Parametric Yield of VLSI Systems under Variability: Analysis and Design Solutions

by

Kian Haghdad

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2011

© Kian Haghdad 2011

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Variability has become one of the vital challenges that the designers of integrated circuits encounter. variability becomes increasingly important. Imperfect manufacturing process manifest itself as variations in the design parameters. These variations and those in the operating environment of VLSI circuits result in unexpected changes in the timing, power, and reliability of the circuits. With scaling transistor dimensions, process and environmental variations become significantly important in the modern VLSI design. A smaller feature size means that the physical characteristics of a device are more prone to these unaccounted-for changes. To achieve a robust design, the random and systematic fluctuations in the manufacturing process and the variations in the environmental parameters should be analyzed and the impact on the parametric yield should be addressed.

This thesis studies the challenges and comprises solutions for designing robust VLSI systems in the presence of variations. Initially, to get some insight into the system design under variability, the parametric yield is examined for a small circuit. Understanding the impact of variations on the yield at the circuit level is vital to accurately estimate and optimize the yield at the system granularity. Motivated by the observations and results, found at the circuit level, statistical analyses are performed, and solutions are proposed, at the system level of abstraction, to reduce the impact of the variations and increase the parametric yield.

At the circuit level, the impact of the supply and threshold voltage variations on the parametric yield is discussed. Here, a design centering methodology is proposed to maximize the parametric yield and optimize the power-performance trade-off under variations. In addition, the scaling trend in the yield loss is studied. Also, some considerations for design centering in the current and future CMOS technologies are explored.

The investigation, at the circuit level, suggests that the operating temperature significantly affects the parametric yield. In addition, the yield is very sensitive to the magnitude of the variations in supply and threshold voltage. Therefore, the spatial variations in process and environmental variations make it necessary to analyze the yield at a higher granularity. Here, temperature and voltage variations are mapped across the chip to accurately estimate the yield loss at the system level.

At the system level, initially the impact of process-induced temperature variations on the power grid design is analyzed. Also, an efficient verification method is provided that ensures the robustness of the power grid in the presence of variations. Then, a statistical analysis of the timing yield is conducted, by taking into account both the process and environmental variations. By considering the statistical profile of the temperature and supply voltage, the process variations are mapped to the delay variations across a die. This ensures an accurate estimation of the timing yield. In addition, a method is proposed to

accurately estimate the power yield considering process-induced temperature and supply voltage variations. This helps check the robustness of the circuits early in the design process.

Lastly, design solutions are presented to reduce the power consumption and increase the timing yield under the variations. In the first solution, a guideline for floorplaning optimization in the presence of temperature variations is offered. Non-uniformity in the thermal profiles of integrated circuits is an issue that impacts the parametric yield and threatens chip reliability. Therefore, the correlation between the total power consumption and the temperature variations across a chip is examined. As a result, floorplanning guidelines are proposed that uses the correlation to efficiently optimize the chip's total power and takes into account the thermal uniformity. The second design solution provides an optimization methodology for assigning the power supply pads across the chip for maximizing the timing yield. A mixed-integer nonlinear programming (MINLP) optimization problem, subject to voltage drop and current constraint, is efficiently solved to find the optimum number and location of the pads.

Acknowledgements

I would sincerely like to thank my advisor, Professor Mohab Anis, for his unconditional support. He has always been a great source of guidance, knowledge, and inspiration for me. Without his input, this research would not have been possible. I would like to express my sincere gratitude and appreciation to the members of the advisory committee, Professor Karim S. Karim, Professor Andrew A. Kennings, Professor El Mostapha Abdoulhamid, Professor Siddharth Garg, and Professor Patricia M. Nieva. Their feedback has been essential to bring my thesis to the current form.

In addition, I am thankful to my M.Sc. advisor, Professor Yehea Ismail at Northwestern University, IL, for his continued support along the way. I am also grateful to Professor Kumaraswamy Ponnambalam, from the Department of Systems Design Engineering for sharing his preliminary Matlab codes of the yield optimization problem. I would like to thank Dr. Yongkui Han from University of Massachusetts Amherst, MA, for providing the interconnect matrix I used in the floorplanning in the presence of temperature variations. Moreover, I thank Dr. Javid Jaffari, for sharing the preliminary Matlab codes of the statistical thermal analyzer.

I had an opportunity to work as a teaching assistant for Professor James Barby, whose guidelines in teaching, are most appreciated. Furthermore, I would like to thank all the faculty members, and also the staff, including Wendy Boles, Lisa Hendels, Annette Dietrich, Susan King, and Phil Regier in the Department of Electrical and Computer Engineering for helping me pursue my Ph.D. program.

In addition, I would like to acknowledge the financial support I received through the postgraduate scholarship award from the National Sciences and Engineering Research Council (NSERC) of Canada.

I extend my thanks to all my friends and colleagues in the VLSI research lab, including Edgar Mateos Santillan, Javid Jaffari, Hassan Hassan, Minoo Mirsaeedi, Hasan Mostafa, Akhilesh Kumar, Ahmed Nour, and many more who helped make the graduate program an enjoyable experience for me.

I especially want to thank my father, my first and foremost teacher, and my mother who has inspired me beyond measure. I also want to thank my sisters and brothers for their invaluable support. Special thanks to my caring and loving wife, thank you for your encouragement and kindness.

Dedication

This is dedicated to my family.

Table of Contents

List of Tables xi

List of Figures xv

1 Introduction 1

1.1 Motivation . . . 1

1.2 Thesis Organization . . . 2

2 Background and Related Work 4

2.1 Introduction . . . 4

2.2 Process Variations . . . 5

2.2.1 Variations Classifications . . . 6

2.2.2 Variations in the Threshold Voltage . . . 7

2.2.3 Interconnect Variations . . . 10

2.3 Environmental Variations . . . 11

2.3.1 Variations in the Supply Voltage . . . 12

2.3.2 Temperature Variations . . . 13

2.4 Other Sources of Variations . . . 14

2.5 Impact of the Variations on Design . . . 15

2.5.1 Impact on Performance . . . 15

2.5.2 Impact on Power Consumption . . . 16

2.5.3 Modeling Variations . . . 17

2.6 Related Work: Design under Variations 18
2.6.1 Variability-Aware Circuit Design 18
2.6.1.1 Power-Performance Trade-Offs 18
2.6.1.2 Threshold Voltage Assignment 21
2.6.1.3 Adaptive Solutions 22
2.6.1.4 Voltage Scaling 24
2.6.2 Design Optimization at Architecture Level 28
2.6.2.1 High Temperature Effect 28
2.6.2.2 Temperature-Aware Floorplanning 29
2.6.3 Effects of Process Variations at the System Level 30
2.6.3.1 Power Grid Verification 31
2.6.3.2 Statistical Static Timing Analyses 32
2.6.3.3 Power and Temperature Estimation 34
2.7 Proposed Analysis and Methodologies 36

3 Designing Robust Integrated Circuits Considering Supply and Threshold Voltage Variations 39
3.1 Introduction 39
3.2 Related Work 40
3.3 Design Metrics for the Yield Estimation 41
3.3.1 Energy Delay Model 41
3.3.2 Incorporating Temperature 43
3.4 Constructing the Feasible Region and Modeling the Design Variable Distribution 46
3.4.1 Minimum Performance Constraint 46
3.4.2 Maximum Temperature and Thermal Reliability Constraint 46
3.4.3 Constructing the Feasible Region 47
3.5 Yield Optimization 48
3.6 Simulation Results and Discussion for 90nm 51
3.6.1 Solving the Optimization Problem 51

3.6.2 Optimizing the Design Metrics and Simulation Results 52
3.6.3 Design Considerations 56
3.7 Impact of Scaling 59
3.7.1 Trend of Variations in the Design Parameters 59
3.7.2 Comparing the Results for Different Technologies 61
3.8 Design Insights for Current and Future Technologies 61
3.9 Conclusions 64

4 Power Grid Analysis and Verification Considering Temperature Variations 66

4.1 Introduction 66
4.2 Related Work 67
4.3 Statistical Thermal Analysis 68
4.3.1 Motivation and Workflow 68
4.3.2 Statistical Thermal Model 69
4.4 Voltage Drop Statistics 76
4.5 Power Grid Verification 78
4.6 Results and Discussion 79
4.7 Conclusions 84

5 Parametric Yield Analysis Considering Process-Induced Temperature and Supply Voltage Variations 85

5.1 Introduction 85
5.2 Related Work 86
5.3 Statistical Profile of Temperature and Voltage Drop 87
5.4 Timing Yield 89
5.4.1 Simulation Results and Discussion 94
5.5 Power Yield 98
5.5.1 Experimental Results and Design Insights 100
5.6 Conclusions 102

6 Design Solutions for VLSI Systems under Variability **104**
6.1 Introduction 104
6.2 Solution 1: Total Power Reduction in the Presence of Temperature Variations 104
6.2.1 Related Work 105
6.2.2 Preliminaries and Understanding the Correlation between the Total Power and Temperature Variations 106
6.2.2.1 Temperature and Power Modeling 106
6.2.2.2 The Correlation between the Total Power and Temperature Variations 107
6.2.3 Proposed Floorplanning Optimization 109
6.2.3.1 Objective Function 109
6.2.3.2 The Optimization Methodology 110
6.2.4 Results and Discussion 112
6.3 Solution 2: Power Supply Pads Assignment for Maximizing Timing Yield . 118
6.3.1 Related Work 118
6.3.2 Voltage Drop and Supply Current Statistics 119
6.3.3 Design Constraints and Yield Optimization 122
6.3.4 Results and Discussion 123
6.4 Conclusions 126

7 Conclusions and Future Work **128**
7.1 Summary of Contributions 128
7.2 Future Research Directions 129

Appendices **130**

A Publications from this Research **131**

B Acronyms **133**

C Variables **135**

References **138**

List of Tables

3.1 Yield optimization for various design metrics 51
3.2 Yield optimization for different technology nodes 65
3.3 Increasing yield by relaxation of the constraints 65

4.1 Runtime of the proposed methodology 83

5.1 Timing yield and statistics of delay, comparing different methods 96
5.2 The error in the delay calculation when the statistical thermal profile is ignored 97
5.3 Runtime of the extraction of the delay statistics 98
5.4 Power yield and statistics of the total power, comparing different methods 102
5.5 Runtime of the extraction of the total power statistics 102

6.1 Runtime of the optimization process for Alpha and MCNC benchmarks . . 114
6.2 Achieving various objectives for four different leakage to total power ratios 116
6.3 Timing yield and statistics of delay, comparing optimal and random supply pad assignment 125
6.4 Runtime of the pad assignment optimization and macromodel extraction . 126

List of Figures

2.1 Device characteristics that are impacted by the imperfect manufacturing process 5

2.2 Imperfect ion implantation 7

2.3 Rapid reduction in the number of dopants 8

2.4 Difference between the lithography wavelength and feature size for the current and future technologies 9

2.5 Roll-off effect in the threshold voltage in respect to the channel length . . . 9

2.6 Pattern dependent problems of dishing and erosion in copper CMP 11

2.7 Metal line resistance as a function of line width for different dishing radius 11

2.8 Temporal variations in supply voltage 12

2.9 Spatial temperature variations of a Pentium M chip running Applu SPEC benchmark 13

2.10 Temporal temperature variations of a Pentium M chip running Applu SPEC benchmark 14

2.11 Impact of variations on delay benchmark 15

2.12 Impact of variations on power benchmark 16

2.13 Large variations in leakage power and performance are due to process variations, at 130nm 17

2.14 The design domains for the parametric yield analysis and optimization . . 18

2.15 Partition of a circuit to model the correlated component of variation 19

2.16 Joint probability distribution function for the bivariate Gaussian distribution for c3540 19

2.17 Optimal operating line and different optimal metrics 20

2.18 Total power as a function of frequency for transistor sizing and dual V_{th} assignment 21
2.19 Power delay curves for 99.9% timing and power yield 22
2.20 Schematic for achieving multiple operating modes 22
2.21 Leakage vs. delay spread due to process variation 23
2.22 Impact of variations on power benchmark 24
2.23 Block diagram of ABB test chip 25
2.24 Comparing the effectiveness of adaptive solutions 25
2.25 Temperature dependent deactivation scheme 26
2.26 Effect of WID process variations on Energy/Operation 27
2.27 Normalized power consumption of three different schemes for the $70nm$ technology 28
2.28 Block temperatures for gcc benchmark 29
2.29 A temperature profile comparison between two floorplanning methods . . . 30
2.30 RC network representing a power grid under variability 31
2.31 Distribution of the upper bound of the supply voltage in respect to a user-defined threshold value 32
2.32 Histogram of C880 circuit delay under supply voltage variations 33
2.33 Probability density function of leakage to active power ratio 34
2.34 Organization of the research conducted in this thesis 36

3.1 Normalized EDP contours and iso-performance curves 43
3.2 The steady-state temperature and power estimation methodology and identifying thermal runaway region 44
3.3 Normalized temperature dependent EDP contours and iso-performance curves 45
3.4 Normalized EDP contours, iso-performance curves, and contours of temperature 48
3.5 The final location of the tolerance box over which the yield is maximized . 49
3.6 Monte Carlo simulations for various voltage pairs 53
3.7 The effect of selecting three different design metrics on the final location of the tolerance box 54

3.8 The simulated circuit consisting of a ring oscillator and a multi-level NAND chain for selecting different activity factors 54

3.9 Normalized EDP contours and iso-performance curves of 90nm CMOS technology for the simulated circuit 55

3.10 Sensitivity of the parametric yield to the activity factor and transistor sizing 56

3.11 Sensitivity of yield to temperature constraint for three cases of variations in the design variables 57

3.12 The average normalized yield loss and the error in the estimated yield for the discrete case where limited number of voltage pairs are available 58

3.13 Area of the feasible region and the tolerance box for different technology nodes 60

3.14 Normalized EDP contours, iso-performance curves, and temperature contour for 32nm technology 62

3.15 Sensitivity of yield to maximum allowed temperature for different technology nodes 62

3.16 Sensitivity of the parametric yield to the activity factor and transistor sizing for different technology nodes 63

3.17 Maximum yield design centers and the shift in the thermal runaway region for different technology nodes 64

4.1 Discretized die with six cores and the package structure [1] (a) Top view (b) Lateral view 70

4.2 Normalized leakage power as a function of temperature 71

4.3 Normalized power consumption as a function of supply voltage 71

4.4 Flowchart of statistical thermal analyzer 73

4.5 Statistical profile of the temperature and voltage drop of a 16642 node power grid across an Alpha 21364 CPU core 80

4.6 PDF of the voltage drop at the maximum expected value (point C) and at the maximum standard deviation (point D 81

4.7 Q-Q plot of the voltage drop for the Monte-Carlo samples and that of the proposed method at the maximum expected value (point C) 82

4.8 Distribution of error between the proposed methodology and Monte-Carlo simulations 82

5.1 Normalized leakage power as a function of temperature for different gates . 93

5.2 Normalized leakage power consumption as a function of supply voltage for different gates . . . 93
5.3 Normalized dynamic power consumption as a function of supply voltage for different gates . . . 93
5.4 statistical profiles for circuit s38584: (a) the profile of temperature expected value, (b) temperature standard deviation, (c) voltage drop expected value (d) voltage drop standard deviation . . . 95
5.5 Probability density function of delay for circuit s38584 . . . 96
5.6 Sensitivity of the yield to the reduction of the temperature coefficients . . . 99
5.7 Probability density function of the total power for circuit s38584 . . . 101

6.1 Normalized increase in the leakage power in respect to the increase in temperature standard deviations, for the Alpha processor running gcc . . . 108
6.2 Normalized deviation from the minimum total power (power increase) as a function of temperature variations . . . 108
6.3 Normalized deviation from the minimum leakage power as a function of the temperature variations . . . 109
6.4 Optimization methodology using the correlation found between the temperature variations and total power of a floorplan . . . 111
6.5 Optimum floorplan with minimum total power and the lowest possible temperature variations of the core of an Alpha processor . . . 113
6.6 Normalized deviation from the minimum total power for the objectives of the existing work in the literature and those of this work . . . 114
6.7 Comparison of the temperature variations for two different objectives . . . 115
6.8 Macromodel Schematic . . . 120
6.9 Discretized die area , the tiles that share critical paths, and the candidate pads . . . 124
6.10 Timing yield sensitivity of two test circuits as a function of the number of supply pads . . . 126

Chapter 1

Introduction

1.1 Motivation

The increasing impact of process and environmental variations on the yield and the complexity of different parameters has made parametric yield an attractive subject for research. Recently, the director of Computer-Aided Design and Test at the Semiconductor Research Corp. (SRC), presented eight hypothetical companies with potential solutions for nanometer variability, and asked the panelists, where they would invest. The winners were startups, offering lithography and process variation modeling, variation-resistant regular fabrics, and variation-tolerant designs [2]. In addition, the industry advocates conducting research on the yield. Anthony Nicoli, from Mentor Graphics Corp points out that "traditionally the path to yield was fairly simple: comply with all the design rules, and yield would follow. In the nanometer era, the game has changed. To succeed in the yield game, we need new ways to incorporate yield functionalities into the newly developed automated design tools" [2].

Ted Vucurevich, CTO at Cadence said, EE times reported, "variability is a first-class design concern. Gate oxides are so thin that a change of one atom can cause a 25 percent difference in substrate current." Although the introduction of high-k materials and metal gates has created opportunities that mitigate the impact of variability on the design, but the challenges remain. Furthermore, Vucurevich said, the modes in which a device operates have become a source of variability. A cell phone chip exhibits different "hot spots" depending on whether it is taking a call, playing a video, or displaying pictures. Handling the challenges, Vucurevich said, will require a next-generation EDA architecture.

The number of transistors in VLSI circuits increases from one generation to the next. The ever increasing demand for more functionalities, as well as the increased density of devices and interconnects, results in more power consumption and more heat generation

within the chip. The supply voltage (V_{dd}) and threshold voltage (V_{th}) are two crucial design variables that directly impact the power consumption and performance of circuits. In addition, temperature plays a significant role in the design of modern VLSI circuits. Leakage power is strongly dependant on the threshold voltage and the operating temperature. Also, the dynamic power depends on the supply voltage. These power components increase as the technology scales [3]. Also, the performance of a circuit depends on these voltages and is degraded at high temperatures. Therefore, variations in these parameters directly impact both power and performance, and consequently, the yield.

In addition to the V_{dd} and the V_{th}, temperature impacts various aspects of VLSI circuit and system design. Technology scaling shrinks the sizes, while the total power consumption increases. The result is an increase in the power density and, therefore, a heat generation which is manifested as the elevated temperature of integrated circuits. High operating temperatures mean that the design will not meet its objectives. For example, performance degradation, a jump in the leakage power, and serious reliability concerns are the consequences of such temperature increases [4][5].

Finally, process variations significantly impact the leakage power, a pivotal parameter in designing a power grid. Because of the strong relationship between the temperature and leakage power, the variations also impose statistical behavior on the operating temperature. In addition, the metal resistivity of a power grid increases with the temperature. Therefore, ignoring the interdependency between leakage and temperature can introduce large errors in the power grid design, and, consequently, an increase in the timing and power yield.

The necessity to address these issues at the system level is the motivation behind this thesis. Although variations in the supply voltage impact the delay of a single gate, the IR drop is addressed at the chip level. In addition, the variations in the threshold voltage impose uncertainty in the leakage current of a single gate. However, the leakage current variations are translated into cross-chip variations in the voltage drop. Moreover, both power and performance of a circuit are functions of the operating temperature. But, an architecture-level decision can impact the chip floorplan, change the thermal profile, and, as a result, increase the power and degrade the circuit performance. Consequently, analyses and optimizations are proposed in this thesis so that the designer can accurately estimate and enhance the parametric yield at different design stages.

1.2 Thesis Organization

The remaining chapters are organized as follows:

Chapter 2 provides an overview of some investigations in the literature, particularly

of the design of VLSI circuits and systems in the presence of process and environmental variations.

Chapter 3 comprises a newly developed design-specific yield optimization, considering variations in the supply and threshold voltage. The work provides a guideline for design centering under voltage variations.

In addition, a scaling analysis of the yield optimization, under the supply and threshold voltage variations is presented in this chapter. Here, the impact of supply (V_{dd}) and threshold voltage (V_{th}) variations on the yield loss for the current and future CMOS technologies is investigated.

Chapter 4 concerns the power grid analysis, considering a statistical thermal profile across the grid. An efficient verification method is developed to ensure the robustness of the power grid in the presence of variations.

Chapter 5 is composed of a statistical analysis of the timing yield by taking into account both the process and the environmental variations. By considering the statistical profile of the temperature and supply voltage, the process variations are mapped to the delay variations across the chip.

Also, this chapter introduces a method to accurately estimate the power yield, considering the process-induced temperature and supply voltage variations. Here, by considering the statistical profile of temperature, and V_{dd}, the power yield is estimated for the chip.

Chapter 6 consists of design solutions to alleviate the process-induced environmental variations and improve parametric yield. First, the floorplanning in the presence of temperature variations is explored. The impact of the temperature variations on different objectives of the floorplanning is examined, and an efficient methodology to achieve the objectives is proposed.

Subsequently, an optimization methodology for assigning power supply pads across the chip for maximizing the timing yield is presented. A mixed-integer nonlinear programming (MINLP) optimization problem subject to voltage drop and current constraint is efficiently solved to find the optimum number and location of the pads.

Chapter 7 concludes the thesis and suggests some future work on the design of integrated circuits and systems in the presence of variations.

Chapter 2

Background and Related Work

2.1 Introduction

Pursuing Moore's law has introduced a broad range of challenges. To accommodate more transistors and to gain a better performance in each new generation of VLSI circuits, the challenges need to be addressed at many levels from circuit to system. As identified in the 2006 International Technology Roadmap for Semiconductors (ITRS), variability is one of the key difficult challenges in scaled technologies. This chapter provides a background for variations and their sources in nanoscale design. Also, some of the techniques used to alleviate the impact of variations are explained. At the end of the chapter, the proposed methodologies for designing under variations are briefly introduced.

Variations in the manufacturing process and design environment impact the performance, power, integrity, and reliability of a design. Traditionally, worst-case scenarios have been used to analyze the impact of variations. However, these corner-based approaches lead to pessimistic results. Guard-banding for parameter variations does not guarantee a reliable design, and is costly and increases the time to market. Since different corners must be studied by different simulations, the process cannot represent a unified result. However, statistical approaches provide a unified mechanism in which the implication of variations for the degradation of different design metrics is taken into account simultaneously. This helps the designer achieve the correct analysis and best optimization for the design metrics and yield.

New generations of ICs exhibit a sharp increase in the magnitude of the variations and are exposed to new sources of variability. Variation in the channel length has almost doubled, from 130nm to 65nm [6]. This indicates that the traditional corner-based methods should be replaced by statistical approaches for both analysis and optimization of the design in the presence of variations.

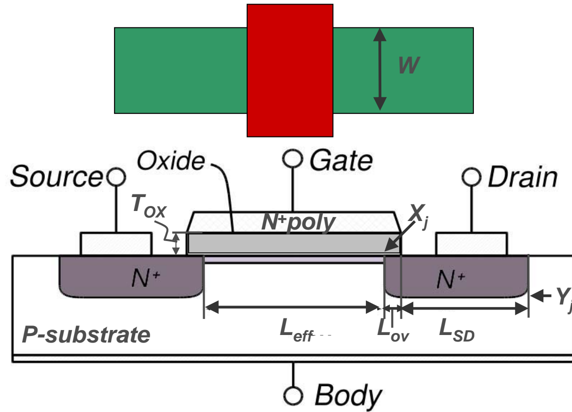


Figure 2.1: Device characteristics that are impacted by an imperfect manufacturing process.

In the traditional corner-based approach, a margin is considered, for the design parameters. Because at the design time, some factors are difficult to model. The margin, however, is growing rapidly due to the increase in variability. For example, such a margin in the delay becomes a large portion of the clock cycle. To understand the unknown factors, the variations must be characterized by their sources.

2.2 Process Variations

The manufacturing process is not perfect. Many fabrication parameters are involved in controlling the process. Imperfect parameters result in a deviation from the nominal values to be met in a design. As depicted in Fig 2.1, design parameters include some device geometry parameters, such as the effective channel length(L_{eff}), oxide thickness (t_{ox}), channel width (W), overlap length (L_{ov}), and junction depth (Y_j). In addition, the dopant concentration (N_a), inter-layer dielectric thickness (t_{ILD}), and interconnect dimensions are other parameters that are affected by process variability. The power and performance of both the device and interconnects are significantly affected by the deviation of the parameters from their nominal values. The variations can impact different designs in different ways. Some designs are more susceptible to variation than others. The variation of the design metrics creates a statistical distribution over a large number of samples. These distributions are used to define the parametric yield as a measure for describing a fraction of the design samples that meet certain criteria. The parametric yield is referred to as a timing yield, when the timing measurement is intended. Similarly, a power yield is employed as a metric for the samples that meet the power requirements.

2.2.1 Variations Classifications

Process variations are categorized into two classes: inter-die and intra-die variations. This classification is useful because they impact the design differently.

- **Inter-Die Variations:** This term refers to die-to-die or wafer-to-wafer or lot-to-lot variations. It is assumed that the process parameter does not vary on a single die. To capture this variation, a shift from the nominal value is chosen. The shift is a random value that is the same for the parameter on the die. Lens aberrations result in a variation in the gate thickness from one die to the next on a wafer. This is an example of inter-die variations. An analysis of these variations is conducted by using the corners of the process parameters. If more than one parameter is studied, the correlation between their variations is also included. This can increase the complexity, when the number of parameters is high.

- **Intra-Die Variations:** These are the variations within a single die, where a random variable is needed for each device or a portion of the circuit to represent such variations. The intra-die variations consist of two patterns.
  Random Variations: These are the deviations from the nominal values due to purely random sources such as random doping fluctuations. However, random uncertainties such as the dopant number and location cannot be predicted and cause all the devices in close proximity to exhibit different, characteristics. Random variations can also have a spatial correlation. For example, variations in the channel length of two neighboring transistors can be very similar. Some other process parameters such as t_{ox} or N_a are usually uncorrelated.

  Systematic Variations: The variability of some process parameters exhibit a systematic behavior which can be predicted. Because of lithographic and etching techniques, layout-dependant variations exist within a die. Such variations can be modeled. For example, an across-chip channel length variation can be predicted by modeling the optical proximity correction (OPC). In addition, the modeling of chemical mechanical polishing (CMP) can be employed to predict the variations in Inter-Layer Dielectric (ILD) [7]. For simplicity and the lack of manufacturing information, the systematic variations can also be threaded as random variations. By using the aforementioned variations, the channel length parameter is expressed as follows:

$$L = L_{nom} + \Delta L_{inter} + \Delta L_{spatial}(x_i, y_i) + \Delta L_{random,i} \ , \tag{2.1}$$

  where $\Delta L_{spatial}(x_i, y_i)$ takes into account the spatial correlation of the channel length.

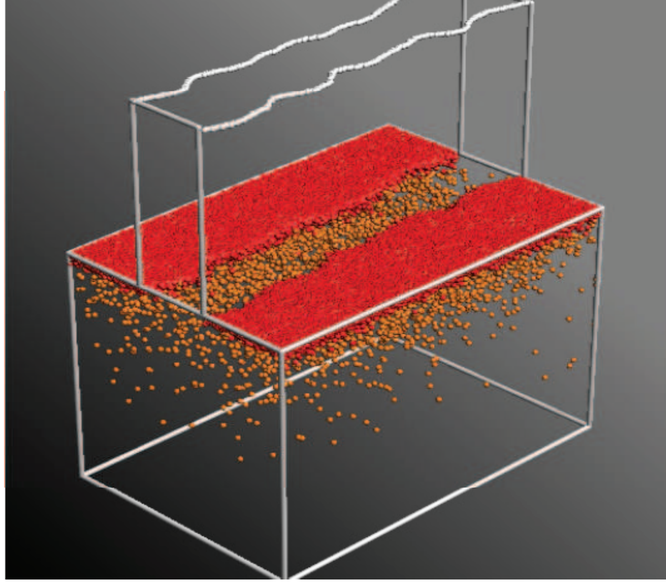


Figure 2.2: Imperfect ion implantation [11].

2.2.2 Variations in the Threshold Voltage

Variations in the threshold voltage occur, because the manufacturing parameters such as the channel length, oxide thickness, and doping density deviate from their nominal values. Such variations can be systematic or random in nature, and fall into the following categories: within-die, die-to-die, inter-wafer, and lot-to-lot. Systematic variations depend on the position of the device on a die and the layout environment surrounding the devices [8]. Lithographic, etching, and layout information are used to model, predict, and compensate for systematic variations [9]. In current technologies, the largest portion of the variations in the threshold voltage is due to the variations in the channel length which is more systematic. However, the share of random fluctuations in the variations is increasing, as they become more significant with scaling [10]. Fig. 2.2 illustrates the imperfect distribution of dopants in a device. The analytical model for the variations due to the random dopant distribution is expressed as follows [13]:

$$\sigma V_{th} = \left(\frac{\sqrt[4]{4q^3 \varepsilon_{Si} \phi_B}}{2}\right) . \frac{T_{ox}}{\varepsilon_{ox}} . \frac{\sqrt[4]{N}}{\sqrt{W_{eff} Leff}} \ , \tag{2.2}$$

where q denotes the elementary charge, ε_{Si} and ε_{ox} are the permittivity of silicon and oxide, T_{ox} is the gate oxide thickness, and W_{eff} and L_{eff} are the respective effective channel width and length. Also, $Q_B = 2K_B T ln(N/ni)$, where K_B represents the Boltzmanns constant, N is the channel dopant concentration, T the absolute temperature, and n_i denotes the

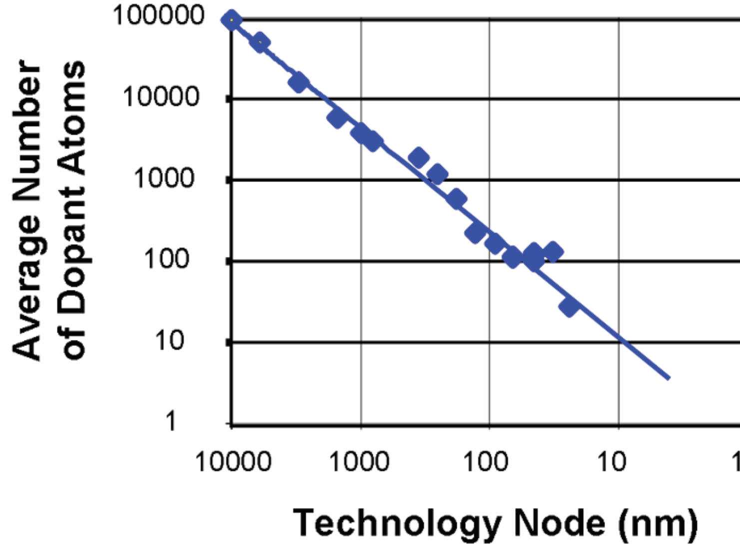


Figure 2.3: Rapid reduction in the number of dopants [12].

intrinsic carrier concentration. Fig. 2.3 demonstrates that the number of dopants is rapidly reduced as technology scales. Therefore, in the current and future technologies adding or removing few atom results in a relative large V_{th} variation. It is evident from (2.2) that the variation is inversely related to the dimensions of the device. Here, increasing the width of a transistor can reduce the variation. However, as discussed in the next chapter, the increase in the transistor size can lead to the parametric yield loss due to the higher power consumption and consequently elevated temperature.

In addition to the random dopant fluctuation, variations in the channel length strongly impacts the V_{th}. As shown in Fig. 2.4, the gap between the feature size of a CMOS device and the wavelength, used in the lithography to manufacture it, is increasing [14]. This diffraction of light, referred to as Optical Proximity Effect (OPE), in turn, results in larger variations in the channel length. In short channel devices, the depletion region extends into the channel and affects the electric field and potential inside the channel. This phenomena is called Short Channel Effect (SCE). Fig. 2.5 demonstrates the impact of channel length variations on the V_{th} roll-off due to the SCE. It is seen that variations in the V_{th} is larger for smaller feature sizes. The following equation expresses the dependency of variations in the V_{th} on the channel length [16]

$$\begin{aligned} V_{th} &= V_{th0} - \lambda_d V_{DS} \\ \lambda_d &= k L_{eff}^{-2.7} \ , \end{aligned} \tag{2.3}$$

where V_{th0} is the long channel threshold voltage, k is a technology dependent parameter, and λ_d is the Drain-Induced Barrier Lowering (DIBL) coefficient. It is evident that, because

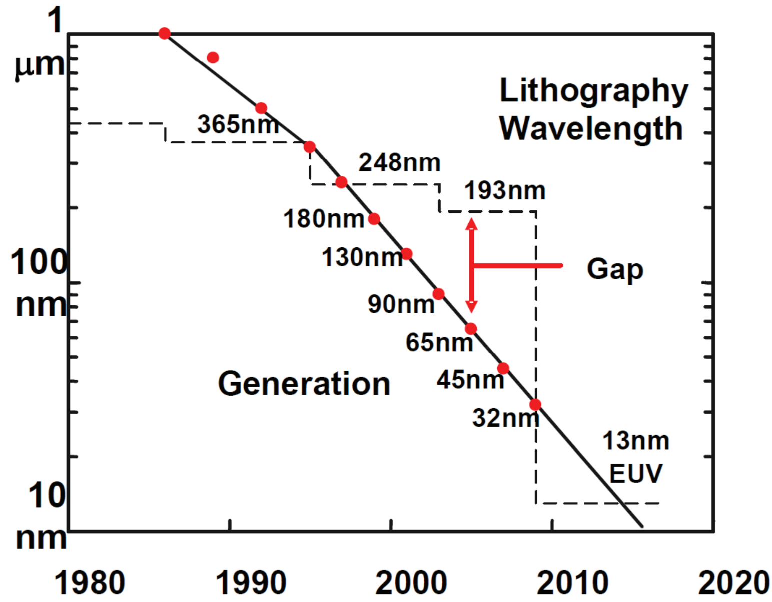


Figure 2.4: Difference between the lithography wavelength and feature size for current and future technologies [14].

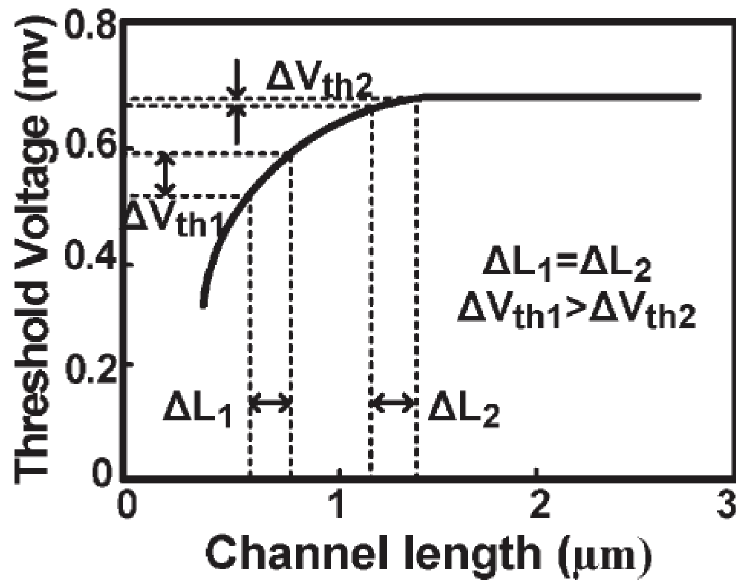


Figure 2.5: Roll-off effect in threshold voltage in respect to the channel length [15].

of DIBL, the threshold voltage is greatly susceptible to the variations in the channel length. Also, the V_{th} is a function of the drain to source voltage (V_{DS}).

There are other sources of variations that cause fluctuation in the V_{th} including Line Edge Roughness (LER), variations in the oxide thickness, channel width, mobility, and oxide charges. LER is associated with the imperfect gate patterning. The uneven poly-gate edges results in large variations in the V_{th} for sub-50nm devices [17]. Variations in the oxide thickness impacts the threshold voltage. However, introducing high-k materials and metal gates has restored the traditional scaling trend of the gate thickness and the variations is well-controlled [12]. Variations in the channel width also change the V_{th} characteristics. The impact is less significant, compared to the fluctuation in the channel length because the width is several times larger. Mobility of charge carriers, in a device, correlates with the drift velocity, the average velocity under the electric field, and the applied electric field. The drift velocity is directly related to the Electric field. Therefore, the variation in the mobility is a function of the variation in the field, temperature, and impurity [18]. Variations in the mobility and threshold voltage are related but the dependency is not significant [19]. Finally, the presence of charges in the oxide affects the mobility and results in the V_{th} mismatch, in different devices. The variation is more pronounced in the interface of high-k materials [12].

2.2.3 Interconnect Variations

Interconnect parameters exhibit large variations. Lithography modifies the line width and line spacing, where they depend on the neighboring pattern in the layout (proximity effect), the location in the layout (lens aberration), and the density of features on the mask (flare) [20]. Variations in the line width and line space impose fluctuations in the line resistance and inter-line capacitance.

In addition, Chemical-Mechanical Polishing (CMP) process is used to remove unwanted metals and have flat topography on the wafer. However, CMP is also layout patter dependent and is subject to variations. Different hardness of interconnect and dielectric material results in the imperfect CMP process. Fig. 2.6 depicts two problems associated with the variations in CMP: dishing and erosion. Dishing results in a thin interconnect and increases for wide interconnects. Erosion causes interconnect and dielectric thinning and is more significant for dense areas. Variations in the line and Inter-Layer Dielectric (ILD) thickness, caused by CMP, lead to larger resistance and thus degrade performance [21]. Fig. 2.7 shows the dependency of the interconnect resistance on the line width and the radius of the dishing. The etching process, used for forming the vias and metal contacts, is also imperfect. Therefore, the uneven etching leads to variations in the thickness of the vias and contacts. This, in turn, imposes variations in the metal resistance [22].

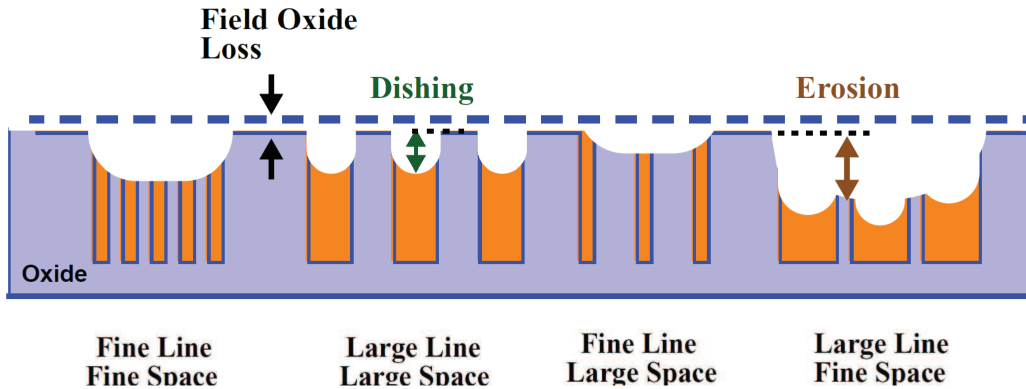


Figure 2.6: Pattern dependent problems of dishing and erosion in copper CMP and the impact on the interconnect height [21].

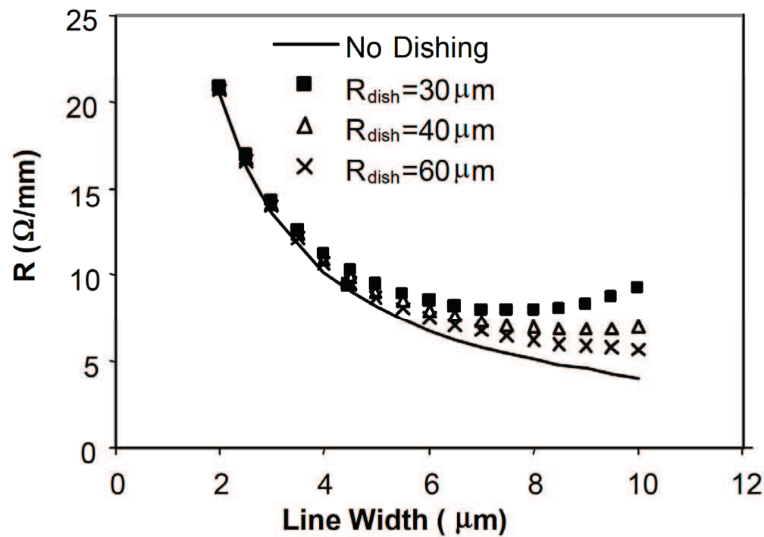


Figure 2.7: Metal line resistance as a function of line width for different dishing radius [22].

2.3 Environmental Variations

Environmental variations are due to fluctuations in the parameters of the environment, where the chip is operating. Variations in the parameters such as the supply voltage, operating temperature, and switching activity impact the performance, power, and reliability of the chip.

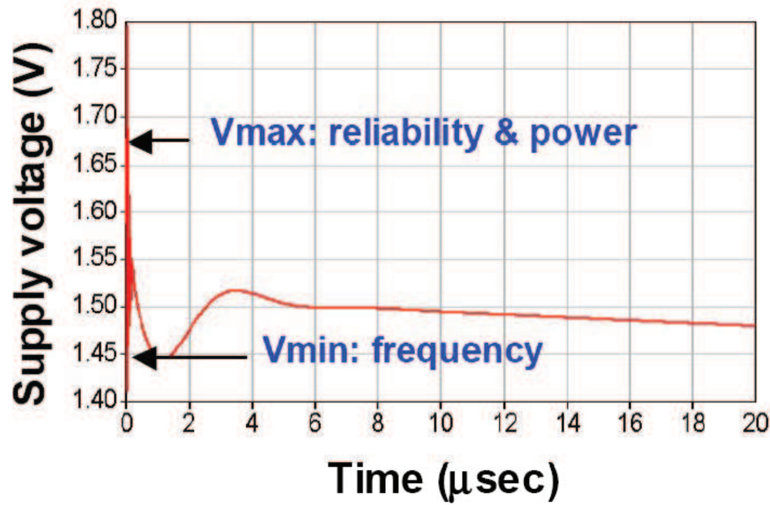


Figure 2.8: Temporal variations in supply voltage [14].

