. Therefore, we propose some simplifications to evaluate Equation 5.4 efficiently.

Figure 5.4: A path example to illustrate the impact of sizing a gate on its neighboring gates in the path.

Figure 5.4(a) shows a path where the size of the gate $i$ is increased in an small step. Figure 5.4(b) shows the change in the mean value and standard deviation of the delay of each gate in the path due to the small increase in the size of the gate $i$. As can be seen, only the timing response of the gates $i - 1$, $i$, and $i + 1$ are significantly affected. Both the mean and standard deviation of the gate $i - 1$ increases due to the larger input capacitance of the sized gate. On the other hand, the mean and standard deviation of the gate $i + 1$ reduces because of its input signal switches faster as gate $i$ becomes stronger. Obviously, the mean value and the standard deviation of the delay of the sized gate are the most reduced when the size of this gate is increased. It should be noted that the change in the standard deviation of the delay of a gate is much smaller than the change in the mean value, as was observed before in Figure 5.3. Based on above observations the following approximations are made:
5.3. PROPOSED GATE SIZING METRICS

Sensitivity of the mean of the path delay to gate sizing

The change in the mean delay of a path is mainly due to the change in the mean delay of the gates in the segment of the gate \(i\). Therefore, we approximate the sensitivity of the mean delay of a path to gate sizing (First term of Equation 5.4) as the sum of the sensitivities of the gates in the segment of the gate:

\[
\frac{\partial \mu_D}{\partial K_i} \approx \frac{\partial \mu_{D,i-1}}{\partial K_i} + \frac{\partial \mu_{D,i}}{\partial K_i} + \frac{\partial \mu_{D,i+1}}{\partial K_i} 
\]

(5.5)

where \(\mu_{D,i-1}, \mu_{D,i}\) and \(\mu_{D,i+1}\) are the aged delays of the gates \(i-1, i\) and \(i+1\) in the path segment of the gate being analyzed, \(CL_{i-1}\) is the load capacitance of the gate \(i-1\), \(C_{in,i}\) is the input capacitance of gate \(i\), \(SRI_{i+1}\) is the signal transition time at input of gate \(i+1\) and \(SRO_i\) is the signal transition time at output of gate \(i\), which is equal to \(SRI_{i+1}\).

Note that by using this approximation only the mean delay of the gate segment needs to be recomputed rather than the mean delay of the entire path.

Sensitivity of the standard deviation of the path delay to gate sizing

The change in the standard deviation of the delay of a path due to sizing a gate \(i\) is mainly due to the change of the standard deviation of the delay of the gate \(i\) and its impact on the covariance with the other path gates. Therefore, we approximate the sensitivity of the standard deviation of the path delay to gate sizing (second term of Equation 5.4) as follows:
5.3. PROPOSED GATE SIZING METRICS

\[
\frac{\partial \sigma_{D,p}}{\partial K_i} = \frac{1}{2\sqrt{\sigma_{D,p}^2}} \cdot \frac{\partial \sum_{i=1}^{N} \sum_{j=1}^{N} \rho_{ij} \cdot \sigma_{Di} \cdot \sigma_{Dj}}{\partial K_i}
\]

\[
\approx \frac{1}{2\sigma_{D,p}} \cdot \left( \frac{\partial \sigma_{Di}^2}{\partial K_i} + 2 \sum_{j \neq i}^{N} \rho_{ij} \cdot \sigma_{Dj} \right) \\
\approx \frac{1}{\sigma_{D,p}} \cdot \left( \frac{\partial \sigma_{Di}^2}{\partial K_i} \sum_{j=1}^{N} \rho_{ij} \cdot \sigma_{Dj} \right)
\]

As can be observed, the sensitivity of the standard deviation of the path delay depends on the spatial correlation that the sized gate \( i \) has with each other of the gates in the path. As shown before, when a gate is highly correlated with the others, its efficiency for reduction of the standard deviation of the path delay increases. Note that Equation 5.6 only depends on the derivative of the standard deviation of the size of the gate itself. Therefore, to evaluate Equation 5.6 only the standard deviation of the gate of interest \( i \) needs to be recomputed.

5.3.2 Proposed Sizing Metrics

Equations 5.5 and 5.6 measure the gate-sizing induced change on the mean and the standard deviation of the delay at a path level. However, the best gates to be sized need to be selected at the circuit level. For this reason, the statistical path sensitivity is combined with other circuit parameters related to the impact of sizing a gate at the circuit level.

Area Impact

Circuit area is a very important constraint in the design of digital circuits because it is proportional to power consumption and it affects the cost of a product. It is desirable to design circuits meeting delay and lifetime constraints at the cost of minimum area overhead. Therefore, it is important to consider the area cost of sizing a gate, in the metrics formulation.
We can define the Area Impact of sizing a gate as the gate area increase per unit of the size of the gate. The area impact of a gate depends on the cell layout. For instance, consider the layout of an inverter gate shown in Figure 5.5. The area of the cell is the product of the height and the width \( (A_i) \) of the cell,

\[
Area_{gate} = K \cdot (W_N + W_P) \cdot (A_i)
\]  

(5.7)

The height of the cell depends on the channel width assigned to NMOS and PMOS devices for minimum size symmetrical cell and the gate size, which proportionally scales the transistors dimensions. While the height of the cell is design-dependent, the width of the cell depends on the layout-rules only, which are given by the technology. As shown in Figure 5.5, the design rules affecting the height of a cell are the Active area extension (AEX), the contacts width (CW), the Poly to Diffusion Contact distance (PDC), and the channel length.

When the size of a gate \( (K) \) is increased, the area increases linearly, as shown in Figure 5.6. The change in circuit area is given by the slope of the lines for each gate, which depends on the minimum channel width assigned to PMOS and NMOS and the width of the cells. As can be observed, increasing the size of an
inverter gate has a lower impact on the area, while increasing the size of NOR gates has the highest area cost. The area impact of sizing a gate should be taken into account in the metrics.

Based on this observation, in the proposed metric the statistical delay sensitivity is normalized with respect to the gate area impact \( A_i \), which is a layout dependent parameter but can be estimated for a pre-designed gate library. By normalizing gate delays sensitivities \( S^D_{K}/A_i \), the metric considers the gate delay reduction per unit increase of area.