2.3.1 Variations in the Supply Voltage

Variations in the supply voltage are primarily due to the non-uniformity of the distribution of the power supply and the changes in the switching activity of the circuit. The current drawn from the power supply varies for different dies. However, recent voltage regulators are less sensitive to the current [23]. Therefore, the die-to-die V_{dd} variations can be small. Historically, designers tend to limit the within-die changes in the V_{dd}, due to the IR and Ldi/dt drops, to a maximum of 10%. Nonetheless, with the scaling of technology, the increase in the current density and rate of switching make it more challenging to retain this traditional bound on the supply voltage noise [24].

Voltage drop reduces the overdrive current, degrading the performance. An increase in the voltage and switching activity increases the dynamic power. As illustrated in Fig. 2.8, variations in the supply voltage are time dependant. Because of changes in the workload and current flowing in the power grid, the supply voltage and activity fluctuate. The change in the current can be caused by the increase in the leakage current, which, in turn, can be the result of the process or temperature variations. Analyzing the corners of the variations is a difficult task due to the fluctuating work load, temperature, and demanding current. However, it is vital to constraint the supply voltage to a given maximum variations within 5%-10%.

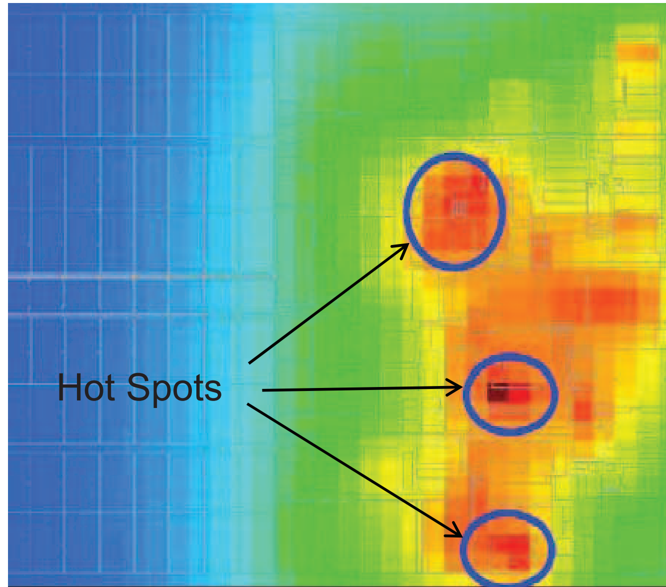


Figure 2.9: Spatial temperature variations of a Pentium M chip running Applu SPEC benchmark [25].

2.3.2 Temperature Variations

Different parts of a chip have different power densities that result in temperature gradients across a chip. Such changes in the thermal map create spatial variations, forming hotspots. This spatial temperature variation can also change over time based on the instruction mixes that are executed, targeted at different functional blocks. Fig. 2.9 indicates the spatial and Fig. 2.10 shows temporal variations in the temperature of a Pentium M chip running the Applu SPEC benchmark [25]. With the scaling of technology, the hotspots move from the points with the highest switching activity to those with the low threshold voltage [26]. This, however, can cause a performance mismatch, increasing the power consumption, and jeopardizing the reliability of a design [27]. To address these issues, a statistical framework is needed to model the variations, and, then, analyze the effect of such variations on circuits and systems.

In addition to the process variations, temperature variations impact the V_{th}. Temperature and leakage power are closely related. Therefore, the within-die temperature fluctuation and, as a result, the average power consumption, varies for different dies. This, in turn, causes inter-die variations in the temperature and V_{dd}. To estimate the spread of the supply voltage, and temperature, the impact of the process variations on these two environmental parameters must be analyzed [28][29]. Higher temperatures exponentially increase leakage power. The interdependency of the subthreshold the leakage and tem-

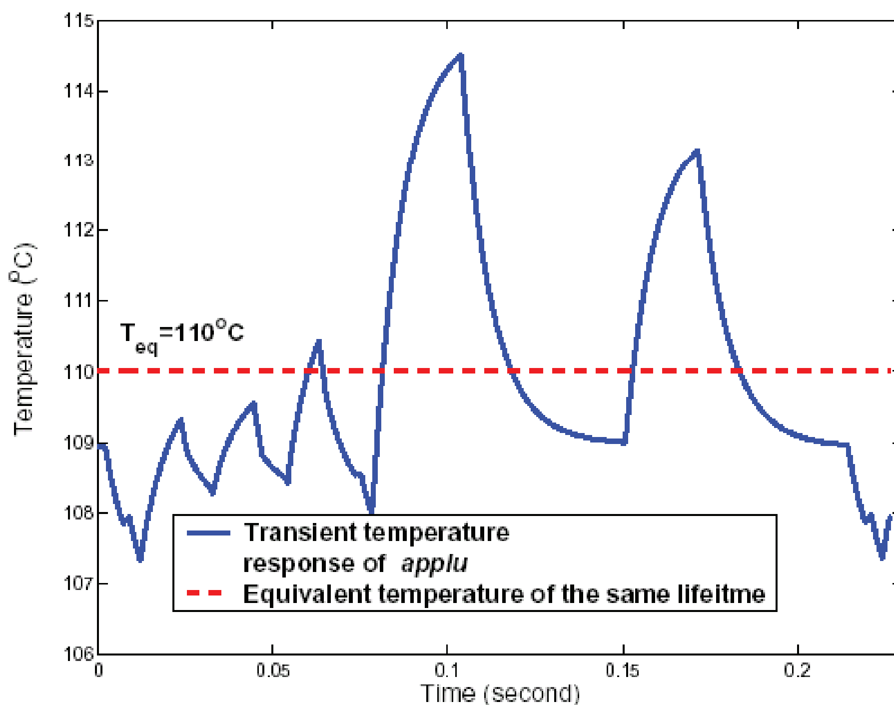


Figure 2.10: Temporal temperature variations of a Pentium M chip running Applu SPEC benchmark [25].

perature can also create a positive feedback that can lead to higher temperatures, and, eventually, thermal runaway. Most mechanisms fail because they depend on the operating temperature and such that the variations pose reliability issues. Therefore, a self-consistent analysis is required to explore the impact of temperature variations [23].

2.4 Other Sources of Variations

In addition to the aforementioned variations, there are other types of variations that occur over a long period of time. Hot carrier injection and negative bias temperature instability (NBTI) cause the threshold voltage to increase over time. Electromigration is another failure mechanism that shrinks the wire width and, in a severe case, leads to open circuits. These variations depend on process and environmental variations. For example, fluctuations in the oxide thickness impact NBTI and hot carrier injection, degrading the performance and leakage power. Also, electromigration depends on the current density and temperature. Thus, the reduction in the width of a wire and increase in the temperature

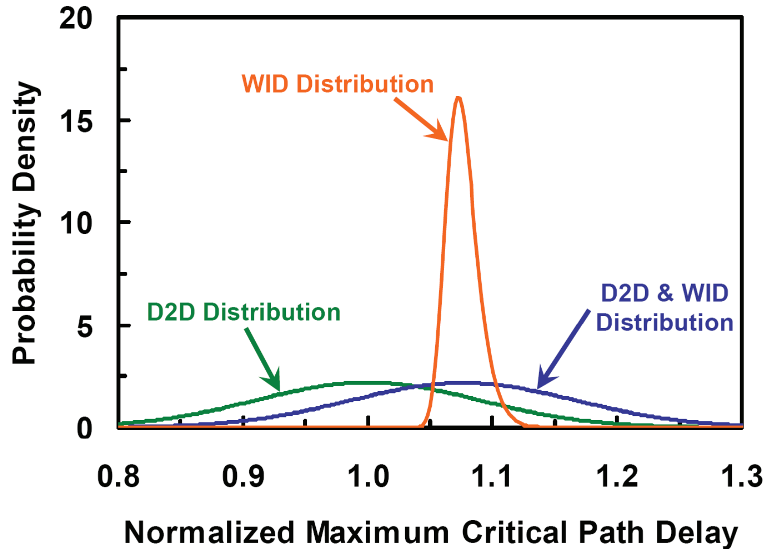


Figure 2.11: Impact of variations on delay [31].

exacerbate electromigration. The time-dependency nature of these variations make it difficult to investigate their impact in a short period of time. Consequently, burn-in tests are employed to accelerate such phenomena. During the burn-in test, which is time consuming and expensive, the chips are placed under current and temperature stress to show their possible vulnerability to the time-dependant failure mechanisms. Sachdev et. al show that the burn-in environment can change due to the increase in the leakage power [30]. Variations in the temperature and threshold voltage can cause such an increase. Therefore, it is crucial to take into account the different sources of variations in a statistical analysis to avoid an unexpected yield loss.

2.5 Impact of the Variations on Design

With the scaling of the CMOS technology, the impact of the V_{dd}, V_{th}, and temperature variations on performance, power, and reliability becomes more significant.

2.5.1 Impact on Performance

Variations in the threshold voltage impose fluctuations in the delay. Fig. 2.11 demonstrates the distribution in the delay for the different types of variations. Device performance is related to the difference between the V_{dd} and V_{th}, referred to as the overdrive voltage

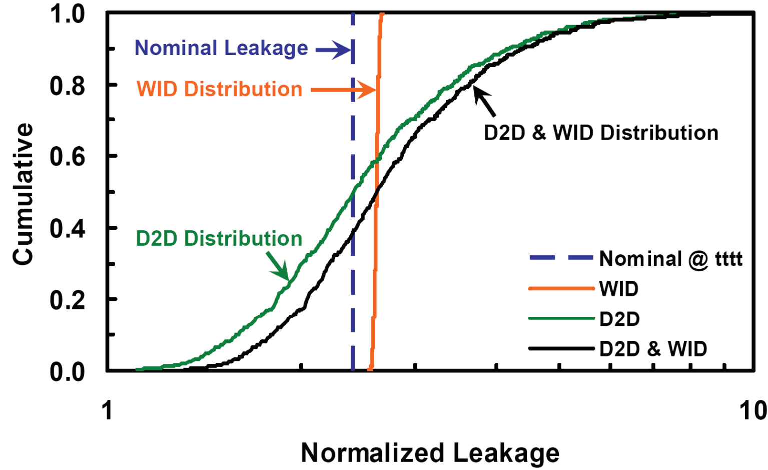


Figure 2.12: Impact of variations on power [33].

(V_{dd} -V_{th}). As technology scales, the V_{dd} is reduced to maintain a constant electric field across the gate-oxide, and to limit the increase in the power density to ensure reliability. The reduction in the V_{dd} causes the gate delay to increase. As a result, the V_{th} is also reduced to maintain an acceptable performance. Although, the values of the V_{dd} and V_{th} decrease, the magnitude of their variations becomes comparable to their nominal values. Consequently, in scaled technologies, the variations in the overdrive are then comparable to the overdrive's nominal value such that the effect of the variations on the performance is more pronounced. As an example, a 10% variation in the V_{dd} can cause a 20% variation in the delay [32]. Moreover, the mobility of the charge carriers and the resistance of the interconnects depend on the operating temperature. Thus, variations in the temperature leads to variations in the device and interconnect delay [18].

2.5.2 Impact on Power Consumption

The subthreshold leakage power, exponentially, depends on the V_{th}. This dependency is clearly illustrated in Fig. 2.12, where the cumulative distribution functions of the leakage power for the different types of process variations are identified. Also, the dynamic power consumption has a quadratic relationship with the power supply. Thus, an increase in the magnitude of the variations in the V_{th} and V_{dd} impacts the principal components of the total power consumption.

The dependency of the mobility and threshold voltage on temperature, makes performance sensitive to the variation in temperature. This dependency, however, is the source of uncertainty in the performance and can lead to a timing yield loss. In addition to performance, the exponential relationship between the subthreshold leakage and temperature

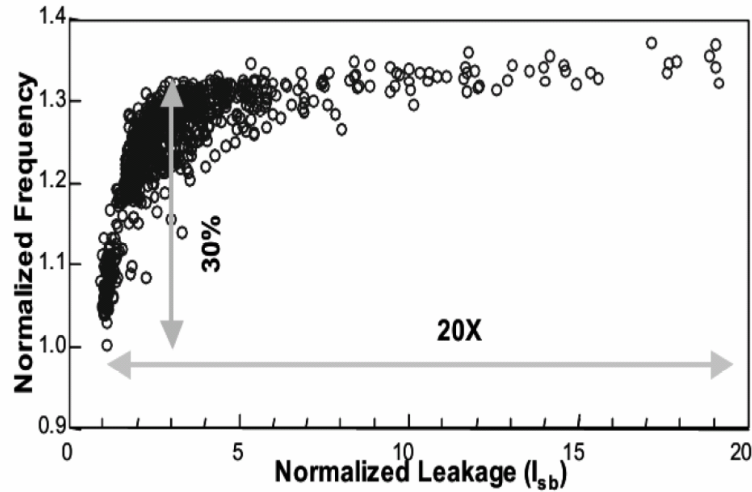


Figure 2.13: Large variations in leakage power and performance are due to process variations, at 130nm [14].

indicates it is crucial to account for the temperature variations in any power analysis. The effect of the temperature variations on design is a multi-dimensional problem that must be solved by including the supply voltage, threshold voltage, delay, leakage and dynamic power, and temperature. These parameters, together, address the electrothermal coupling in the design.

Fig. 2.13 depicts that for a 30% variation in frequency, the leakage current varies up to 20 times, due to the process variations. As it can be seen, many design samples at the two extreme sides fail to meet either the timing or leakage power constraints. It is a clear example that shows the effect of the growing variations and their impact on the different aspects of a design.

2.5.3 Modeling Variations

Different models at different stages of a design are used to model power and performance. However, these models are not perfect and result in different error magnitudes. Some of the models are more aggressive where as others are more conservative. The former models lead to a yield loss, and the latter ones reject many designs, because they do not meet the constraints. The impact of aggressive models is a huge overhead and in some cases, overdesign. Therefore, the conservative models are preferred, while taking into account the modeling variations.

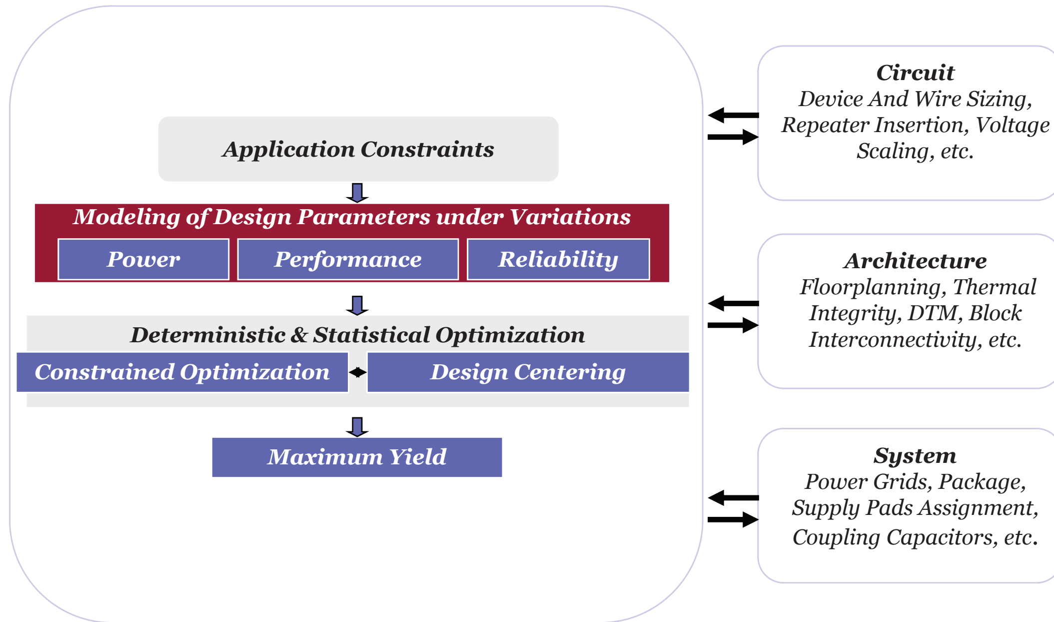


Figure 2.14: The existing approaches for yield analysis and optimization are categorized into different design domains.

2.6 Related Work: Design under Variations

Several researchers have addressed the variability-aware design and optimization of power, performance, and yield under variations [34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 14]. Fig. 2.14 depicts different design domains at which the parametric yield has been studied in the literature. Some focus on circuit-level analyses, where as others study the impact of variations at the architecture, and system level. CAD methodologies have also been examined to increase the robustness of the design.

2.6.1 Variability-Aware Circuit Design

2.6.1.1 Power-Performance Trade-Offs

Sylvester et al. attain a gate-level parametric yield estimation by obtaining the correlation between performance and leakage power [34]. Initially, the inter-die and intra-die threshold voltage variations, mostly due to the variations in the channel length and dopant fluctuation, are modeled. Spatial correlation is also taken into account by partitioning a circuit (Fig. 2.15) and assigning random variables. Then, the joint probability distribution

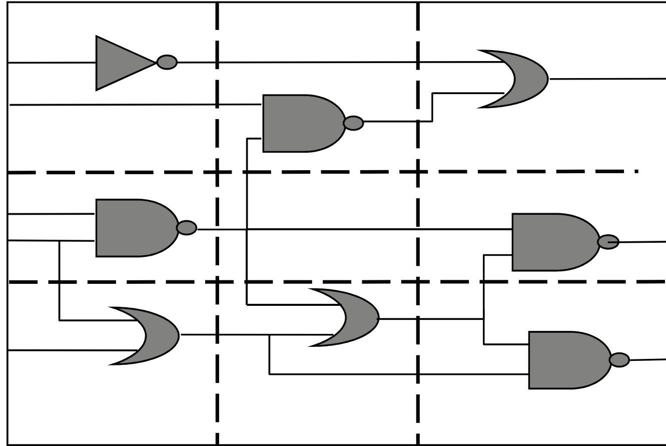


Figure 2.15: Partition of a circuit to model the correlated component of variation [34].

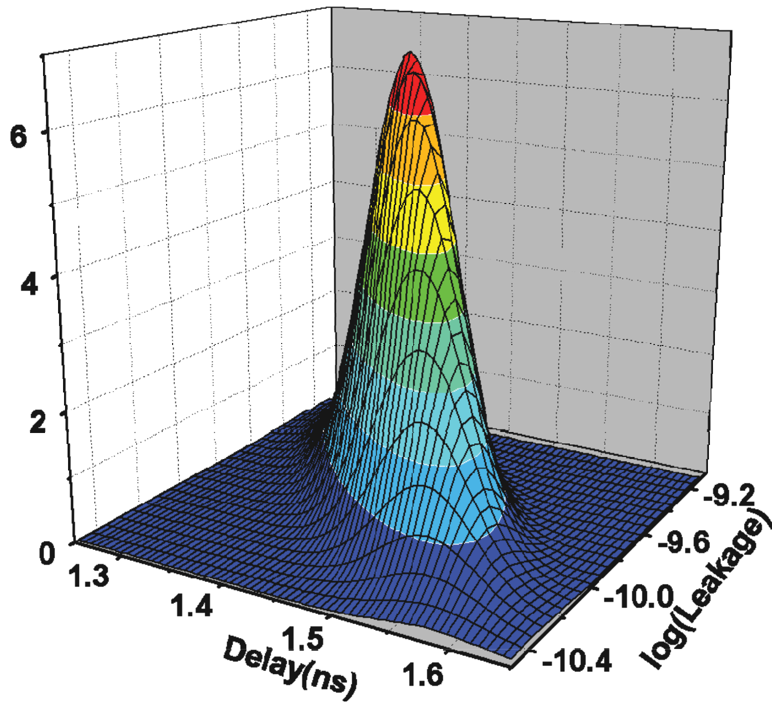


Figure 2.16: Joint probability distribution function for the bivariate Gaussian distribution for c3540 [34].

function is used to estimate the yield to avoid the error that is imposed, if the performance and leakage are modeled independently. The distribution is illustrated in Fig. 2.16.

More specifically, some analyses have been carried out in the literature to investigate

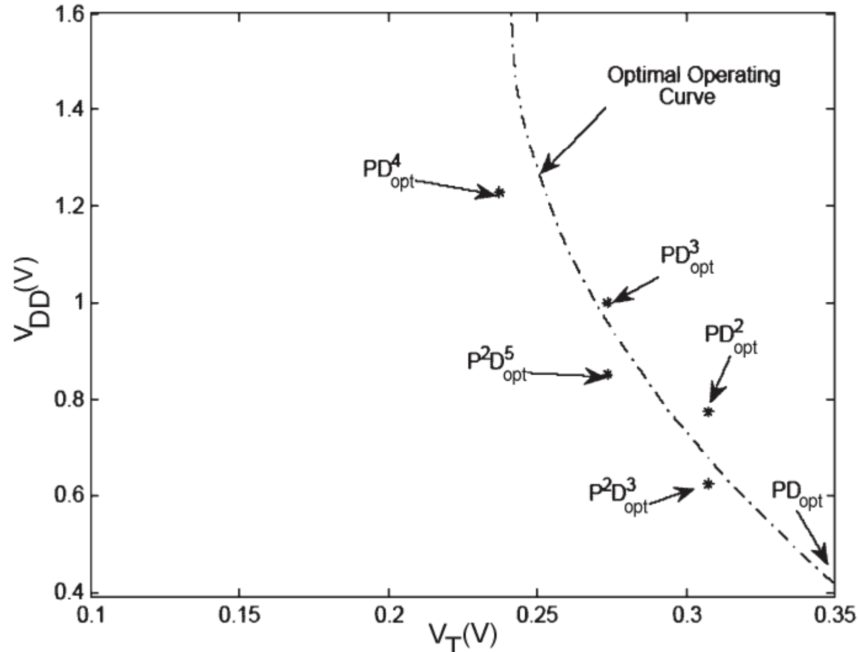


Figure 2.17: Optimal operating line and different optimal metrics [46].

the impact of the variations in the V_{dd} and V_{th} on the parametric yield. The impact of leakage on the parametric yield is analytically shown, and the sensitivity of the yield to the V_{dd}, power and performance is examined by Rao et al. [47]. However, this conservative analysis is based on the minimum and maximum channel lengths and the V_{th}. In addition, the effect of temperature on the leakage power is ignored. Gonzalez et al. have investigated the effect of scaling the V_{dd} and V_{th} on energy and delay [48]. Also, they have demonstrated that optimizing both voltages saves energy, and increases performance. The effect of the V_{dd} and V_{th} variations on their design metric, Energy Delay Product (EDP), has also been addressed, but the EDP, under uncertainty, is obtained by the multiplication of the energy and the delay at the four corners of the supply voltage and temperature. The electrothermal effects have been incorporated in the optimization by Banerjee et al. to account for the high temperature impact on performance, power, and reliability [49].

The variations in the V_{th} have a larger impact on the EDP, if the electrothermal coupling is taken to account. In particular, Sengupta and Saleh have proposed more general metrics ($P^m D^n$ and PT^μ) to give priority either to the power, or the delay for specific applications [46][50]. It has been demonstrated how a design metric such as the EDP changes with the variations in the V_{dd} and V_{th}. From Fig. 2.17, it is evident that all the optimal points for the trade-off between the power and performance, measured using $P^m D^n$ metric, lie along the trajectory of the best operating point (unconstrained optimal).

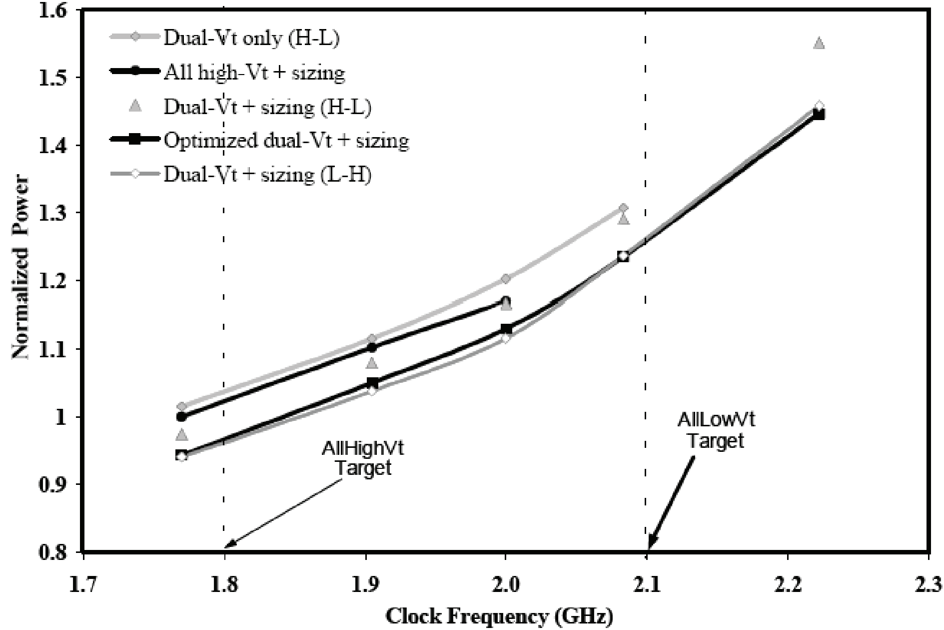


Figure 2.18: Total power as a function of frequency for transistor sizing and dual V_{th} assignment [51].

2.6.1.2 Threshold Voltage Assignment

Dual V_{th} assignment is an effective leakage power reduction technique [51], where high V_{th} transistors are used in circuits with delay slack. Fig. 2.18 compares different methods for the threshold voltage assignment and transistor sizing. It can be seen that iterative sizing and dual V_{th} assignment leads to best results. The dual V_{th} technique has been revisited by Agarwal et al. to account for all leakage components and the variations in the delay and leakage to maximize the leakage power saving [41]. The simultaneous sizing and dual V_{th} design are applied, where the impact of the halo profile on the variations of the threshold voltage is considered. Since a high threshold voltage (high-V_{th}) has larger variations due to the high halo doping concentration, a device-aware dual-V_{th} is needed to minimize the leakage, while guaranteeing the yield. It is claimed that a 10%-20% extra leakage power is saved, compared to the conventional dual-V_{th} design [41].

The total power has been minimized under timing yield constraints by Devgan et al. [35]. The power reduction is achieved by simultaneous gate sizing and the V_{th} assignment. Based on the power-delay sensitivity, time slacks are assigned to the certain gates in a circuit to reduce the power under variations in the channel length and threshold variations. As demonstrated in Fig. 2.19, the statistical optimization results in a better trade-off between power and performance.

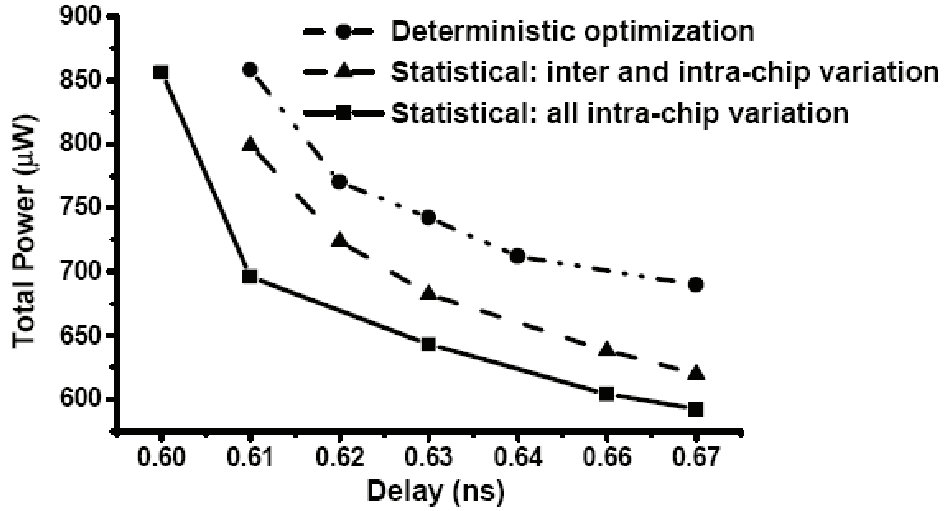


Figure 2.19: Power delay curves for 99.9% timing and power yield [35].

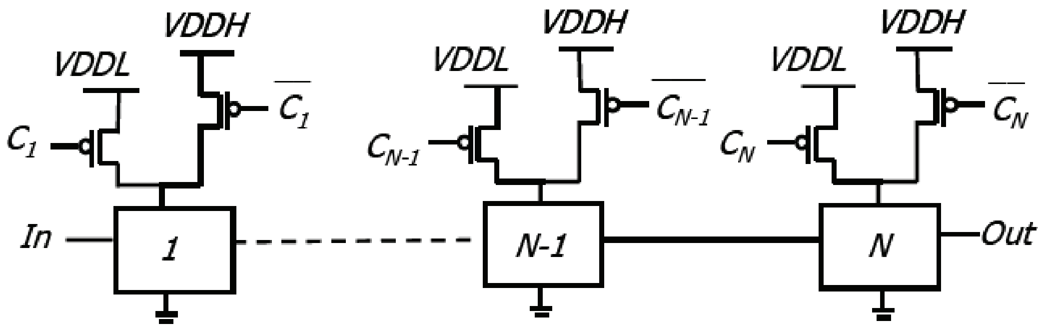


Figure 2.20: Schematic for achieving multiple operating modes [40].

2.6.1.3 Adaptive Solutions

Adaptive techniques have been reported to alleviate the impact of variations on power and performance of a circuit. Agarwal and Nowka have proposed an adaptive technique to reduce the spread in the delay [40]. As depicted in Fig. 2.20, the circuit technique comprises a combination of two supply voltage levels, VDDH and VDDL. When computationally intensive jobs are executed, the entire logic operates at a high level. Performance is traded or power, when a high performance is not needed for some logic runs at the VDDL. The authors compare their method with Clustered Voltage Scaling (CVS) and dynamic voltage scaling (DVS), and argue that, unlike CVS, the adaptive switching capability, between the different modes, renders it a dynamic approach. Also, by using static voltage levels, no

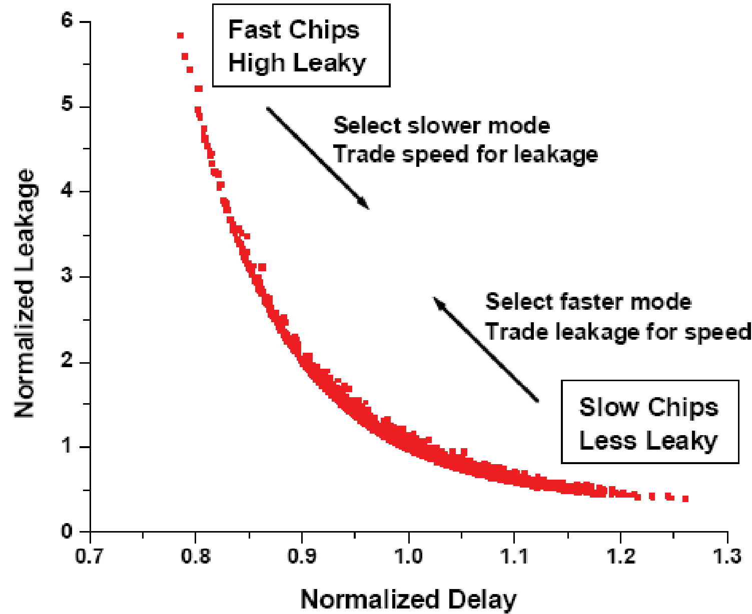


Figure 2.21: Leakage vs. delay spread due to process variation [40].

voltage converter is needed, and, therefore, it does not have the overhead of the DVS. In addition, it is claimed that, unlike the DVS at the system level, the method is applicable at a finer granularity and has the potential to save more energy. Tightening the distribution of the delay is achieved by switching to the opposite mode if necessary. For example, the circuit switches to the slow mode for operating at a high performance, where the leakage is high, and it switches to the fast mode when the leakage is very low and the delay is high. Comparing Fig. 2.21 and Fig. 2.22 demonstrates the effectiveness of the adaptive solution in tightening the power and performance distributions.

An Adaptive Voltage Scaling (AVS) is proposed by Elgebaly and Sachdev [43]. The AVS emulates the actual critical path under various conditions of the process. Tracking the critical path helps to avoid a large margin, required for the delay to ensure error-free operation. A customized path delay is programmed to track the critical path on the chip. This is to reduce the margin required by conventional circuits under voltage and temperature variations. Such tracking across different processes and interconnect parasitic corners achieves more energy efficiency, compared to open-loop or closed-loop systems.

Tschanz et al. study bidirectional Adaptive Body Bias (ABB) and, as a result, show a reduction in the frequency variations by a factor of seven [52]. The threshold voltage, and, therefore, leakage power and performance, can be controlled by applying non-zero voltage to the device body in respect to its source. Reverse Body Bias (RBB) increases the V_{th}, and, thus, reduces leakage power. Forward Body Bias (FBB) reduces the V_{th} and improves

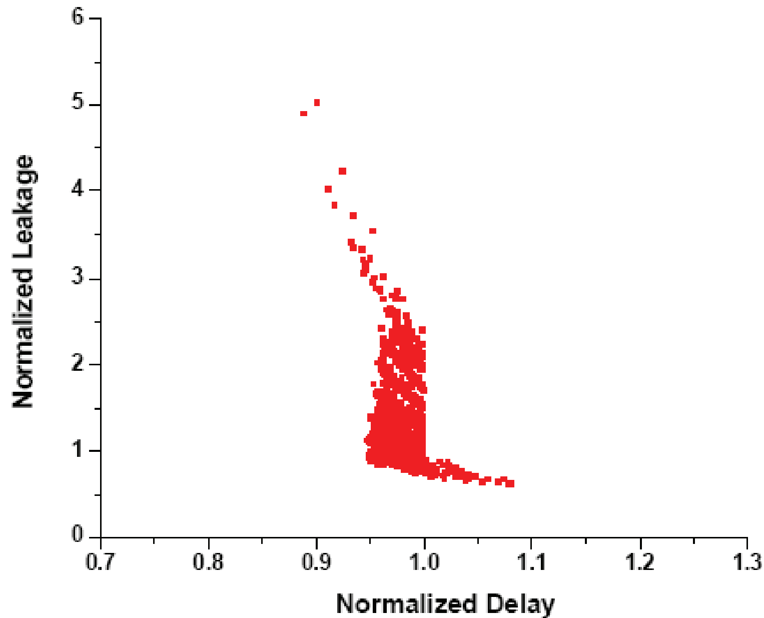


Figure 2.22: Impact of variations on power [40].

its roll-off, and, therefor, enhances the performance. Fig. 2.23 shows the block diagram of the test chip. The phase detector compares the target frequency with the frequency of the critical path. The counter and the bias selector provides a suitable body bias according to the frequency difference. This mechanism can be utilized to reduce the impact of process variations and have more dies in the highest frequency bin. Fig. 2.24 demonstrates that utilizing both ABB and AVS yield better performance, where a large number of dies ends up in higher frequency bin.

Chen and Naffziger compare the effectiveness of adaptive body bias (ABB) and adaptive supply voltage (ASV) in reducing the variability and improving power and performance [37]. In the post-silicon tuning, ABB changes the threshold voltage by either a forward or a reverse body bios. This tightens the distribution of the maximum frequency and power. In addition, the ASV has similar effect on power and performance. It is argued that both the ABB and ASV are effective in trading performance for power and visa versa. A little difference exists between the two methods, attributed to their physical complexity, requirements for voltage regulation, and silicon overhead.

2.6.1.4 Voltage Scaling

Sylvester et al. first compare two existing V_{dd} assignments, CVS and Extended Clustered Voltage Scaling (ECVS) [44]. Both the CVS and ECVS traverse from the Primary Output

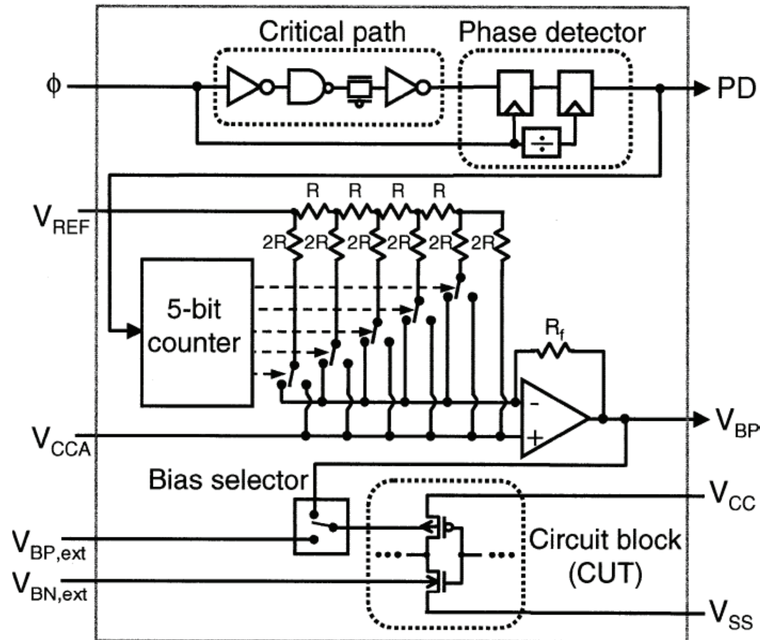


Figure 2.23: Block diagram of ABB test chip [52].

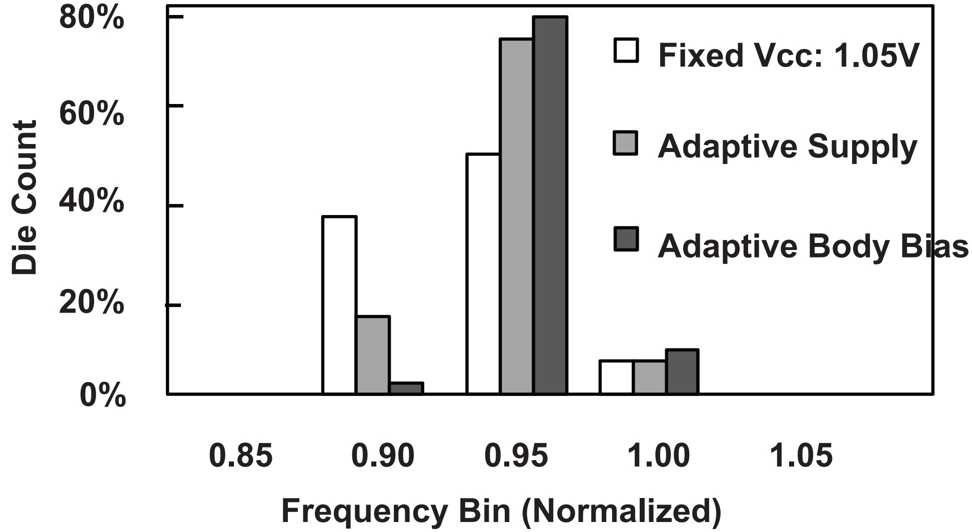


Figure 2.24: Comparing the effectiveness of adaptive solutions [53].

(PO) to the Primary Input (PI) and assign voltages to the gates in a levelized manner. However, the CVS cells with the V_{ddL} cannot drive those with the V_{ddH}. This is due to the imperfect switching of the driven cell, causing a huge leakage current in the gate. For the

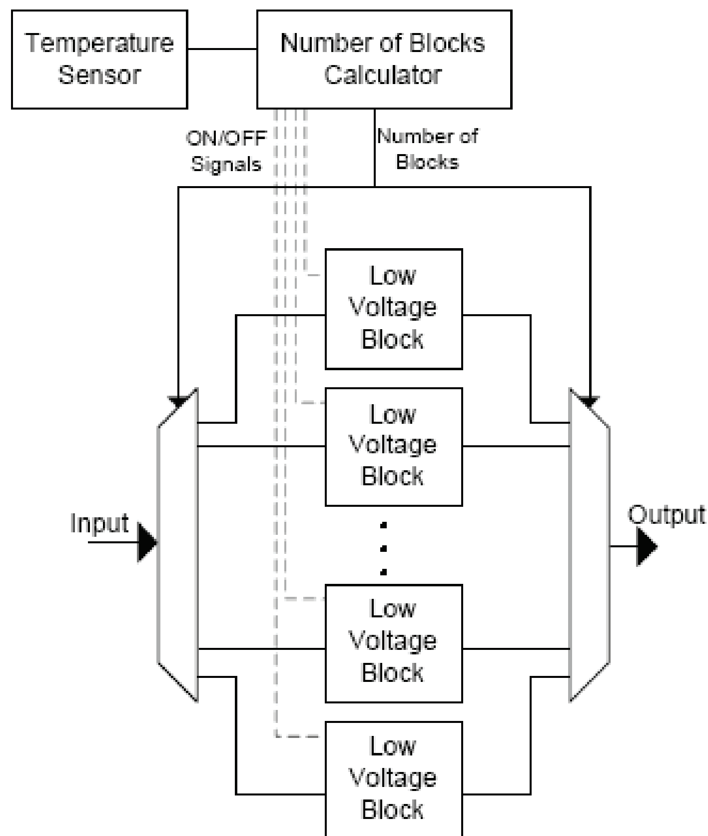


Figure 2.25: Temperature dependent deactivation scheme [42].

CVS, the level conversion occurs at the output of sequential elements, called synchronous level conversion. Extra power saving is achieved in the ECVS by using Asynchronous Level Converters (ACL). They facilitate a level conversion at the output of a V_{ddL} driven cell. A greedy algorithm, presented by the authors, makes it possible to remove the limitation of visiting the cells for a voltage assignment in a legalized manner. It is also observed that the level converters significantly impact the system-level power consumption.

Najm et al. report a methodology to minimize the power consumption of a parallel system design, considering within-die process variations [42]. Their results show that the optimum supply voltage is higher than that obtained when no within-die variation is taken into account. It is also observed that changes in the temperature can have a substantial impact on the selection of the optimum supply voltage and on the power consumption of a parallel system. Although body bios has been used in the literature to reduce the impact of temperature variations, the work proposes a Temperature Dependant Deactivation Scheme (TDDS), illustrated in Fig. 2.25. It relies on a temperature sensor to calculate the number

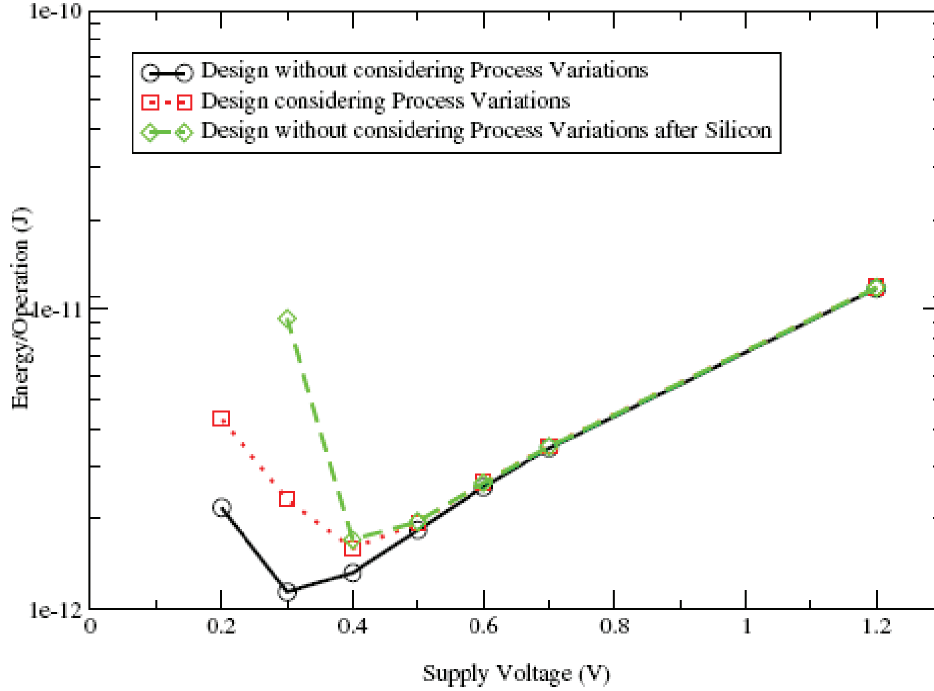


Figure 2.26: Effect of WID process variations on Energy/Operation [42].

of required parallel blocks, based on the operating temperature. This is claimed to reduce the power variation by lowering the temperature variations, and, consequently, lets the designer select a lower supply voltage and thus save more power. Fig. 2.26 shows that if process variations are not considered, the difference between the expected energy (solid line) and the circuit energy after silicon (top curve) is large.

Alioto and Palumbo evaluate the delay sensitivity to the variations in the supply voltage [38]. For this, several full adders with different topologies are examined. It is demonstrated that the delay sensitivity is reduced with the increase in the V_{dd}. Therefore, the increase in the supply voltage not only increases the delay, traded for a higher power, but also leads to a lower sensitivity in the delay. It is also pinpointed that the technology scaling results in a higher sensitivity in the delay in respect to the V_{dd} which highlights the importance of the supply voltage scaling in recent technologies.

Jha et al. propose a two-phase approach for DVS and ABB to deterministically optimize both the dynamic and leakage power consumption for distributed real-time embedded systems [39]. The objective is to perform the DVS and ABB to optimize the trade-off between the power and execution time. Therefore, for a given clock frequency, an optimal supply voltage and body bias voltage are obtained. Then, slacks are allocated for a set of tasks for a precedent relationship and real-time constraints. The results, shown in

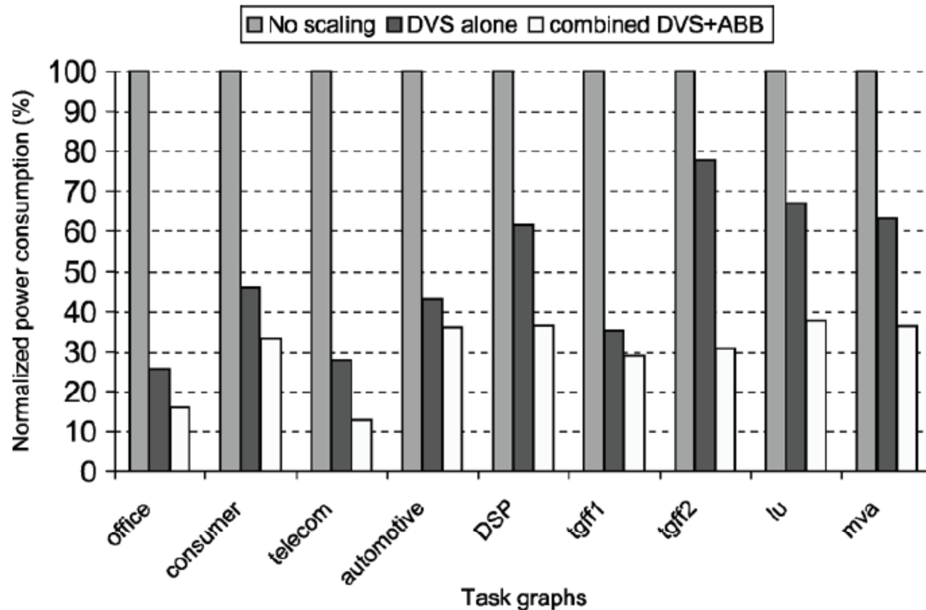


Figure 2.27: Normalized power consumption of three different schemes for the $70nm$ technology [39].

Fig. 2.27, indicate a large reduction in the total power when both DVS and ABB reduce dynamic and leakage power.

2.6.2 Design Optimization at Architecture Level

Elevated temperatures introduce a wide range of issues in IC design. On one side of the spectrum there is a reduction in the carrier mobility of a single device, and, on the other side the heat transfer and thermal reliability of the chip.

2.6.2.1 High Temperature Effect

Several studies have been conducted on the high temperature effects on the reliability and performance of integrated circuits. Pedram et. al, have conduct a survey on thermal analysis and management in VLSI circuits [56]. Elevated temperatures, in this report, are identified as a key source of power consumption that impacts the reliability and performance of CMOS circuits. Ajami et al. have investigate the effect of a non-uniform temperature distribution on the substrate regarding the performance of the interconnects, clock skew, and *IR* drop [57, 58]. It is found that high temperature gradients degrade

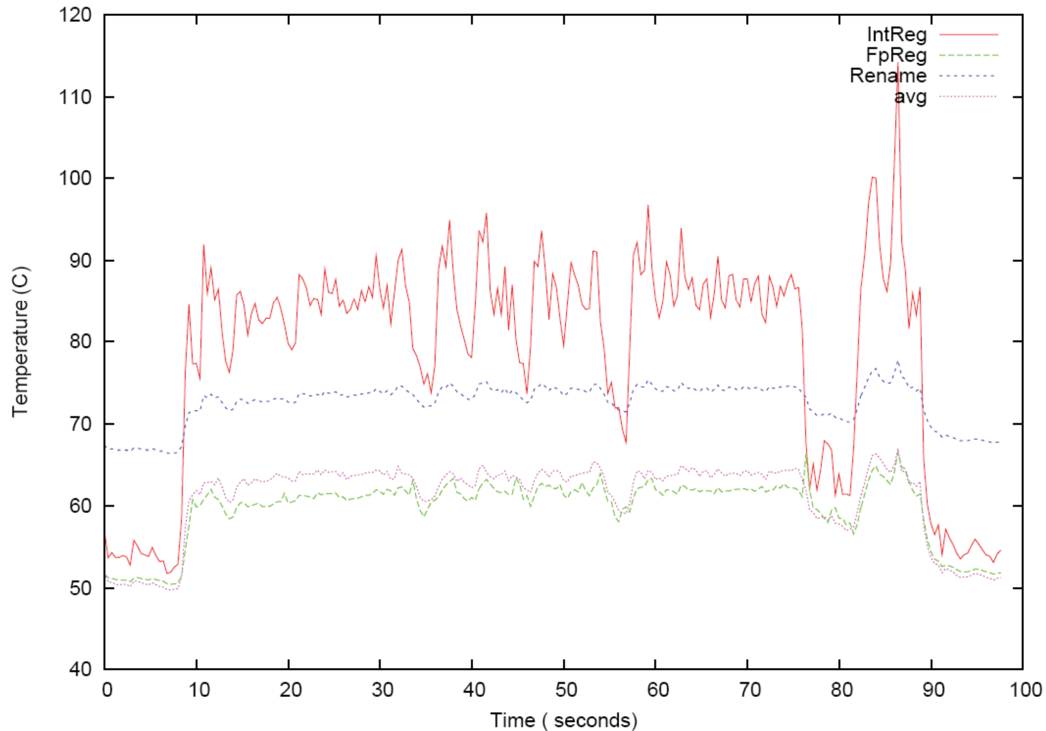


Figure 2.28: Block temperatures for gcc benchmark [54].

the interconnect performance and clock skew. To alleviate such impact, Srivastava et al. propose an electrothermal analysis tool for managing the hot-spots [59]. As seen from Fig. 2.28, these hot spots can move overtime from one architectural block to the next.

2.6.2.2 Temperature-Aware Floorplanning

Floorplanning has proved to be effective in reducing the peak temperature [60]. At the architecture level, the designer has information regarding the interconnection of the architectural blocks, block activities, and DTM policies [61]. Subsequently, the available information can be utilized to address such issues as the temperature variations more effectively.

Most of the existing work on floorplanning focus on optimizing the area and wire length. An extensive survey on floorplanning algorithms, most of which use simulated annealing, has been presented [62]. Architectures with different floorplans have also been compared to meet the performance and thermal constraints [55, 63]. Fig. 2.29 exhibits the difference in temperature variations for two different floorplanning techniques. Considering the thermal coupling, with the neighboring block around the multiple cores results in lower average

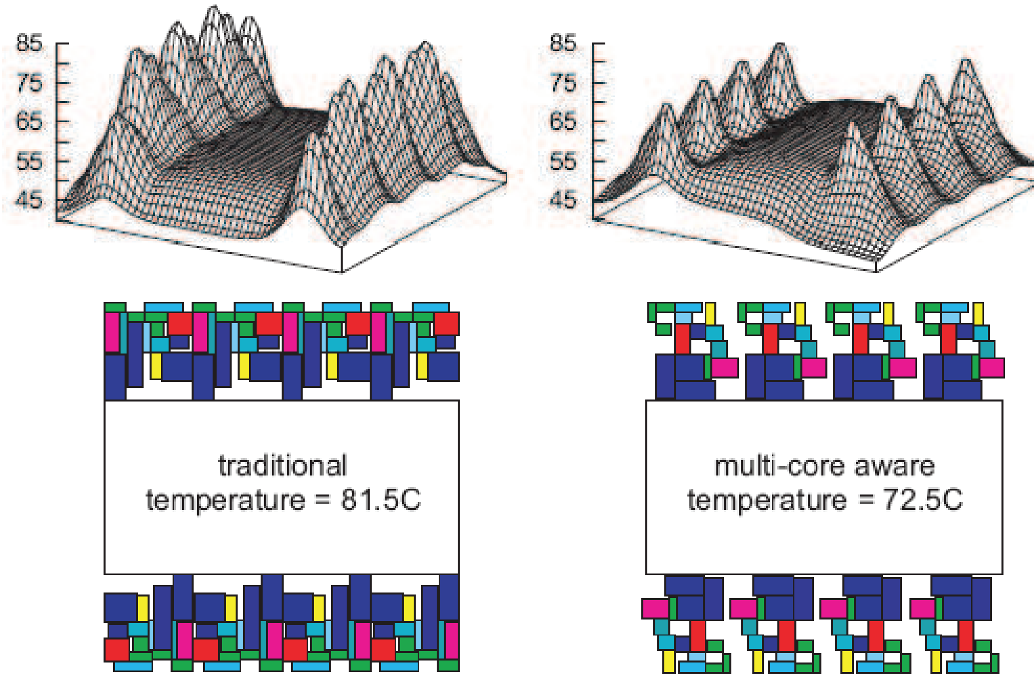


Figure 2.29: A temperature profile comparison between two floorplanning methods [55].

temperature.

HotFloorplan has been presented to minimize the peak temperature [60]. Zhou et al. propose a temperature-aware floorplanner for three dimensional integrated circuits [64]. A temperature-dependent leakage model is also included to connect the feedback loop between the temperature distribution and the leakage power consumption. The idea in the optimization process is to reduce the area, white space, wire length, and via count. Also, the layer assignment and global optimization are integrated. In addition to these studies on thermal integrity and minimizing the number of hot-spots, [65][66] have investigated the impact of leakage power in a system on chip. Gupta et al. have provided an optimization guideline for leakage-aware floorplanning [65]. Mogal and Bazargan have proposed an algorithm for leakage reduction by modeling the temperature dependant leakage on the thermal profile [66].

2.6.3 Effects of Process Variations at the System Level

Variations in the supply voltage are related to several global components such as the power distribution network, package, and coupling capacitors. Fig. 2.30 illustrates a RC network

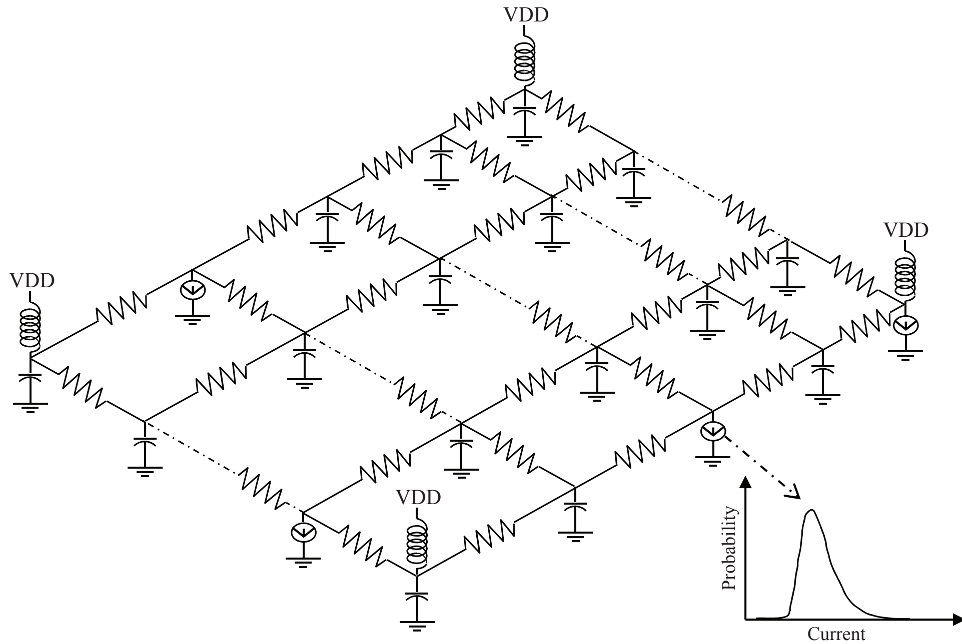


Figure 2.30: RC network representing a power grid under variability.

representing a power grid under variability, where the variations impose statistical measures on the current flowing into the underlying circuits. To address the variations, some system level information such as the network resistivity, block power consumptions, and thermal profile, is needed. Therefore, some analyses have been undertaken and some optimization methods have been studied at the system level to deal with the variations.

2.6.3.1 Power Grid Verification

Ferzli and Najm provides a statistical model for voltage drop due to the current noise induced by leakage variations [36]. With this model, a verification method is developed to find the parts of the power distribution network that are susceptible to such variations. Fig. 2.31 demonstrates the distribution of the difference between the nodes upper bound voltage and a user-defined threshold value. This difference is a figure of merit for the statistical verification procedure and determines if a node on the power grid is safe. The methodology avoids pessimistic conclusions that can arise due to worst-case studies by analyzing the within-die leakage variations.

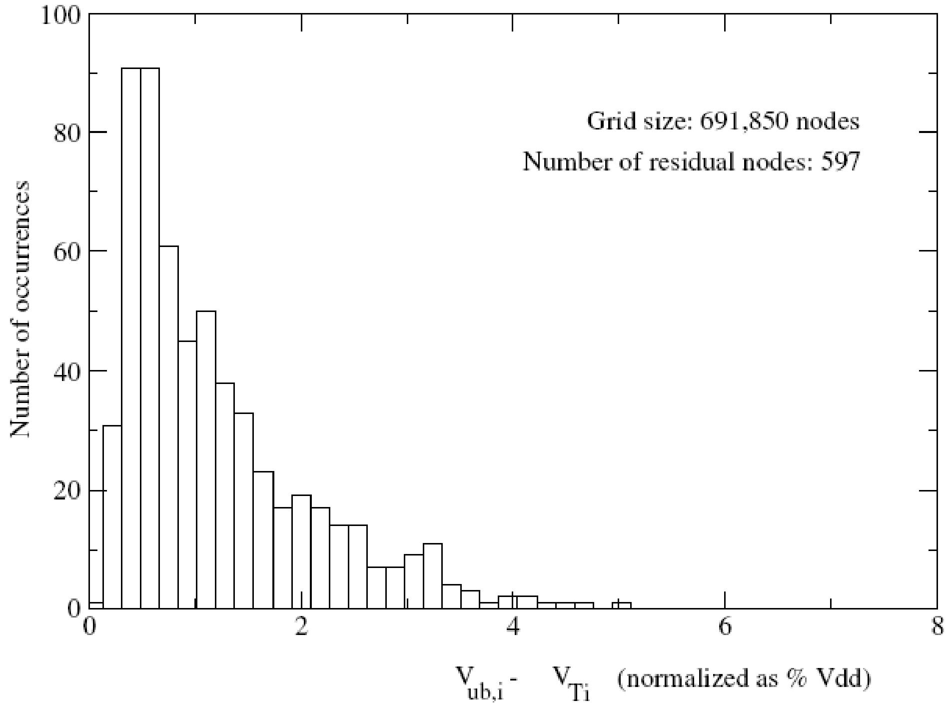


Figure 2.31: Distribution of the upper bound of the supply voltage in respect to a user-defined threshold value [28].

2.6.3.2 Statistical Static Timing Analyses

Maurine et al. propose a sensitivity-based timing analysis technique to capture the impact of the temperature and supply voltage variations on the timing of combinational circuits [45]. By using the methodology, the performance of a design can be computed under different temperature and voltage conditions. Since performance is strongly temperature and supply voltage dependant, to alleviate the complexity of the verification step, an analytical timing model is developed to relax the pessimistic margin suggested by a corner-based approach.

Najm et al. propose a Static Timing Analysis (STA) methodology, where the maximum delay of a circuit is obtained by considering the mismatch between the power supply of successive gates on a path [67]. In conventional STA approaches, the maximum delay of the circuit is computed at the worst corners of the supply voltage. However, in the presented timing verification technique, the dependency of the circuit delay on the local gate delay is not sufficient, such that a global approach, considering the voltages on the power grid, is needed. Instead of solving a computationally expensive vector-based grid analysis, a

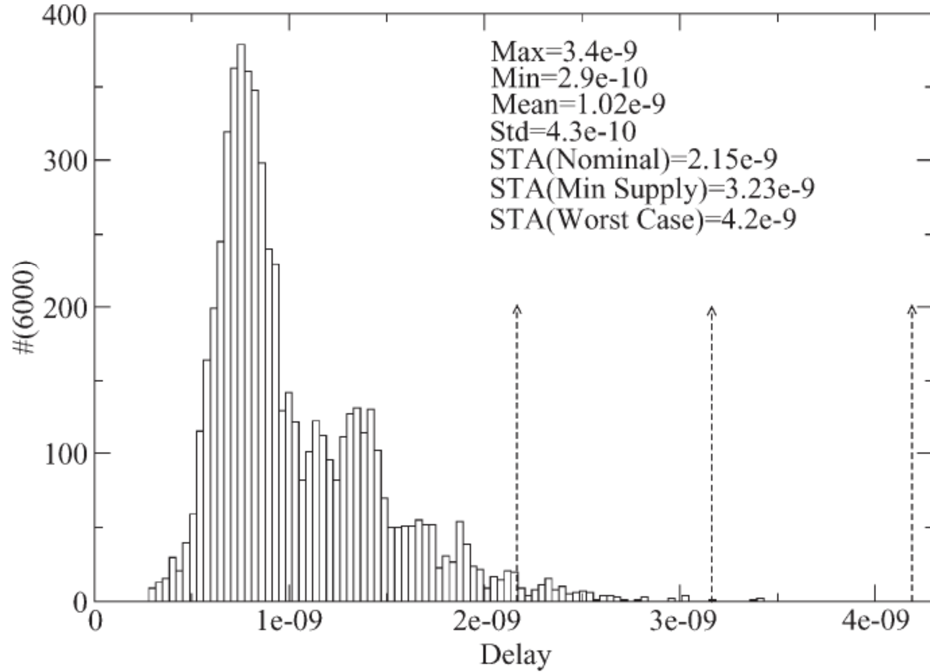


Figure 2.32: Histogram of C880 circuit delay under supply voltage variations [67].

vectorless technique is used in which the upper bounds of the circuit currents are employed as constraints on the current sources of the grid. The delay of the circuit is expressed as a function of the supply voltage variations. Fig. 2.32 shows the histogram of the delay. As seen from the figure, applying minimum supply voltage to all the gates does not lead to the worst case delay. The actual worst case delay is calculated by considering the voltage variations and supply mismatch.

Ferzli et al. propose a full-chip model that models die-to-die and within-die process variations by a generic parameter model [68]. The generic nature of the model provides the ease of use for taking the impact of the variations into account for analyzing the static timing before layout. Although the worst case device file setting calls for a setting of $\triangle L = +3\sigma L$, the generic model uses $\triangle L = +\delta\sigma L$. The δ is modeled as a function of the timing yield. This generic formulation helps to continue the use of existing STA tools early in the design, where the layout information is not available. The timing analysis is performed by utilizing a generic critical path by examining the statistical properties of such a path. The yield loss is estimated by a statistical model for the process variations.

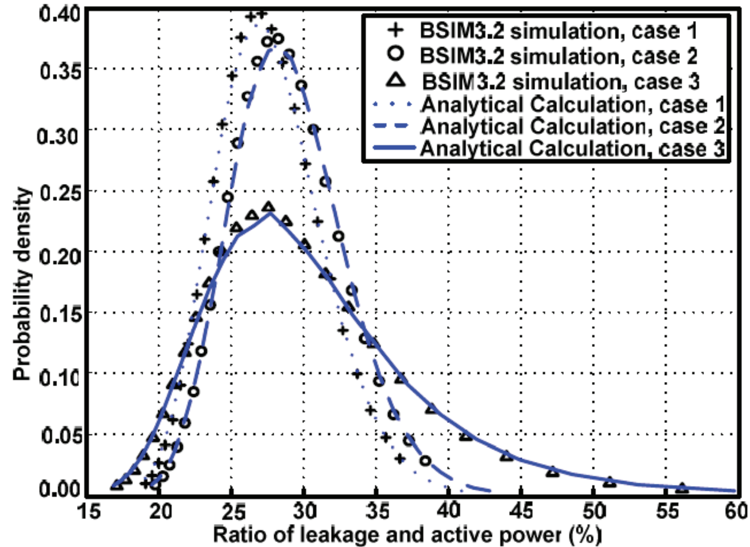


Figure 2.33: Probability density function of leakage to active power ratio for three cases: die-to-die (case 1), adding within-die (case 2), and considering die-to-die temperature variations (case 3) [23].

2.6.3.3 Power and Temperature Estimation

Full-chip leakage estimation in the presence of supply voltage and temperature variations has been preposed by Nassif et al. [5]. By employing an iterative method, the voltage and temperature profile of a chip are obtained, and a closed-form model is applied to attain the leakage profile of the chip.

Process, voltage, and temperature variations, and their impact on the circuit and microarchitecture are examined by Borkar et al. [14]. When the number of critical paths increases, the mean of the frequency distribution is reduced. In addition, the variations in the delay increases with the reduction of the logic depth. Therefore, the microarchitecture designs that increase the critical path or reduce the logic depth also reduce the probability of meeting the target performance. The authors suggest the ABB, as well as the temperature and supply voltage control techniques to increase the tolerability of the design to variations.

A probabilistic framework is presented for full-chip subthreshold leakage estimation under process, voltage, and temperature (PVT) variations [23]. In addition, a sensitivity analysis of the PVT variations is conducted for the leakage power. The impact of the variations on the yield is also examined. It is demonstrated that ignoring the die-to-die and within-die PVT variations results in a significant error in estimating the yield. Fig. 2.33 illustrates the probability density function of leakage to active power ratio for three

cases. In case 1, only die-to-die variations are considered. Within-die variations are added in case 2, and die-to-die temperature variations are taken into account in case 3. It is evident that ignoring within-die and temperature variations impose significant error in the spread of the leakage power.

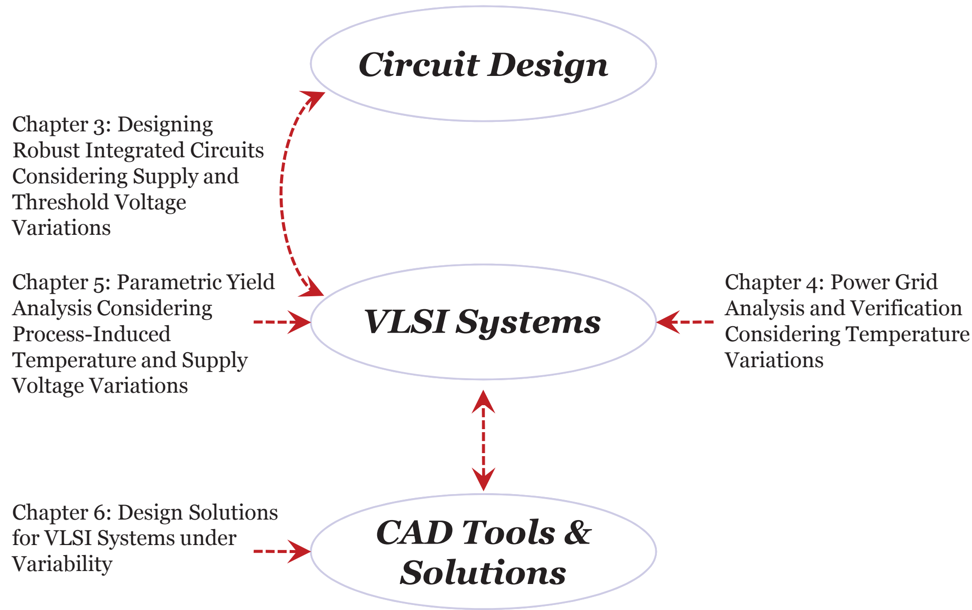


Figure 2.34: Organization of the research conducted in this thesis.

2.7 Proposed Analysis and Methodologies

This thesis proposes analyses and solutions for addressing the impact of process and environmental variations on VLSI systems. The results of the analyses can be used at various design stages. This allows the designer to utilize the flexibility of the early phases or the detailed knowledge of the later design stages for minimizing the impact of the variations. An overview of the research conducted in this thesis is shown in Fig. 2.34.

First, to get some insight into the circuit design under process and environmental variations, parametric yield is investigated at the circuit level. Here, a robust design is identified so that the power and performance of the circuits are most immune to the variations in the V_{dd} and V_{th}. The existing work focus on the trade-off between the power and performance for a nominal design. Their proposed extensions to include the uncertainty in their studies, do not provide a clear guideline for designing under variations. In many cases, the design is subject to constraints, and, therefore, an unconstrained optimum point, suggested by these analyses, does not satisfy the constraints, and, thus, is not acceptable. This part of the thesis proposes a statistical methodology for optimizing the V_{dd} and V_{th} under variations to maximize the parametric yield. The study comprises the following contributions at the circuit level.

- A design center in the V_{dd}-V_{th} plane is identified by a statistical design methodology, where the center has the highest probability of meeting the constraints in the presence

of variations.

- A two-level optimization method is proposed that maximizes the yield, and, as a secondary objective, achieves the best possible (near-optimal) trade-off between the power and performance, specific to a given application.
- A guideline for a variability-aware V_{dd} and V_{th} scaling optimization is developed. This helps a designer to take into account the effect of the switching activity, transistor sizing, and design constraints on the voltage scaling schemes, while maximizing the targeted yield.

It is also important to see how the design center moves in the V_{th}-V_{dd} plane. Here, the trend for the shift in the design center is given to maximize the yield for different technologies. The designer can predict the impact of the variations on the yield and foresee the trade-off between power and performance, and develop an effective scheme for the voltage scaling. The details of the proposed work are presented in Chapter 3.

Motivated by the results of the investigation, found at the circuit level, the impact of variations is examined at the system level. Process variations significantly impact leakage power, a pivotal parameter in designing a power grid. Because of the strong relationship between the temperature and leakage power, the variations also impose statistical behavior on the operating temperature. In addition, the metal resistivity of a power grid increases with temperature. Therefore, ignoring the interdependency between the leakage and temperature can introduce large errors in the power grid design. In this part of the thesis, initially in Chapter 4, a power grid analysis is proposed that considers a statistical thermal profile across the grid. Then, this analysis is employed, in Chapter 5, to analyze the timing and power yield. Here, by considering the statistical profile of the temperature and supply voltage, the process variations are mapped to the delay variations across a die. Moreover, the statistical behavior of the temperature and voltage drop, imposed by the process variations, affects the power consumption. The existing research forgo to consider the interdependency between the voltage and temperature variations. To avoid inaccurate results, the power yield analysis is proposed to address the interdependency and check the robustness of the circuits early in the design process.

Finally, in Chapter 6, CAD and design solutions are proposed to alleviate the variations and their impact on the design. First, the correlation between the total power consumption and the temperature variations across a chip is examined. As a result, floorplanning guidelines are proposed that utilize the correlation to efficiently optimize the chip's total power and takes into account the thermal uniformity. Research in the literature, discussing high temperature effects on IC design, focus on either thermal integrity or on the leakage power reduction. In this thesis, not only is a given maximum temperature constraint guaranteed, but also the sensitivity of the total power consumption to variations in the

temperature of the different blocks is minimized. This research presents thermal-aware floorplanning to efficiently minimize the total power consumption of a chip in the presence of temperature variations. The thesis provides answers to the following questions.

- Is there any correlation between the total power consumption of a chip and the temperature variations of the floorplan?
- How should a designer optimize a floorplan to minimize the total power efficiently?
- How can the thermal variations on the chip, subject to a given small deviation from the minimum total power, be reduced?

Finally, a solution is proposed to maximize the timing yield. Here, an optimization methodology is presented for assigning the power supply pads across the chip. A mixed-integer nonlinear programming (MINLP) optimization problem, subject to voltage drop and current constraints, is efficiently solved to find the optimum number and location of the pads. The details of the proposed solutions are provided in Chapter 6.

Chapter 3

Designing Robust Integrated Circuits Considering Supply and Threshold Voltage Variations

3.1 Introduction

A preliminary research is conducted to get some insights into what parameters are involved in designing robust integrated circuits. The circuit level investigation and the respective simulation results are discussed in this chapter. Understanding the results is critical to analyze and optimize the parametric yield at the system level, in the subsequent chapters.

Process and environmental variations cause design variables to deviate from their nominal values. These variations are on the rise, and thus, the robustness of the design of integrated circuits emerges as one critical challenge [3]. The supply voltage (V_{dd}) and threshold voltage (V_{th}) are two significant design variables which directly impact the power consumption and performance of the circuits. The scaling of these voltages has become a popular option for designers in order to reduce switching and leakage power, and manage the operating temperature [69],[70]. Authors in [69] and [71] employ a dynamic voltage scaling scheme using a feedback control to scale the supply voltage, and therefore, control the dynamic power. In addition, scaling the threshold voltage can significantly change the subthreshold leakage power. Adaptive Body Bias (ABB) can effectively control the body bias voltage and thus vary V_{th} dynamically over a continuous range at run time [72][73]. Threshold voltage can also be scaled in a discrete fashion [74],[70]. In the scaling of V_{dd} and V_{th}, the voltages must be selected so that the design constraints are met and the circuits operate with the least sensitivity to the variations of these two parameters, V_{dd} and V_{th}.

3.2 Related Work

Several analyses have been carried out in the literature to investigate the impact of the variations in V_{dd} and V_{th} on the parametric yield. In [47], the impact of leakage on the parametric yield has been analytically shown, and the sensitivity of the yield to the V_{dd}, power and performance has been examined. However, this conservative analysis is based on the minimum and maximum channel length and the V_{th}. In addition, the effect of temperature on the leakage power is not taken into account. Gonzalez et al. have investigated the effect of scaling V_{dd} and V_{th} on energy and delay [48]. It has been demonstrated that optimizing both voltages can save energy, and increase performance. The effect of V_{dd} and V_{th} variations on their design metric, Energy Delay Product (EDP), has also been addressed, but the EDP under uncertainty has been obtained by the multiplication of the energy and the delay at the four corners of the supply voltage and temperature. In [49], the electrothermal effects have been incorporated in the optimization to account for the high temperature impact on power, performance, and reliability. Variations in the V_{th} have a larger impact on the EDP, if electrothermal coupling is taken to account. In particular, two papers [46][50] have proposed more general metrics (P^mD^n and PT^μ) to give priority either to the power or the delay for specific applications. It has been demonstrated how a design metric such as the EDP changes with the variations in V_{dd} and V_{th}. However, these work focus on the trade-off between the power and performance for a nominal design. Their extension to include the uncertainty in their studies do not provide a clear guideline for designing under variations. In addition, in many cases, the design is subject to constraints, and therefore, an unconstrained optimum point, suggested by these analyses, does not satisfy the constraints, and thus, is not acceptable.

This chapter describes the work on design-specific yield optimization considering variations in supply and threshold voltage. The contributions of the work are as follows:

- A design center in the V_{dd}-V_{th} plane is identified, by a statistical design methodology, where the center has the highest probability of meeting the constraints in the presence of variations.
- A two-level optimization method is proposed that maximizes the yield and, as a secondary objective, it achieves the best possible (near-optimal) trade-off between the power and performance specific to a given application.
- A guideline for variability-aware V_{dd} and V_{th} scaling optimization is developed. This helps a designer to take into account the effect of switching activity, transistor sizing, and design constraints on the voltage scaling schemes while maximizing the targeted yield.

To achieve these contributions, a feasible region is initially constructed by using the constraint contours in the V_{dd}-V_{th} plane. A tolerance box that represents the variations in the voltages ($\mu \pm 3\sigma$) is placed in the design space. The final location of the box and its center indicate the optimum value of the design variables; i.e., the V_{dd} and V_{th}. This design center maximizes the parametric yield, and optimizes the application-specific design metric such as the EDP, as the secondary objective, with respect to the maximum yield. The minimum performance and maximum temperature are the design constraints in this research work. The method proves to be reliable, efficient, and converges in polynomial time.

3.3 Design Metrics for the Yield Estimation

To measure the goodness of the trade-off between power and performance, designers select a metric. The type of the metric depends on the importance of power versus performance in the application. Energy delay product is a popular metric in high performance applications. Here, initially, EDP is used as the chosen metric, and, then, the impact of the selecting other metrics on the design is discussed.

3.3.1 Energy Delay Model

The impact of variations in V_{dd} and V_{th} on power and performance can be demonstrated using EDP. There are several gates on the critical path, each with a different delay. Nevertheless, the changes in the V_{dd} and V_{th} impact all the gates in a similar way. Therefore, the delay of each gate is almost proportional to the delay of an inverter [48]. By using the alpha power model, the delay of an inverter is expressed as

$$T_g = \frac{C.V_{dd}}{I_D} \tag{3.1}$$

and the maximum clock frequency of the chip is given by [75]:

$$f = \frac{1}{T_g.L_d} \tag{3.2}$$

where C is the load capacitance, I_D is the drain current, and L_d is the logic depth. The drain current is expressed as [76]

$$I_D(T) = K\nu_{sat}(T)(V_{dd} - V_{th}(T))^a \tag{3.3}$$

where K, ν_{sat}, and a are a technology constant, saturation velocity, and saturation velocity index respectively. Note that stacked gates can show more sensitivity to variations in

threshold voltage. The sensitivity is a function of circuit design-style, input pattern, and sizing [77]. However, forced stacks are usually used in non-critical path to avoid the delay penalty. Hence, we also use the delay model provided by Gonzalez et al. to illustrate the proposed methodology.

In addition, the power consumption in VLSI circuits consists of two major components: dynamic and leakage power. The short circuit power is less significant, and thus, is neglected for simplicity [78][79]. The total dynamic power per operation of the chip, dissipated due to the switching activity, is given by

$$P_{dynamic} = \frac{1}{2}\alpha C_{eff} V_{dd}^2 f \tag{3.4}$$

where $0 \leq \alpha \leq 1$ is the node dynamic transition activity factor, f is the clock frequency and C_{eff} is the total load capacitance of the output node. In addition, the static current, including the subthreshold, gate leakage, and drain-induced barrier lowering current, contribute to the dissipation of static energy. The gate leakage in the 90nm technology where SiO_2 is used as the gate dielectric can be significant. However, starting from 45nm due to the use of high-k dielectrics and specifically for the high performance applications running at high temperature, subthreshold leakage power is dominant [80][81]. Therefore, the focus of this work is on the subthreshold leakage component:

$$P_{static} = I_s W_{eff} V_{dd} e^{\frac{-V_{th}}{\gamma V_0}} (1 - e^{\frac{-V_{th}}{\gamma V_0}}) \tag{3.5}$$

I_s is the zero-threshold leakage current, γ is the subthreshold slope factor, V_0 is the subthreshold slope, and W_{eff} is the effective width. In addition, a design metric is required to determine which value of the V_{dd} and which of the V_{th} give the desired performance for a given energy budget. The EDP is an example of such metrics and is expressed as follows [49]:

$$EDP = \frac{K^2 I_s L_d V_{dd}^3}{\nu_{sat}(V_{dd} - V_{th})^\alpha} \left[\frac{\alpha C_{eff}}{I_s K L_d} + \frac{e^{\frac{-V_{th}}{\gamma V_0}} (1 - e^{\frac{-V_{th}}{\gamma V_0}})}{\nu_{sat}(V_{dd} - V_{th})^\alpha} \right] \tag{3.6}$$

where K is proportionality constant specific to a given technology. As similarly illustrated in [48] for $0.25\mu m$ technology, Fig. 3.1 depicts the reversed normalized EDP contours and iso-performance curves at an average of 40 °C for $90nm$ CMOS technology. The optimal value shown here is the nominal EDP value of an unconstrained minimization of the EDP equation. Also, the contours are obtained by solving (3.6) by using numerical methods for different values given to EDP. Note that the contours are normalized by dividing the minimum EDP by the calculated EDP for any pair of the (V_{th} and V_{dd}). For example, the EDP value of the contour, identified as 0.5, is twice as large as that of the minimum EDP. Also, points on the Iso-performance curve identified as 1.3 outperform those on the curve identified as 1 by 1.3 times. It is also assumed that the chip does not operate in the

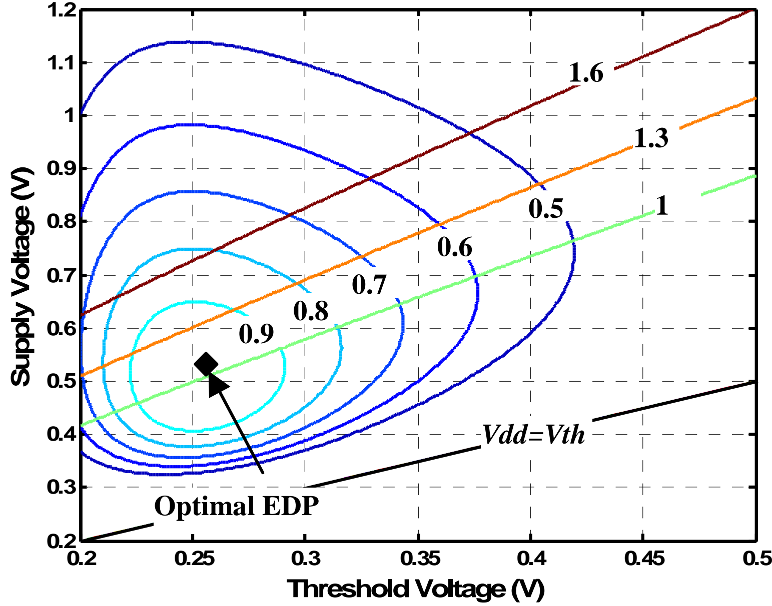


Figure 3.1: Normalized EDP contours and iso-performance curves for velocity saturation index $\alpha = 1.3$ at 40 ° C.

subthreshold region. Therefore, the valid pairs are the ones that lie above the $V_{dd} = V_{th}$ line.

3.3.2 Incorporating Temperature

It is essential to take into account the effect of high temperatures on the power and performance. The threshold voltage is expressed as a function of temperature as follows:

$$V_{th} = V_{th0} - k(T_j - T_{amb}) \tag{3.7}$$

where V_{th0} is the V_{th} at the ambient temperature, k is the temperature coefficient of the V_{th}, and T_j is the junction temperature. ν_{sat} is another parameter that changes with the temperature [82] and is expressed as

$$\nu_{sat} = \nu_{sat0} - \eta(T_j - T_{amb}) \tag{3.8}$$

where ν_{sat0} and η are the saturation velocity at the ambient temperature and the saturation velocity temperature coefficient.