**Gate Criticality**

Gate criticality can be considered as the number of critical paths passing through a gate [142]. Critical paths of a circuit, are those paths that do not meet the timing constraint. Let us consider the circuit with two critical paths \( P_1 \) and \( P_2 \) shown in Figure 5.7. If we resize a non-shared gate in one path (e.g., gate G1), it would be required to resize a gate in the other path. In this case, two gates would be resized to meet timing constraint. However, if a shared gate (e.g., G4 or G5) is resized, only one gate impacts the area cost. Shared gates between paths present a higher criticality and are topologically preferred for size optimization.
5.3. PROPOSED GATE SIZING METRICS

Path Slack

The aged delay of a critical path depends on its number of gates and their rate of degradation [8]. The amount of path delay degradation and its number of gates determine the severity of a path to violate the timing constraint. Therefore, we use aged path slack time to take into account these effects. The slack time of a path $p$ is defined as the difference between the path delay and the delay constraint:

$$ slack_p = D_{cons} - (\mu_{D,p} + 3 \cdot \sigma_{D,p}) $$  \hspace{1cm} (5.8)

where $D_{cons}$ is the delay constraint, $\mu_{D,p}$ is the mean path delay and $\sigma_{D,p}$ is the standard deviation of the path delay. Then, the path slack is positive if the path violates the timing constraint and negative if it meets the desired constraint.

Metrics Formulation

The statistical delay sensitivity of a path, reveals which gate in the path has a larger impact on the $\mu + 3\sigma$ delay of the path. This parameter is combined with other important information of the gates to form the proposed sizing metrics.

Two gate sizing metrics are proposed to guide the optimization process: One that measures the benefit of sizing-up a gate, and other that measures the benefit of sizing-down a gate. For each gate $i$, the two following sizing metrics (See Equation 5.9) are evaluated:
5.3. PROPOSED GATE SIZING METRICS

\[ M_{SU,i} = \frac{S_{Ki,AVG}^D \cdot |\text{Slack}_{Ki,AVG}^+| \cdot N_i}{\Delta A_i} \quad M_{SD,i} = \frac{\text{Slack}_{Ki,AVG}^- \cdot \Delta A_i}{S_{Ki,AVG}^D \cdot N_i} \] (5.9)

where \( M_{SU,i} \) and \( M_{SD,i} \) are the sizing-up and sizing-down metrics, respectively. \( S_{Ki,AVG}^D \) is the average delay sensitivity of the \( N_i \) paths passing through the gate to change in the gate size \((Ki)\), \( \text{Slack}_{Ki,AVG}^- \) is the average slack of these paths, and \( \Delta A_i \) is the area impact of sizing the gate, which depends on the cell layout. Each metric is evaluated for a different path set. For sizing-up, the metric is evaluated within the paths with positive slack (violating the delay constraint). Thus, \( \text{Slack}_{Ki,AVG}^- \) takes a positive value. On the other hand, \( \text{Slack}_{Ki,AVG}^- \) takes a negative value for the sizing-down metric. The value of \( N_i \) is also different depending on the metric that is being evaluated.

The parameters used in the metric provide topological, timing and area-related information. The metric score determines the delay-area trade-off of sizing a gate. The sizing-up metric score increases for gates influencing many paths since they allow to improve various paths at a time. The sizing-up metric score also increases for those gates in slow paths with large negative slacks as those paths should be optimized with higher priority. A large delay sensitivity also increases the sizing-up metric score as a large delay reduction can be obtained by increasing the gate size. Finally, the sizing-up metric score reduces for gates with a high area impact because increasing the size of those gates is area costly. A similar interpretation of the parameters is made for the sizing-down metric. In this case, the sizing-down metric score increases for those gates affecting few paths and with low delay sensitivity to gate sizing (low impact on delay) and large positive slack. Also, gates with a large area impact are preferred due to potential area savings when down-sizing the gate.
5.4 Methodology for Efficient Selection and Sizing of Critical Gates

The proposed framework for efficient critical gate selection and sizing consists of the three principal steps shown in Figure 5.8. In the first step, the Potential Critical Paths (PCPs) of the circuit are identified using Worst-case aging assumptions. A PCP is a path whose delay may exceed the nominal delay of the circuit due to the combined effect of aging and process variations. These paths impose some guardband for reliable circuit operation. The second-step performs a multiple workload-aware aging analysis of the PCP set. In this step, a realistic maximum delay degradation of the PCPs is estimated for a finite number of Signal Probability combinations at main inputs (representing a possible executed workload). In such way, uncertainty about the specific signal probability profile of the circuit is addressed. In the third step, the delay of the PCPs is improved using a metric-guided statistical size optimization technique. The result of this flow is an optimized design that meets a given target guardband with low area overhead.

Figure 5.8: Flow of the proposed gate sizing optimization methodology.
5.4. METHODOLOGY FOR EFFICIENT SELECTION AND SIZING OF CRITICAL GATES

5.4.1 PCP Identification Under Worst-Case Aging

Aging-Aware Statistical Static Timing Analysis (SSTA) is run assuming worst-case aging conditions. The devices in the circuit are assumed under near-static stress ($\alpha \approx 1$) and operating under a high temperature ($T = 120^\circ C$). Those paths whose $\mu + 3\sigma$ corner of the aged delay distribution is greater than the nominal (without aging and PV) delay of the circuit are identified as PCP. These paths impose some guardband for reliable operation.

The identification of PCPs under worst-case aging conditions allows focusing our optimization flow in only a restricted path set rather than in the entire circuit, allowing to minimize computational effort.

5.4.2 Multiple Workloads Aging Analysis

One major challenge to be addressed for aging-aware circuit design is that the exact workload that the circuit will experience over the lifetime is hardly known in advance at circuit design phase. The workload impacts on the stress duty cycle ($\alpha$) of each device and on their operating temperature [140], [124], which in turn influence BTI degradation, complicating circuit reliability analysis and optimization.