There is a positive feedback between the subthreshold leakage and the temperature in which an increase in the temperature results in an increase in the subthreshold leakage. The higher power dissipation generates more heat which, in turn, results in the higher

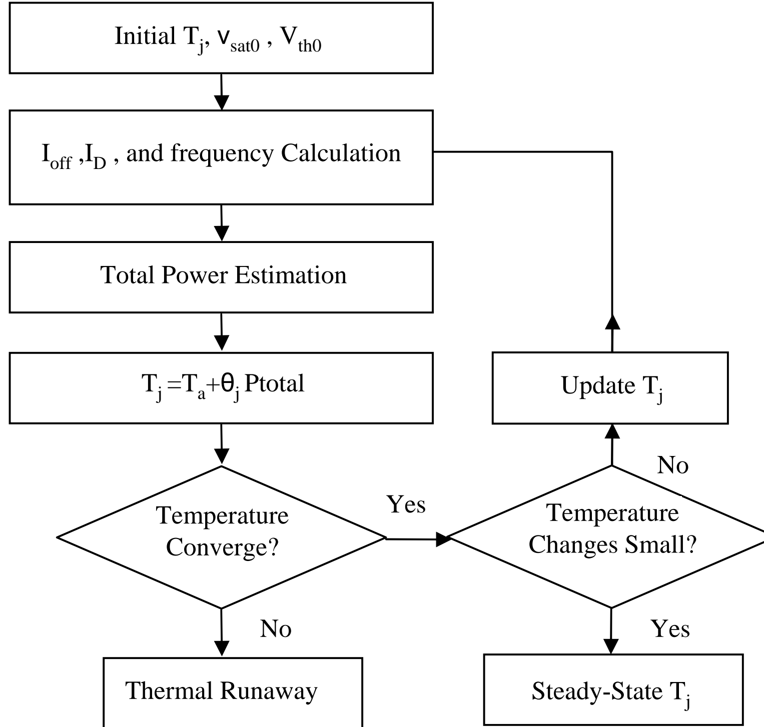


Figure 3.2: The steady-state temperature and power estimation methodology and identifying thermal runaway region.

temperature. This closed loop either converges, or a thermal runaway occurs, possibly leading to a thermal breakdown.

Power density of different parts of a chip varies and therefore, temperature may change from one point to the next. Hence, accurate 3-D thermal modeling is needed for precise temperature estimation. This requires detailed information about the cooling system, the application that is executed, and the chip floorplan which is not included at this level. Therefore, as similarly done in [49], the 1-D thermal model is employed to estimate the temperature for a given pair of V_{dd} and V_{th}. The following shows the average junction temperature as a function of the total power consumption:

$$T_j = T_{amb} + \theta_{ja} P \tag{3.9}$$

where T_{amb}, θ_{ja}, and P are the ambient temperature, thermal impedance of the junction to the ambient, and the total power consumption respectively. Fig. 3.2 shows the methodology for estimating the steady-state temperature. To obtain the temperature, the system is initiated at room temperature, and the total power is estimated. Then, (3.9) is employed to calculate a new junction temperature for the power consumption. At each iteration, the convergence is checked to obtain the final temperature, or the thermal runaway is identified; i.e., the chip temperature does not converge to a stable temperature. As long as the

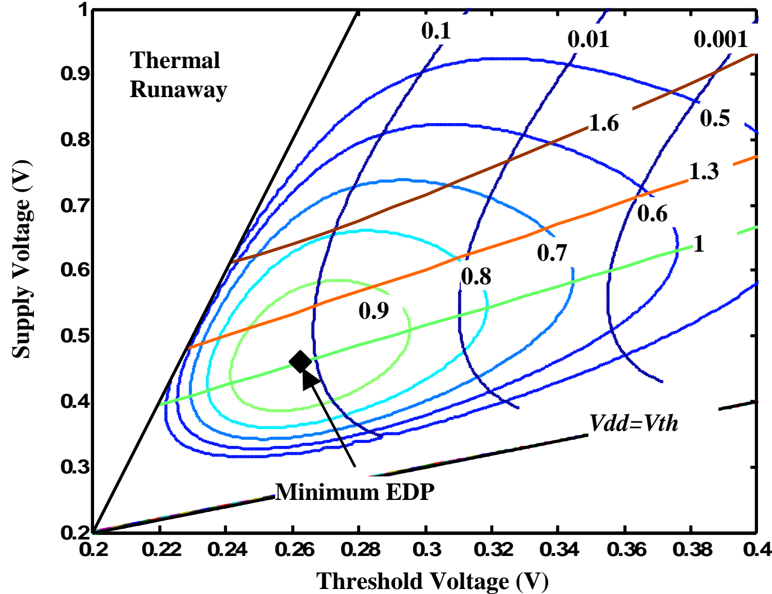


Figure 3.3: Normalized EDP contours and iso-performance curves where the effect of temperature on power and performance is taken into account. The contours of subthreshold leakage to total power ratio are also shown and thermal runaway region is specified.

cooling system can remove enough heat, or the total power does not steadily increase, the chip is thermally stable.

The steady-state EDP and iso-performance contours for a $90nm$ design are depicted in Fig. 3.3. These contours are similar to that of reported in [49] for $130nm$ technology. All the contours, drawn in Fig. 3.3, are obtained for the steady-state temperature, and after being updated in several iterations. Once the steady-state temperature is found, the power and frequency are updated accordingly. In this process, the methodology, shown in Fig. 3.2, is applied to several pairs of (V_{dd} and V_{th}). In addition, area in which the loop does not converge, and therefore, a thermal runaway occurs, is specified. The thermal runaway usually occurs, where the V_{th} is small and the V_{dd} is large, resulting in large values for the leakage and dynamic power. For example, if $V_{th} = 0.294\ V$ and $V_{dd} = 0.5\ V$, the normalized EDP is 0.9 (the actual $EDP = (1/0.9) \times\ EDP_{mininum} \approx 1.11\ EDP_{mininum}$). In addition, for this example, $f/f_{EDPmin} = 1$ where f_{EDPmin} is the frequency at the minimum EDP point. However, for $V_{th} = 0.22\ V$ and $V_{dd} = 0.6\ V$ the temperature does not converge; subthreshold leakage monotonically increases and the thermal runaway is identified. In addition, the contours of subthreshold leakage to total power ratios are also indicated. Towards the low threshold voltages, this ratio exponentially increases. However, the effectiveness of the leakage control techniques depends on the ratio of the subthreshold leakage to the total power [83]. Therefore, a designer can limit this ratio to achieve an improved power saving mechanism. As a result, this can impose a constraint on the selected

pair of V_{dd} and V_{th}.

3.4 Constructing the Feasible Region and Modeling the Design Variable Distribution

The design constraints are application dependent. In high-performance applications, attaining the targeted performance is the main requirement and most design decisions are made to deliver this performance. However, in a mobile application, necessary steps are taken to save as much power as possible. In addition, the design constraints can also change over time [84][85]. For example, in a real-time applications the minimum performance constraint can change to manage the deadlines. Hence, depending on the application, V_{dd} and V_{th} can be selected to fulfill the design constraints. To illustrate the methodology proposed in this work, three design requirements are considered to construct the design space.

3.4.1 Minimum Performance Constraint

The first constraint in the design is the minimum performance or maximum delay. This ensures that the design delivers the guaranteed minimum performance. From the circuit design perspective, the circuit clock frequency must exceed a given minimum value (f_{min}). Therefore,

$$f \geq f_{min} \tag{3.10}$$

To meet this constraint, the design should shift towards the higher supply voltages or lower threshold voltages in order to increase the overdrive voltage (V_{dd}-V_{th}). However, this shift is bound by another constraint, the maximum temperature.

3.4.2 Maximum Temperature and Thermal Reliability Constraint

As discussed in Section 3.3, for the close loop between the subthreshold leakage and temperature, it is essential that the circuit operates at a stable steady-state temperature; i.e., the design space must not overlap the thermal runaway region. Any point in this region does not have a known steady-state temperature. Thus having just the maximum temperature constraint is not sufficient to exclude this region from the design space. The region is identified using the iterative method shown in Fig. 3.2.

In addition, a limit on the temperature that the circuit can reach is also essential. Constraining the maximum operating temperature is pivotal for reducing the chip failure due to electromigration, Time Dependant Dielectric Breakdown (TDDB), thermal cycling,

and other temperature dependent failure mechanisms on a chip [86]. As a result, the following constraint must also be satisfied:

$$T_j \leq T_{max} \tag{3.11}$$

In practice, the maximum temperature, T_{max}, is obtained so that the design meets the ten years of Mean Time To Failure (MTTF) requirement. Due to the increase in the activity of a circuit, the junction temperature can exceed this maximum value. In such cases a Dynamic Thermal Management (DTM) triggers a policy such as dynamic voltage scaling to reduce the power and temperature [61]. Note that, in this work with the focus on high-performance circuits, maximum temperature is selected as a design requirement. However, in a power-limited application where a limited power budget must be met, the maximum power can replace this constraint.

Finally, as mentioned in Section 3.3, it is assumed that the chip does not operate in the subthreshold region such that

$$V_{dd} \geq V_{th} \tag{3.12}$$

This constraint is not usually active; i.e., it does not intersect with the feasible region. This is due to the high overdrive voltage needed to meet the minimum performance, and thus, the V_{dd} must be relatively higher than the V_{th}.

3.4.3 Constructing the Feasible Region

The feasible region, F_s, is formed by the previous constraints: minimal performance, maximum temperature, and thermal reliability. Any point in F_s satisfies the constraints and is expressed as

$$F_s = \{x \in \Re^2 \mid f_i(x) \geq 0, \ i = 1, 2, 3\} \tag{3.13}$$

where x is the design parameter vector ($[V_{th} \ ; V_{dd}]$), and $f_i(x)$ is i^{th} facet plane, defined by the constraints given in (3.10), (3.11), and the thermal runaway region respectively. These constraints can assume different values from one application to the next. Here, for the purpose of illustration, it is assumed that the minimum frequency is 30% greater than the frequency at the point with the minimum EDP in the V_{th}-V_{dd} plane. In high-performance ICs, temperatures as high as 120 °C exist [87]. Here, 100 °C is chosen as an example of the maximum temperature and the feasible region in Fig. 3.4 is constructed by considering these values of the constraints. To exclude the thermal runaway region from the design space, methodology illustrated in Fig. 3.2 was used where the thermal resistance of the junction to ambient was assumed to be 0.9 K/W.

So far, the design space has been constructed. For the purpose of the yield optimization, distribution of the design variables is modeled. Simulation results indicate that the V_{th} and

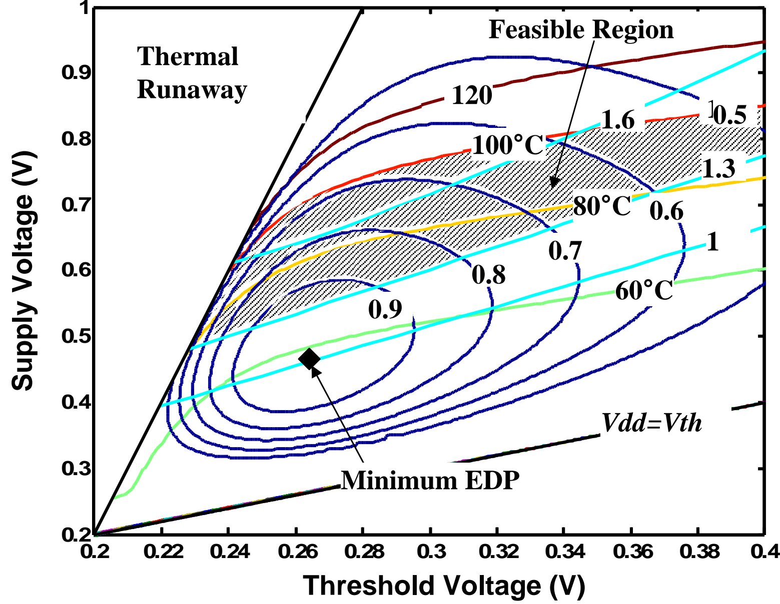


Figure 3.4: Normalized EDP contours, iso-performance curves, and contours of temperature are shown. The feasible region used as an example in this work is also shaded.

V_{dd} variations can be modeled as normal distributions [23][88]. Nonetheless, the normal distribution does not have a closed form Cumulative Distribution Function (CDF) which is necessary for the yield estimation. Therefore, Kumaraswamy's distribution [89], Double-Bounded-Probability Density Function (DB-PDF) with the following form is used instead:

$$f(z) = abz^{a-1}(1 - z^a)^{b-1}$$
$$z = \frac{x - x^{min}}{x^{max} - x^{min}},\ and\ x^{min} \leq x \leq x^{max} \tag{3.14}$$

where x^{min} and x^{max} are the lower and upper bounds of the design variables. By assigning different values to a and b, the PDF can take a variety of shapes, including a truncated Gaussian distribution. The closed form CDF of this distribution is given by

$$F(z) = 1 - (1 - z^a)^b \tag{3.15}$$

This is used for the yield estimation in the following section.

3.5 Yield Optimization

To maximize the yield, as the primary objective, a two-level optimization is employed. First, the design metric, (EDP which is PT^2, PT^3, PT^4), is deterministically optimized

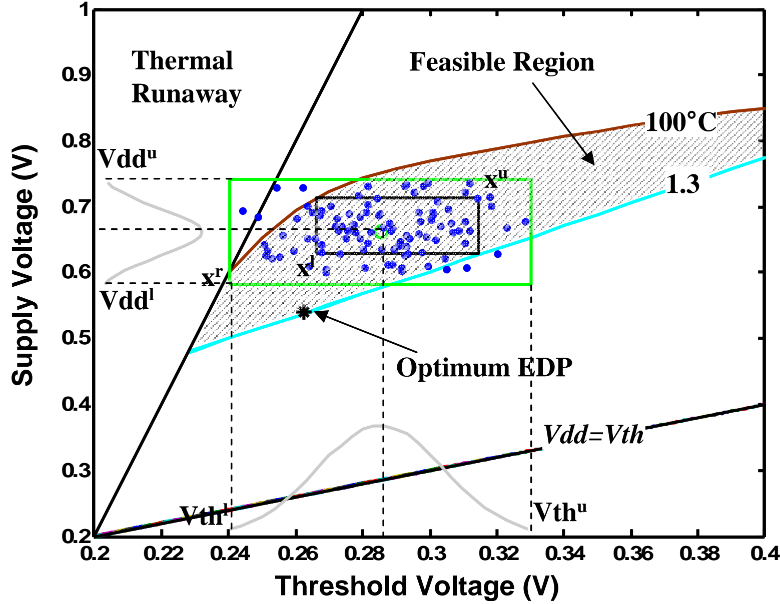


Figure 3.5: The final location of the tolerance box over which the yield is maximized.

to find the optimum point in the feasible region. At the second level, a tolerance box, initially its center located on the deterministic optimum point, is moved over the design space to find the best box's location in order to maximize the yield. Between the two locations of the tolerance box and for the same yield, the one with the minimum deviation from the deterministic optimum point is desirable and is selected. The tolerance box is rectangular and related to the probability distribution of the design variables, stated in (3.16). These variables are assumed to be independent. Fig. 3.5 depicts the final location of the optimum tolerance box for $90nm$ technology, and the center at which the immunity of the design to the variations is maximal ($3\sigma_{Vth} = 0.045$ and $3\sigma_{Vdd} = 0.08$ [3]). This center is as close as possible to the optimum EDP in the feasible region. The minimum performance is assumed to be $f_{min} = 1.3f_{EDPmin}$, and the maximum allowed temperature is 100 °C for this example. The outer box represents the tolerance box and the inner box is the tolerance box with the maximum yield in the feasible region. At the second optimization level, the optimal tolerance box is bound by the tolerance range and indicates a design with the largest yield for that range. If the variations in the design variables can be controlled so that the size of the outer box is reduced to that of the inner box, the yield becomes 100%. The final location and the center of the box, indicate a design that has the highest immunity to the variations in V_{dd} and V_{th}. The yield is expressed as a function of the lower and upper bounds of the variables [90], and a reference point (referring to the

location of the optimum box), and is obtained by

$$
\begin{aligned}
Yield(x^r, x^l, x^u) &= \prod_{i=1}^{2} Pr\{x_i^l \leq x_i \leq x_i^u\} \\
&= \left[F\left(\frac{V_{th}^u - V_{th}^r}{t_{Vth}}\right) - F\left(\frac{V_{th}^l - V_{th}^r}{t_{Vth}}\right)\right] \\
&\times \left[F\left(\frac{V_{dd}^u - V_{dd}^r}{t_{Vdd}}\right) - F\left(\frac{V_{dd}^l - V_{dd}^r}{t_{Vdd}}\right)\right] \qquad (3.16)
\end{aligned}
$$

where x^r or $[V_{th}^r, V_{dd}^r]$ is the bottom left corner of the outer box, x^l and x^u are the bottom left and upper right corners of the optimum box in respect to the design variables. Also t_{Vdd} and t_{Vth} represent the range of the distribution of V_{dd} and V_{th} and F denotes the CDF of the variables. By using the model in (3.16), the yield is maximized as follows:

$$
\begin{aligned}
& max \;\; Yield(x^r, x^l, x^u) \\
& subject\ to: \\
& \quad R(x^l, x^u) \subseteq F_s \\
& \quad x^r \geq x^{min} \\
& \quad x^l \geq x^r \\
& \quad x^u - x^l \leq t \\
& \quad x^r + t \leq x^{max} \qquad (3.17)
\end{aligned}
$$

where R is the inner optimum tolerance box contained in the feasible region. Therefore,

$$
R(x^l, x^u) = \{x \in \Re^2 \mid x^l \leq x \leq x^u\} \qquad (3.18)
$$

The previous optimization is implemented iteratively. The final location of the tolerance box is a function of the yield and the shape of the feasible region. To better clarify the two-level optimization, assume that the feasible region is large enough to surround the entire tolerance box. As long as the tolerance box is inside the feasible region, no matter where the design center is located, the yield is 100%. Although the yield is maximal in all of those locations, the center must be as close as possible to the deterministic optimum point. This is to also provide the best possible trade-off between power and performance. The two-level optimization is to meet these two objectives. By minimizing the objective function λ in (3.19), the tolerance box is moved over the feasible region to maximize the yield and minimize the deviation from the deterministic optimum point.

$$
\begin{aligned}
min \;\; \lambda &= [(x - x^c)(x - x^c)^T]^{\frac{1}{2}} \\
subject\ to: \;\; & g_i(x) = 0 \qquad (3.19)
\end{aligned}
$$

Table 3.1: Yield optimization for two different cases of variations in supply and threshold voltages and for various design metrics (estimated and monte carlo results for 90nm technology).

Parameters		Design Metrics					
		PT^2 (EDP)		PT^3		PT^4	
6σ (%)	V_{th}	30	15	30	15	30	15
	V_{dd}	20	10	20	10	20	10
V_{th}	Nominal (V)	0.286	0.272	0.292	0.283	0.293	0.284
	STD	0.018	0.090	0.018	0.090	0.018	0.090
V_{dd}	Nominal (V)	0.668	0.620	0.683	0.653	0.685	0.695
	STD	0.036	0.018	0.036	0.018	0.036	0.018
Normalized PT^{μ}	Nominal	0.782	0.830	0.920	0.940	0.890	0.920
	STD	0.041	0.026	0.048	0.027	0.052	0.031
Yield (%)	Estimated	94	100	94	100	94	100
	Monte-Carlo	96	100	96	100	96	100

where superscript T stands for the transpose of a vector, x is a point on the surface of constraint $g_i(x)$ which has the shortest distance from the center of the tolerance box (x^c).

3.6 Simulation Results and Discussion for $90nm$

3.6.1 Solving the Optimization Problem

The optimization problem is solved by using Sequential Quadratic Programming (SQP) in MATLAB. The MOSFET model for $90nm$ technology is adapted from the BSIM4 model [91]. A closed loop between the leakage power and temperature is used to update the temperature iteratively, to find the steady-state temperature for any pair of V_{th} and V_{dd}, and to identify the thermal runaway region. As stated in the previous section, in addition to identifying the thermal runaway region, the other constraints for this specific design are assumed to be $f_{min} = 1.3 f_{EDPmin}$ and $T_{max} = 100\ °C$. The subscript EDP_{min} indicates the parameter's value at the solution point of the unconstrained EDP minimization in the V_{th}-V_{dd} plane. This point is obtained by minimizing EDP given in (3.6). Note that the optimum values, here, calculated for V_{th} and V_{dd} , are not in the feasible region and are used for normalization only.

Table 3.1 shows the yield, nominal, and standard deviation of the design variables and for the different design metrics. Note that, due to the variations in the design variables, power, performance, and EDP have statistical measures. Mean and standard deviation of

EDP, at the maximum yield point, are given for different standard deviations in V_{th}-V_{dd}. To increase the yield, the process and environmental variations must be controlled so that the size of the tolerance box (the outer box) is as close as possible to that of the 100% yield box (the inner box). This is managed by increasing the precision of the equipment in the fabrication process, at a higher cost, or by controlling the noise sources such as the IR drop. To attain a good trade-off between the increase in the yield and the cost of the design and manufacturing, the financial data must also be evaluated.

Although a designer can gain some limited control over the manufacturing results through a litho-friendly layout design [92][93], usually, there is little control over reducing the variations in the design variables, and for the designer, these variations are considered fixed. Therefore, to increase the yield, the constraints must be relaxed. For example, the minimum allowed frequency can be lowered to achieve a 100% yield. This occurs when the feasible region is expanded with a new f_{min} constraint so that the yield loss is zero. The relaxed minimum frequency is calculated by adding a few extra iterations. Fig. 3.6 reflects the Monte Carlo simulation for the relaxed and tight constraints. It is evident that several designs fail to satisfy either constraint. If the frequency is high, the probability of violating the maximum temperature increases due to the high dependency of the subthreshold leakage on the operating temperature.

3.6.2 Optimizing the Design Metrics and Simulation Results

Any pair of V_{dd} and V_{th} in the feasible region satisfies the design constraints. However, for the optimum pair not only is the yield maximal, but also, it pinpoints the best possible trade-off between the power consumption and performance. For this, a design metric (PT^{μ}) is initially selected, based on the priority that needs to be given to the power as opposed to the performance in a given application. The respective metric is optimized, subject to the constraints, and the solution of the constrained optimization is adopted as the initial solution for the yield optimization. While maximizing the yield, λ in (3.19) is minimized. This ensures that the center of the tolerance box is as close as possible to the optimum value of the design metric in the feasible region. Consequently, the probability that more design samples can meet the constraints, while yielding a better efficiency, is higher. Fig. 3.7 reflects how the different design metrics (PT^2 (EDP), PT^3, and PT^4) cause the design center to move within the feasible region, leading to different designs in the presence of uncertainty in the design variables. To better illustrate the difference, the minimum frequency constraint is relaxed to $f_{min} = f_{EDPmin}$, and the variations in the design variables are assumed to be ($3\sigma_{Vth} = 0.023$ and $3\sigma_{Vdd} = 0.04$). This expands the region, thus, the tolerance box has more flexibility to move within the region. Therefore, the center of the tolerance box has a chance to get as close as possible to the deterministic optimum point, while maintaining its maximum yield.

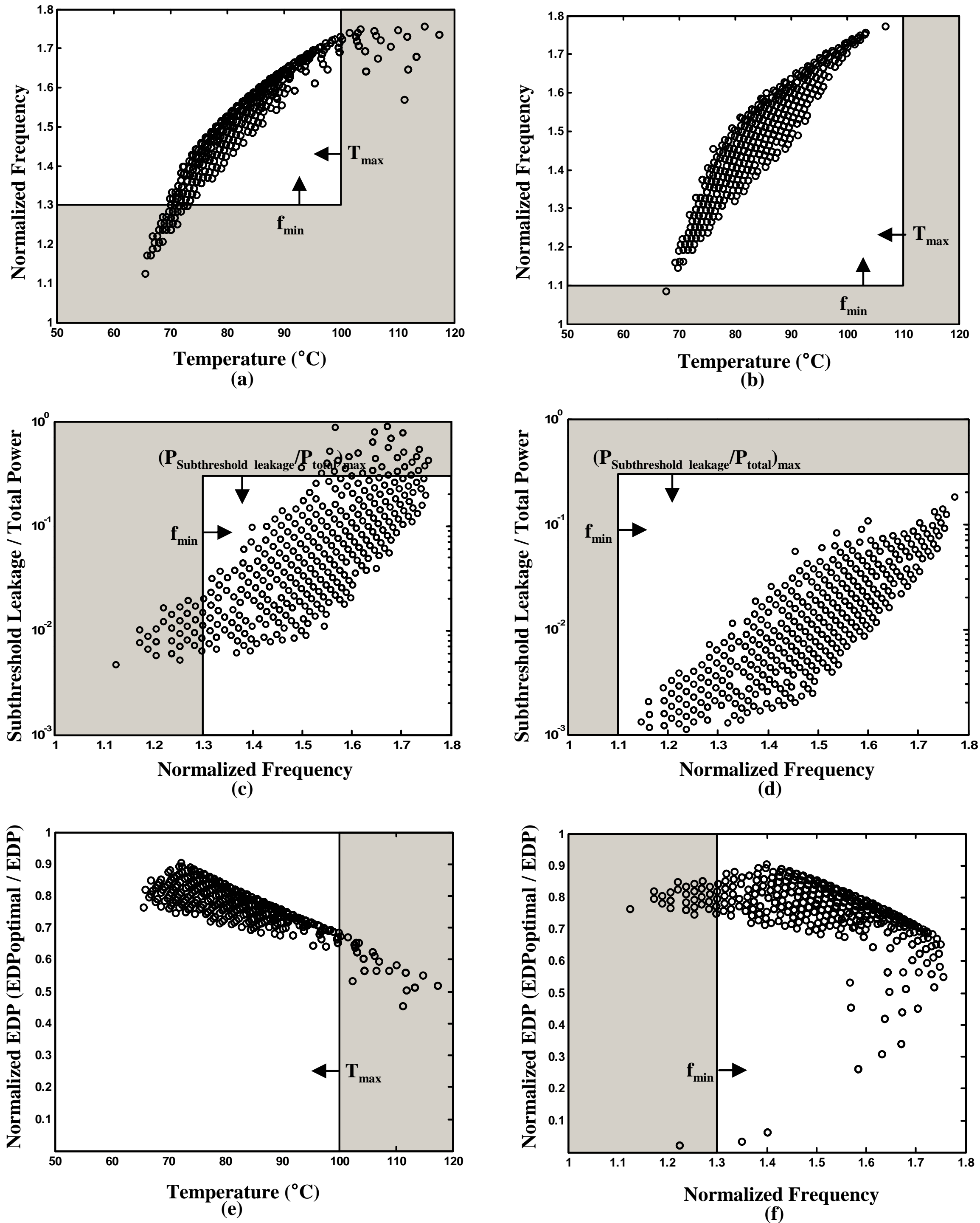


Figure 3.6: Monte Carlo simulations for various voltage pairs: (a),(c) for the tight constraints (yield = 87%); (b),(d) for the relaxed constraints (yield = 99%), (e),(f) normalized EDP (tight constraints yield = 87%).

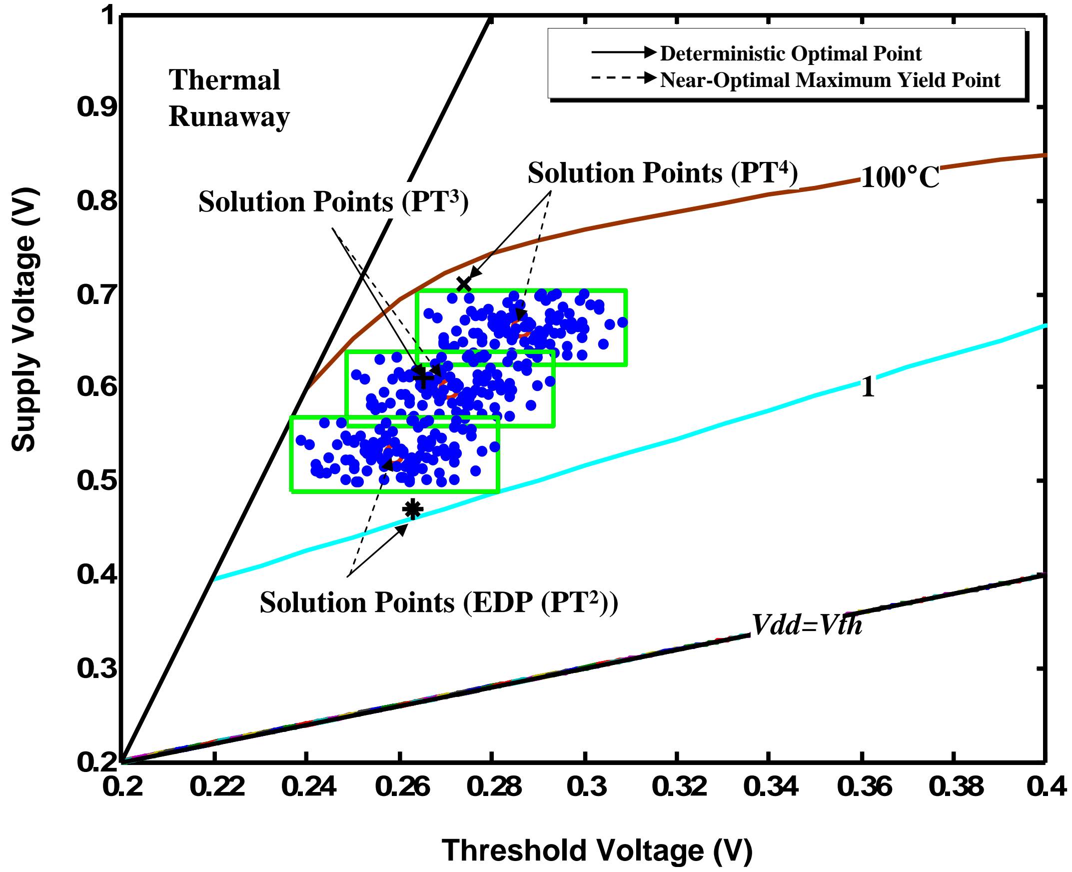


Figure 3.7: The effect of selecting three different design metrics on the final location of the tolerance box and their respective design centers.

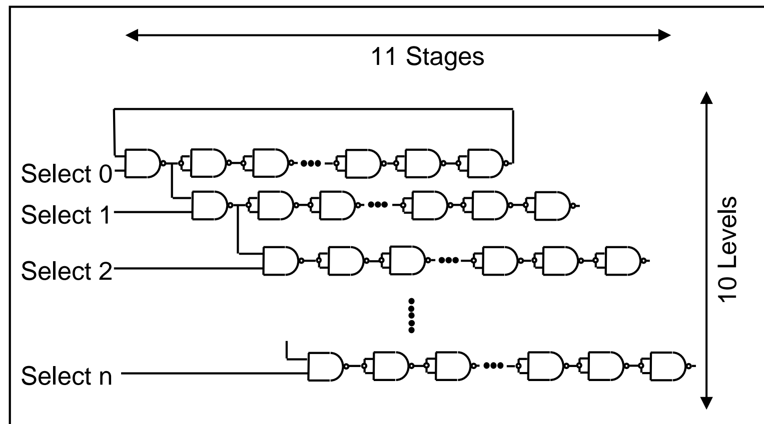


Figure 3.8: The simulated circuit consisting of a ring oscillator and a multi-level NAND chain for selecting different activity factors.

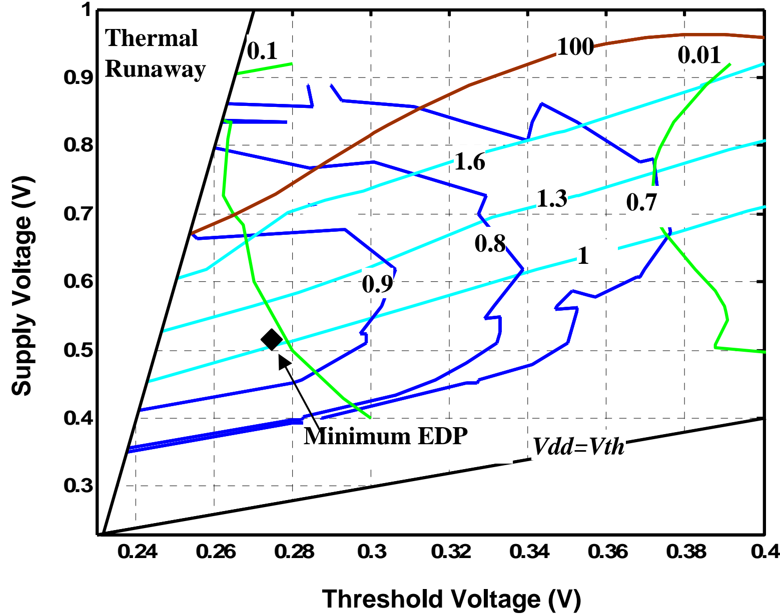


Figure 3.9: Normalized EDP contours and iso-performance curves of $90nm$ CMOS technology for the simulated circuit (Fig. 3.8) for $\alpha = 0.1$. The contours of temperature and subthreshold leakage to total power ratio are also shown.

To verify the methodology, Cadence SPECTRE simulations are performed for the circuit adapted from [46]. This circuit, in (Fig. 3.8), consists of 11 stages and 10 levels. The first level is a NAND ring oscillator where as the other levels are NAND chains. The activity factor of the circuit is controlled by assigning '0' or '1' to the Select inputs. By assigning '1' to Select0 and '0' to the subsequent select inputs, the circuit becomes a ring oscillator (activity factor of 0.1). The circuit is simulated by using $90nm$ CMOS technology. To choose the V_{th} for the different simulations, the device model is modified and new models are generated. The temperature and V_{dd} are swept in order to extract the frequency and power consumption for vectors of V_{th} and V_{dd}.

Fig. 3.9 denotes the normalized EDP contours, iso-frequency curves and contours of the subthreshold leakage to the total power ratio for the simulated circuit with an activity factor of 0.1. The contours are very similar to those, obtained by using the power and delay model. The distortion in the contours is due to the variations of the V_{th} around the value given to V_{th} parameter in the model and its effect on the frequency and power.

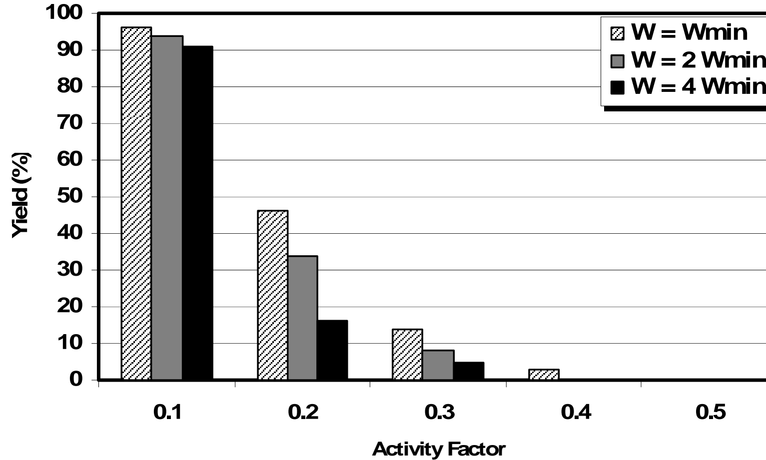


Figure 3.10: Sensitivity of the parametric yield to the activity factor and transistor sizing.

3.6.3 Design Considerations

Fig. 3.10 conveys the effect of the activity and transistor sizing on the yield. As more select inputs are set at '1', the number of chains that switch increases (activity increases). The frequency of the circuit is the same as that of the ring oscillator ($f_{min} = 1.3 f_{EDPmin}$ for all activity factors). However, the switching probability of the internal nodes increases, and consequently, the power consumption increases. As a result, the temperature of the circuit is elevated. Therefore, in the V_{th}-V_{dd} plane, the contours of the temperature shift towards the lower supply voltages and have a smaller shift towards higher V_{th} values. For example, at a higher activity, the circuit reaches 100 °C at a lower V_{dd}. This causes the feasible region to shrink, and results in a yield loss.

The variations in the threshold voltage is slightly reduced by the increase in the device size, that in turn, enhances the yield if considered independently [94]. However, in an application for which temperature is constrained by a maximum value, the increase in the size of the transistors reduces the yield. Because the transistors parasitic capacitance, and therefore, their power consumption are increased by increasing their sizes. Consequently, similar to those of the higher activity, the temperature contours shift towards the lower V_{dd}s and have a smaller shift towards higher V_{th}s where it results in the yield loss. The simulation results demonstrate that when the transistors sizes are scaled up the contours of frequency are steeper and slightly shifted towards lower V_{dd}s and have a small shift towards higher V_{th}s which indicate a small performance enhancement. However, the increase rate in the power consumption is larger than that of the performance. Consequently, the shift of the temperature contours is greater than that of the performance contours. The overall result is that the feasible region shrinks and the yield is reduced. As shown in Fig. 3.10, for the given constraints, this rapid change of the feasible region makes the yield more

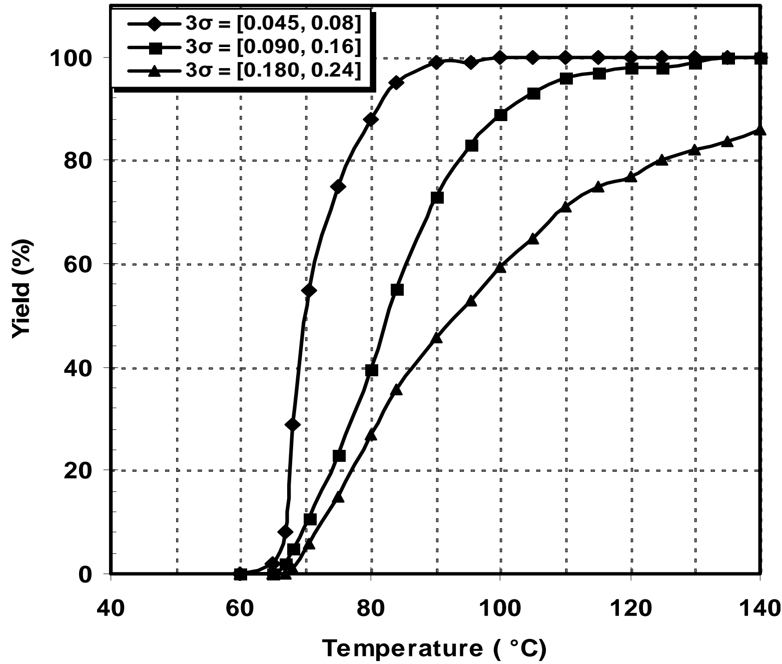


Figure 3.11: Sensitivity of yield to temperature constraint for three cases of variations in the design variables

sensitive to the sizing than to the activity where the activity factor is 0.2. Because at 0.2, the feasible region is small enough that the power increase due to scaling up the sizes has a big impact on the yield and imposes a sharp yield loss.

In many cases, the yield is not 100% for the given set of constraints. To increase the yield, a designer can relax the design constraints. This is specifically the case when the variations in the design variables are fixed such that there is no control over the size of the tolerance box. By employing a DTM system with a faster policy response, a designer can select the maximum temperature that is less conservative. As depicted in Fig. 3.11, the parametric yield is very sensitive to the temperature constraint. Such a sensitivity is more pronounced when the variations is lower. This provides the motivation to investigate the trade-offs between using more expensive DTMs and lower yields.

As technology scales, the design center that is most immune to the variations, for the same set of constraints, tends to shift towards the smaller values of the V_{dd} and higher values of the V_{th}. This is due to the fact that by increasing the power density, leakage, and temperature, the thermal runaway region is expanded towards the higher V_{th}, and the temperature contours shift towards the lower V_{dd}s. That is, in spite of the need for a higher overdrive voltage, to achieve an acceptable performance. Thus, variations slow down the

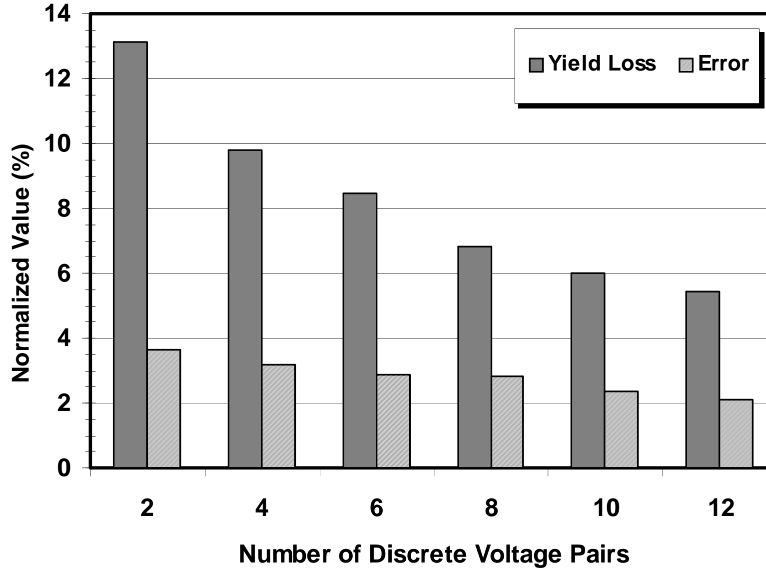


Figure 3.12: The average normalized yield loss and the error in the estimated yield for the discrete case where limited number of voltage pairs are available.

rate of the V_{th} scaling in new technologies. It is also noteworthy that some points in the tolerance box fall in the thermal runaway region. Obviously, the location of the design center affects not only the parametric yield loss, but also the risk of chip failure.