To address unpredictability of the circuit workload at the design phase, we refine the aging conditions at which the delay of each PCP is evaluated during design optimization by performing a *Multiple Workload Aware Maximum Aging Analysis*. The main idea of this step is to determine a realistic maximum upper bound for the delay degradation of each PCP. This is done by evaluating the delay degradation of the previously selected PCPs for various workload profiles. In order to speed-up the computational time, the following actions are made:

1. Only the mean value of the aged delay of each PCP is computed for each workload.
5.4. METHODOLOGY FOR EFFICIENT SELECTION AND SIZING OF CRITICAL GATES

2. If the cumulative maximum mean aged delay of a path does not increase for a given number of consecutive workloads, the actual path delay is taken as maximum and the path is no longer updated for the next workloads.

Algorithm 2 describes the proposed approach. For a user-defined number of workloads profiles, a set of signal probabilities at main circuit inputs are generated and propagated to internal nodes. A uniform random number generator between 0 and 1 is used to obtain the signal probability assigned to each input. Then, the stress duty cycle ($\alpha$) of each transistor in the circuit is computed. The operating temperature of each cell is also computed as it strongly influences BTI mechanism. For each path, only the mean value of the delay degradation is computed. This is done to avoid the excessive computational cost of updating the PCPs timing degradation statistically for each workload profile. If the obtained delay degradation of a PCP $j$ ($\mu_{\Delta D_{j,WL}}$) is greater than the maximum obtained with previous workload profiles ($\mu_{\Delta D_{j,MAX}}$), the maximum aged delay of the path is updated and the conditions of duty cycle and temperature of path’s devices causing such maximum degradation for the path are stored. If the maximum delay degradation of a path does not increase within a consecutive user-defined number (N) of workload profiles, a flag signal (PCP[j].MAX) is activated, indicating that the currently stored delay degradation for the path is a good enough estimation for the maximal path delay degradation. Then, this path is no longer evaluated for the subsequent workload profiles. This is done to mitigate the computational cost associated with this step. It is important to note that the signal probability profile that causes maximum path delay degradation can be different for each path.

Once the conditions that cause maximum degradation for each path have been identified, the standard deviation of the path's delay is computed, and the set of PCPs is reduced by discarding those paths whose maximum aged at the $\mu + 3\sigma$ corner do not exceed the nominal circuit delay. This process mitigates the computational effort required for design optimization.
5.4. METHODOLOGY FOR EFFICIENT SELECTION AND SIZING OF CRITICAL GATES

**Algorithm 2**: Multiple Workload-Aware Maximum Aging Analysis

**Input**: Potential Critical Paths, $Max_{WL}$

**Output**: $\alpha$ and $T$ of PCP gates causing worst aging

for $WL = 1$ to $WL = Max_{WL}$ do

1. generate propagate SP()
2. compute duty cycle()
3. compute temperature()
4. compute $\Delta V_{th_{BTI}}$

for $j = 1$ to $j = Num.PCPs$ do

1. if $PCP[j].MAX == 0$ then
2. Compute $\mu_{\Delta D_j,WL}$
3. if $\mu_{\Delta D_j,WL} > \mu_{\Delta D_j,MAX}$ then
4. $\mu_{\Delta D_j,MAX} = \mu_{\Delta D_j,WL}$
5. Save $\alpha$ and $T$ of each device in PCP $j$
6. else
7. if $\mu_{\Delta D_j,MAX}$ has not increased in the last N profiles then
8. $PCP[j].MAX = 1$
9. $PCP j$ is not evaluated for next WLs
10. end
11. end
12. end
13. end
14. end

**Figure 5.9**: Maximum delay degradation of some paths as function of the number of tested workload profiles.

Figure 5.9 shows the behavior of the cumulative maximum delay degradation obtained for some paths of circuit C1908 as a function of the number of tested (sampled) workload profiles. As can be seen, after some workload profiles are tested, the maximum delay degradation obtained for all the paths tend to saturate.
This behavior suggests that only a moderated number of workload profiles need to be analyzed to get a good estimation of the maximum aged delay that a path can take.

Once identified the realistic conditions for maximal delay degradation of each path, the circuit is optimized to guarantee reliable operation of the PCP set even under such conditions.

5.4.3 Selection and Sizing of Critical Gates

Iterative selection and sizing of critical gates is done to optimize the circuit design for a reduced guardband constraint. In such way, lifetime reliability can be assured with low area overhead.

Guardband Computation

The first step in the gate sizing procedure (See Figure 5.8) is to compute the actual guardband required for reliable circuit operation. The time-guardband that each PCP impose over the nominal circuit delay ($GB_p$) is given by,

$$GB_p = (\mu_{Dp} + 3\sigma_{D,p}) - D_{nom}$$  (5.10)

where $\mu_{Dp}$ and $\sigma_{D,p}$ are the mean value and the standard deviation of maximum aged delay of PCP $p$, and $D_{nom}$ is the nominal circuit delay without aging. Due to the combined impact of aging and process variations, these guardbands may be too large, and if one path exceeds the target guardband $GB_t$, it may perform a faulty operation before the lifetime ends.

Identification of Fast and Slow PCPs

The PCPs are then separated in two different subsets depending on the corresponding guardband imposed by each path, as illustrated in Figure 5.10: a) Slow-PCPs subset, which has negative slack ($GB_t - GB_p < 0$); and b) Fast-PCPs
5.4. METHODOLOGY FOR EFFICIENT SELECTION AND SIZING OF CRITICAL GATES

subset, which has some positive slack ($GB_t - GB_p > 0$). This classification is done to exploit the fact that different design actions can be taken over each PCP subset: Some gates are sized-up in the Slow-PCPs to improve their delay, while some gates are sized-down in the Fast-PCPs to take advantage of their slack to mitigate area overhead.

Gate Selection and Sizing Heuristic

The gate selection metrics are evaluated to identify the best critical gates to be sized in each critical path subset. The sizing-up metric is evaluated within the Slow-PCPs to identify the most beneficial gates to improve their timing efficiently. The sizing-down metric, which identifies those gates with little impact on delay and belonging to paths with enough slack, is evaluated within the Fast-PCPs to mitigate area overhead with little negative impact on delay.

Algorithm 3 summarizes the sizing procedure using the following heuristics to select the best gates to size up/down using the sizing-metrics scores.

Sizing-Up Selection: The obtained sizing-up metric score $M_{SU,i}$ reflects the benefit of Slow-PCPs delay reduction vs. area trade-off established by each gate. Thus, the N gates with the highest $M_{SU,i}$ are picked and size-up proportionally to their respective score: $\Delta K = step \cdot M_{SU,i}$. Where N is a user-defined number of gates that are sized at each iteration and step is the maximum size change that a
gate can take at an iteration.

**Sizing-Down Selection:** The obtained sizing-down metric score $M_{SD,i}$ reflects a trade-off between Fast-PCPs delay increase to area reduction. However, the interdependence between Fast-PCPs and Slow-PCPs must be considered for sizing-down gates. This is because sizing-down a gate having a high $M_{SD,i}$ may negatively impact on Slow-PCPs if the gate also has a high $M_{SU,i}$ score. Therefore, the two following conditions are applied to pick the gates to size-down:

- Gates sized-up is not allowed to be sized-down in the same iteration.
- Gates that have a sizing-up metric score ($M_{SU,i}$) larger than a constraint ($C_{M_{SU}}$) are not allowed to be sized-down.

The constraint in $M_{SU,i}$ is used to limit the negative impact on the slow-PCPs delay of sizing-down gates. The value of the constraint is dynamically changed along the sizing process. It is initially set to 1 (maximum) to maximize area savings as any gate is allowed to be sized-down, but it is gradually reduced if the delay of the Slow-PCPs is not being improved in a given iteration, so that delay converges towards the desired target delay. The N gates with the highest $M_{SD,i}$ score fulfilling the aforementioned conditions are sized-down according to the following rule: $\Delta K = -\text{step} \cdot (1 - M_{SU,i}) \cdot M_{SD,i}$. Thus, the amount of size reduction of a selected gate reduces (increases) if the gate has a high (low) $M_{SU,i}$ ($M_{SD,i}$) score.

In general, the size-down procedure is useful when the initial design has oversized gates due to a non-optimal design. Also, it becomes beneficial when a gate in a Fast-PCP is driven by a gate in a Slow-PCP. This may occur if the gate in the Fast-PCP was sized-up at the beginning of the optimization procedure (i.e., the gate was critical first), but then its influence to the remaining Slow-PCPs decreases.

Once the selected gates are sized, the PCPs timing information is updated (See Algorithm 3). The delay of each PCP is evaluated under the conditions


5.5. RESULTS FROM APPLICATION TO ISCAS BENCHMARK CIRCUITS

of temperature and duty cycle of the devices that caused maximum aged path delay. The gate selection and sizing process is iteratively repeated until the target guardband is achieved. *It is worth to mention that the parameters $N$ and step impact on the speed of the sizing process and the accuracy of the results.*

**ALGORITHM 3:** Sizing Heuristic

<table>
<thead>
<tr>
<th>Input:</th>
<th>Gates Metrics Scores ($M_{SU,i}$ and $M_{SD,i}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output:</td>
<td>Selected Gates with Updated Size</td>
</tr>
</tbody>
</table>

Set $C_{MSU} = 1$

Rank gates in Slow-PCPs according to $M_{SD,i}$

Rank gates in Fast-PCPs according to $M_{SD,i}$

// Size-up gates in Slow-PCPs:

for $i = 1$ to $N$ do

\[
K_i = K_i + \text{step} \cdot M_{SU,i}
\]

end

// Size-Down gates in Fast-PCPs:

for $i = 1$ to $N$ do

if $M_{SU,i} < C_{MSU}$ and $i$ was not sized-up then

\[
K_i = K_i - \text{step} \cdot (1 - M_{SU,i}) \cdot M_{SD,i}
\]

end

end

Update PCPs Timing Information

if $\max(\text{GB}_p)$ is not reduced then

\[
C_{MSU} = C_{MSU} - 0.1
\]

end

5.5 Results from Application to ISCAS Benchmark Circuits

The proposed gate sizing optimization technique for guardband reduction has been implemented in C++ code and applied to some ISCAS benchmark circuits designed using a 32nm Synopsys Generic Technology [65]. The original design of each circuit was of minimum area, where all the gates are of minimum size.

5.5.1 Sensitivity Approximation

Let us first analyze the accuracy of the proposed path statistical sensitivity approximation. For this analysis, the impact of sizing each gate in the $\mu + 3\sigma$ delay
5.5. RESULTS FROM APPLICATION TO ISCAS BENCHMARK CIRCUITS

Figure 5.11: Statistical delay sensitivity of a Path with respect to sizing each gate in the path. Longest path of C1908 circuit.

Data is shown for both the sensitivity obtained with the proposed derivative approximation and the exact derivative calculation, where the statistical delay of the path is completely re-computed when the size of each gate is perturbed. As can be observed, the proposed approximation follows well the derivative obtained with the exact computation. Therefore, the proposed approximation measures well the impact that sizing a gate has on the delay of a path.

5.5.2 Optimization Results

Table 5.1 shows detailed results obtained from the application of the proposed methodology to ISCAS 85/89 circuits. The second and third columns give the total number of paths and gates in the circuits. Circuits of different size and complexity were considered. Columns 4-7 show results related to the multiple workload-aware degradation analysis step. The column labeled as PCPs correspond to the number of Potential Critical Paths, which are those paths whose $\mu + 3\sigma$ delay can become greater than the nominal delay of the circuit. These paths are the ones considered during selection and sizing of the gates. As can
be observed, the number of PCPs does not depend on the total number of paths (i.e., the number of PCPs in c7552 and s1423 is very different, but these circuits have a similar number of paths). The number of PCPs changes depending on the susceptibility of each circuit to aging and the circuit topology. Column 5 gives the number of gates that compose the selected set of PCPs. These gates are called Critical Gates. The proposed heuristic identifies which critical gates are more beneficial to be sized accordingly to the obtained metrics score. Column 6 shows the initial guardband that would have to be added to the nominal delay to assure reliable operation of the non-optimized circuit under the combined effect of aging and process variations. As can be seen, the percentage of guardband needed can be up to 45% of the nominal delay, which may be unacceptably large for high-performance state-of-art designs. Column 7 shows the CPU time spent in the multiple workload-aware degradation analysis. This corresponds to the time for evaluating the PCP set for multiple workload profiles. However, it should be noted that the number of times each path is evaluated may be different depending on when it is detected that the maximal delay obtained for a path does not further increase as more workload profiles are analyzed.