So far, both V_{dd} and V_{th} have assumed any values in their ranges. However, in some voltage scaling schemes such as V_{th} hopping, the available nominal voltage levels are limited [71][95]. Thus, in order to maximize the yield, neighbor-mapping is required. Therefore, the closest available voltage pair in respect to the center of the tolerance box is selected. As depicted in Fig. 3.12, deviating from the original maximum yield center degrades the yield. The figure shows the average yield loss and the error imposed when a limited number of threshold and supply voltage levels is available. This yield loss is much higher when the available number of voltage pairs is low. Therefore, relaxing the constraints must be considered, for this cases, in order to improve the yield. In addition, neighbor mapping can introduce error in obtaining the maximum yield for the discrete case. This is due to the fact that the feasible region is not symmetric. Therefore, two points with the same distance from the center of the tolerance box can have two different yield values. However, the error decreases with the increase in the number of available voltage pairs. These average yield loss and error numbers are obtained using Monte-Carlo (MC) simulations for the example depicted in Fig. 3.5 where the voltage pairs are randomly generated in the feasible region. Then, the pair representing the maximum yield is identified among the available pairs. This yield is compared with the maximum yield obtained from MC, for the continuous case, to report the yield loss. The maximum yield, from the discrete case, is also compared with

the one obtained using the methodology and the error is calculated. The MC simulation is executed for 1000 times and the average numbers are reported on the figure. The low complexity and high efficiency of the presented methodology makes it appealing even for the discrete cases where high accuracy is not required or the number of available pairs is high. Moreover, the high sensitivity of the subthreshold logic circuits to temperature and process variation is a good motivation to extend the method in order to increase the robustness of sub-threshold MOS logic families [96]. However, the subthreshold designs are low power/ performance applications and the thermal runaway constraint would not be useful. Instead, the maximum power and minimum performance can form the feasible region and PDP can be employed as a more appropriate design metric.

3.7 Impact of Scaling

Because of the increase in process and environmental variations, the physical characteristics of a device are even more prone to uncertainties in future technology nodes. Understanding the trend of the changes in the parametric yield, as transistor dimensions are scaled, helps designers to make informed decisions, in regards to voltage scaling, transistor sizing, power-performance trade-offs, and thermal management.

Researchers from Intel® have argued that process variations are not an "insurmountable barrier" to Moores Law, but is simply another challenge to be overcome [12]. They provide evidences regarding ICs designed by using $45nm$ technology, where the use of HiK+MG has been a significant factor in variation management.

A key objective of this section is to quantify the changes in the parametric yield of the current and future nodes, and provide designers with insight into designing robust integrated circuits in scaled technologies. In addition, design centering is used to propose a design with the maximum immunity to the variations for each technology node. Also, the change in the design feasible region, illustrated in this work, is interesting from a designer's point of view. This helps to balance the trade-off between relaxing the design constraints and the cost of maintaining or increasing the yield in the future generations of CMOS technologies.

3.7.1 Trend of Variations in the Design Parameters

Variations in the threshold voltage occur because the manufacturing parameters deviate from their nominal values. There are several sources of variations, some of which are critical. Random dopant fluctuation (RDF) is becoming more significant as the number of dopant atoms in the device channel decreases to less than 100 beyond $45nm$. Consequently, RDF has a high impact on the V_{th}. Also, the line-edge and line-width roughness,

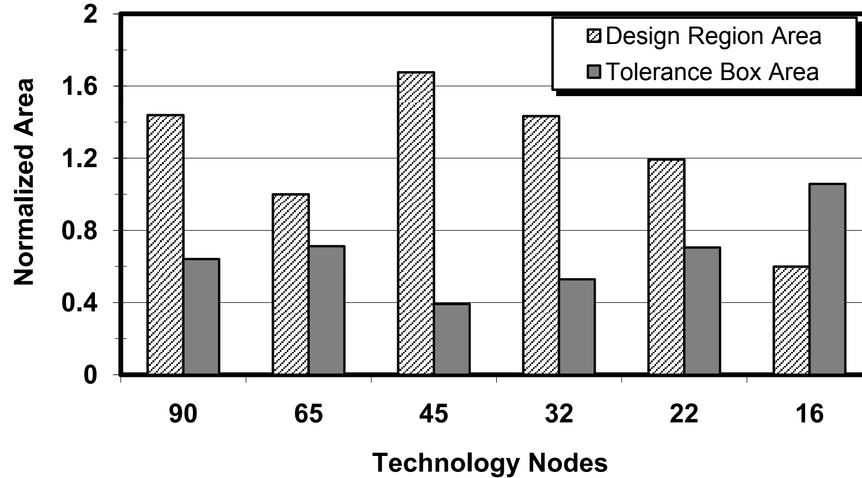


Figure 3.13: Area of the feasible region and the tolerance box for different technology nodes.

associated with poly-gate patterning, increase the subthreshold current and degrade the V_{th} characteristics. The introduced HiK+MG are also impacted by the variations in the gate dielectric, including oxide thickness, interface traps and fixed charges. The variations in the V_{th}, due to these sources are becoming comparable to the RDF. To alleviate the effect of the variations on the threshold voltage, the process and design techniques, or a combination of both, are utilized. The square poly endcap, patterning, and dielectric trap improvement are among the process techniques, while input chopping is an example of the design techniques in $45nm$ [12]. However, despite these temporary improvements, the fluctuations in the V_{th} increase beyond $45nm$ [97], [3].

The variations of the supply voltage are primarily due to the non-uniformity in the distribution of the power supply and the changes in the switching activity of the circuit. Historically, designers tend to limit the within-die changes in the V_{dd} due to the IR and Ldi/dt drops, to a maximum of 10%. Nonetheless, with the scaling of technology, the increase in the current density and rate of switching make it more challenging to retain this traditional bound on the supply voltage noise [24].

In this research, the interest is in the total variations in the V_{dd} and V_{th} due to the aforementioned sources. The variation in the V_{dd} is assumed to be 10% and the variation in the V_{th} for different nodes, have been estimated according to our simulation results and others in the literature [12], [98], and [99]. These values are listed in Table 3.2.

Fig. 3.13 compares the area of the feasible regions from $90nm$ to $16nm$. These are the regions, bound by the design constraints for the given technology node. To find the design centers, the yield optimization methodology, explained in Section 3.5, is applied to different technologies. The dimensions of the tolerance box, moved over the design space, depend

on the probability distribution of the design variables, V_{th} and V_{dd}, for a given technology. The smaller the tolerance box (lower variations), the higher the chance is for the box to be embedded within the design space, where more designs fall in the feasible region. This results in a larger yield. As seen in Fig. 3.13, the relative area of the tolerance box to that of the feasible region, increases by scaling the devices beyond $45nm$.

3.7.2 Comparing the Results for Different Technologies

To compare the different nodes, Cadence SPECTRE simulations are performed on the circuit, adapted from [46]. For technologies from $45nm$ to $16nm$, HSPICE simulations are carried out on the Predictive Technology Models (PTM) [100]. The circuit in Fig. 3.8 consists of 11 stages and 10 levels. The first level is a NAND ring oscillator whereas the other levels are NAND chains. The activity factor of the circuit is controlled by assigning '0' or '1' to the Select inputs. By assigning '1' to Select 0 and '0' to the subsequent select inputs, the circuit becomes a ring oscillator (an activity factor of 0.1).

The 1D thermal model, $T_j = T_{amb} + \theta_{ja}P$, is used iteratively to estimate the steady state temperature for a given pair of V_{dd} and V_{th}. Here, T_{amb}, θ_{ja}, and P are the ambient temperature, thermal impedance of the junction to the ambient, and the total power consumption, respectively. If the model does not converge for a voltage pair, the corresponding design is considered to be in the thermal runaway region. θ_{ja} is assumed to be 0.9 K/W, for all the nodes to have a fair comparison. Fig. 3.14 illustrates the normalized EDP contours, iso-performance curves, and temperature contours for one of the nodes in $32nm$ technology. The distortion of the contours is caused by variations of the V_{th} around the value set in the model file and its impact on power and performance.

3.8 Design Insights for Current and Future Technologies

If the constraints that form the feasible region are tight, the region shrinks, and, therefore, the yield loss increases. Fig. 3.15 illustrates the yield as a function of the maximum allowed temperature (relaxing the temperature constraint). As seen from the figure, when the design tolerates higher temperatures, a higher yield is achieved for all the technologies. However, the sensitivity of the yield varies for each case. It can also be observed that when the variations increase, the sensitivity decreases, where further relaxation of the temperature constraint does not gain as much yield enhancement. In addition, when the variations increase, no matter what the maximum temperature constraint is, a 100% yield is not achievable for future nodes. Furthermore, a designer can gain some limited

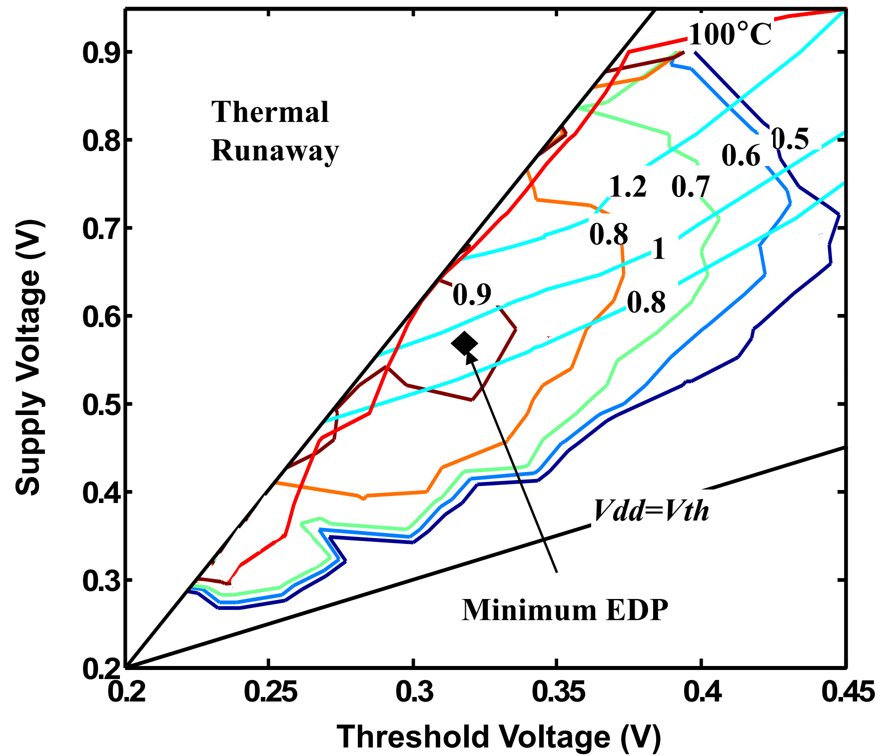


Figure 3.14: Normalized EDP contours, iso-performance curves, and temperature contour for 32nm technology.

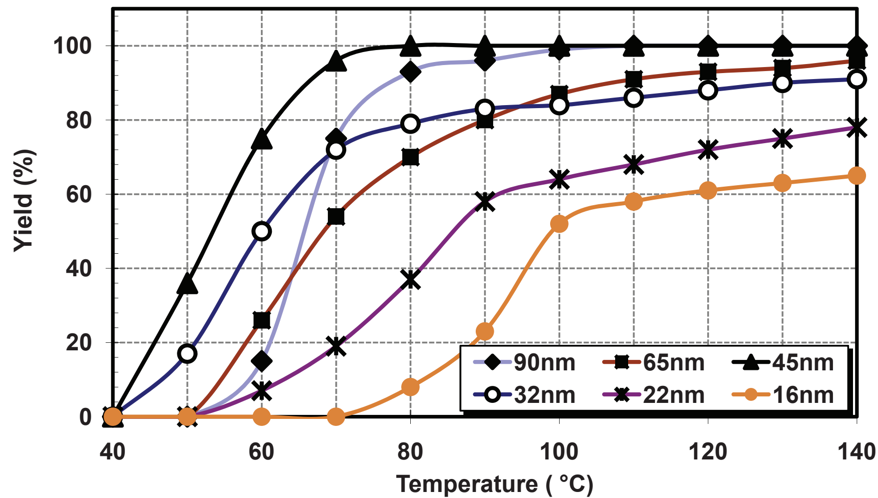


Figure 3.15: Sensitivity of yield to maximum allowed temperature for different technology nodes (relaxing the temperature constraint).

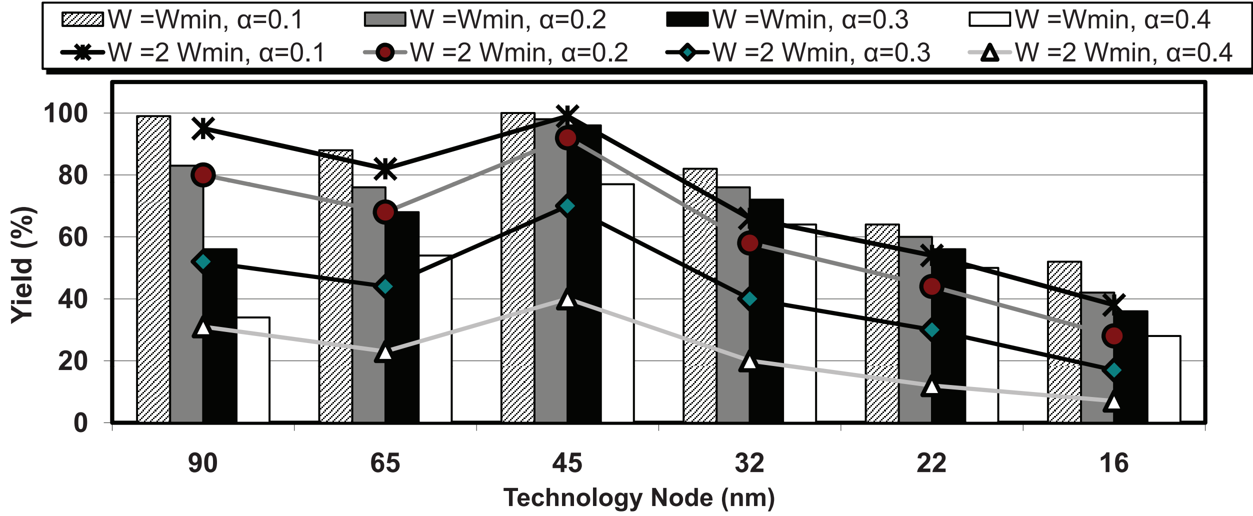


Figure 3.16: Sensitivity of the parametric yield to the activity factor and transistor sizing for different technology nodes.

control over the manufacturing results through a litho-friendly layout design [92][93] or other techniques, discussed in Section 3.7.1. Usually, the designer has little control over reducing the variations in the design variables. Therefore, in such cases, relaxing the performance constraint is the last alternative.

Fig. 3.16 depicts the impact of the activity factor and gate sizing on the parametric yield for different technology nodes, where W_{min} is the width of a minimum size inverter in the simulated circuit. By setting more select signals in the simulated circuit at '1', the number of chains which switch increases. Here, more internal nodes switch, and, thus the dynamic power consumption increases. Consequently, the temperature of the circuit increases, resulting in an increase in the yield loss. The larger gate sizes reduce the variations in the threshold voltage due to the relative lower fluctuations in the number of dopants [94]. However, for an application where the maximum temperature is a constraint, bigger gates have the same impact as the most activities and result in a higher yield loss. By using HiK+MG in $45nm$, temporarily, alleviates the impact of the activity and gate sizing on the yield degradation. The sensitivity of the yield loss to the activity and gate size is larger for the technologies that are more sensitive to a power increase such as $16nm$.

In Fig. 3.17, the result of yield optimization is depicted, indicating the robust design centers for different technology nodes. Also, the shift in the thermal runaway region is illustrated, where their region borders are identified by lines. The area to the left of each line represents the region, where no steady state temperature is found for the specified technology. Note how the design centers move towards the higher V_{dd} and V_{th} values, and

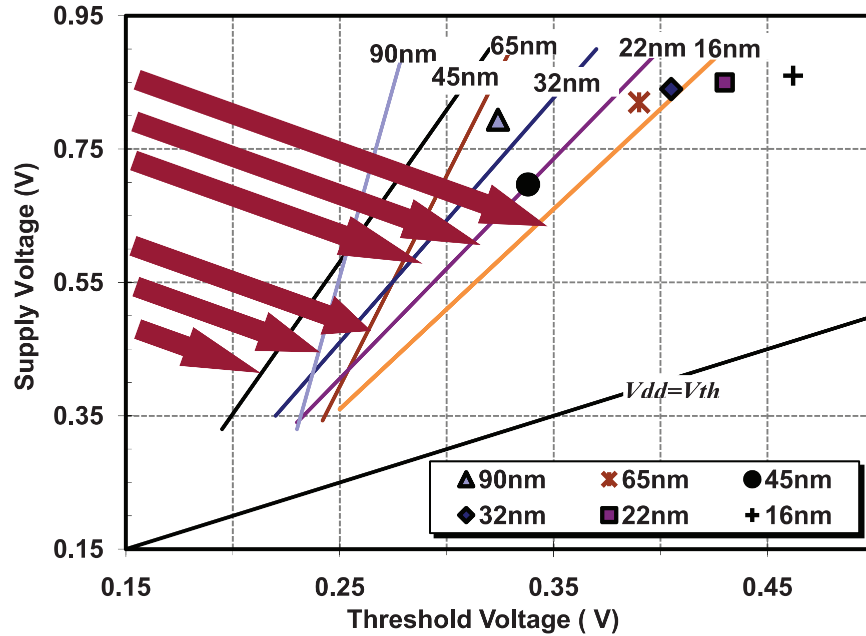


Figure 3.17: Maximum yield design centers and the shift in the thermal runaway region for different technology nodes.

the thermal runaway region expands to the right for the technologies beyond 45nm. This slows down dynamic voltage scaling, a vital power saving mechanism, and the V_{th} scaling that attains a performance enhancement as a substantial scaling motivation. Therefore, the thermal reliability concern manifests itself as another obstacle in designing robust integrated circuits.

Table 3.2 shows the results of the yield maximization for the specified technologies. The mean and standard deviation of the design metrics, at the maximum yield point, are given for each node. As shown in Table 3.3, relaxing the performance, Case 1, and temperature, Case 2, by 30% can significantly increase the yield. However, allowing either a lower performance or a higher temperature in many applications is not an option. In such cases to guarantee a 100% yield, a better control over the voltage noise, as well as an increase in the precision of manufacturing equipment are needed.

3.9 Conclusions

Robustness is becoming a top priority that drives most design decisions for nanometer CMOS technologies. A statistical methodology is proposed in this work to achieve the robustness and the best possible trade-off. To increase the yield, a designer can relax the

Table 3.2: Yield optimization for different technology nodes.

Patameters			Technology Nodes					
			90	65	45	32	22	16
6σ (V)	V_{th}		0.108	0.120	0.066	0.099	0.149	0.223
	V_{dd}		0.200	0.200	0.200	0.180	0.160	0.160
Deterministic Center	V_{th}	Nominal (V)	0.271	0.376	0.265	0.324	0.331	0.350
	V_{dd}	Nominal (V)	0.510	0.647	0.458	0.584	0.584	0.584
Maximum Yield Center	V_{th}	Nominal (V)	0.324	0.390	0.338	0.405	0.430	0.462
		STD	0.020	0.020	0.011	0.021	0.031	0.037
	V_{dd}	Nominal (V)	0.794	0.820	0.697	0.840	0.850	0.860
		STD	0.026	0.037	0.033	0.027	0.036	0.027
	Normalized EDP	Nominal	0.560	0.436	0.680	0.441	0.412	0.464
		STD	0.051	0.038	0.046	0.032	0.027	0.060
	Normalized Performance	Nominal	1.538	1.334	1.384	1.357	1.270	1.392
		STD	0.133	0.133	0.126	0.109	0.109	0.175
Yield (%)			99	88	100	84	64	52

Table 3.3: Increasing yield by 30% relaxation of performance, Case 1, and temperature, Case 2 constraints

		Technology Nodes					
		90	65	45	32	22	16
Yield (%)	Case 1	100	100	100	100	96	88
	Case 2	100	96	100	91	78	65

constraints. In a design with tighter constraints, the chosen metric (the given priority to the delay as opposed to the power) is less important. It is the yield that dictates where the design center should be, not the power-performance trade-off. Using this methodology, a pair of optimal V_{dd} and V_{th} is obtained for which the yield is maximized and a near-optimal trade-off between power and performance is achieved (as a secondary objective).

Introducing high-k materials and metal gates creates opportunities that mitigate the impact of variability on the design. However, it is demonstrated in this thesis that enhancing performance, reducing power consumption, and attaining an acceptable level of reliability beyond $45nm$ technology requires close attention to the trend of changes in the parametric yield. The maximum yield design centers for future nodes are also proposed in this chapter, and the impact of switching activity and device sizing on the parametric yield is discussed. However, expanding the feasible region, where, for example, higher temperatures can be tolerated, should be considered to enhance the design robustness in scaled technologies.

Chapter 4

Power Grid Analysis and Verification Considering Temperature Variations

4.1 Introduction

The experimental results, at the circuit level, in the previous chapter reveals two important points. First, temperature significantly impacts the parametric yield, especially for high performance applications. It is found that the yield is degraded in a circuit with a high activity factor and large devices, where the increase in the temperature is responsible for the yield loss. Second, the yield is very sensitive to the value of supply and threshold voltage and their respective variations. In fact, based on the simulation results at $22nm$, about 20% of all designs lead to thermal runaway and fail. This is due to the increase in the magnitude of the variations. Therefore, the interdependency between temperature and voltage variations should be taken into account to accurately estimate and optimize the yield.

Motivated by the results of the study at the circuit level, the impact of process variations on the voltage drop is studied in this chapter. Subsequently, the interdependency of process and environmental variations and the impact on the system parametric yield are investigated in the next chapter. The global nature and spatial variations of process, temperature, and supply voltage call for the analysis at the system level to achieve more accurate results.

Reducing power supply (V_{dd}) and threshold voltage (V_{th}) in modern integrated circuits have increased the sensitivity of the circuits to the voltage variations. To guarantee the appropriate functionalities for the circuits, a power grid must deliver a proper V_{dd} to each node over the chip. One of the principal sources of variations in the node voltages is the variation of the leakage current, drawn from a node. This is a result of process variations

manifesting themselves as large V_{th} variations in new technologies [101]. This chapter presents an analysis for power grid verification in the presence of such variations.

4.2 Related Work

Traditionally, power grids have been designed, based on the data available from previous experience. The lack of data in the early stage makes it challenging to start with what was known as "safe" and "over-designed grids" in former designs. Several efforts have been reported in the literature to analyze power grids when there is limited information about the current density of underlying circuits. A recent work presents a deterministic power grid verification method that takes the local current constraints as the user input in order to find the worst voltage drop on the grid [102]. The method uses a sparsity technique that reduces the size of the optimization problem. In [36], a method is proposed for estimating the voltage variance of a power grid. Here, a lognormal distribution for the voltage drop is obtained, considering random leakage currents. In [103], a stochastic method is provided for analyzing voltage drop variations of an on-chip power grid network. In the work, the data retrieved from normal transient simulation of a circuit is employed to estimate the variance of node responses. The authors in [104] derive an upper bound for a worst-case voltage drop on a grid by using the information of chip power consumption. Here, a random-walk algorithm is presented for the heuristic search of the worst voltage drop.

The aforementioned work do not account for high temperature impact in their power grid analyses. Although the random selection of current sources in existing works simplifies the analysis of a power grid, such a selection does not provide realistic results. Power grid verification is carried out as early as the floorplanning stage, where the design is flexible enough to allocate adequate resources to a portion or all of the grid. However, changing the floorplan can significantly alter the chip thermal profile. The use of independent leakage numbers cannot capture the impact of such changes on the final results.

Few work have included high temperature impacts on leakage power and discussed the implication for *IR* drop in the power supply network [105], [106]. However, these deterministic analyses do not address variability in the design of modern ICs. Su et al. [5] have estimated the leakage power across the chip considering uneven temperature and voltage profile. They have also used a polynomial model to include the impact of these non-uniformities in the leakage estimation. But they forgo to account for the variability of process parameters. Therefore, their method only provides a rough estimate of the nominal leakage power.

The subthreshold leakage power is exponentially related to the transistor operating temperature. Also, the resistivity of interconnects linearly increases with increased tem-

peratures. Consequently, ignoring the temperature variations leads to significant underestimations in the power grid design.

The contributions of this chapter of the thesis are as follows:

Due to the strong interdependency of leakage power and temperature, the process variations impose statistical behavior on the operating temperature. Also, the metal resistivity of a power grid increases with temperature. This work generates a statistical thermal profile across a power grid. Then the close loop between the temperature and leakage power consumption is used to map the process variations to the voltage drop statistics. It is shown, in this chapter, that the IR drop is significantly impacted by the statistical thermal profile.

In addition, motivated by the results of the analysis, it is strongly advocated that any verifications of a power grid should account for process-induced statistical thermal profile. A power grid verification method is also presented to efficiently find any voltage violation in the microprocessor. Since both dynamic and leakage power depend on the supply voltage, the iterative method, in this work, ensures a high accuracy for the extracted voltage drop moments.

4.3 Statistical Thermal Analysis

4.3.1 Motivation and Workflow

The strong correlation between the process variations and the statistical thermal profile has been demonstrated in the early version of the thermal analyzer [107]. It has been shown that process variations can impose significant variations in the temperature across the chip. In fact, based on the results in [107], using just the nominal temperature map and ignoring the statistical thermal profile leads up to 30 °C underestimation of the on-die temperature in $90nm$ technology. This temperature difference will be even larger for smaller feature sizes where variations are more significant. As illustrated later in this section, Fig. 4.2, our simulation of a ring oscillator shows that this, in turn, results in up to 80% error in the estimation of the leakage power. The impact of such inaccuracy on the power grid verification and timing yield is significant. Ignoring statistical measures of the operating temperature leads to underestimation of IR drop and consequently can cause performance degradation.

In this section, the thermal analyzer is extended to include the interdependency between the statistical thermal profile and supply voltage variations.

The power grid analysis, in this work, consists of two stages. First, statistical measures of the power and temperature are extracted across the die. Next, as discussed in the next

section, the statistical moments of the voltage drop are obtained by using the extracted temperature and power moments. This two stages are executed sequentially in a loop (outer loop). This is to ensure high accuracy by taking into account the impact of the voltage drops, while the temperature and power statistics are computed. The complete algorithm is discussed in Algorithm 4.1.

The voltage drop on the nodes of a power grid are a function of the currents drawn off the nodes. The power grids are modeled as RC networks with current sources, representing the drawn currents, connected to the nodes. These currents, in turn, depend on the power consumption of the underlying circuits. The power consumption composes of two primary components, dynamic power and leakage power. Process variations result in significant variations in leakage power across a die. Subthreshold, gate leakage, and drain-induced barrier lowering, all, contribute to the total leakage power. However, due to the use of high-k dielectrics specifically for high performance applications, subthreshold leakage is the dominant component [80]. In addition, subthreshold leakage is exponentially related to the operating temperature of the circuits. Because of this strong interdependency, variations in leakage power impose statistical behavior on the operating temperature that must be captured to accurately analyze the power grid.

4.3.2 Statistical Thermal Model

This section models and extracts the statistics of the power and temperature for the analysis of the power grid in the subsequent sections.

The following heat transfer equation governs the chip steady-state temperature [23]:

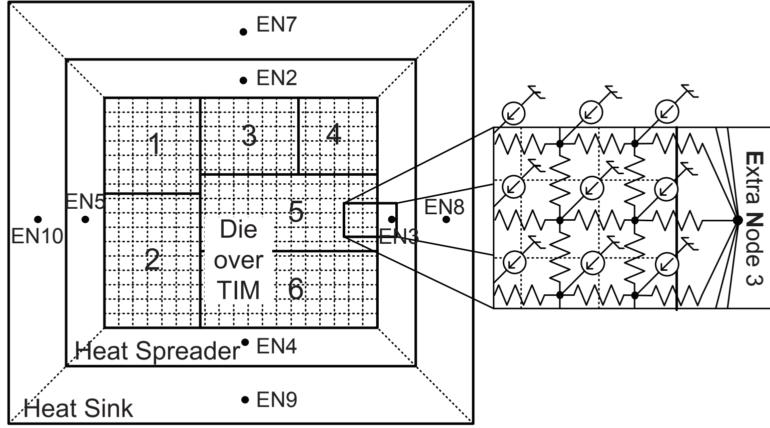
$$k(x, y, z).\nabla^2 T(x, y, z) + p(x, y, z) = 0 \ , \tag{4.1}$$

where k is the thermal conductivity, T is the temperature, and p is the power density. Leakage component of the power consumptions is spatially correlated across the die [108]. Therefore, a deterministic solution for the heat transfer equation does not capture the variability impact as well as the interdependency between the leakage and temperature.

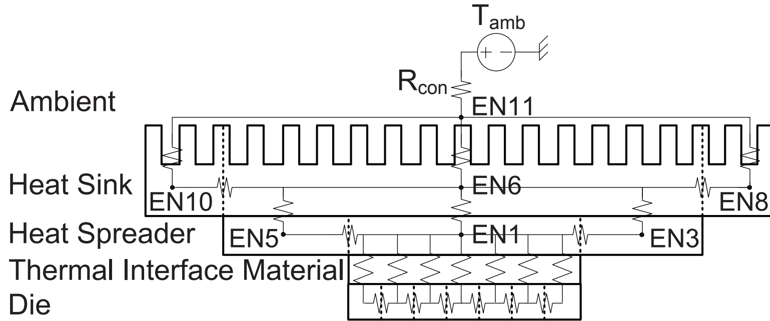
To accurately estimate the static measures of temperature distribution over the die, a statistical thermal analyzer is adopted from [107]. Initially, as illustrated in Fig. 4.1, the die area is discretized into n grids and modeled as an equivalent circuit network [1]. Thus, the grids' temperature are obtained by the following matrix multiplication:

$$t_{m\times 1} = A_{m\times m} \times p_{m\times 1} \ , \tag{4.2}$$

where t is the vector of grid temperatures. A is the inverse of the admittance matrix of the equivalent circuit, for modeling the heat transfer. Here, m is the total number of



(a)



(b)

Figure 4.1: Discretized die with six cores and the package structure [1] (a) Top view (b) Lateral view

thermal grids and other grids for modeling the packaging components. Also, p_m is the current source used to model the chip to ambient removing power, where $p_m = T_{\text{amb}}/R_{\text{con}}$ and R_{con} is the heat resistance from the heat sink to the air. Note that this equivalent circuit is used to accurately take into account the heat transfer from the die to the thermal interface, heat spreader, heat sink and subsequently to ambient. In this 3D thermal model, the heat is transferred to the heat sink by heat conduction. The heat, then, is removed from the heat sink by the cooling system through heat convection. R_{con} represents the thermal resistance of the heat sink to ambient. All these thermal resistors are used to form matrix A.

We consider inter-die and spatially correlated intra-die variations in the gate length (L_g) and oxide thickness (T_{ox}), as the sources of the variations. Note that the Random Dopant Fluctuation (RDF) was also included in the primary investigation. However, including large

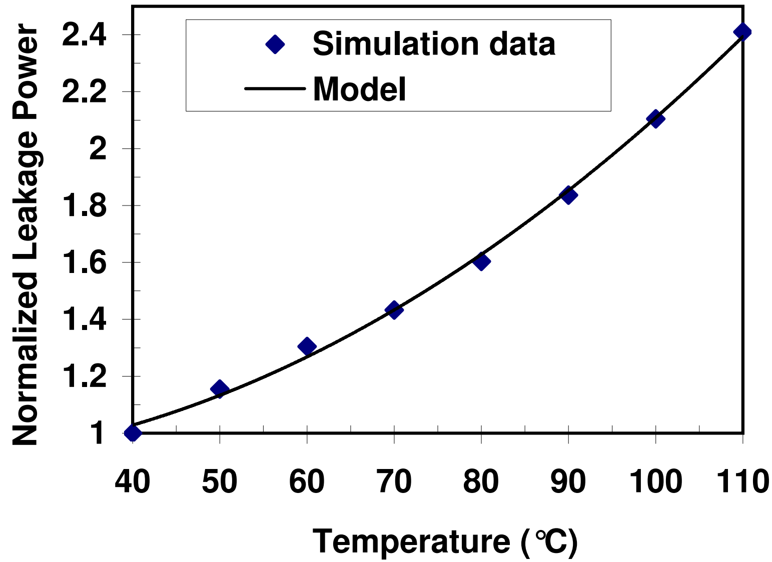


Figure 4.2: Normalized leakage power as a function of temperature for the circuit in Fig. 3.8.

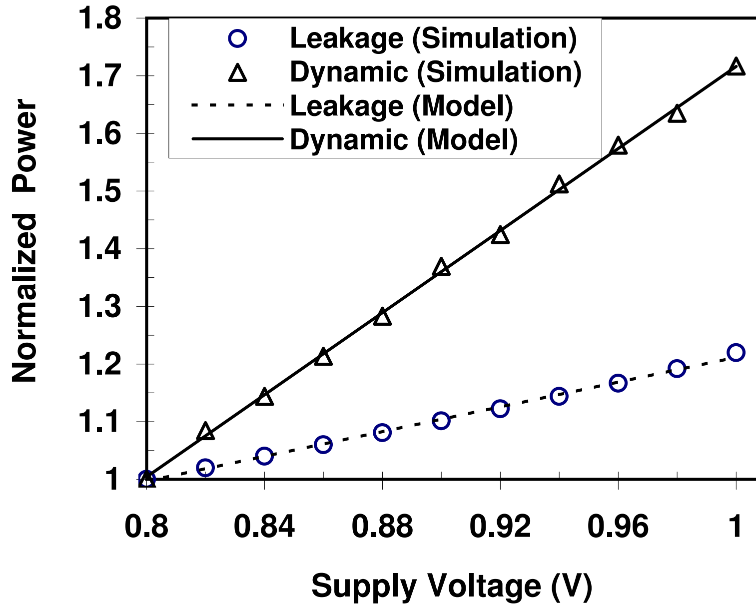


Figure 4.3: Normalized power consumption as a function of supply voltage for the circuit in Fig. 3.8.

number of uncorrelated random variables results in a random variable with zero standard deviation (law of large numbers), and, thus, RDF is ignored. The following expresses the

variability in a single grid i, hereafter is referred to as the thermal grid to distinguish it from the power grid:

$$X_i = X_0 + \triangle X_i \ , \tag{4.3}$$

where X_0 is the nominal value of the physical parameter X, and $\triangle X_i$ is the deviation from the nominal value. $\triangle X_i = \{\triangle L, \triangle T_{ox}\}$ is assumed to have a Gaussian zero-mean random distribution with the standard deviation of $\{\sigma_L, \sigma_{T_{ox}}\}$.

The subthreshold leakage current is exponentially related to the variations in the gate length and the gate leakage current has an exponential dependency on the oxide thickness [101]. Therefore, the normalized total leakage current in thermal grid i can be approximated by computing

$$\begin{aligned} \hat{I}_{leak-i} &= e^{\beta_{L_i}.\triangle L_i + \beta_{T_{ox_i}}.\triangle T_{ox_i}} \\ \beta_{X_i} &= \left.\frac{\partial(\ln \hat{I}_{leak-i})}{\partial X}\right|_{X=X_0} \ , \end{aligned} \tag{4.4}$$

where β_{X_i} is the first-order derivative of the leakage current logarithm in thermal grid i.

To model the leakage power, a polynomial regression around the temperature and supply voltage nominal values is chosen for each thermal grid. In addition the dynamic power is modeled using a quadratic function of the supply voltage. As shown in Fig. 4.2 and 4.3, the models are in good agreement with the data obtained by simulating the circuit in Fig. 3.8 using 65nm technology. Therefore,

$$\begin{aligned} P_{dyn-i} &= b_{1_i}\left(1 + b_{2_i}(V_i - V_{dd}) + b_{3_i}(V_i - V_{dd})^2\right) \\ P_{leak-i} &= \hat{I}_{leak-i} \times c_{1_i}\big(1 + c_{2_i}(T_i - T_{\text{ref}}) + c_{3_i}(T_i - T_{\text{ref}})^2 \\ &\quad + c_{4_i}(V_i - V_{dd}) + c_{5_i}(V_i - V_{dd})^2 + c_{6_i}(T_i - T_{\text{ref}})(V_i - V_{dd})\big), \end{aligned} \tag{4.5}$$

where P_{leak-i}, P_{dyn-i}, and V_i are random variables that represent the leakage power, dynamic power, and supply voltage, respectively, in thermal grid i, considering the process variations. b_{2_i}-b_{3_i} and c_{1_i}-c_{6_i} are the fitting parameters, and b_{1_i} is the dynamic power at the ideal supply voltage $V_i = V_{dd}$. Also, c_{1_i} denotes the nominal total leakage power in thermal grid i at reference temperature T_{ref} and V_{dd}. The statistical behavior of the power consumption leads to variations in the operating temperature of each thermal grid which is modeled by the following random variable:

$$T_i = \sum_{j=1}^{n} a_{ij}.P_j + a_{im}.p_m \ , \tag{4.6}$$

where P_j is the total power in thermal grid j and a_{ij} is an element in the inverse of thermal admittance matrix (A).

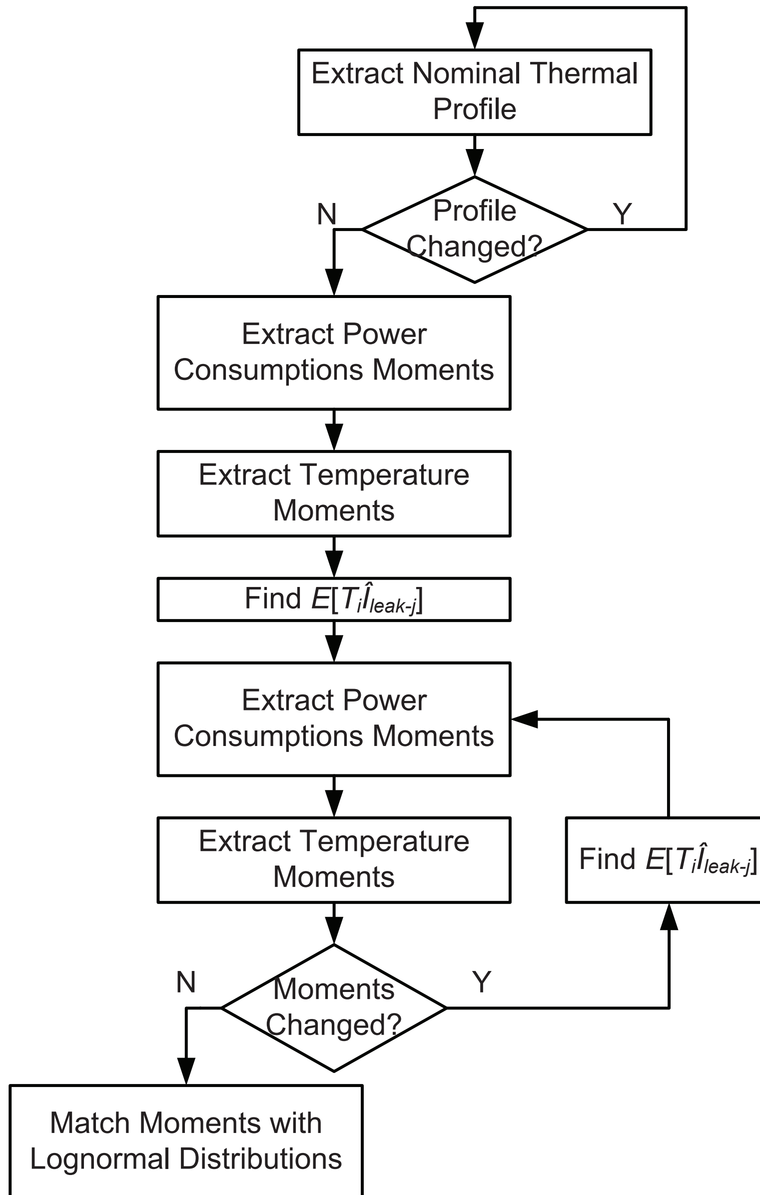


Figure 4.4: Flowchart of statistical thermal analyzer [107].

The aforementioned power/temperature models are employed to obtain the statistical measures of power/temperature in the statistical thermal analyzer, depicted in Fig. 4.4. In the first step, to consider the closed-loop effect between leakage and temperature, the following must be evaluated iteratively:

$$
\begin{aligned}
P_{dyn-j} &= b'_{1_j}\left(1 + b'_{2_j}V_j + b'_{3_j}V_j^2\right) \\
T_i^{(k+1)} &= \sum_{j=1}^{n} a_{ij}\left(p_{dyn-j} + p_{leak-j}^{(k)}\right) + a_{im}p_m \\
p_{leak-j}^{(k)} &= c'_{1_j}.\hat{I}_{leak-i}\times \\
&\quad \left(1 + c'_{2_j}T_j^{(k)} + c'_{3_j}T_j^{(k)^2} + c'_{4_j}V_j + c'_{5_j}V_j^2 + c'_{6_j}T_j^{(k)}V_j\right) \ ,
\end{aligned} \tag{4.7}
$$

where b'_{1_j}-b'_{3_j} and c'_{1_j}-c'_{6_j} are obtained from b_{1_j}-b_{3_j} and c_{1_j}-c_{6_j}, T_{ref}, and V_{dd}. After the nominal power/temperature are calculated for each thermal grid, the uncertainty, due to process variations, is captured in the second step. Note that the spatial correlation between the process variations in the two grids are taken into account in this step. Here, the expected value and covariance of power are extracted for all the thermal grids by evaluating

$$
\begin{aligned}
E[P_j] &= E[P_{dyn-j} + P_{leak-j}] \\
\text{Cov}(P_i, P_j) &= E[P_iP_j] - E[P_i]E[P_j] \ .
\end{aligned} \tag{4.8}
$$

The expected value of the power in thermal grid j is given by

$$
\begin{aligned}
E[P_j] = b'_{1_j}&(1 + b'_{2_j}E[V_j] + b'_{3_j}E[V_j^2]) \\
&+ c'_{1_j} \times (E[\hat{I}_{leak-j}] + c'_{2_j}E[T_j\hat{I}_{leak-j}] \\
&+ c'_{3_j}E[T_j^2\hat{I}_{leak-j}] + c'_{4_j}E[V_j\hat{I}_{leak-j}] \\
&+ c'_{5_j}E[V_j^2\hat{I}_{leak-j}] + c'_{6_j}E[T_jV_j\hat{I}_{leak-j}]) \ ,
\end{aligned} \tag{4.9}
$$

and

$$
\begin{aligned}
E[P_iP_j] = E[&P_{dyn-i}P_{dyn-j}] + E[P_{dyn-i}P_{leak-j}] \\
&+ E[P_{dyn-i}P_{leak-j}] + E[P_{leak-i}P_{leak-j}] \ .
\end{aligned} \tag{4.10}
$$

Adopted from [107], the following properties are utilized to find the components of (4.9), (4.10), and consequently extract the moments of power in (4.8).