Columns 8 to 13 of Table 5.1 show the results obtained from applying the proposed methodology for selection and sizing of critical gates to reduce the initial
5.5. RESULTS FROM APPLICATION TO ISCAS BENCHMARK CIRCUITS

guardband to more acceptable sizes of 20% and 10%. The number of PCPs in the initial design that violates the corresponding target guardband (Slow-PCPs), the area overhead, and the CPU time for design improvement are given. When the guardband constraint is of 20% the area overhead for most of the circuits remains low because only some slow-PCPs out of the whole PCP set need to be improved. However, when the guardband constraint becomes stringent, the number of slow-PCPs significantly increases for most of the circuits. This increase depends on how balanced are the delays of the PCPs. The area overhead and the corresponding CPU time also increases for stringent guardband constraints as further optimization is needed to achieve the target guardband. It is worth to mention that the area overhead for gate sizing depends on the circuit topology as it significantly changes from one circuit to another. Therefore, designers should trade-off the desired target guardbands and the corresponding area budget for optimization.

5.5.3 Benefit of the Multiple Workload-Aware Aging Analysis

Tables 5.2 and 5.3 show the results for the cases when worst-case conditions and when only one single workload condition are assumed for aging analysis, respectively. When only a representative single workload profile is used, the number of PCPs, the number of Critical Gates and the estimated guardband for the circuit are smaller than those obtained with our proposed multiple workload degradation analysis approach. This is because, in our approach, at least one of the tested workload profiles caused more aging in the PCPs than the single workload profile. Consequently, the area overhead when designing circuits using the single duty cycle assumption is lower than the area overhead obtained with our proposal. Also, the CPU time for design optimization is slightly lower. However, it should be noted that the exact workload profile that a circuit will experience over the
5.5. RESULTS FROM APPLICATION TO ISCAS BENCHMARK CIRCUITS

Lifetime is not well known in advance at the design phase. If the workload profile that the optimized circuit experiences over the lifetime are different than the used during design, some of the paths may degrade enough to cause a failure to time specifications. When worst-case aging is assumed, the number of PCPs, Critical Gates and estimated GB for the circuit significantly increase, which results in significant area overhead and extra CPU time since the PCPs receive more sizing than needed. For instance, consider circuit c2670, where 12.68% of the area is saved with respect to worst-case design for a 20% of target guardband when our approach is used. The area saved increases to 49.19% for tighter 10% guardband constraint. A similar comparison can be made for the other circuits.

<table>
<thead>
<tr>
<th>Circuit</th>
<th>PCPs</th>
<th>CGs</th>
<th>GB(%)</th>
<th>( GB_t = 20% )</th>
<th>( GB_t = 10% )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>slow-PCP</td>
<td>Area (%)</td>
<td>CPU (sec)</td>
<td>slow-PCP</td>
<td>Area (%)</td>
</tr>
<tr>
<td>c432</td>
<td>6773</td>
<td>12.49</td>
<td>1651.70</td>
<td>25005</td>
<td>36.52</td>
</tr>
<tr>
<td>c880</td>
<td>359</td>
<td>10.91</td>
<td>47.78</td>
<td>791</td>
<td>35.77</td>
</tr>
<tr>
<td>c1908</td>
<td>1370</td>
<td>7.14</td>
<td>228.39</td>
<td>4322</td>
<td>30.85</td>
</tr>
<tr>
<td>c2670</td>
<td>143</td>
<td>3.47</td>
<td>35.16</td>
<td>372</td>
<td>14.55</td>
</tr>
<tr>
<td>c3540</td>
<td>3476</td>
<td>0.85</td>
<td>2918.81</td>
<td>27636</td>
<td>2.97</td>
</tr>
<tr>
<td>c5315</td>
<td>637</td>
<td>2.16</td>
<td>187.42</td>
<td>2066</td>
<td>8.63</td>
</tr>
<tr>
<td>c7552</td>
<td>622</td>
<td>0.27</td>
<td>120.83</td>
<td>3047</td>
<td>1.07</td>
</tr>
<tr>
<td>s298</td>
<td>22</td>
<td>5.70</td>
<td>1.02</td>
<td>45</td>
<td>15.39</td>
</tr>
<tr>
<td>s838</td>
<td>67</td>
<td>4.23</td>
<td>2.59</td>
<td>139</td>
<td>10.18</td>
</tr>
<tr>
<td>s1423</td>
<td>166</td>
<td>1.00</td>
<td>522.07</td>
<td>1110</td>
<td>4.90</td>
</tr>
<tr>
<td>s5378</td>
<td>88</td>
<td>0.99</td>
<td>12.59</td>
<td>376</td>
<td>4.25</td>
</tr>
<tr>
<td>s9234</td>
<td>196</td>
<td>2.27</td>
<td>115.29</td>
<td>887</td>
<td>8.32</td>
</tr>
<tr>
<td>s13207</td>
<td>1337</td>
<td>0.84</td>
<td>4602.98</td>
<td>21395</td>
<td>2.91</td>
</tr>
<tr>
<td>s35932</td>
<td>978</td>
<td>6.00</td>
<td>114.42</td>
<td>1711</td>
<td>18.31</td>
</tr>
</tbody>
</table>

An interesting observation from Table 5.3 is that the computational for sizing the circuit assuming under worst-case assumptions can be significantly longer than the computational time for sizing the circuit using the proposed approach because more sizing iterations are required to reduce the overestimated delay degradation. In this case, it is evident that the extra computational time used to evaluate the timing response of the paths for various workload profiles is justified.