Property 1 *Given a normal random variable X with mean and variance of (μ, σ^2), if $Y = e^{\beta . X}$, then the expected value of Y can be calculated as*

$$E[Y] = exp\left\{\beta\mu + \frac{\beta^2\sigma^2}{2}\right\} \tag{4.11}$$

From this property,

$$E[\hat{I}_{leak-i}] = e^{\frac{\beta^2_{L_i}\sigma^2_{L_i} + \beta^2_{Tox_i}\sigma^2_{Tox_i}}{2}} \tag{4.12}$$

Property 2 *Given a vector of lognormal correlated random variables, $X_{n\times 1} = [X_1, X_2, ..., X_k]$. If m_{Xi} and s_{Xi} are the expected value and standard deviation of the $ln(X_i)$, respectively, and $\rho_{X_iX_j}$ is the correlation coefficient between the $ln(X_i)$ and $ln(X_j)$, the random variable $Y = \prod_{i=1}^{k} X_i^{ni}$ is lognormal with the expected value of*

$$E[Y] = e^{\sum\limits_{i=1}^{k} n_i m_{X_i} + \sum\limits_{i=1}^{k-1}\sum\limits_{j=i+1}^{k} n_i n_j s_{X_i} s_{X_j} \rho_{X_iX_j} + \frac{\sum\limits_{i=1}^{k} n_i^2 s_{X_i}^2}{2}} \tag{4.13}$$

and $\rho_{X_iX_j}$ is obtained from the following property:

Property 3 *Assume two correlated lognormal random variables $X_1 = e^{Z_1}$ and $X_2 = e^{Z_2}$ with given $E[X_1X_2]$, where Z1 and Z2 have the mean and standard deviation of (μ_1, σ_1) and (μ_2, σ_2), respectively, then*

$$\rho_{Z_1Z_2} = \frac{ln(E[X_1X_2]) - \left(\mu_1 + \mu_2 + \frac{\sigma_1^2\sigma_2^2}{2}\right)}{\sigma_1\sigma_2} \tag{4.14}$$

By using the above properties, all components in (4.9), (4.10), and subsequently the moments of power in (4.8) are determined. The extracted statistic measures of power in all the thermal grids form matrices of expected values and covariance ($M_{P_{n\times 1}}$, $S_{P_{n\times n}}$). Using the matrix notation, the matrices of the expected values and covariance of temperature over the die are restated as

$$\begin{aligned} M_{T_{n\times 1}} &= A_{n\times n} \times M_{P_{n\times 1}} + p_m.a_{n\times 1} \\ S_{T_{n\times n}} &= A_{n\times n} \times S_{P_{n\times n}} \times A^T_{n\times n}\ , \end{aligned} \tag{4.15}$$

where $A_{n\times n}$ is the first left/upper $n \times n$ sub-matrix of the inverse admittance, and $a_{n\times 1}$ is the vector of the ambient temperature coefficients (a_{im}).

The extracted moments of temperature, are fed back and the moments of power are updated. The new power moments are used to obtain the temperature statistical measures. The iterative process (inner loop) is continued until the closed loop converges. By concluding this stage, the moments of temperature and power are used in the next stage for calculating the statistics of the voltage drop.

4.4 Voltage Drop Statistics

The objective of this section is to integrate the statistical thermal profile, extracted in Section 4.3, into the power grid model. This is to accurately map process variations to the statistics of the voltage drop across the power grid.

We consider a RC network that is distributed over the die in multiple metal layers. Each branch of the grid is modeled with a resistor, and all the nodes have a capacitor to the ground. Also, an ideal current source is assumed to be connected to the nodes in the first metal layer (M1). The Modified Nodal Analysis (MNA) governs the relationship between the current and voltage of every node, and is expressed as

$$GV(t) + C\frac{V(t)}{dt} = -i(t) + GV_{dd} \ , \tag{4.16}$$

where V(t) is the vector of voltages at each node, and i(t) is the vector of the current sources. G, C are the conductance and capacitance matrices, respectively. By setting $v(t) = V_{dd} - V(t)$,

$$Gv(t) + C\frac{v(t)}{dt} = i(t) \ . \tag{4.17}$$

An AC analysis of the power grid provides more detailed information, regarding the voltage drop wave form. However, in the early stage of the design, including transient data requires assumptions that can lead to inaccurate results. This occurs because there are various modes that the circuits can operate in. In addition, the thermal response time is several magnitudes larger than the clock speed. Therefore, a DC analysis is considered in this work. Nevertheless, the methodology is flexible enough to take the trace of the dynamic power as input and generate the transient data. The matrix format of (4.17) is

$$GV = I \ , \tag{4.18}$$

where V is the vector of voltage drop, and I is the vector of currents drawn off the power grid. Since the Power grid nodes are distributed over the die, several power grid nodes exist in each thermal grid. This is due to the fact that the power grid has a higher resolution

(thousands of nodes) in respect to the thermal grid (e.g., a 50×50 grid). Let v_l denote the voltage drop on node l of the power grid, located over thermal grid i. For all the N nodes,

$$[GV]_l = I_l, \quad \forall l = 1, ..., N \ . \tag{4.19}$$

To solve for voltage drop, (4.18) is rewritten as

$$V = G^{-1} I \tag{4.20}$$

In addition, the resistivity of power grid wires is a linear function of temperature variations expressed by

$$R = r_0(1 + c(T_l - T_{ref})) \ , \tag{4.21}$$

where r_0, is the resistivity at the reference temperature, and c is the temperature coefficient of the resistance. Therefore (4.20) is restated as

$$V = G_0^{-1} \Phi I \tag{4.22}$$

$$\Phi = \text{diag}(1 + c(T - T_{ref})) \ , \tag{4.23}$$

here G_0 is the conductance matrix at the ambient temperature. By applying the $E[.]$ and $Var[.]$ operators, as similarly done in [36], the statistical moments of the voltage drop is extracted by

$$\begin{aligned} E[V] &= G_0^{-1} E[\Phi I] \\ \text{Var}(V) &= G_0^{-1(2)} \text{Var}(\Phi I) \ , \end{aligned} \tag{4.24}$$

where $G_0^{-1(2)}$ denotes a matrix whose elements are the square of each element in the inverse of G_0.

Property 4 *Let $[X \sim (\mu_X, \sigma_X)]$ and $[Y \sim (\mu_Y, \sigma_Y)]$ denote two dependent random variables. The expected value and the variance of $E[XY]$ are given as in [109]*

$$E[XY] = Cov(X, Y) + E(X)E(Y) \tag{4.25}$$

and

$$Var(XY) = \mu_Y^2 \sigma_X^2 + \mu_X^2 \sigma_Y^2 + 2\rho \mu_X \mu_Y \sigma_X \sigma_Y \ . \tag{4.26}$$

The moments of (ΦI) are extracted by using (4.25), and (4.26). Note that ρ is obtained, based on the dependency of I and Φ on temperature. To verify the new model, 10000 Monte Carlo simulations are performed on (4.19) for a power grid with 16642 nodes. This model, on average, introduces a 0.9% and 3.8% error in the mean and standard deviation

of the estimated voltage drops, respectively. The error is reduced for larger power grids due to the increase in the number of nodes located over each thermal grid.

By using this property, (4.24) is evaluated. The calculated moments of the voltage drops are used in the next iteration, where dynamic and leakage power are updated and subsequently new values for temperature statistics are determined.

Considering the lognormal distribution of I, the voltage drop also has a lognormal distribution [110]. This occurs because of the linear relationship between the current sources and nodes' voltage drop. The extracted voltage drop statistics are employed for the verification and mapping of the chip voltage drop, depicted in the next sections.

4.5 Power Grid Verification

Maintaining an acceptable noise margin at the reduced supply voltage in modern technologies requires a restrict upper bound on voltage drop. Consequently, a power grid verification is necessary to ensure that the voltage drop, at each node, does not exceed the given threshold. Given the lognormal distribution of voltage drop V_l, with the expected value and variance of (m_l, v_l), a normal random variable W_l, with a mean and standard deviation of (μ_l, σ_l), exists such that $W_l = \ln(V_l)$. The power grid is said to be robust, if

$$\bar{v}_l = e^{\mu_l + r\sigma_l} \leq v_t, \quad \forall l = 1, ..., N \ , \tag{4.27}$$

where $\bar{v}_l$ is the upper bound of the voltage drop, v_t is the maximum acceptable voltage drop, and r is closely related to the confidence level $100 \times (1 - \gamma)$. γ is a small positive number and r is found by using the normal inverse cumulative distribution function, written as

$$r = F^{-1}\left(P|_{\substack{\mu=0 \\ \sigma=1}} = 1 - \frac{\gamma}{2}\right) \ . \tag{4.28}$$

For example, to attain 99.73% of the nodes meet the maximum voltage drop constraint, it is necessary to use 3σ measure ($r = 3$). In addition, the mean and standard deviation of W_l are expressed as [111]

$$\mu_l = \ln\left(\frac{m_l^2}{\sqrt{v_l + m_l^2}}\right)$$
$$\sigma_l = \sqrt{\ln\left(\frac{v_l}{m_l^2} + 1\right)} \ . \tag{4.29}$$

From (4.28) and (4.29), (4.27) can be evaluated for all the nodes in order to check the robustness of the power grid.

Algorithm 4.1 MAXIMUM_VOLTAGE_DROP

Input: μ_X, σ_X, ρ_X, A, P_{dyn}, P_{leak}, G, γ, δ {Process variations data, package thermal admittance, initial power conductance, confidence, and tolerance}

Output: $\bar{v}$

repeat

1: Extract T_i {nominal temperature}

  repeat

  Compute $T_i^{(k)}$ using (4.7)

  until $|T^{(k+1)} - T^{(k)}| <$ tolerance

2: Extract M_P, S_P, M_T, S_T {statistical moments of power and temperature}

  repeat

  Compute $E[P_j]$, $\text{Cov}(P_i, P_j)$ using (4.8)

  Extract M_T, S_T using (4.15)

  until moments change < tolerance

3: Estimate $E[\Phi I]$, $\text{Var}(\Phi I)$

  Compute $E[\Phi I]$ using (4.25)

  Extract $\text{Var}(\Phi I)$ using (4.26)

4: Compute $\bar{v}$ {voltage drop upper bound}

  Obtain $E[V]$ and $\text{Var}(V)$ using (4.24)

  Obtain μ_l, σ_l, r using (4.28), (4.29)

  Compute $\bar{v}$ using (4.27)

until changes of V moments < tolerance

For a very large power grid, calculating the inverse of conductance matrix G_0, consumes large CPU and memory resources. Therefore, the inverse matrix can be approximated using a method such as SPAI [102], [112] to speed up the verification process.

The statistics of the voltage drop are extracted from (4.24). Subsequently, the upper bound of the voltage drops are obtained, and the robustness of the power grid is checked from (4.27). Algorithm 4.1 depicts the verification process by taking the power numbers, package information as the input and, provides the maximum voltage drop across the die. The algorithm runtime is discussed in the next section.

4.6 Results and Discussion

To implement the methodology, an Alpha 21364 microprocessor running a MCF application is used for modeling the power and temperature. The power consumptions are assigned to the microarchitectural blocks in an ev6 floorplan according to [113], [114]. The package,

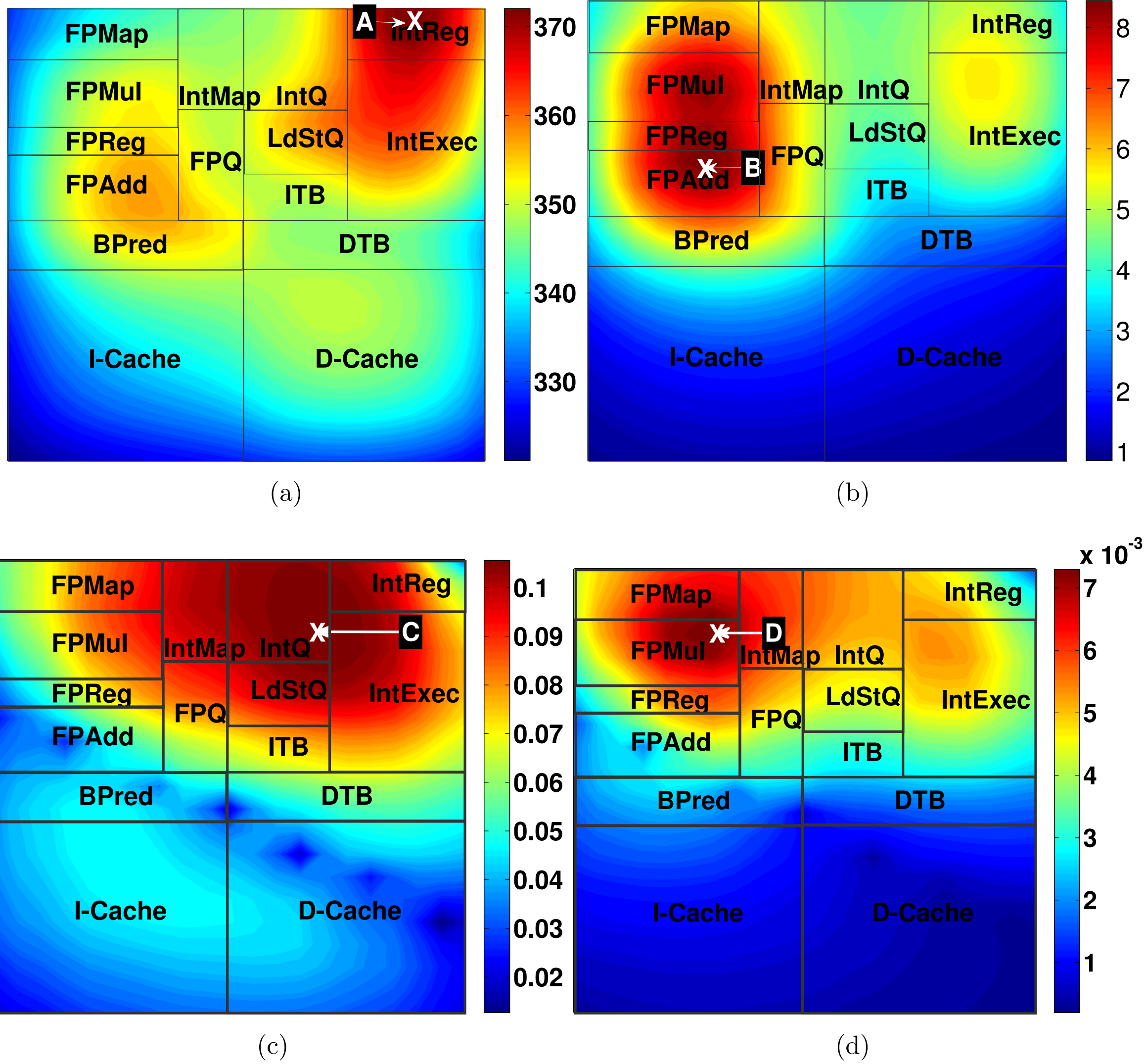


Figure 4.5: Statistical profile of the temperature and voltage drop of a 16642 node power grid across an Alpha 21364 CPU core.

illustrated in Fig. 4.1(b), is assumed to have a 50μm thermal interface at the top of a die whose thickness is 300μm. The package also consists of a $30 \times 30 \times 1$ mm heat spreader, as well as a $60 \times 60 \times 6.9$ mm heat sink. The ambient temperature is assumed to be 35 °C. The $3\sigma_L$ and $3\sigma_{T_{ox}}$ are set to 12% and 5%. The diminishing rate (ρ_X) for considering the spatially correlated variations follows that of [115]. Also, the floorplan area is discretized to $n = 50 \times 50 = 2500$ thermal grids. In regards to the power grid, we select metal layers, pitch, and width per layer based on IBM benchmarks [116], scaled for 65nm technology.

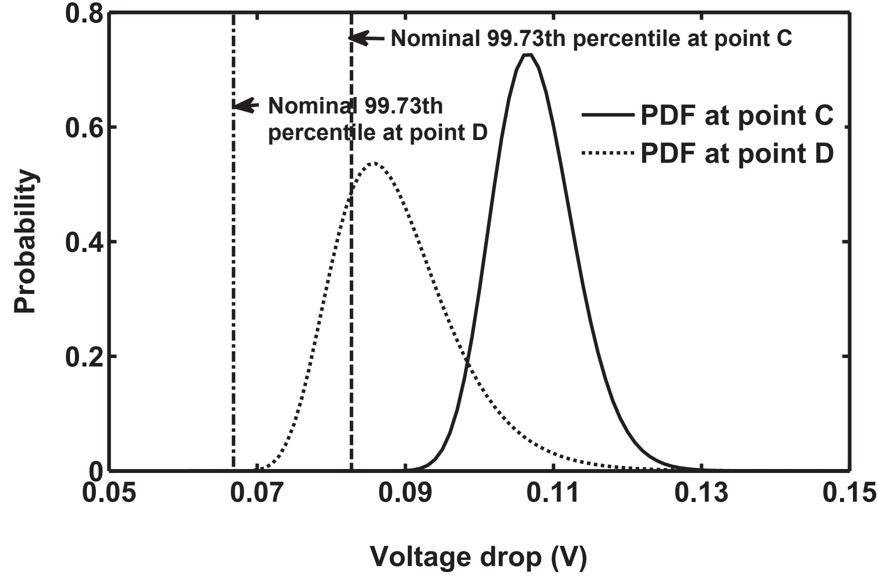


Figure 4.6: PDF of the voltage drop at the maximum expected value (point C) and at the maximum standard deviation (point D). The nominal upper bound of the voltage drops are also depicted at these points, where the nominal thermal profile is used.

We also assume a uniform distribution of C4s taking $V_{dd} = 1V$. Also, the temperature coefficient is assumed to be 0.0032 considering the thin barrier effect for the copper wires [106].

The proposed method is implemented in MATLAB and executed on a 3.4 GHz Pentium-4 PC with 2GB RAM. 2 iterations for the outer loop and 3 iterations for the inner loop (temperature-leakage closed loop) are enough to reduce the average error in the expected value of the voltage drops to less than 0.1%. Fig. 4.5 shows the profile of the temperature and voltage drop statistics across the core floorplan for a power grid with 16642 nodes and 100 C4s. In Fig. 4.5(a) and 4.5(b), the expected value and standard deviation of the temperature are mapped respectively. Also in Fig. 4.5(c), the expected value and in Fig. 4.5(d), the standard deviation of the voltage drop are demonstrated. Nodes A, B, C, and D represent the largest values of the temperature expected value, the highest temperature standard deviation, the maximum voltage drop, and the largest voltage drop standard deviation respectively.

Fig. 4.5(d), depicts how the voltage drop changes from one fabricated chip to the next under process variations. Node D exhibits the maximum standard deviation in the voltage drop among different fabricated sample chips. Usually, blocks with a high activity (e.g., the blocks close to node A) and those with a high performance (e.g., the blocks close to node B), both exhibit high voltage drop variations. However, from (4.8), the variations in

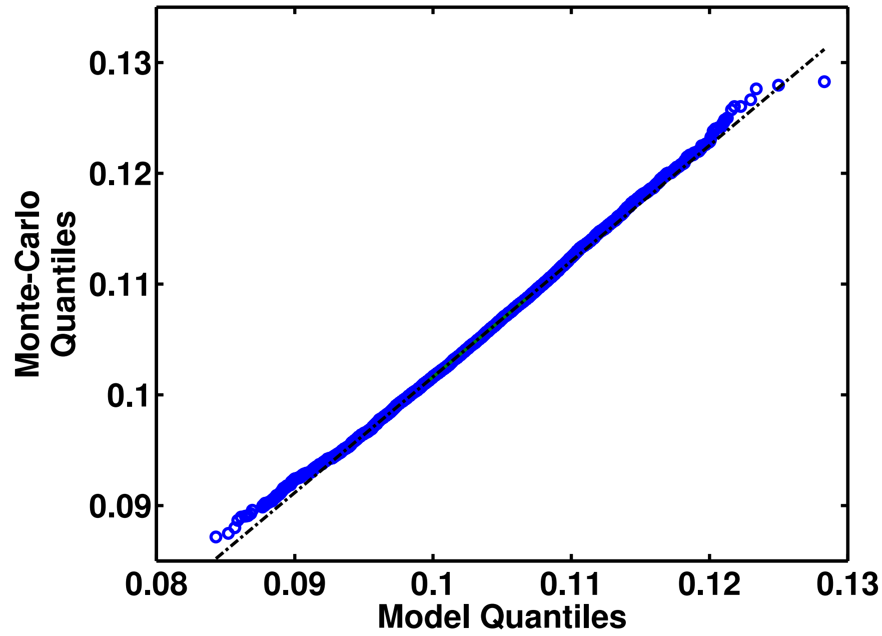


Figure 4.7: Q-Q plot of the voltage drop for the Monte-Carlo samples and that of the proposed method at the maximum expected value (point C)

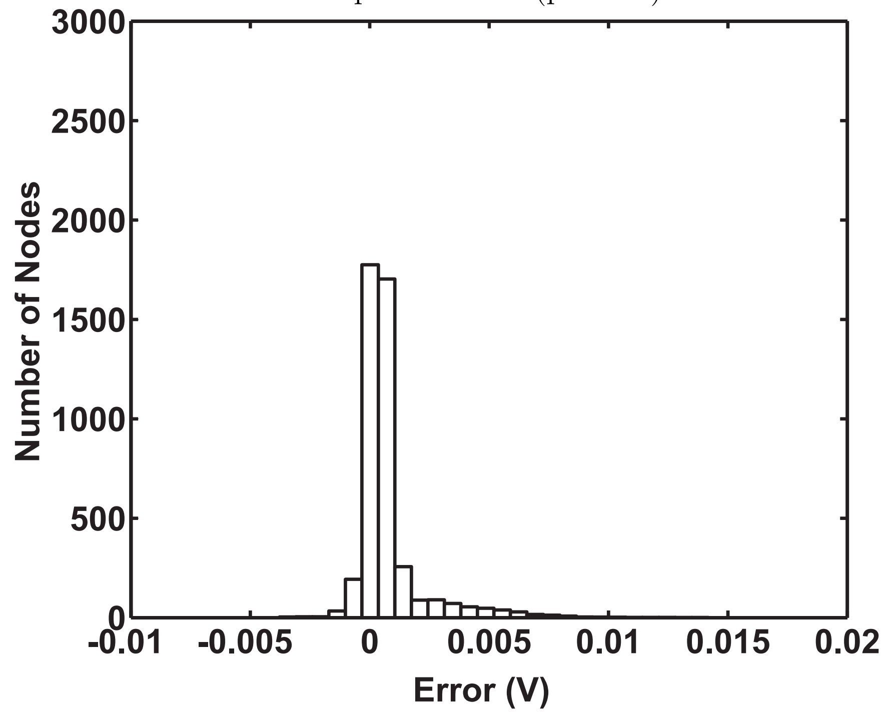


Figure 4.8: Distribution of error between the proposed methodology and Monte-Carlo simulations

Table 4.1: Runtime of the proposed methodology.

Number of Nodes	Runtime
2,026	5.8 min
4,762	6.1 min
16,642	6.8 min
43,682	7.3 min
82,370	8.3 min
128,882	10.1 min
434,282	16.8 min

the power consumption, and therefore, the current drawn off the power grid are functions of the leakage and temperature of the thermal grid. Consequently, the largest expected value and variation might not occur on the same node. This, in turn, can lead to a design failure that does not take into account the variations in temperature and power due to the process variations.

Fig. 4.6 shows the Probability Density Function (PDF) of the voltage drop at node C and D of Fig. 4.5. Also the upper bound, 99.73^{th} percentile, of the voltage drops are identified, where the nominal thermal profile is used. Here, the nominal thermal profile is extracted and utilized to estimate the statics of the voltage drop. Then 99.73^{th} percentile of the voltage drop is calculated. To have a fair comparison, the temperature-leakage loop is also considered for the nominal case. It is evident that depending on the nominal thermal profile leads to ignoring many unsafe power grid nodes with large voltage drops. In addition, note that the mean of the voltage drop at node C is relatively higher than that of the node D. However, because of the larger voltage drop variation at node D, the upper bound of the voltage drop for both nodes are very close. This figure pin points the possibility of the design failure. Without extracting the statistical thermal profile, node C may be highlighted as the worst corner/ upper bound for the voltage drop. Nevertheless, in reality, a sample chip may experience a higher voltage drop at the less anticipated point, node D. Therefore, the large variations can easily undermine any noise margin a designer sets aside to guarantee a reliable functionality.

For the verification of large power grids, the direct computation of G^{-1} takes large memory resources. Therefor, the proposed methodology, depicted in Algorithm 4.1, extracts the temperature/power statistics, utilizes the sparsity of the G matrix, and then estimates the voltage drop upper bound. Fig. 4.7 compares the quantiles of the voltage drop distribution obtained using the proposed method with that of the Monte-Carlo sim-

ulations. The comparison is illustrated for the node with the largest expected value of the voltage drop, point C. In addition, Fig. 4.8 depicts the distribution of the total error, introduced by the newly developed methodology, in respect to the Monte-Carlo simulation results. As shown in the figure, the average error is small and the maximum error is less than 1% of V_{dd}.

The available techniques for sparse matrix operations in Matlab are employed to manage the memory. Then, Algorithm 4.1 is executed for different power grids. Table 4.1 denotes the runtime of the verification methodology. Note that the low runtime of the presented method is due to the availability of statistical information. The extraction of the power statistics across the die, in the first step, eliminates the need for solving expensive Linear Problems (LP), as done in some existing work ([102],[117]). In addition to the efficiency and accuracy, unlike the corner-based verification methods, the presented methodology leads to more realistic upper bounds for the voltage drops in the presence of correlated variations in the early stage of the design.

4.7 Conclusions

In this chapter, a power grid analysis and a verification method are proposed, where the statistical behavior of power and temperature in the presence of the process variations are taken into account. Variations in the gate length and oxide thickness are mapped to the upper bounds of the voltage drop across a die, considering fluctuations in temperature-dependent power consumption. The statistics of power and temperature are extracted by modeling the inter-die and spatially correlated interadie variations. By using these statistics and the locality of the power grid problem, the voltage drop is modeled over the die. Finally, the results indicate a small error in respect to Monte-Carlo simulations. This is achieved by feeding back the voltage drop statistics for accurate computation of temperature and power statistical moments in subsequent iterations.

Chapter 5

Parametric Yield Analysis Considering Process-Induced Temperature and Supply Voltage Variations

5.1 Introduction

This chapter studies the parametric yield by taking into account the process and environmental variations. Process variations introduce uncertainty in the design parameters. The leakage power significantly varies from one manufactured chip to the next. This is due to the process-induced variations in the threshold voltage (V_{th}). In addition, the voltage of the nodes on a power grid is a function of the current drawn by the underlying circuits. Therefore, fluctuation in the leakage current leads to variations in the supply voltage (V_{dd}). Here, the role of the operating temperature across the die is two-folded: first the temperature has a strong interdependency with the leakage power. In addition, the resistivity of the power grid's wires increases as the temperature increases. Consequently, for an accurate estimation of the variations in the delay, leakage and dynamic power, the variations in the V_{th}, V_{dd}, and temperature must be taken into account.

Parametric yield is a metric that indicates what percentage of all the manufactured chips meet the design constraints. The constraints and the priority of the parameters are application-specific. In high-performance applications, the ultimate goal is to deliver a guaranteed minimum performance. Therefore, most design decisions are made to deliver such a requirement. However, in some mobile applications, both the power and performance are vital. Consequently, the design utilizes a set of power reduction techniques to maintain

the power consumption within the budget and to deliver the functionality in a timely manner. Moreover, for a hearing device, all efforts are centered on minimizing the power. As a result, for many applications, the estimation of the timing and power yield, early in the design phase, is critical. Here, the flexibility of the design in the initial stages can be employed to address any timing or power violations.

5.2 Related Work

Traditionally, the impact of variations is addressed at the device, circuit, and logic level. An optimization methodology is reported in [35] to minimize the total power by considering the process variations. The work utilizes the transistor down-sizing and multi-threshold assignment are used to reduce the power. In [118], a gate-level leakage model is extended to obtain the distribution of the leakage power. The distribution of the gate lengths for all the gates in a circuit are combined to analytically extract the mean and standard deviation of the leakage power for the circuit. In [119], a high level synthesis framework where the power yield of the function units is enhanced. The impact of the process variations is taken into account for the module selection, scheduling, and resource sharing.

Recently, due to the increase in the variability, more studies have been focused on the analysis of variations at the system level. Some of these work reviewed and discussed in [120]. Some variation reduction techniques such as adaptive body bias and variation-tolerant microarchitecture mechanisms are considered in [121]. Here, Variation-tolerant techniques for coping with the variability and estimating the power for a SOC at the system level are also described. A two-sided yield window for power and performance is discussed in [122, 123] to provide a constraint-based analysis for parametric yield. A full-chip leakage estimation is reported in [124], where some information such as the cell library, cell usage, and dimensions of the layout are applied to find the statistics of the leakage power.

More specifically on the timing yield, a static timing analysis (STA) considering statistical voltage noise is presented in [125]. The work utilizes an orthogonalization method to transform the correlated variables involved in the power noise to a subset of uncorrelated variables, and estimates the delay by using STA. To facilitate the current timing verifications that use corners and timing margins, a generic approach for determining the timing yield is suggested in [126]. Here, the design corners are selected, and the impact of the variations on the existing cell/transistor corners is examined. In some early work such as in [127], the timing analysis relies on the use of uncorrelated variations. The authors in [128], model the variations as the global sources to account for the correlation, and evaluate the impact on the timing yield. In [129], [130], Principal Component Analysis (PCA) and Independent Component Analysis (ICA) are developed, respectively, in order to simplify

the complexity of including the correlation on Statistical STA (SSTA). A nonlinear delay optimization is proposed and solved in [131] where the impact of voltage noise on the timing yield is accounted for by placing constraints on currents. A two-sided yield window is discussed in [122] to provide a constraint-based analysis for the timing yield. Here, the impact of the variations on the power and performance is investigated. The authors in [132] include a simple temperature model for their power/timing model, and discuss V_{th} and V_{dd} scaling to maximize the timing yield under variations.

Process variations impose the statistical behavior on the temperature that is missing in existing reports. It is understood that timing verification must be performed deterministically to be completely accurate. However, a statistical analysis on the timing yield early in the design process predicts the impact of the variations on the robustness of the design. Altering the floorplan can significantly change the thermal profile, resulting in timing violations.

In addition, the impact of temperature variations on power has been studied in [133], where a Monte Carlo analysis is chosen to capture the variations in the power. In [132], a simple temperature model is included in the power/timing model and maximizes the parametric yield under variations.

This part of the thesis generates a statistical thermal profile and maps its associated voltage drops across the power grid. Then the extracted statistics of temperature and voltage drop are used to accurately estimate the chip timing and power yield. This allows designers to utilize the flexibility of early design stages to prevent any possible over-design.

5.3 Statistical Profile of Temperature and Voltage Drop

In this section, the statistical models of temperature and voltage drop variations are restated. The models are discussed in detail in Section 4.3 and Section 4.4.

The strong correlation between the process variations and the statistical thermal profile was discussed in Section 4.3. It has been shown that process variations can impose significant variations in the temperature across the chip. Ignoring statistical measures of the operating temperature leads to underestimation of IR drop and consequently can cause performance degradation.

The interdependency of power and temperature is modeled to obtain the statistical measures of power/temperature in the statistical thermal analyzer, depicted in Fig. 4.4. In the first step, to consider the closed-loop effect between leakage and temperature, the following must be evaluated iteratively:

$$
\begin{aligned}
P_{dyn-j} &= b'_{1_j}\left(1 + b'_{2_j}V_j + b'_{3_j}V_j^2\right) \\
T_i^{(k+1)} &= \sum_{j=1}^{n} a_{ij}\left(p_{dyn-j} + p_{leak-j}^{(k)}\right) + a_{im}p_m \\
p_{leak-j}^{(k)} &= c'_{1_j}.\hat{I}_{leak-i}\times \\
&\left(1 + c'_{2_j}T_j^{(k)} + c'_{3_j}T_j^{(k)^2} + c'_{4_j}V_j + c'_{5_j}V_j^2 + c'_{6_j}T_j^{(k)}V_j\right) \ , \qquad (5.1)
\end{aligned}
$$

where b'_{1_j}-b'_{3_j}/c'_{1_j}-c'_{6_j} are the fitting parameters. After the nominal power/temperature are calculated for each thermal grid, the uncertainty, due to process variations, is captured in the second step. Note that the spatial correlation between the process variations in the two grids are taken into account in this step. Here, the expected value and covariance of power are extracted for all the thermal grids by evaluating

$$
\begin{aligned}
E[P_j] &= E[P_{dyn-j} + P_{leak-j}] \\
\text{Cov}(P_i, P_j) &= E[P_iP_j] - E[P_i]E[P_j] \ . \qquad (5.2)
\end{aligned}
$$

By using the properties explained in Section 4.3, all components and subsequently the moments of power in (5.2) are determined. The extracted statistic measures of power in all the thermal grids form matrices of expected values and covariance ($M_{P_{n\times1}}$, $S_{P_{n\times n}}$). Using the matrix notation, the matrices of the expected values and covariance of temperature over the die are restated as

$$
\begin{aligned}
M_{T_{n\times1}} &= A_{n\times n} \times M_{P_{n\times1}} + p_m.a_{n\times1} \\
S_{T_{n\times n}} &= A_{n\times n} \times S_{P_{n\times n}} \times A_{n\times n}^T \ , \qquad (5.3)
\end{aligned}
$$

where $A_{n\times n}$ is the first left/upper $n \times n$ sub-matrix of the inverse admittance, and $a_{n\times1}$ is the vector of the ambient temperature coefficients (a_{im}).

The extracted moments of temperature, are fed back and the moments of power are updated. The new power moments are used to obtain the temperature statistical measures. The iterative process (inner loop) is continued until the closed loop converges. By concluding this stage, the moments of temperature and power are used in the next stage for calculating the statistics of the voltage drop.

In addition, the power grid model is restated. The details of the model are elaborated on in Section 4.4. The objective, here, is to integrate the statistical thermal profile, extracted in the beginning of this section, into the power grid model. This is to accurately map process variations to the statistics of the voltage drop across the power grid.

The Modified Nodal Analysis (MNA) governs the relationship between the current and voltage of every node and its matrix format for the DC analysis is given by

$$GV = I \ , \tag{5.4}$$

where V is the vector of voltage drop, and I is the vector of currents drawn off the power grid. To solve for voltage drop, (5.4) is rewritten as

$$V = G^{-1}I \tag{5.5}$$

In addition, the resistivity of power grid wires is a linear function of temperature variations expressed by

$$R = r_0(1 + c(T_l - T_{ref})) \ , \tag{5.6}$$

where r_0, is the resistivity at the reference temperature, and c is the temperature coefficient of the resistance. Therefore (5.5) is restated as

$$V = G_0^{-1}\Phi I \tag{5.7}$$

$$\Phi = \mathrm{diag}(1 + c(T - T_{ref})) \ , \tag{5.8}$$

here G_0 is the conductance matrix at the ambient temperature. By applying the $E[.]$ and $Var[.]$ operators, as similarly done in [36], the statistical moments of the voltage drop is extracted by

$$\begin{aligned} E[V] &= G_0^{-1}E[\Phi I] \\ \mathrm{Var}(V) &= G_0^{-1(2)}\mathrm{Var}(\Phi I) \ , \end{aligned} \tag{5.9}$$

where $G_0^{-1(2)}$ denotes a matrix whose elements are the square of each element in the inverse of G_0. The moments of (ΦI) are extracted by using the properties (4.25), and (4.26). The calculated moments of the voltage drops are used in the next iteration, where dynamic and leakage power are updated and subsequently new values for temperature statistics are determined.

5.4 Timing Yield

Delivering the right functionality in an acceptable timing interval is the primary objective in designing high performance ICs. However, process variations impact the design parameters and may lead to timing violations. Therefore, a number of fabricated chips may not adhere to the designs' set objectives. The percentage of these failed chips represents the timing yield loss. An accurate estimation of the yield, specifically in the early stage of the design

gives the designer the flexibility to change small or large portions of the circuits and their parameters to maximize the yield.

The delay of the gates and interconnects depends on their operating temperature. In addition, the delays are functions of the supply voltage. With the increase in the manufacturing process variations, the fluctuations in the threshold voltage, increases where they are directly affected by the parameters such as gate length and oxide thickness. The exponential relationship between leakage current and V_{th} and the closed-loop between the leakage power and temperature makes the temperature a significant parameter in the variability-aware design. Moreover, the power delivery network consists of thousands of nodes in various metal layers connected to the underlying circuits. Since the IR drop is a function of the current drawn by the circuits, the node voltage indirectly depends on the operating temperature. Consequently, ignoring the impact of the spatially correlated process variations on the temperature undermines the accuracy of the estimated yield.

In STA, a Program Evaluation and Review Technique (PERT)-like circuit graph is employed for the delay estimation [134]. In this commonly used technique, the primary inputs and outputs are connected to a virtual source and a virtual sink, respectively. The timing graph is then traversed in a topological order from the source to the sink by using a sum or a max operation to calculate the delay of each edge and to find the longest path. Process and environmental variations, however, affect the delay of the gates and interconnects and therefore, introduce uncertainty in identifying the longest path with the largest delay. In SSTA, the problem is formulated as finding the distribution of the maximum delay of all paths from all the primary inputs to all the primary outputs. The delay of each gate or a wire segment is represented by a correlated Gaussian Random Variable (RV) [$d_k \sim N(\mu_k, \sigma_k)$].

To extract the statistics of the circuit delay, PCA is chosen to decorrelate the RVs by discretizing the circuit area [130]. The delay of a gate or an interconnect k is expressed as a function of a linear combination of the PCs,

$$d_k = \mu_k + \sum_{j=1}^{m} a_{kj} pc_j \tag{5.10}$$

μ_k is the mean of d_k, and a_{kj} are the coefficients, associated with pc_j. Because the PCs are orthogonal, the variance is obtained from

$$\sigma_{d_k}^2 = \sum_{j=1}^{m} a_{kj}^2 \tag{5.11}$$

The available thermal grids are used to extract the circuit delay statistics. Each PC, pc_j, associated with thermal grid j must capture the impact of the process variations on delay

statistics. Also, the electrothermal coupling must be taken into account. The threshold voltage is expressed as a function of the temperature variations as follows:

$$V_{th} = V_{th0} - k(\Delta T) \ , \tag{5.12}$$

where V_{th0} is the V_{th} at the ambient temperature, and k is the temperature coefficient of V_{th}. In addition, from (5.6), the interconnect delay linearly increases with the increase in the temperature. Also, the gate delay is directly related to the temperature variations through the carrier saturation velocity (ν_{sat}) that is a function of ΔT [82] such that

$$\nu_{sat} = \nu_{sat0} - \eta(\Delta T) \ , \tag{5.13}$$

where ν_{sat0} and η are the saturation velocity at the ambient temperature and the saturation velocity temperature coefficient, respectively. Moreover, the supply voltage variations are closely related to the variations in the neighboring leakage currents drawn off the power grid. Because of the interdependency of the leakage and temperature, and also the locality of power grid problem ([135]), the voltage noise is a function of ΔT. By considering the aforementioned dependencies, the delay component in thermal grid i can be stated as

$$d_i{}' = d_i + a\Delta V_{thi}(\Delta T_i) + b\Delta T_i + c\Delta V_i(\Delta T_i) \ , \tag{5.14}$$

where d_i is the delay RV, representing the decorrelated delay component due to the variations in the gate length and oxide thickness. The mean and variance of $d_i{}'$ are obtained by solving

$$\begin{aligned} \mu_{d_i{}'} &= \mu_{d_i} + a\mu_{V_{thi}} + b\mu_{T_i} + c\mu_{V_i} \\ \sigma^2_{d_i{}'} &= \sigma^2_{d_i} + a^2\sigma^2_{V_{thi}} + b^2\sigma^2_{T_i} + c^2\sigma^2_{V_i} \\ &\quad + 2[ab\text{Cov}(V_{thi}, T_i) + bc\text{Cov}(T_i, V_i) + ac\text{Cov}(V_{thi}, V_i)] \ , \end{aligned} \tag{5.15}$$

where the coefficients of a,b, and c are obtained from the sensitivity analysis, and the covariance between the parameters are easily obtained from their discussed dependencies (5.12) and (5.13).

To perform the PERT-like traversal algorithm, the sum and max operation are used to find the longest path in the graph. For the sum operation, $d_{sum} = \sum_{i=1}^{l} d_i{}'$, the mean and variance of d_{sum} are the sum of the mean and variance of $d_i{}'$. Here, l denotes the number of gates and interconnect segments in the path. For the max operation, however, the distribution of the path delay is not necessarily Gaussian. But the delay can be approximated as a Gaussian distribution, represented by an RV $[d_{max} \sim N(\mu_{max}, \sigma_{max})]$. To capture the correlation, two paths are taken at a time, such that

$$\begin{aligned} d_{max} &= max\{d_1, d_2, ..., d_{l-1}, d_l\} \\ &= max\{d_1, d_2, ..., max\{d_{l-1}, d_l\}\} \\ &= max\{d_1, d_2, ..., max\{d_{l-2}, d_{N-1,l}\}\} \\ &= max\{d_1, d_2, ..., d_{l-2,l}\} = d_{1,l} \ , \end{aligned} \tag{5.16}$$

where $max\{d_{l-1}, d_l\}$ is approximated by $d_{l-1,l}$ with a normal distribution, whose mean($\mu_{l-1,l}$) and standard deviation ($\sigma_{l-1,l}$) are extracted as follows [136]:

$$\mu_{l-1,l} = \mu_{l-1}\Phi(\beta) + \mu_l\Phi(-\beta) + \alpha\varphi(\beta) \tag{5.17}$$

and

$$\begin{aligned}\sigma_{l-1,l}^2 = &(\mu_{l-1}^2 + \sigma_{l-1}^2)\Phi(\beta) + (\mu_l^2 + \sigma_l^2)\Phi(-\beta) \\ &+ (\mu_{l-1} + \mu_l)\alpha\varphi(\beta) - \mu_{l-1,l}^2 \ ,\end{aligned} \tag{5.18}$$

where

$$\alpha = \sqrt{(\sigma_{l-1}^2 + \sigma_l^2 - 2\rho_{l-1,l}\sigma_{l-1}\sigma_l)} \tag{5.19}$$

and

$$\beta = \frac{(\mu_{l-1} - \mu_l)}{\alpha} \tag{5.20}$$

and Φ, φ are, respectively, the Cumulative Distribution Function (CDF) and the Probability Density Function (PDF) of the normal RV [$\sim N(\mu = 0, \sigma = 1)$], and $\rho_{l-1,l}$ is the correlation coefficient between d_{l-1} and d_l. After estimating the statistics of $d_{l-1,l}$, the same process is used to find the distribution of $d_{l-2,l} = max\{d_{l-2}, d_{l-1,l}\}$. However, for this, the required correlation coefficient between d_{l-2} and $d_{l-1,l}$ must, first, be calculated by [136]

$$\rho = \frac{\sigma_{l-1}\rho_{l-2,l-1}\Phi(\beta) + \sigma_l\rho_{l-2,l}\Phi(-\beta)}{\sigma_{l-1,l}} \tag{5.21}$$

By continuing the recursive process, the distribution of d_{max} can be estimated. For sequential circuits, the setup and hold time must be evaluated for possible violations. For the setup, the distribution of the maximum required arrival time at the latches is obtained by using the discussed method. For the hold time, however, an analysis of the short-paths is needed. This, in fact, is a special case of the max operation, where $d_{min} = -max(-d1, ..., -dl)$.

With the estimated distribution of the circuit delay, the timing yield is expressed as [137]

$$Y_t = P(delay < d_t) = \Phi\left(\frac{d_t - \mu_{delay}}{\sigma_{delay}}\right) \ , \tag{5.22}$$

where μ_{delay} and σ_{delay} are the mean and standard deviation of the circuit delay, and d_t is the target delay.

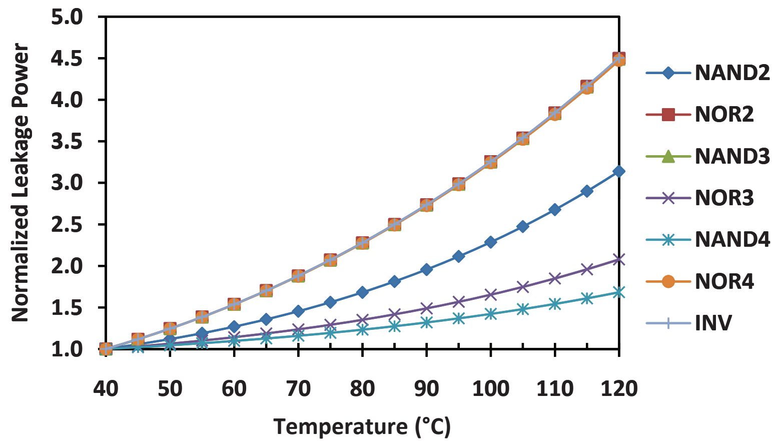


Figure 5.1: Normalized leakage power as a function of temperature for different gates

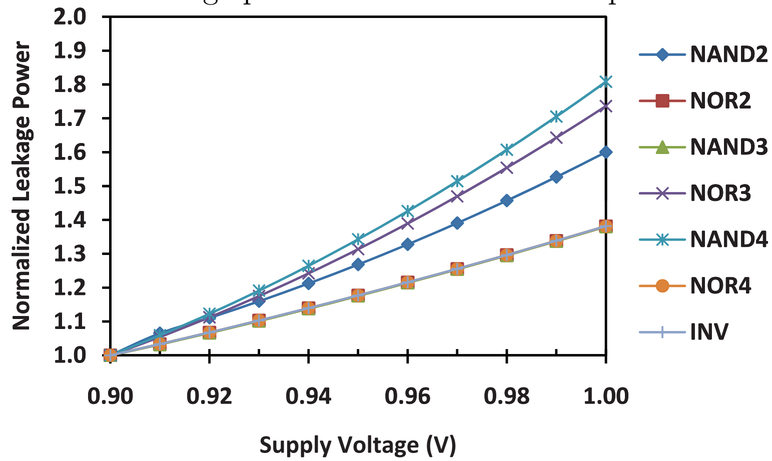


Figure 5.2: Normalized leakage power consumption as a function of supply voltage for different gates

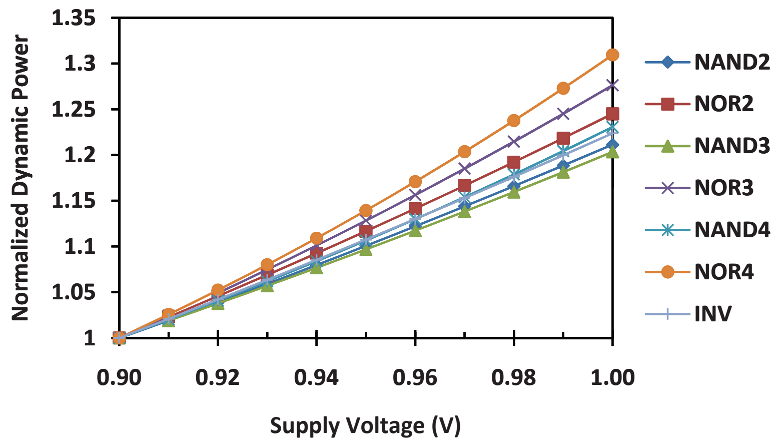


Figure 5.3: Normalized dynamic power consumption as a function of supply voltage for different gates

5.4.1 Simulation Results and Discussion

The proposed method is implemented in C++ and MATLAB and executed on a 3.4 GHz Pentium-4 PC with 2GB RAM. The experiments are performed on edge-triggered ISCAS89 benchmarks by using 65nm technology parameters. The process parameters are adapted from [115]. The package is assumed to have a 50μm thermal interface at the top of a die, whose thickness is 300μm. The package also consists of a $30 \times 30 \times 1$ mm heat spreader, as well as a $60 \times 60 \times 6.9$ mm heat sink. In regards to the power grid, the metal layers, pitch, and width per layer are selected according to the IBM benchmarks [116], scaled for 65nm technology. A power grid with 4762 nodes is used and a uniform distribution of C4s is assumed where $V_{dd} = 1.0V$. 2 iterations for the outer loop and 3 iterations for the inner loop (temperature-leakage closed loop) are enough to reduce the average error in expected value of the voltage drops to less than 0.1%.

The circuits are initially placed by using Capo [138], and a global routing is performed for all the nets. Then the die area is discretized into a number of grids, used for both the delay calculation and extracting the temperature, and the voltage drop statistics. The number of grids are selected according to the size of the circuits. The power of each grid is calculated, based on the placement of the gates and interconnects and their power consumption. Subsequently, statistical thermal profile and the statistics of the voltage drop in each grid are extracted. Here, based on the simulation results for the gates, shown in Fig. 5.1, 5.2, and 5.3, lookup tables are constructed for both dynamic and leakage power. To be more accurate, different fitting factors are used for different gates. This is to update dynamic and leakage power at various channel lengths, oxide thicknesses, temperatures, and voltage drops in each thermal grid.

Fig. 5.4(a) and 5.4(b) illustrate the expected value and standard deviation of the temperature across the s38584 circuit, respectively. In addition, Fig. 5.4(c) and 5.4(d) exhibit the profile of the expected value and the standard deviation of the voltage drop for the circuit, respectively. Here, point A denotes a location on a sample die that, on average, exhibits the largest temperature. Point B represents a location on the die that shows the largest temperature variations, comparing all manufactured dies. Similarly, C1 is a point on the die where the expected value of the voltage drop is maximal and D1 pinpoints a location on the die with maximum voltage drop variations. In addition, C2 and D2 denote the nodes on the power grid with ideal supply voltage.

Fig. 5.4 depicts how the temperature and voltage drop changes, for the s38584 circuit, from one fabricated chip to the next under process variations. For example, in all fabricated chips, the upper left area of the circuit experiences the maximum changes in the temperature and voltage drop. Note that, here, for both the temperature and voltage drop profiles, the position of the hot spots for the expected value and standard deviation are close to each other. However, this might not be the case for a die with multiple microar-

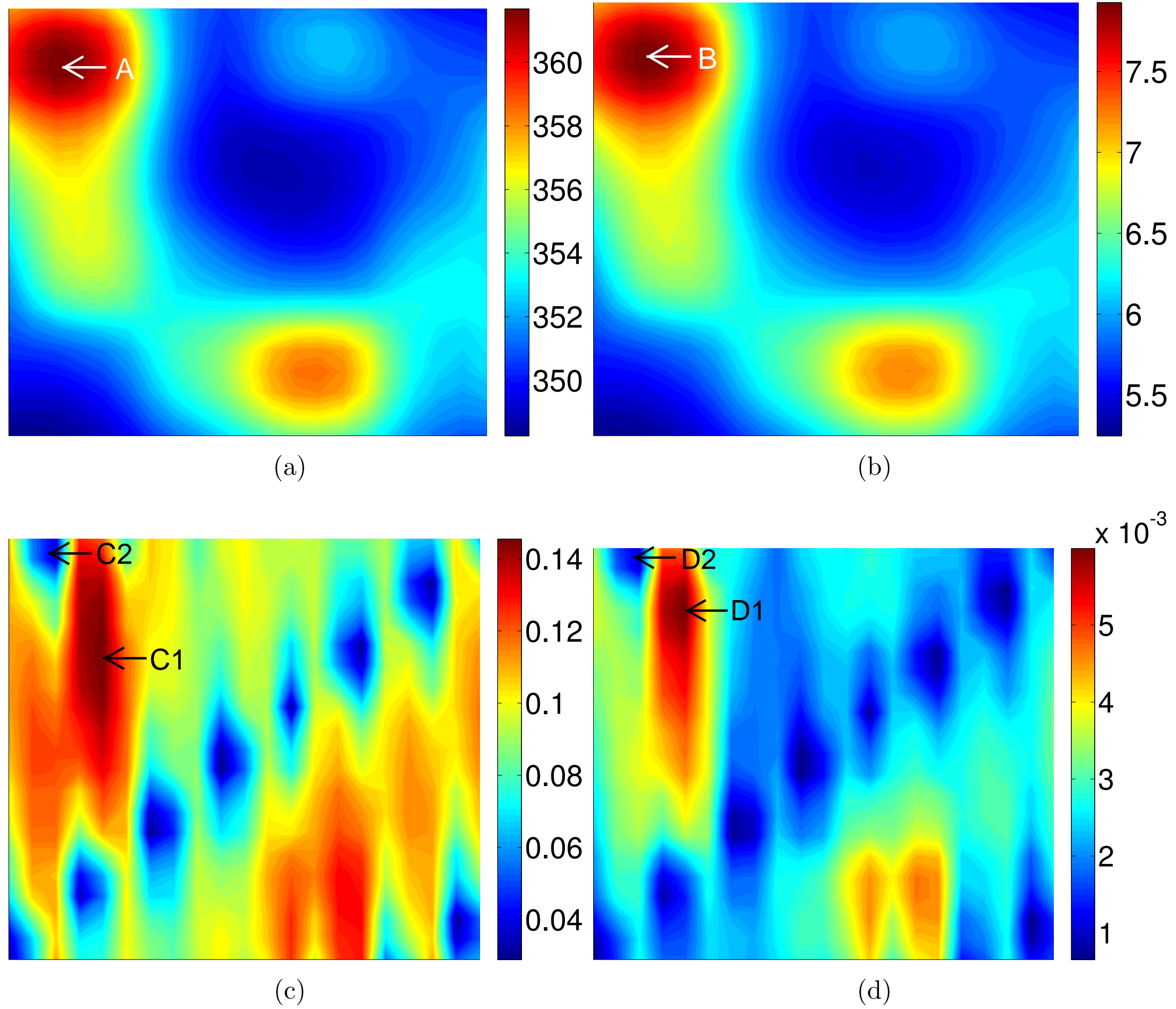


Figure 5.4: statistical profiles for circuit s38584: (a) the profile of temperature expected value, (b) temperature standard deviation, (c) voltage drop expected value (d) voltage drop standard deviation

chitectural blocks. Usually, the blocks with both high activities and those blocks with high performance both exhibit high temperature and voltage drop variations. However, from (5.2), the variations in the power consumption, and therefore, the current drawn off the power grid are functions of the leakage and temperature of the thermal grid. Consequently, the largest expected value and variation might not occur on the same grid. This, in turn, can impact the statistics of the delay and lead to the failure of a design that does not take into account the variations in the temperature and power under process variations.

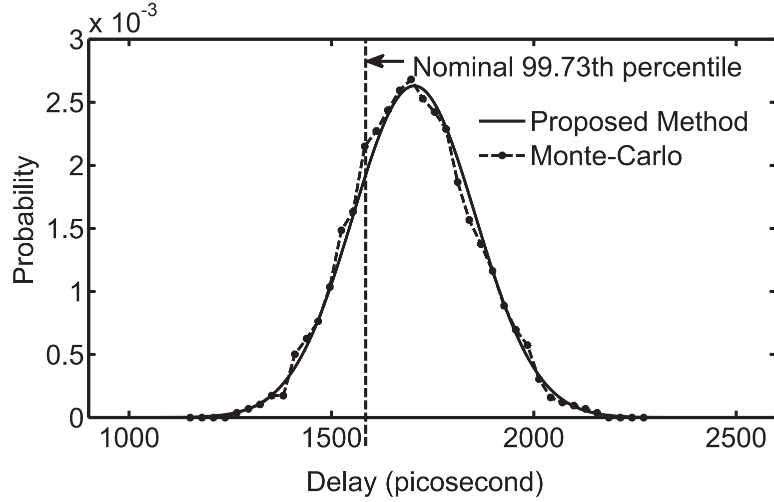


Figure 5.5: Probability density function of delay for circuit s38584

Table 5.1: Timing yield and statistics of delay, comparing different methods.

Circuit		Our Approach				SSTA		Target Delay (picosecond)	Yield (%)
		Model		Monte-Carlo					
Name	Number of Cells	Mean (picosecond)	Std (picosecond)	Mean (picosecond)	Std (picosecond)	Mean (picosecond)	Std (picosecond)		
s27	13	107.1	8.8	105.5	8.6	89.5	8.8	115.8	83.8
s1196	547	501.4	48.1	499.4	47.9	435.1	41.7	560.2	88.9
s5378	2958	444.8	40.0	443.1	39.6	413.1	38.4	528.3	98.1
s9234	5825	717.0	73.5	715.9	73.2	600.9	56.3	769.7	76.3
s13207	8260	1271.0	135.0	1270.4	133.8	1073.1	106.5	1392.4	81.6
s15850	10369	1451.7	143.9	1446.7	143.1	1212.4	116.2	1561.1	77.6
s35932	17793	1136.7	96.9	1135.5	95.5	1074.7	95.9	1362.4	99.0
s38417	23815	1009.7	92.6	1005.9	90.6	874.0	87.9	1137.7	91.7
s38584	20705	1704.6	151.7	1701.1	149.9	1530.8	143.1	1960.0	95.4

Finally, the statistical measures of the temperature and voltage drop are taken into account to find the expected value and standard deviation of the circuit delay. Fig. 5.5 is a comparison of the probability density function of the delay, obtained by the proposed method. Also the PDF of the delay, estimated from 10,000 Monte-Carlo simulation runs, is shown in the figure. Moreover the upper bound, 99.73^{th} percentile, of the delay is identified, where the nominal thermal profile is used. Here, the nominal thermal profile is extracted and utilized to estimate the statics of the voltage drop. Then 99.73^{th} percentile of the delay is calculated. To have a fair comparison, the temperature-leakage loop is also considered for the nominal case. It is evident that depending on the nominal thermal profile leads to significant delay underestimation.

Table 5.1 summarizes the delay statistics of the benchmark circuits of the proposed

Table 5.2: The error in the delay calculation when the statistical thermal profile is ignored.

Circuit	99.73^{th} Percentile of Delay (ps)		Error (%)
	Statistical Thermal Profile	Nominal Thermal Profile [45]	
s27	131.3	124.0	5.9
s1196	643.1	585.8	9.8
s5378	561.9	478.5	17.4
s9234	935.5	867.2	7.9
s13207	1671.8	1474.9	13.4
s15850	1876.0	1661.2	12.9
s35932	1422.0	1349.2	5.4
s38417	1277.7	1159.4	10.2
s38584	2150.8	1980.3	8.6

method, obtained from the model and the Monte-Carlo simulations. To calculate the statistics of the delay by the Monte-Carlo simulations, the 10,000 samples are generated, where channel lengths, oxide thicknesses, temperatures, and voltage drops, is selected randomly according to their distributions. Then the delay of the circuit is computed for each sample. The average and standard deviation of the samples' delay are compared with respective values that are calculated by the Statistical Static Timing Analyzer (SSTA) of [129]. It is seen that by ignoring the impact of the statistical thermal profile, and as a result, the inaccurate estimation of the expected values and standard deviations of the voltage drop, results a significant underestimation of both the mean and sigma of the circuit delay. The timing yields, listed in Table 5.1, are estimated in relation to the target delay and the new statistics of the delay. Assume that a designer puts a constraint on the delay, and uses the SSTA upper bound ($\mu + 3\sigma$) as the target delay. In such a case, the increase in the delay mean and sigma values results in a yield loss. In the case of the s9234 circuit, almost 23.7% of the designs do not meet the maximum delay constraint, and fail. Table 5.2 compares the upper bound of the circuits' delay, 99.73^{th} percentile, utilizing the nominal and statistical thermal profiles [45]. It is evident that a large error occur in the estimation of timing yield by ignoring the statistical measures of the temperature and the respective impact on the voltage drop.

A number of factors impact the yield, including the number and placement of the critical paths, location of the temperature hot spots, and the location of the power grid C4s. These parameters affect the sensitivity of the yield to the variations in the temperature and voltage drop. Fig. 5.6 demonstrates the sensitivity of the yield as a function of the temperature coefficient. Note that the coefficients are inversely normalized. By reducing the temperature coefficient, the increase of the timing yield varies significantly for the three

Table 5.3: Runtime of the extraction of the delay statistics.

Circuit		Runtime		
Name	Grid Count	Extracting Temperature and Voltage Statistics (second)	Estimation of Delay Statistics (second)	Total (second)
s27	4	364.0	0.1	364.1
s1196	16	365.0	4.6	369.6
s5378	64	366.0	32.3	398.3
s9234	64	366.0	71.1	437.1
s13207	100	370.6	146.1	516.7
s15850	100	370.6	164.9	535.5
s35932	100	370.6	563.8	934.3
s38417	100	370.6	386.7	757.3
s38584	100	370.6	427.5	798.1

examined circuits. Therefore accurate extraction of the thermal and voltage drop profile is crucial for the yield estimation.

The runtime of the proposed approach is provided for the benchmark circuits in Table 5.3. Increasing the grid resolution increases the accuracy of the statistical thermal profile and node voltage drops. Nevertheless, a 10×10 resolution keeps the error lower than 3% for the s38584 circuit, and the yield is estimated in less than 10 mins.

5.5 Power Yield

As discussed in this chapter, the statistical moments of the temperature and voltage drop statistics are updated iteratively. The entire process is repeated until the system reaches a steady state. To determine the distribution of the chip total power at the steady state, the following relationship between the total power and the average chip's temperature is utilized:

$$T_{avg} = T_a + R_\theta . \frac{P_{tot}}{A} , \tag{5.23}$$

where T_{avg} is the average temperature of the chip, T_a, R_θ, and A are the ambient temperature, the junction to the ambient thermal resistance, and the chip area, respectively. The moments of T_{avg} can be found from the statistics of the temperatures in all the thermal

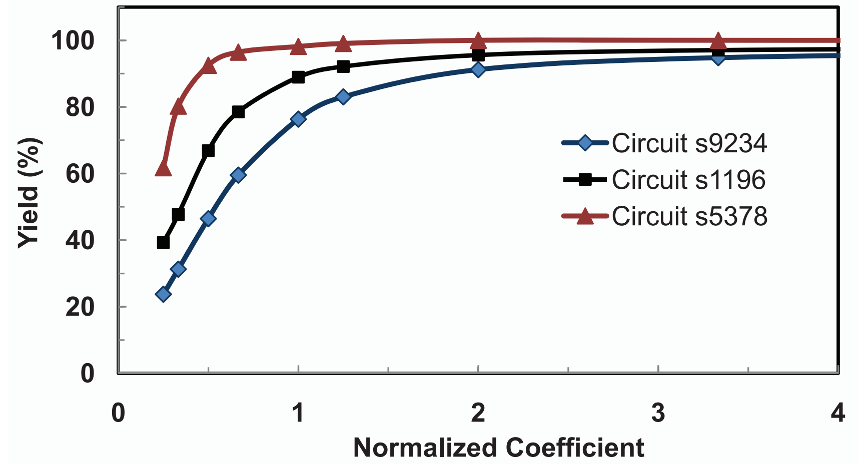


Figure 5.6: Sensitivity of the yield to the reduction of the temperature coefficients

grids by computing

$$E[T_{avg}] = \frac{\sum_{i=1}^{n} M_T(i)}{n}$$
$$\mathrm{Var}(T_{avg}) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} S_T(i,j)}{n^2} \quad . \tag{5.24}$$

From (5.23), we have

$$E[P_{tot}] = (E[T_{avg}] - T_a).\frac{A}{R_\theta}$$
$$\mathrm{Var}(P_{tot}) = \mathrm{Var}(T_{avg}).\left(\frac{A}{R_\theta}\right)^2 \tag{5.25}$$

Finally, the power yield can be found by using the moments of the chip's total power as follows [137]:

$$Y_p = P(P_{tot} < P_b) = \Phi\left(\frac{ln(P_b) - \mu_{P_{tot}}}{\sigma_{P_{tot}}}\right) \quad , \tag{5.26}$$

where P_b is the power budget that must be met, $\mu_{P_{tot}}$ and $\sigma_{P_{tot}}$ are the mean and standard deviation of a normal distribution respectively, associated with the lognormal distribution

of P_{tot}, given by

$$\mu_{P_{tot}} = \ln\left(\frac{m^2}{\sqrt{v+m^2}}\right)$$
$$\sigma_{P_{tot}} = \sqrt{\ln\left(\frac{v}{m^2}+1\right)}\ . \tag{5.27}$$

Here, the mean (m) and variance (v) of P_{tot} are obtained from (5.25).

5.5.1 Experimental Results and Design Insights

As similarly done for the timing yield analysis, the proposed method is implemented in C++ and MATLAB, and executed on a 3.4 GHz Pentium-4 PC with 2GB RAM. The experiments are performed on edge-triggered ISCAS89 benchmarks by using 65nm technology parameters. The process parameters are adapted from [129]. The package is assumed to have a 50μm thermal interface at the top of a die, whose thickness is 300μm. Also, the package consists of a $30 \times 30 \times 1$ mm heat spreader, as well as a $60 \times 60 \times 6.9$ mm heat sink. In regards to the power grid, the metal layers, pitch, and width per layer are selected according to the IBM benchmarks [116], scaled for the 65nm technology. A power grid with 4762 nodes is chosen, and a uniform distribution of C4s is assumed, where $V_{dd} = 1.0V$. Two iterations for the outer loop, and three iterations for the inner loop (the temperature-leakage closed loop) are enough to reduce the average error in the expected value of the voltage drops to less than 0.1%. Initially, the circuits are placed by using Capo [138], and a global routing is performed for all the nets. Then, the die area is discretized into a number of grids, for the power calculation and extracting the temperature, and the voltage drop statistics. The number of grids are selected according to the size of the circuits. The power of each grid is calculated, based on the placement of the gates and interconnects and their power consumption. Subsequently, the statistical thermal profile and the statistics of the voltage drop in each grid are extracted. Here, based on the simulation results for different gates, lookup tables are constructed for both the dynamic and leakage power. To be more accurate, different fitting factors are used for different gates. This is to update the dynamic and leakage power at various channel lengths, oxide thicknesses, temperatures, and voltage drops in each thermal grid. Fig. 5.7 depicts a comparison of the probability density function of the power, obtained by the proposed method. Also the PDF of the power, estimated from 10,000 Monte-Carlo simulation runs, is shown in the figure.

The statistics of the total power consumption are depicted in Table 5.4 for the benchmark circuits by using the proposed method, obtained from the model and the Monte-Carlo simulations. The results are compared with the respective values that are calculated by the Statistical Power Analyzer (SPA). In SPA, the statistics of the total power are obtained

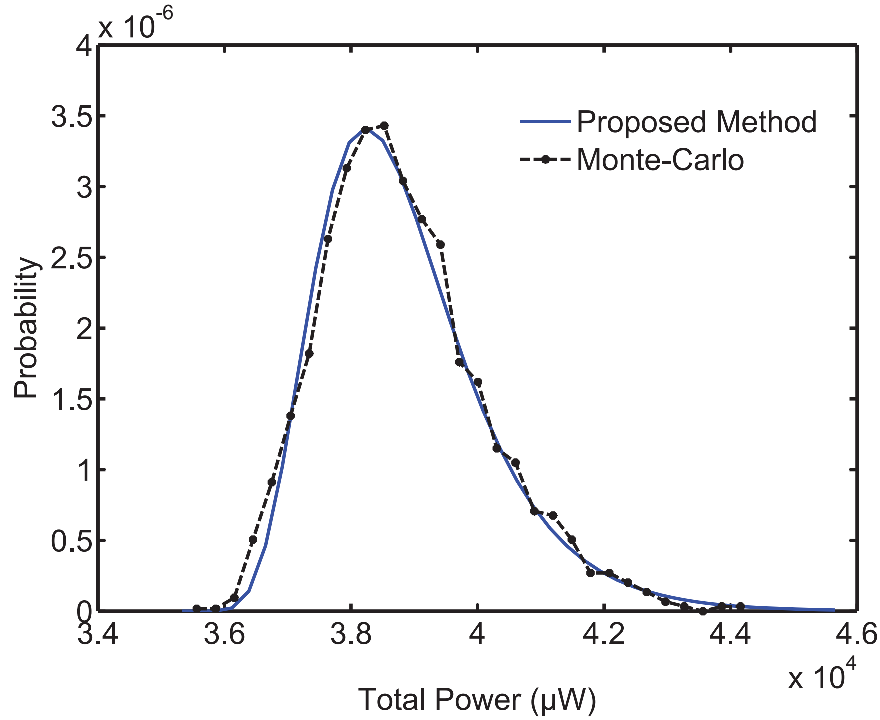


Figure 5.7: Probability density function of total power for circuit s38584.

by considering the process variations but ignoring the impact of the statistical thermal profile. The power yields, listed in Table 5.4, are estimated from the target power and the statistics of the power obtained from the presented model. Assume that a designer puts a constraint on the power, and uses the SPA upper bound ($\mu + 3\sigma$) as the target power. In such a case, the increase in the power mean and sigma values results in a yield loss. In the case of the s9234 circuit, almost 33.9% of the designs do not meet the maximum power constraint, and fail.

Table 5.5 lists the runtime of the proposed approach for the benchmark circuits. Increasing the grid resolution increases the accuracy of the statistical thermal profile, and node voltage drops. However, a 10×10 resolution keeps the error lower than 1.3% for the s38584 circuit, and the yield is estimated in less than 5 mins. The analysis presented in this chapter can be used to enhance the system architecture or to improve the power management policies.

Table 5.4: Power yield and statistics of the total power, comparing different methods.

Circuit		Our Approach				SPA		Target Power (μW)	Yield (%)
		Model		Monte-Carlo					
Name	Number of Cells	Mean (μW)	Std (μW)	Mean (μW)	Std (μW)	Mean (μW)	Std (μW)		
s27	13	32.9	2.0	32.9	2.0	29.9	1.0	32.9	50.4
s1196	547	1426.3	81.2	1438.2	83.2	1320.8	49.8	1470.2	64.4
s5378	2958	6986.3	277.8	7154.7	283.6	6795.3	189.1	7362.7	81.1
s9234	5825	8150.9	305.0	8243.5	311.7	7681.6	223.5	8352.1	67.1
s13207	8260	13574.4	659.9	13898.0	637.3	12610.5	369.2	13718.1	56.3
s15850	10369	13960.4	488.8	14196.2	514.5	13185.3	365.0	14280.4	66.9
s35932	17793	24433.2	827.9	24277.9	840.8	23546.7	656.7	25516.9	80.4
s38417	23815	33967.9	1106.1	33961.1	1109.7	32785.6	945.1	35621.0	83.5
s38584	20705	38863.2	1357.6	38975.2	1365.8	37638.5	1037.2	40750.0	81.8

Table 5.5: Runtime of the extraction of the total power statistics.

Circuit		Runtime		
Name	Grid Count	Extracting Temperature and Voltage Statistics (second)	Estimation of Power Statistics (second)	Total (second)
s27	4	105	2	107
s1196	16	112	14	126
s5378	64	114	129	243
s9234	64	114	129	243
s13207	64	126	129	255
s15850	100	126	154	280
s35932	100	126	154	280
s38417	100	126	154	280
s38584	100	126	154	280

5.6 Conclusions

A higher complexity of the statistical static timing analysis is tolerated to avoid over-design. Therefore, it is crucial to include the effective factors in the estimation of the timing yield. This work proposes a comprehensive method for the timing and power yield analysis by taking into account the statistical thermal profile and its associated voltage drop across the die. Here, the impact of the intra-die process variations and the spatial correlation is considered to extract the statistical thermal and voltage drop profile. Then, these profiles are used to estimate both timing and power yield. The results indicate that if the variation of the temperature and its related voltage drop are ignored, a significant yield loss occurs. Finally, the additional overhead, with respect to a conventional statistical

analysis, is linear in the number of grids.

Chapter 6

Design Solutions for VLSI Systems under Variability

6.1 Introduction

As discussed in the pervious chapters, there are challenges in designing robust VLSI systems. However, there exist some opportunities to be explored in order to reduce the sources of variations as well as their impact on the parametric yield. This chapter proposes two optimization mechanisms and shows that they can be effective in reducing the variations. First, floorplanning under temperature variations is studied, where the total power is minimized. Subsequently, optimum supply pad assignment is presented to maximize the timing yield.

6.2 Solution 1: Total Power Reduction in the Presence of Temperature Variations

There are several components, each with a different functionality, on a VLSI chip. These components have various power consumptions, and consequently, operate at different temperatures. Temperature variations, as high as 50 °C, can exist in a chip due to aggressive dynamic power management (DPM), clock gating, and nonuniform switching activities of the various blocks. High temperature gradients significantly impact the performance and reliability of VLSI circuits [139],[57]. Changing the floorplan of a chip creates a different heat transfer path and can be very effective in addressing the non-uniformity of the temperature distribution across a chip. In addition, the total power of a chip is a function of the wire length and the temperature distribution of the chip's floorplan.

Therefore, it is very crucial to take into account both the total power consumption and the temperature variations in the floorplanning, as proposed in this work.

6.2.1 Related Work

The related work on floorplanning are categorized into three main categories. The primary focus of the works in the first category is to compare the floorplanning algorithms in regards to their complexity and efficiency for area and wire length optimization. In [140], slicing algorithm and in [141] non-slicing floorplans are optimized for area and wire length. Also, in [142] a modified convex formulation is presented and solved to minimize the area of the floorplan. An extensive survey on floorplanning algorithms, most of which use simulated annealing, has been presented in [62]. These work do not include the important thermal or power effect in their floorplanning methodology. In the second category, the operating temperature is minimized. Architectures with different floorplans, in order to meet the performance and thermal constraints, are compared in [55] and [63]. HotFloorplan is presented in [60] to minimize the peak temperature. Lowering the peak temperature of an architecture has also been studied in [143] and [144], where a communication profile is explored. A genetic algorithm is proposed in [145] to facilitate the search for a floorplan that has a smaller area as well as a lower maximum temperature.

Although the studies in this category propose thermal-aware floorplanning solutions, power consumption is not considered in their work. Minimizing the leakage power is the focus of the third category. In [146], [65], and [66], the implication of leakage power on floorplanning in a system on chip is examined. Authors in [146] present an active subthreshold leakage reduction using task migration where the computation of hot modules are migrated to reduce leakage by reducing the chip temperature. An optimization guideline for leakage-aware floorplanning is provided in [65] and an algorithm for leakage reduction is presented in [66] by modeling the temperature-dependant leakage for a thermal profile. Although the studies in the last category include leakage power and temperature in their work, total power consumption and non-uniformity of thermal profile are missing.

The focus of this work is to simultaneously reduce total power consumption and gain from the benefits associated with a more uniform thermal profile. We consider the nonlinear relationship between subthreshold leakage and the operating temperature of the blocks of a floorplan. This work examines our intuition that temperature variations of the hot blocks have higher impact on leakage variations than temperature variations of the cold blocks. As a result, reducing the thermal variations may lead to lowering the number of high leakage blocks and consequently saving power.

- This chapter intends to provide guidelines for optimizing a floorplan in order to minimize the total power, as a primary objective, while achieving the minimum

possible temperature variations across a chip as a secondary goal. To realize these two objectives, a correlation is established between the total power consumption of a chip and the thermal uniformity of the floorplan. Subsequently, the impact of the temperature variations of the chip thermal profile on leakage and variations power is analyzed. The results show that lower thermal variations can lead to a significant savings in total power.

- In addition, it is demonstrated that despite the aforementioned correlation, the most uniform thermal profile does not necessarily correspond to a maximum power reduction. Therefore, it is illustrated that a small deviation from the minimum total power can be traded for a significant increase in the uniformity of the temperature distribution.

This work utilizes Parquet floorplanner [147], a fixed-outline/ non-slicing floorplanning tool, in its core engine where the search is performed using simulated annealing algorithm. However, the presented floorplanning methodology is not limited to these types of floorplans and other tools/algorithms can also be employed to gain from the proposed benefits.

6.2.2 Preliminaries and Understanding the Correlation between the Total Power and Temperature Variations

In this section, the thermal and power models are detailed and the correlation between the total power consumption and the temperature variations, associated with a floorplan, is investigated.

6.2.2.1 Temperature and Power Modeling

There are two types of leakage power, active leakage current accounts for the leakage power consumed during the circuit active mode and standby leakage, dissipated when the circuit is in the standby mode. Because of process variations, up to 20 times variations in the leakage power and 30% variations in performance have been reported in the literature [14]. In addition, the temperature distribution and power consumption are interdependent due to the dependency of both active and standby leakage power on temperature [146]. Therefore, this dependency needs to be modeled to accurately obtain the thermal and power profile. Of the available methods, HotSpot, the compact thermal model developed and verified by Skadron et. al. has proved to be efficient and accurate. The model uses the existing duality between the thermal and electrical phenomena to construct a thermal RC network [148]. To obtain the profile of the total power for each floorplan, the dynamic and leakage power of the floorplan are added for each block. Any deviation from the

original floorplan impacts the dynamic power through the changes in the total wire length. Therefore, the relative changes in the dynamic power need to be captured. The well-known Half Perimeter Wire Length (HPWL) is used in the literature to account for such changes in the dynamic power. Here, the wire length between two blocks is expressed as

$$WireLength = |x1 - x2| + |y1 - y2| \tag{6.1}$$

where x1, y1, x2, and y2 are the coordinates of the center of the blocks. The interconnect matrix, representing the number of interconnects between every two blocks, is then used in the floorplanning to provide the total HPWL for the generated floorplan. The leakage power of a floorplan comprises two components: gate leakage and subthreshold leakage. Gate leakage current is due to the direct tunneling of electrons through the gate oxide which is independent of temperature [149]. Therefore, the gate leakage is insensitive to temperature. However, subthreshold leakage power has a supperlinear interdependency with the operating temperature. To estimate the total leakage for a generated floorplan, initially, the leakage of the original floorplan is obtained at room temperature. Subsequently, the subthreshold leakage power of each block is updated according to the new temperature of the block. The subthreshold leakage of each block can be modeled by the following superlinear function, as similarly done in [88]

$$P_{Lj} = K_j A_j \{c_1(T_j - T_a)^3 + c_2(T_j - T_a)^2 + c_3(T_j - T_a) + c_4\} \tag{6.2}$$

where K_j is the block leakage factor, A_j is the block area, T_j is the block average temperature, and c1-c4 are the fitting constants obtained for each block. These constants are calculated by plugging some known leakage and temperature numbers in (6.2).

6.2.2.2 The Correlation between the Total Power and Temperature Variations

Thousands of generated floorplans, represented by dots, in Fig. 6.1 illustrate the correlation between the leakage power and the temperature standard deviation for the Alpha 21264 processor, scaled down to the 90-nm technology running gcc application. The increase in the leakage is normalized in respect to the minimum leakage power. Every floorplan has different temperature and power statistics. Note that the variation in temperature, here, is represented by the standard deviation. As seen in Fig. 6.1, when the temperature variations increase, the probability of attaining a higher leakage power also increases. This is due to the nonlinear dependency of the subthreshold leakage power on temperature. For a floorplan with high temperature variations, the probability of having hot blocks is higher. Consequently, the probability of having blocks that leak more increases with the temperature variations. In addition, dynamic power randomly fluctuates with temperature variations. As a result, as demonstrated in Fig. 6.2, there is a correlation between the total power and the temperature variations. Note that this correlation is even more pronounced

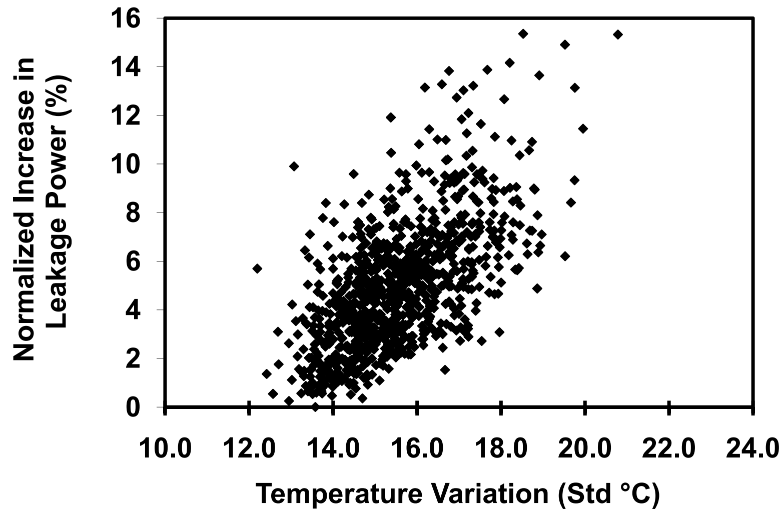


Figure 6.1: Normalized increase in the leakage power in respect to the increase in temperature standard deviations, for the Alpha processor running gcc.

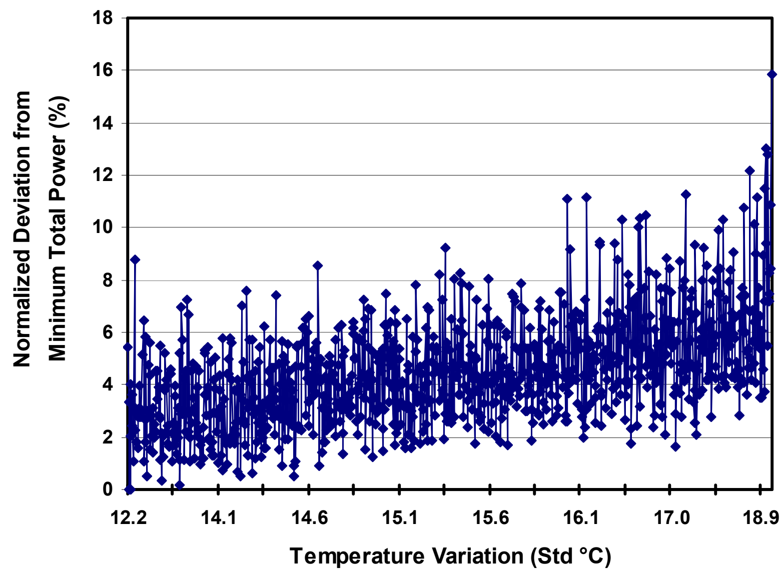


Figure 6.2: Normalized deviation from the minimum total power (power increase) as a function of temperature variations.

for high thermal variations. This is due to the fact that the spread of the leakage power increases with the spread in temperature distribution.