The robustness of the optimized designs for 20000 random generated workload
5.5. RESULTS FROM APPLICATION TO ISCAS BENCHMARK CIRCUITS

Table 5.3: Results using worst-case aging analysis

<table>
<thead>
<tr>
<th>Circuit</th>
<th>PCPs</th>
<th>CGs</th>
<th>GB(%)</th>
<th>( \mu + 3\sigma ) delay</th>
<th>Area (%)</th>
<th>CPU (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>c432</td>
<td>53800</td>
<td>152</td>
<td>50.97</td>
<td>26176</td>
<td>31.47</td>
<td>3740.57</td>
</tr>
<tr>
<td>c880</td>
<td>2360</td>
<td>158</td>
<td>59.96</td>
<td>1004</td>
<td>54.33</td>
<td>210.47</td>
</tr>
<tr>
<td>c1908</td>
<td>11276</td>
<td>204</td>
<td>52.75</td>
<td>5409</td>
<td>38.12</td>
<td>4793.69</td>
</tr>
<tr>
<td>c2670</td>
<td>837</td>
<td>131</td>
<td>51.52</td>
<td>461</td>
<td>16.67</td>
<td>99.286</td>
</tr>
<tr>
<td>c3540</td>
<td>204056</td>
<td>459</td>
<td>47.17</td>
<td>26432</td>
<td>2.69</td>
<td>10172.12</td>
</tr>
<tr>
<td>c5315</td>
<td>9331</td>
<td>551</td>
<td>53.22</td>
<td>3007</td>
<td>11.43</td>
<td>788.71</td>
</tr>
<tr>
<td>c7552</td>
<td>12866</td>
<td>1070</td>
<td>53.87</td>
<td>3615</td>
<td>1.30</td>
<td>449.40</td>
</tr>
<tr>
<td>s298</td>
<td>88</td>
<td>38</td>
<td>55.18</td>
<td>54</td>
<td>21.95</td>
<td>3.28</td>
</tr>
<tr>
<td>s838</td>
<td>501</td>
<td>135</td>
<td>68.67</td>
<td>251</td>
<td>27.62</td>
<td>23.96</td>
</tr>
<tr>
<td>s1423</td>
<td>10885</td>
<td>408</td>
<td>48.13</td>
<td>2054</td>
<td>6.39</td>
<td>2973.07</td>
</tr>
<tr>
<td>s5378</td>
<td>1287</td>
<td>264</td>
<td>55.96</td>
<td>512</td>
<td>6.88</td>
<td>82.18</td>
</tr>
<tr>
<td>s9234</td>
<td>4795</td>
<td>335</td>
<td>49.53</td>
<td>861</td>
<td>6.00</td>
<td>376.15</td>
</tr>
<tr>
<td>s13207</td>
<td>200000</td>
<td>510</td>
<td>49.71</td>
<td>21606</td>
<td>2.78</td>
<td>15752.84</td>
</tr>
<tr>
<td>s35942</td>
<td>11121</td>
<td>5021</td>
<td>54.61</td>
<td>1388</td>
<td>11.94</td>
<td>726.19</td>
</tr>
</tbody>
</table>

Figure 5.12: Histograms of the \( \mu + 3\sigma \) corner of the circuit aged delay obtained for an exhaustive number (20000) of multiple signal probability groups at main circuit inputs.

profiles was analyzed. For each workload profile, the corresponding duty-cycle and operating temperature of the devices were computed, and SSTA was performed to obtain the corresponding \( \mu + 3\sigma \) delay of all the PCPs. Then, the maximum \( \mu + 3\sigma \)delay among all the PCPs was identified, since this value corresponds to the maximum delay that the circuit can take for the given signal probability profile. Figures 5.12(a) and 5.12(b) show histograms of the \( \mu + 3\sigma \) delay of circuits c880 and s298 for both the optimized design using our proposes multiple workload-
Table 5.4: Percentage of WLs for which 10% of guardband may be violated.

<table>
<thead>
<tr>
<th>Circuit</th>
<th>Multiple WLs</th>
<th>Single WL</th>
</tr>
</thead>
<tbody>
<tr>
<td>c432</td>
<td>0.40</td>
<td>70.50</td>
</tr>
<tr>
<td>c880</td>
<td>4.79</td>
<td>55.11</td>
</tr>
<tr>
<td>c1908</td>
<td>0.11</td>
<td>32.24</td>
</tr>
<tr>
<td>c2670</td>
<td>3.52</td>
<td>53.62</td>
</tr>
<tr>
<td>c3540</td>
<td>3.05</td>
<td>7.62</td>
</tr>
<tr>
<td>c5315</td>
<td>4.17</td>
<td>29.7</td>
</tr>
<tr>
<td>c7552</td>
<td>0.05</td>
<td>24.77</td>
</tr>
<tr>
<td>s298</td>
<td>3.21</td>
<td>18.08</td>
</tr>
<tr>
<td>s838</td>
<td>0.33</td>
<td>9.71</td>
</tr>
<tr>
<td>s1423</td>
<td>0.29</td>
<td>0.48</td>
</tr>
<tr>
<td>s5378</td>
<td>4.57</td>
<td>51.98</td>
</tr>
<tr>
<td>s9234</td>
<td>1.58</td>
<td>31.88</td>
</tr>
<tr>
<td>s13207</td>
<td>0.24</td>
<td>15.10</td>
</tr>
<tr>
<td>s35932</td>
<td>2.04</td>
<td>79.05</td>
</tr>
<tr>
<td><strong>Avg.</strong></td>
<td><strong>2.02</strong></td>
<td><strong>34.79</strong></td>
</tr>
</tbody>
</table>

Aware aging analysis and the optimized design using only one single workload profile for aging analysis. As can be seen, there are some workloads for which the $\mu + 3\sigma$ delay of the circuit may violate the allowed 10% of guardband. However, it is clear that the optimized design with the proposed approach may violate the guardband for a significantly lower number of workloads, which demonstrates the benefit of the proposed approach. Table 5.4 shows the percentage of signal probability profiles for which the $\mu + 3\sigma$ delay of the circuits violated the specified guardband constraint of 10%. For most of the circuits, the robustness of the optimized design using the multiple workload-aware degradation analysis is significantly better than the optimized designs using only one single workload profile. Therefore, the obtained designs with the proposed approach are more reliable.

In the case that the coverage of possible workloads want to be improved, designers can trade-off the degree of circuit reliability and the computational cost of performing a more exhaustive duty-cycle-aware aging analysis step.
5.5. RESULTS FROM APPLICATION TO ISCAS BENCHMARK CIRCUITS

5.5.4 Gate Sizing Optimization Comparison

The efficiency of the proposed gate sizing optimization metrics and the heuristic was compared against other aging-aware metrics proposed in [142] and [135], which are given in Equation 5.11

\[ M_{i,[142]} = \frac{N_i \cdot \Delta D_i}{\max(N_i, \Delta D_i)} + \delta \quad M_{i,[135]} = S_{D_i K_i}^D \cdot \sum_p^N \Delta D_i \]  

(5.11)

where \( M_{i,[142]} \) is the metric proposed in [142], \( N_i \) is the number of paths (PCPs) passing through the gate \( i \), \( \Delta D_i \) is the delay degradation of the gate, and \( \delta \) is a parameter that takes the value of 1 if the gate is in the slowest path of the circuit (the path with the largest negative slack). \( M_{i,[135]} \) is the metric in [135], \( S_{D_i K_i}^D \) is the delay sensitivity of the gate to changes in its size, \( N \) is the number of paths passing through the gate and \( S_{D_i K_i}^D \) is the delay degradation of the gate.

Note that the metrics chosen for comparison are based on different characteristics of the gates. The metric in [142] focuses on identifying the gates suffering largest degradation and affecting many paths. A similar metric has been proposed in [133] to improve theaged performance of critical paths in an ALU. The metric in [135] considers not only the gate delay degradation and the number of paths affected by the gate but also the delay sensitivity on gate sizing. This metric was shown to perform better than that proposed in [44]. The metrics in Equation 5.11 were used in the proposed metric-guided design flow. However, only the sizing-up heuristic was applied since the approaches in [142], and [135] do not consider a metric for sizing-down gates.

Table 5.5 shows the results for 20% and 10% of guardband constraint. For comparison purposes, the area overhead of the proposed approach is also given. It can be observed that our proposal gives designs with lower area overhead than those obtained using the metrics of [142] and [135]. This is because the proposed metrics includes important parameters not taken into account in the others such as the area impact and the paths slack. Furthermore, the proposed metric uses
a statistical sensitivity that takes into account the impact of sizing a gate on the nominal delay, delay deterioration and variability due to process variations. Among the other metrics, the one in [142] is less efficient for gate sizing. This is because this metric only considers the delay degradation and the number of paths impacting by the gate. However, it does not consider the path delay sensitivity to gate sizing. Therefore, it does not measure the potential delay improvement of sizing a gate. Although the metric in [135] includes the delay sensitivity parameters, this sensitivity does not consider aging or process variations effects. Therefore, it may fail to indicate the gates more beneficial for delay improvement.

Table 5.5: Percentage of area overhead for target guardbands of 10% and 20% of using three selection and sizing methods

<table>
<thead>
<tr>
<th>Circuit</th>
<th>GBt = 20%</th>
<th>GBt = 10%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Our [142]</td>
<td>[135]</td>
</tr>
<tr>
<td>c432</td>
<td>12.49</td>
<td>29.40</td>
</tr>
<tr>
<td>c880</td>
<td>15.15</td>
<td>21.85</td>
</tr>
<tr>
<td>c1908</td>
<td>9.76</td>
<td>17.84</td>
</tr>
<tr>
<td>c2670</td>
<td>4.09</td>
<td>10.63</td>
</tr>
<tr>
<td>c3540</td>
<td>0.85</td>
<td>2.76</td>
</tr>
<tr>
<td>c5315</td>
<td>2.38</td>
<td>5.07</td>
</tr>
<tr>
<td>c7552</td>
<td>0.33</td>
<td>1.38</td>
</tr>
<tr>
<td>s298</td>
<td>6.26</td>
<td>10.61</td>
</tr>
<tr>
<td>s838</td>
<td>4.92</td>
<td>8.33</td>
</tr>
<tr>
<td>s1423</td>
<td>1.25</td>
<td>4.12</td>
</tr>
<tr>
<td>s5378</td>
<td>1.14</td>
<td>2.67</td>
</tr>
<tr>
<td>s9234</td>
<td>2.27</td>
<td>3.74</td>
</tr>
<tr>
<td>s13207</td>
<td>0.84</td>
<td>2.22</td>
</tr>
<tr>
<td>s35932</td>
<td>6.00</td>
<td>8.61</td>
</tr>
<tr>
<td>Avg.</td>
<td>4.83</td>
<td>9.23</td>
</tr>
</tbody>
</table>

Figure 5.13 shows the number of iterations performed when using each of the metrics for gate sizing. An iteration corresponds to the process of performing SSTA over all the PCPs to determine the current guarband required for the circuit and the Slow- and Fast- PCP subsets, the evaluation of the sizing metric for each
gate in the PCPs, and the application of the sizing heuristic. It can be observed that the proposed metrics imply a larger number of iterations. This is because the proposed metrics select the gates giving an efficient delay-area trade-off, which are not necessarily the ones improving quicker the circuit delay. On the other hand, the metric in [142] gives a higher priority to those gates in the longest PCP of the circuit, which results in a quick delay reduction but with increased area overhead.

![Graph](image)

(a) $GB_t = 20\%$ of nominal delay

(b) $GB_t = 10\%$ of nominal delay

Figure 5.13: Number of iterations to achieve target guardband.

### 5.6 Conclusions

A gate sizing optimization methodology for guardband reduction in the presence of aging due to BTI and process variations have been presented. Since the workload profile that a circuit experience over the lifetime if unknown at the design
5.6. CONCLUSIONS

phase, the proposed methodology calculates the maximum aged delay of the circuit paths for various workload profiles at main inputs, which define the duty cycle of the devices. In such way, traditional worst-case aging or unreliable single workload assumptions have been avoided. It has been shown that a reasonable number of signal probability profiles is sufficient to obtain a good estimation of the maximum degraded delay of the circuit paths. For delay optimization towards the desired guardband, gate metrics and a sizing heuristic have been proposed to select the best gates for both sizing-up to improve delay and sizing-down to mitigate area overhead. An approximation for the statistical sensitivity of a path delay has been proposed to mitigate computational effort of statistical timing analysis and speed-up metrics evaluation. The application of the proposed methodology on ISCAS benchmark circuits has shown that gate sizing using the proposed approach to estimate the maximum aged delay of the circuit paths results in significant area savings compared to gate sizing under worst-case aging assumptions. Furthermore, it has been shown that the obtained designs can operate reliably for a different usage profile than those used during design optimization. The results using the proposed metrics has been compared against the results using other gates metrics in the literature, and it has been shown that the proposed approach provides a better area-delay trade-off.
Chapter 6

Conclusions

Aging effects have an increasing impact on the lifetime reliability and performance of modern digital integrated circuits designed with deeply scaled CMOS technologies. This thesis has focused on reliability analysis and improvement of digital integrated circuits considering aging due to Bias Temperature Instability (BTI), which is the dominant aging mechanism in digital circuits. Performance degradation due to BTI is highly dependent on operating conditions such as the executed workloads by the circuits, and process-induced parameters variations.

An aging-aware statistical static timing analysis tool has been developed in C++ code to analyze the timing performance of large digital circuits under the combined effects of BTI and Process Variations. The impact of the operating workload of the circuit on devices aging has been taken into account as it strongly influences the stress duty-cycle and operating temperature of the devices, and consequently, the overall circuit degradation. It has been shown that conventional clock guardband to assure reliable circuit operation may be too large (more than 30% of nominal delay), resulting in significant performance loss.

This thesis has addressed the design of reliable digital circuits from two different perspectives: 1) Aging Monitoring of Critical Paths; and 2) Efficient Design for Guardband Reduction.
For aging monitoring, it has been proposed a methodology to select a reduced set of critical paths. Those paths should be monitored for reliable on-line prediction of possible circuit failures due to aging. This allows to minimize the initial guardband of the circuit and to dynamically react against aging by gradually increasing the clock period to keep reliable operation. A statistical selection condition has been proposed to identify those paths that may eventually become the slowest path of the circuit for a given process and aging condition. The selection condition allows to trade-off the degree of circuit reliability and the corresponding size of the critical paths set to be monitored. The proposed methodology first reduces the overall set of topological path to a smaller set of Potential Critical Paths from which detailed Critical Path selection is performed. The selection algorithm considers that different paths may become critical depending on the operating workload. Furthermore, it considers that paths that can become slower than the initial LCP of the circuit may be covered by other paths. The proposed methodology is suitable to be applied at early design phases as it uses a spatial correlation approximation, which has been proposed based on previous analysis of the delay difference between paths using spatial correlation extracting from circuits layouts. It has been shown that using the proposed correlation approximation for path selection results in a much more reduced set of critical paths than using conventional correlation bounding approaches. The proposed methodology considers various possible operating workloads for the circuit during path selection process to avoid conventional worst-case aging assumptions. Designers should trade-off the degree of circuit reliability with the computational effort of analyzing a larger number of workloads. However, it has been shown that for most of the circuits considering a moderate number of workloads is sufficient to identify most of the critical paths. Moreover, it was shown that critical path selection using various possible executed workloads results in a much more reduced set of paths to be monitored, compared to classic worst-case workload assumption. The proposed path selection method using the spatial correlation approximation was compared
with path selection using spatial correlation extracted from circuit layout, and it has been shown that most of the critical paths are effectively selected. Moreover, it has been shown that those paths not selected with the proposed approach have a low probability to become critical.

For reliability-aware circuit design, it has been proposed a gate size optimization technique based on new statistical gate sizing metrics. Unlike other works that only considers sizing-up gates to compensate aging and process variations, the proposed method also considers sizing-down gates to mitigate area overhead. The proposed metrics consider various parameters to measure the benefit of sizing-up/down a gate in the circuit to reduce the required guardband for reliable operation with low area overhead. The parameters taken into account in the gate metrics are: 1) the statistical path delay sensitivity to gate sizing, which measures how much impact has sizing a gate on the $\mu + 3\sigma$ delay of a path; 2) The area impact of sizing the gate; 3) The number of paths passing through the gate; and 4) The slack time of the paths passing through the gate. This thesis improved the classic concept of delay sensitivity by taking into account a nominal delay sensitivity component, a delay degradation sensitivity component, and a standard deviation sensitivity component. A fast approximation of the statistical delay sensitivity of path delay to gate sizing has been proposed based on an analysis of the impact of sizing a gate on the mean delay and standard deviation of the adjacent gates in the path. A multiple workload profiles analysis has been proposed to estimate, in a more realistic way, the maximum delay degradation of the paths. This allows avoiding pessimistic worst-case aging assumptions. Furthermore, this allows reducing the set of paths where size-optimization is focused. It has been shown that a moderate number of workload profiles is sufficient to obtain a good estimation of the maximum degraded delay of the circuit paths. The proposed methodology provides an efficient trade-off between the desired guardband for the circuit and the area overhead. The results have shown that the proposed approach significantly reduces the area overhead and CPU time for size-optimization
compared to circuit design using worst-case aging assumptions. Furthermore, by considering multiple operating workload conditions during circuit design, a more reliable design is obtained with lower area overhead, compared to circuit design considering a single workload trace. The results have also shown that by using the proposed metrics, timing constraints can be met with lower area overhead than that obtained using other gate metrics available in the literature.

The contributions of this thesis provide new aging-aware design techniques to help designers to approach the complex problem of the design of reliable digital integrated circuits under the combined effects of aging due to BTI and process variations.
Published Papers


5. F. Forero, **A. Gomez** and V. Champac, “A methodology for NBTI circuit reliability at reduced power consumption using dual supply voltage,” 2016 17th Latin-American Test Symposium (LATS), Foz do Iguacu, 2016, pp. 81-86.


References


REFERENCES


REFERENCES


[93] Song Jin, Yinhe Han, Huawei Li, and Xiaowei Li. Statistical lifetime reliability optimization considering joint effect of process variation and aging. *Integration, the {VLSI} Journal*, 44(3):185 – 191, 2011.


[97] Maksim Jenihhin, Giovanni Squillero, Thiago Santos Copetti, Valentin Tihomirov, Sergei Kostin, Marco Gaudesi, Fabian Vargas, Jaan Raik, Matteo Sonza Reorda, Letícia Bolzani Poehls, Raimund Ubar, and Guilherme Cardoso Medeiros. Identification and