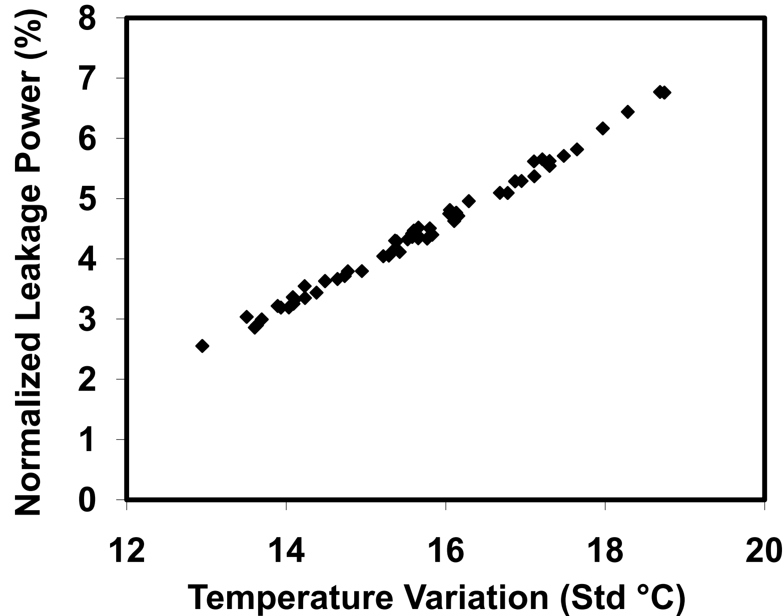


Figure 6.3: Normalized deviation from the minimum leakage power as a function of the temperature variations at constant average temperature of 83 °C.

6.2.3 Proposed Floorplanning Optimization

6.2.3.1 Objective Function

In this research, the objective is to not only minimize the total power, but also to address the thermal integrity by using an efficient and accurate approach. The correlation between the leakage power and temperature variations, found in the previous section, is used to efficiently optimize the floorplan for minimizing the total power and taking into account the thermal integrity. Both the average temperature and temperature variations must be minimized to ensure that the leakage power is also at its minimum. Also, the HPWL component should be included in the objective function to minimize the dynamic power. The leakage power substantially depends on the average temperature of the thermal profile. Variations around an average temperature, as shown in Fig.6.3, lead to variations in the leakage power. As a result, ideally, a floorplanner should minimize the average temperature, temperature variations, and HPWL not only to minimize the total power, but also to increase the thermal uniformity. Therefore, the weighted sum of these components is minimized. The objective function of this work is mathematically expressed as

$$
\begin{aligned}
& min \;\; \{\alpha T_{avg} + \beta \sigma_T + \gamma HPWL\} \\
& subject\ to: \\
& \quad a^i = a^i_{orig} \\
& \quad A \leq A_{orig} \\
& \quad WS \leq WS_{max} \\
& \quad AR = C_{AR}
\end{aligned}
\tag{6.3}
$$

Here, σ_T is the standard deviation of the blocks temperature, and α, β, and γ are the weighting factors. As discussed in the next section, the weighting factors depend on the priority of the objectives that are to be realized. Minimizing this objective function with different weighting factors can lead to achieving various objectives. However, the advocated objective, in this work, is to primarily minimize the total power and as the secondary objective have the maximum possible thermal uniformity across the chip. Since the blocks are "soft blocks", the area of each block (a^i) in the floorplan is the same as that of the original floorplan. However, the height and width of each block can change. In addition, for this fixed-die floorplanning, the total area of the floorplan (A) is constrained to be less than that of the original Alpha floorplan. Since, the increase of whitespace reduces the average temperature, it impacts the thermal resistance of the chip. To establish a fair comparison between the newly attained results and the ones in the literature, the total whitespace in the floorplan is also constrained to a given value (WS_{max}). Finally, the aspect ratio (AR) of the core of the processor takes a given constant value (C_{AR}).

6.2.3.2 The Optimization Methodology

Fig. 6.4 depicts the optimization methodology. In Step 1, the initial values for dynamic power and leakage power are obtained for the given micro-architectural blocks. In the initial moves to have a better starting point, the maximum die temperature is estimated based on the heat diffusion of the hottest blocks to their neighboring blocks. For that, initially, a one dimensional heat transfer equation, $T = Ta + \theta P_{total}$, is utilized in Step 1 to obtain the primary thermal profiles where P is the block's total power and θ is the die to ambient thermal resistance. The reason for using a one dimensional model at this stage is just to find a starting-point floorplan. In Step 2, it does not need high accuracy to identify and start with a floorplan that has a low temperature variation. Initially, in Step 2 the heat diffusion is integrated in the floorplanner's objective function as $Obj = \gamma HPWL - \lambda D$ where γ and λ are the weighing factors for the wire length and heat diffusion. Also $D = \sum shared_length(T1 - T2)$, the total heat diffusion of the die, is

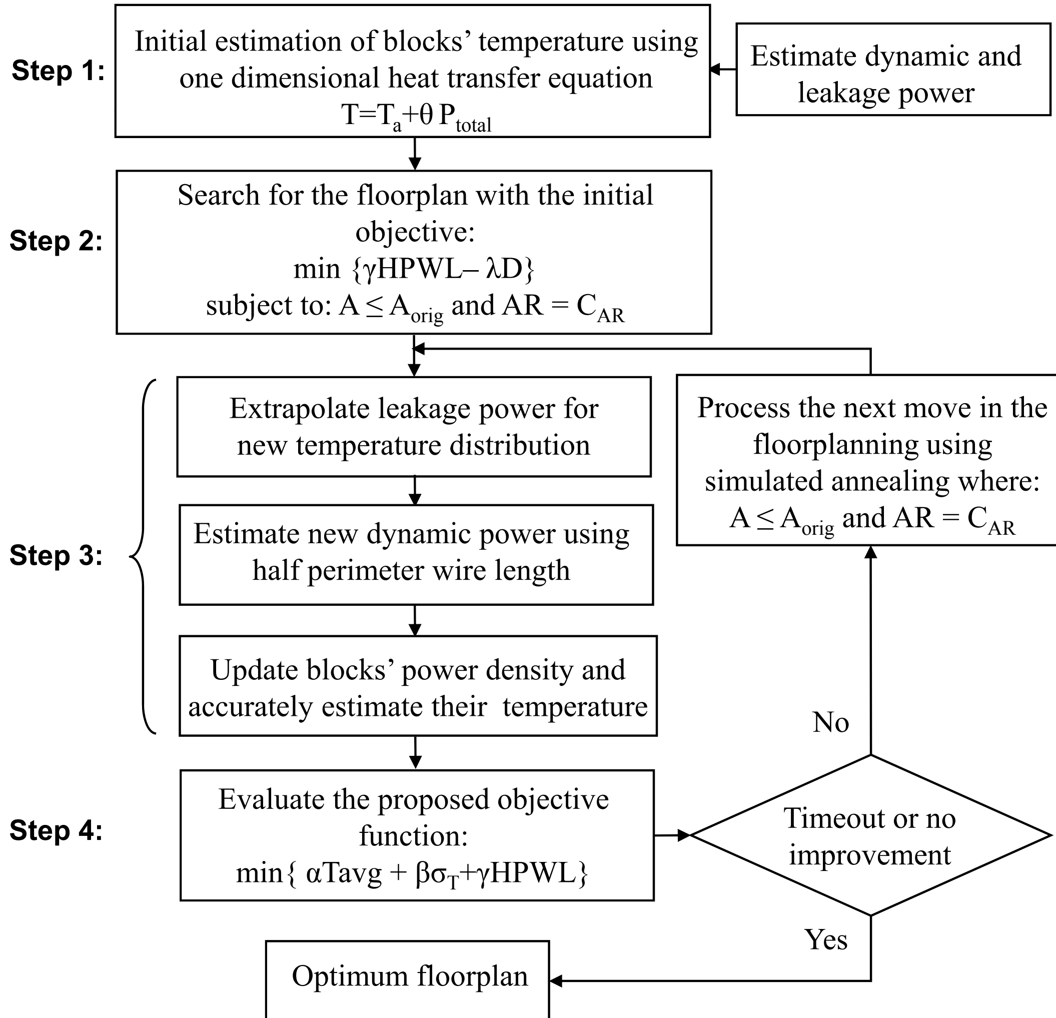


Figure 6.4: Optimization methodology using the correlation found between the temperature variations and total power of a floorplan.

estimated by using the temperature of the blocks and the length of the edges they share [150]. The negative sign of the heat diffusion component in the objective function is to maximize the diffusion and therefore to have a lower maximum temperature. Subsequently, more accurate temperature estimation is employed in Step 3, where leakage numbers are updated for all the blocks according to equation 6.2. In this step, the closed loop between leakage power and temperature is taken into account to find the steady state temperature. The dynamic power is also updated based on the new HPWL. The leakage and dynamic power are added to find the power density and hence extract the new thermal profile. It is more reliable to perform this step in a self-consistent manner where the leakage power is

updated iteratively and thermal runaway is declared if the temperature does not converge i.e. no steady-state temperature is available. Note that the statistics of temperature, T_{avg} and σ_T, are obtained for each generated floorplan by using the temperature of each block as one data sample. Here, the weighted mean and standard deviation of the floorplan is obtained by giving a weighting factor to the temperature of every block. The weight is determined based on the area of the block. In Step 4, in order to obtain the optimum floorplan, (6.3) is evaluated and the next moves are suggested until a timeout is reached or no improvement in the reduction of the objective function is observed. Note that if one is to optimize the floorplan only for power consumption, the total power has to be directly minimized in the cost function. In addition, for high efficiency, the extraction of the thermal profile and the evaluation of the respective cost function have to be moved to the simulated annealing engine.

6.2.4 Results and Discussion

To set up the optimization engine, the thermal profile, dynamic power, and leakage power are estimated by using HotSpot, Wattch [151], and the extended HotLeakage [152], respectively. Also, Parquet is used as the core engine of the floorplanner , and the blocks are set to be "soft blocks" where the area of a block is fixed, but its aspect ratio can change. Wattch is a popular architecture-level power simulator that estimates the switching power for a given application. The tool takes into account the activity of each block, and how many times they are accessed in a given time interval. When executing a given application and based on the estimation of the unit capacitances for the block, the dynamic power is estimated. Before executing Wattch, the system must be configured for the given processor. The extended version of HotLeakage is employed to estimate the gate and subthreshold leakage for all microarchitecture structures. Note that, for Alpha processor, the ev6-like floorplan and for the MCNC the same benchmarks are used for all the applications. The optimization is performed on an Intel® 3.4 GHz CPU with 2GB RAM. Fig. 6.5 signifies the final floorplan with a minimum total power and the lowest possible temperature variations for the core of the Alpha processor running the gcc application. To compare the temperature statistics, the temperature of each block in a given floorplan is taken as a data sample. Then using the weighted mean and standard deviation, the statistics of temperature for the entire thermal profile is obtained [153]. Here, the block area is used as the weight of the block.

Various objectives in floorplanning have been targeted in the literature. Fig. 6.6 offers a comparison of the normalized deviation from the minimum total power for the objectives of the existing work (minimum T_{max}, minimum leakage, and minimum HPWL) with those of this research. The provided data for the existing work has been obtained by regenerating their work using the proposed model. This is done by modifying the cost function to

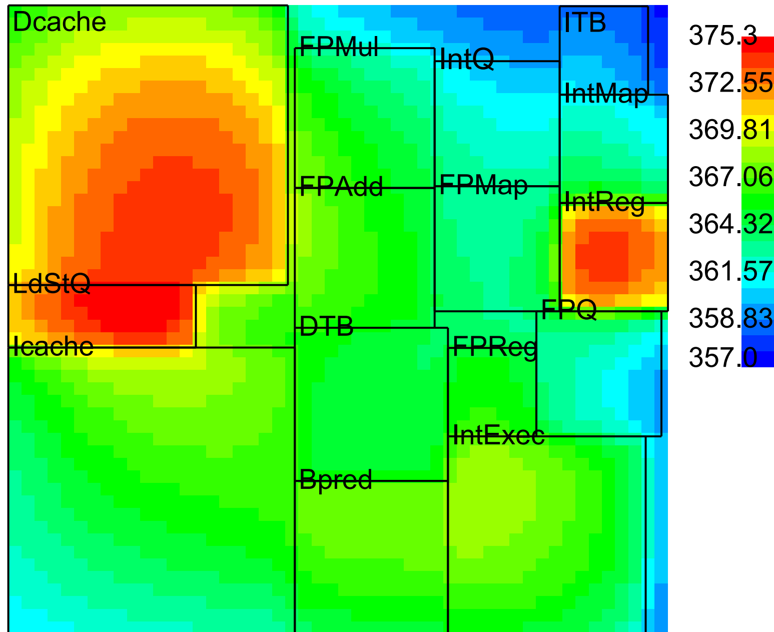


Figure 6.5: Optimum floorplan with minimum total power and the lowest possible temperature variations of the core of an Alpha processor.

replicate their models. The increase in the total power is also shown, when the maximum thermal uniformity is targeted. As Fig. 6.6 shows, the objectives of the existing work can dramatically increase the total power. For example, when the processor executes the gcc program, if the target is to minimize the maximum temperature, the total power is 9% more than the minimum value. For the swim program, minimizing the leakage power as the only objective can result in almost a 10% increase in the total power.

As seen in Fig. 6.2, despite reducing the probability of having a high leakage, the most uniform thermal profile does not necessarily yield the highest total power savings. This is due to the longer wire length, caused by moving the hot blocks away from each other to increase the uniformity in the thermal profile. In order to minimize the total power, these two counter-effects must be balanced. As seen in Fig. 6.6, the deviation from the minimum total power for "Proposed Objective Function" is small for all programs.

Table 6.1 depicts the runtime of the optimization process for Alpha processor and MCNC benchmarks. Note that the majority of the execution time is spent on running HotSpot as an external tool for more accurate temperature estimation. Thus, temperature estimation, internally, in the core engine will improve the runtime.

The secondary objective is to minimize the temperature variations. This is to benefit from the gains that a more uniform thermal profile provides such as a better thermal

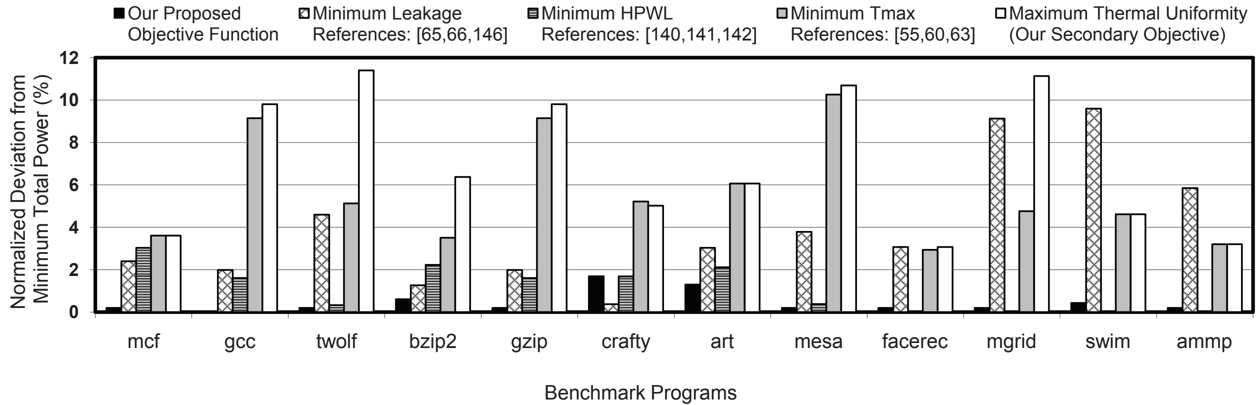


Figure 6.6: Normalized deviation from the minimum total power for the objectives of the existing work in the literature and those of this work.

Table 6.1: Runtime of the optimization process for Alpha and MCNC benchmarks.

Circuit	Block Count	Net Count	Runtime (s)
Alpha	15	32	397
ami33	33	123	490
ami49	49	397	662
apte	9	94	115
hp	11	65	167
xerox	10	181	144

reliability and avoiding performance degradation due to the lower number of hotspots. Fig. 6.7 conveys that by allowing only a 2% increase in the total power, as much as a 25% increase in the thermal uniformity can be achieved. For example, when mcf is the target application, if one optimizes the floorplan to have minimum total power, the thermal uniformity will deviate by 15% from its maximum. The same floorplan can be optimized, using a higher value for β, where the deviation from the maximum thermal uniformity drops to just 2.2%. This uniformity improvement is achieved by allowing only 2% deviation from the minimum total power. For many applications, such as high performance designs where the power budget is not a strict constraint, relaxing the power constraint by 2% is acceptable.

The increase in the thermal uniformity not only enhances the performance and reli-

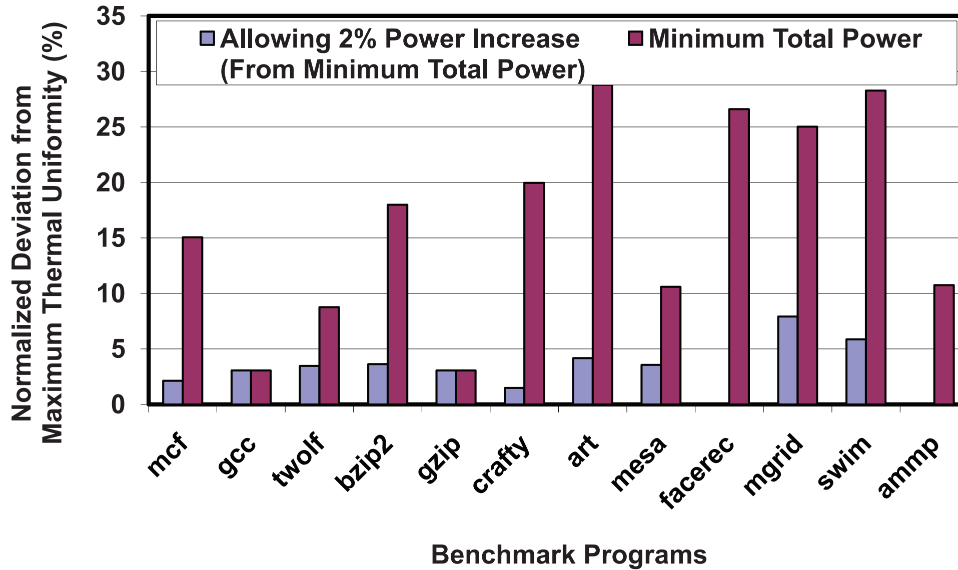


Figure 6.7: Temperature variations of two cases: when the minimum total power is the objective of the floorplanning and the case where a 2% power increase is traded for lower temperature variations.

ability, but also is financially beneficial. This is because a more uniform thermal profile reduces the number of hotspots and eliminates the needs for an expensive cooling systems or a sophisticated Dynamic Thermal Management (DTM) [154].

By selecting the appropriate weighting factors, a different objective can be achieved. In high-performance applications, attaining the targeted performance is the main requirement and most design decisions are made to deliver this performance. In such cases, hotspots are easily formed because some architectural blocks exhibit high activity and large power density. These hotspots impact both the performance of the devices as well as the delay of the interconnects [139], [57]. Assigning a larger weight to β increases the thermal uniformity and reduces the number of hotspots. On the other hand, for a hearing device application, where the power budget is limited, the weight factors are balanced to minimize the total power. When the objective is to guarantee the reliability of a chip, selecting a large value for β addresses two reliability concerns, thermal runaway and electromigration. The hotspots are susceptible to the possibility of a thermal break down when the close loop between the temperature and leakage power does not converge. In addition the 10 years time to failure, due to electromigration, exponentially depends on the operating temperature [155]. Therefore, in order not to violate this requirement the thermal uniformity must be increased accordingly.

Table 6.2 shows how different objectives can be achieved by utilizing the proposed

Table 6.2: Achieving various objectives for four different leakage to total power ratios by selecting the appropriate weighting factors.

Objective	Benchmark Programs	Average Pleak/Ptotal	α	β	γ	Area (mm²)	Wire Length (m)	Deviation from Minimum Leakage (%)	Tmax (°C)	Deviation from Minimum Total Power (%)	Thermal Uniformity (Std °C)
Minimum Leakage	gcc	0.50	0.80	0.20	0.00	250.92	22.00	0.00	105	1.99	13.58
	art	0.41	0.80	0.20	0.00	248.67	22.38	0.00	99	3.04	16.78
	ammp	0.33	0.80	0.20	0.00	241.07	23.19	0.00	95	5.88	13.60
	swim	0.19	0.80	0.20	0.00	256.41	23.32	0.00	67	9.80	4.78
Minimum HPWL	gcc	0.50	0.30	0.30	0.40	248.67	16.50	6.18	109	1.62	14.58
	art	0.41	0.20	0.20	0.60	253.18	16.50	9.94	105	2.12	20.92
	ammp	0.33	0.30	0.30	0.40	245.40	16.50	4.07	99	0.05	13.40
	swim	0.19	0.30	0.30	0.40	244.62	16.50	1.26	68	0.10	5.04
Minimum Tmax	gcc	0.50	0.10	0.90	0.00	251.65	24.24	10.25	99	9.40	12.20
	art	0.41	0.10	0.90	0.00	242.70	24.90	1.24	93	6.07	14.78
	ammp	0.33	0.10	0.90	0.00	256.41	18.74	5.67	88	3.20	12.10
	swim	0.19	0.10	0.90	0.00	246.46	19.55	2.02	64	4.81	3.93
Maximum Thermal Uniformity	gcc	0.50	0.00	1.00	0.00	251.65	24.24	10.25	99	9.80	12.20
	art	0.41	0.00	1.00	0.00	242.70	24.90	1.24	94	6.07	14.78
	ammp	0.33	0.00	1.00	0.00	256.41	18.74	5.67	87	3.20	12.10
	swim	0.19	0.00	1.00	0.00	246.46	19.55	2.02	64	4.81	3.93
Minimum Total Power & Maxim-um Possible Uniformity	gcc	0.50	0.36	0.36	0.28	250.91	18.53	0.97	103	0.00	12.57
	art	0.41	0.36	0.36	0.28	253.18	18.40	4.41	98	1.30	16.24
	ammp	0.33	0.36	0.36	0.28	245.40	16.50	4.07	94	0.20	13.40
	swim	0.19	0.36	0.36	0.28	244.62	16.70	0.99	67	0.43	4.80

methodology. Four cases, for four applications, with different leakage to total power ratios (for the original Alpha floorplan at room temperature) are illustrated where appropriate weighting factors are set to optimize the floorplanning in according to the set objectives. For example, assume a gcc application is executed where average leakage to total power is 0.5 for the original floorplan. To optimize the floorplan for minimum leakage using the proposed methodology, the weighting factors must be selected as follows: $\alpha = 0.8$, $\beta = 0.2$, and $\gamma = 0$. The large value for α is because of the high dependency of leakage power on the average temperature. However, to optimize for minimum HPWL, a larger weight must be given to γ. note that For some objectives such as HPWL, selecting a very high value, for the respective weighting factor, may not be necessary. For the above example, $\gamma > 0.4$ minimizes HPWL. Consequently, in such cases, giving a reasonable weight to other factors leads to relatively realizing other objectives. For minimum T_{max} and maximum thermal uniformity, a larger value is assigned to β (0.9 and 1 respectively for the gcc case). And finally, for the presented objective function, the weight is close for the three factors. This is to achieve a minimum total power and maximum possible thermal uniformity. Also note that these factors are usually set equally for all the benchmarks. But, the setting may be application specific.

It is noteworthy to know that blindly following the maximum thermal uniformity objec-

tive is not beneficial and also may lead to unnecessary power increase. When the objective is to minimize Tmax, the designer gives large values to β for reducing the temperature variation. But the designer does not forego the reduction of the average temperature by applying a non-zero value to α. On the other hand, the objective of absolute possible uniformity in the thermal profile enforces assigning maximum value (1.00) to β. This one-sided optimization is not wise even though it guarantees the lowest number of hotspots.

Meeting the minimum performance constraint is crucial for many applications. To accurately account for performance, it is necessary to model the changes in the delay due to changes in the original floorplan. The use of the sum of weighted latencies is one way to address performance in floorplanning [156]. This is where a weighting factor is assigned to the critical busses between the different blocks to capture the impact on performance. Here, the weighted interconnect matrix in [150] is adopted in the experiments. By minimizing the HPWL in the objective function, the minimum performance degradation is ensured.

6.3 Solution 2: Power Supply Pads Assignment for Maximizing Timing Yield

Power distribution networks must be designed with great care to ensure the delivery of the correct functionalities within a limited time. The increasing process variations, lower noise margins, and high electromigration in modern technologies make it even more challenging to design a robust power distribution network. The high number of power supply pads in such a design and the significant impact of their assignments to different nodes calls for a robust solution. This section proposes an optimum supply pads assignment to bound the variations in supply voltage and maximize the timing yield.

6.3.1 Related Work

The analysis and verification of power grids start from the early design phases [104, 157]. In recent years, the deign of power grid has been investigated in different areas and design stages. A wire sizing optimization is discussed in [158] where the locality of the power grid is utilized to optimize the partitioned grid. The authors in [159] explain their method for floorplanning and power network cosynthesis. Here, the area and wire length are minimized, while fixing the *IR* drop violations.

The authors in [160, 161] optimize the topology of the power distribution network by including the routing congestion and area in their cost function. The optimization of multi-layer topology is studied in [162, 163]. The objective is to minimize the *IR* drops by optimizing the wire width and reducing the mesh layer impedance. The authors in [164, 165, 166] optimize the number and location of the supply pads in order to bound the *IR* drop within its constraint. The optimization of the wire width in a power grid is discussed in [167], while addressing the reliability and current density of the circuits.

The deterministic approaches in the aforementioned work are no longer reliable under process variations. These variations impose statistical measures on the voltage drop due to the variations in the power consumption of the chip blocks. The variations in the supply voltage (V_{dd}), then, introduces new sources of variations in the circuits' delay. The impact of the *IR* drop on the circuits timing is examined in [168, 131]. Here, the timing analysis is carried out to investigate the effect of the supply voltage variations on the critical path delay.

The location of the power supply pads, pins, and voltage regulators, hereafter referred to as pads, substantially affect the design. The magnitude of the voltage drops and their variations depend on the number and location of the pads. Consequently, the delay of the gates on a critical path can change with the pads location. Therefore, it is critical to take into account the chip timing yield while assigning the pads on the power grid. Ignoring

such a dependency, under process variations, leads to unreliable designs that do not meet their timing objectives.

This research formulates the pad assignment optimization as a Mixed-Integer Non-Linear Programming (MINLP) problem. Here, the objective is to minimize the timing yield loss. The constraint on the voltage drop ensures a sufficient noise margin and correct functionality. Also, the maximum current constraints limit the current of the voltage regulator and guarantees the Mean Time To Failure (MTTF) due to electromigration. The well-known macromodeling technique [169] is chosen in our method to increase the efficiency, specifically for large power grids.

The proposed optimization methodology can be employed for finding either the best pad number and location at the chip-level or for the pad/pin assignment at the block-level. In addition, both wire-bond and flip-chip packages can use the methodology for locating the optimal pads on the peripheral power ring in the former case, or for the over the block supply pins in the latter.

6.3.2 Voltage Drop and Supply Current Statistics

The objective of this section is to model the supply voltage variations across the power grid. This is to accurately map the process variations to the statistics of the voltage drop across the power grid.

In this research, as similarly discussed in the previous chapters, a RC network that is distributed over the die in multiple metal layers is considered. Each branch of the grid is modeled with a resistor, and all the nodes have a capacitor to the ground. Also, an ideal current source is assumed to be connected to the nodes in the first metal layer (M1). The Modified Nodal Analysis (MNA) governs the relationship between the current and voltage of each node, and is expressed as

$$GV(t) + C\frac{V(t)}{dt} = -i(t) + GV_{dd} \ , \tag{6.4}$$

where V(t) is the vector of voltages at each node, and i(t) is the vector of the current sources. G, C are the conductance and capacitance matrices, respectively. An AC analysis of the power grid provides more detailed information, regarding the voltage drop wave form. However, including transient data requires assumptions that can lead to inaccurate results. This occurs because there are various modes that the circuits can operate in. Therefore, a DC analysis is considered in this work. Nevertheless, the methodology is flexible enough to take the trace of the dynamic power as the input and use the transient data for the AC analysis. The matrix format of (6.4) is

$$GV = -I + GV_{dd} \ , \tag{6.5}$$

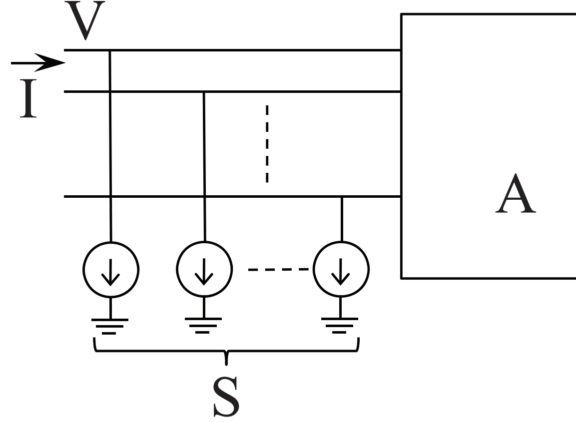


Figure 6.8: Macromodel schematic

where V is the vector of supply voltages, and I is the vector of currents drawn off the power grid.

However, solving (6.5), as a part of an optimization problem, is very expensive due to large number of nodes. Therefore, the macromodel idea in [169] is applied to reduce the number of nodes to a subset of selected nodes, referred to as ports. The transfer characteristic of the macromodel, shown in Fig. 6.8, is given by

$$I = A.V + S, \quad I, V, S \in R^m, \quad A \in R^{m \times m} \ , \tag{6.6}$$

where A is the port conductance matrix, S is the vector of all current sources connected to the ports, I is the vector of currents flowing into the model through the ports, V is a vector representing the port voltages, and m is the number of ports. In the macromodel, the internal nodes of the local grids are abstracted, and the current sources connected to these nodes are moved to the ports. In the proposed optimization, all the ports of the multi-port model are considered as candidates for assigning the supply pads. The relationship between the components of the macromodel and those of the modified nodal system are determined by rearranging the original equations [169] such that

$$\begin{bmatrix} G_{11} & G_{12} \\ G_{12}^T & G_{22} \end{bmatrix} \begin{bmatrix} U \\ V \end{bmatrix} = \begin{bmatrix} J_1 \\ J_2 + I \end{bmatrix} \ , \tag{6.7}$$

where G_{ij} is the submatrix of the conductance matrix, U is the vector of the internal nodes' voltage, and V is the vectors of the ports' voltage. Also, J_1 is the vector of current sources connected to the internal nodes and J_2 denotes the current sources connected to the ports. A and S are obtained from

$$A = G_{22} - G_{12}^T G_{11}^{-1} G_{12} \quad S = G_{12}^T G_{11}^{-1} J_1 - J_2 \tag{6.8}$$

To avoid computing the inverse of a large matrix in the previous equation, as explained in [169], the submatrices of the Cholesky factors are used. The ports of the macromodel are partitioned to the assigned ports and observation ports. It is assumed that the assigned ports have ideal voltage of V_{dd}. By rearranging the matrices, (6.6) is restated as follows:

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{12}^T & A_{22} \end{bmatrix} \begin{bmatrix} V \\ V_{dd} \end{bmatrix} + \begin{bmatrix} S_1 \\ S_2 \end{bmatrix} = \begin{bmatrix} 0 \\ I \end{bmatrix} , \tag{6.9}$$

where A_{ij} are the submatrices of A, V is the vector of voltages at the observation ports, and S_1 and S_2 are the vectors of current sources connected to the observation ports and the supply pads respectively. In addition, I is the vector of the currents flowing into the model through the supply pads. Note that no current flows into the model through the observation ports. Let $\nu = V_{dd} - V$ be the vector of the voltage drops at the observation ports. From (6.9),

$$\begin{aligned} A_{11}\nu &= A_{11}V_{dd} + A_{12}V_{dd} + S_1 \\ I &= A_{12}^T V_{dd} + A_{22}V_{dd} + S_2 - A_{12}^T\nu, \ . \end{aligned} \tag{6.10}$$

By applying the $E[.]$ and $Var[.]$ operators, as similarly done in [36], the statistical moments of the voltage drop is extracted by

$$\begin{aligned} E[\nu] &= A_{11}^{-1}\big((A_{11} + A_{12})V_{dd} + E[S_1]\big) \\ \mathrm{Var}(\nu) &= A_{11}^{-1(2)}\mathrm{Var}(S_1) \ , \end{aligned} \tag{6.11}$$

where $A_{11}^{-1(2)}$ denotes a matrix whose elements are the square of each element in the inverse of A_{11}. Also, from (6.10), the statistical moments of the currents, flowing into the model through the supply pads, are given by

$$\begin{aligned} E[I] &= A_{12}^T V_{dd} + A_{22}V_{dd} + E[S_2] - A_{12}^T E[\nu] \\ \mathrm{Var}(I) &= \mathrm{Var}(S_2 - A_{12}^T\nu) \ . \end{aligned} \tag{6.12}$$

The locality of the grid problem ([135]) states that the voltage drop is closely related to the variations in the neighboring current sources. Therefore, it is safe to assume that the two components in (6.12), S_2 and $A_{12}^T\nu$, are not highly correlated. Here, S_2 is the vector of the current sources connected to the power pads while ν denotes the voltage drops at the observation ports. Consequently,

$$\mathrm{Var}(I) = \mathrm{Var}(S_2) + A_{12}^{T(2)}\mathrm{Var}(\nu) \ . \tag{6.13}$$

The statistics of the voltage drops and currents are used in the next section to put an upper bound on their respective constraints.

6.3.3 Design Constraints and Yield Optimization

Assigning the power pads changes the voltage drop profile across the chip. The delay of the gates on the critical paths is sensitive to these voltage drops. Under process variations, the voltage drops have statistical measures, which, in turn, introduce new sources of uncertainty in the delay. Therefore, the pad assignment impacts the delay, such that a non-critical path can become critical.

To optimally assign the supply pad to the correct ports, the die area is discretized into tiles. The candidate pads are selected in each tile, where the selection depends on the package design. For example, in a flip-chip package, the c4 bump locations determine the candidate power pads. In a wire-bound package, the location of the terminals, routed to the peripheral power ring, determines where the candidate pads should be located.

To capture the impact of the voltage drop variations on the delay, the objective function is formulated by the linear regression,

$$f = \sum_{j=1}^{n} a_j w_j (\mu_{\nu_j} + r\sigma_{\nu_j}) \ , \tag{6.14}$$

where n is the number of tiles, w_j is the weighting factor, a_j is the delay sensitivity to the voltage drop in tile j, μ_{ν_j} is the expected value, and σ_{ν_j} is the standard deviation of the voltage drop distribution in tile j. Also, r is determined by the distribution confidence level in the ν upper bound. For minimizing the delay, all the potential critical paths are taken into account. Many of them pass through different tiles. Assigning a supply pad in a tile with a larger number of critical paths has a larger effect on the delay. Therefore, in the cost function, a weighting factor, given to each tile, is used to take this point into consideration.

In regards to the constraints, for the limited number of pads, the power grid should meet the maximum voltage drop constraint ($\mu_{\nu_j} + r\sigma_{\nu_j} <= \nu_t$). This ensures valid functionalities, as well as acceptable noise margins across the chip.

In addition, the current that flows through the supply pad must not exceed the threshold value ($\mu_{I_j} + r\sigma_{I_j} <= I_t$). This constraint guarantees that the Mean Time To Failure (MTTF), due to electromigration meets the requirement. In addition, the upper bound on the supply current leads to a more uniform current density profile, and, thus, more reliable circuits [106].

Moreover, the expected values and variance of the voltage drops, and the currents must satisfy (6.11) and (6.12). Finally, the number of pads is limited.

In the optimization problem, the MINLP problem is formed:

$$
\begin{aligned}
& min \quad \sum_{j=1}^{n} a_j w_j (\mu_{\nu_j} + r\sigma_{\nu_j}) y_j, \quad y_j \in \{0,1\} \\
& subject\ to: \\
& \quad \mu_{\nu_j} + r\sigma_{\nu_j} <= \nu_t (1 - y_j) \ , \\
& \quad \mu_{\nu_j} = \left[A_{11}^{-1} \big((A_{11} + A_{12}) V_{dd} + \mu_{S_1} \big) \right]_j (1 - y_j) \ , \\
& \quad \sigma_{\nu_j}^2 = \left[A_{11}^{-1(2)} \sigma_{S_1}^2 \right]_j (1 - y_j) \ , \\
& \quad \mu_{I_j} + r\sigma_{I_j} <= I_t y_j \ , \\
& \quad \mu_{I_j} = \left[A_{12}^T V_{dd} + A_{22} V_{dd} + \mu_{S_2} - A_{12}^T \mu_\nu \right]_j y_j \ , \\
& \quad \sigma_{I_j}^2 = \left[\sigma_{S_2}^2 + A_{12}^{T(2)} \sigma_\nu^2 \right]_j y_j \ , \\
& \quad \sum_{j=1}^{n} y_j <= N \ , \ and \\
& \quad \mu_{\nu_j}, \sigma_{\nu_j}, \mu_{I_j}, \sigma_{I_j} >= 0 \ ,
\end{aligned}
\tag{6.15}
$$

where N is the number of candidate pads. The output of the MINLP problem consists of the values for the continuous variables, $\mu_{\nu_j}, \sigma_{\nu_j}$, as well as the integer variable, y_j for $\forall j = 1, ..., N$.

The use of y_j and $1 - y_j$ in the respective constraints enforces the port partitioning, to candidate and to the observation pads, and ensures the feasibility of the solution.

6.3.4 Results and Discussion

The proposed method is implemented in C++ and MATLAB, and executed on a 3.4 GHz Pentium-4 PC with 2GB RAM. The experiments are performed on edge-triggered ISCAS89 benchmarks by using 65nm technology parameters. The process parameters are adapted from [115]. In regards to the power grid, the metal layers, pitch, and width per layer are selected according to the IBM benchmarks [116], scaled for 65nm technology.

To set up the MINLP problem, and test it on the benchmark circuits, the following steps are adhered to.

Step 1: Place and route the circuits. They are initially placed by using Capo [138], and a global routing is performed for all the nets. Then, the die area is discretized into a number of tiles for both the delay calculation and pad assignment. The number of tiles is selected according to the size of the circuits.

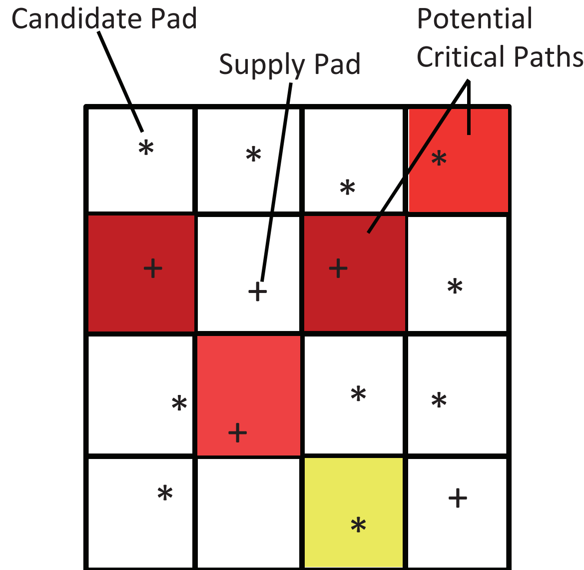


Figure 6.9: Discretized die area , the tiles that share critical paths, and the candidate pads. The darker color represents a larger weighting factor for the respective tile where it shares a larger number of critical paths.

Step 2: Estimate the statistical moments of the current sources in each tile. The simulation data are chosen for the benchmark circuit, instead of random current sources ([131]). This makes the result more accurate. This step is performed by using an in-house tool set.

Step 3: Extract the potential critical paths. Monte-Carlo simulations are executed for each circuit under process variations, and the critical paths are stored in each run. In the end, the tiles' weighting factors are calculated, based on the total number of the critical paths in all the runs, for each tile. Fig. 6.9 reflects an example of the discretized die area, the tiles that share critical paths, and the candidate pads. The darker color represents a larger weighting factor for the respective tile, where it shares a larger number of critical paths.

Step 4: Reduce the size of the problem by employing the macromodel technique, where the components of the macromodel are extracted according to (6.8).

Step 5: Solve the optimization problem with additional partitioning, as described in the previous section. TOMLAB® is used as the optimization engine, where its minlpBB solver utilizes the branch-and-bound search scheme.

Step 6: Estimate the timing yield for the optimal supply pad assignment. Here, the statistical timing analysis is carried out for the circuits by taking into account the

Table 6.3: Timing yield and statistics of delay, comparing optimal and random supply pad assignment.

Circuit			Power Grid		Our Optimization Method			Random Assignment		Target Delay (picosecond)	Yield Improvement (%)
Name	Number of Cells	Number of Candidate Pads	Number of Nodes	Maximum Number of Pads	Delay Mean (picosecond)	Delay Std (picosecond)	Optimal Number of Pads	Delay Mean (picosecond)	Delay Std (picosecond)		
s1196	547	16	4,762	8	480.26	48.56	6	490.81	52.21	436.58	3.5
s5378	2958	64	16,642	16	381.59	33.27	15	406.72	37.05	363.51	17.2
s9234	5825	64	16,642	16	679.74	57.02	16	746.61	59.60	602.36	8.0
s13207	8260	64	16,642	24	1073.23	96.87	24	1171.74	129.12	1058.08	24.9
s15850	10369	64	82,370	24	1286.26	111.98	23	1383.60	121.89	1194.87	14.6
s35932	17793	64	82,370	24	1081.22	88.47	20	1104.10	90.32	1046.10	8.5
s38417	23815	100	434,282	32	886.02	77.02	27	985.55	101.85	875.50	30.6
s38584	20705	100	434,282	32	1534.06	130.98	32	1618.38	134.05	1530.94	23.3

statistical moments of the voltage drop in each tile.

The previous steps are followed for different circuits and power grids. Table 6.3 demonstrates the delay statistics for the optimal and random pad assignment. The number of candidate pads is chosen according to the size of the circuit. In addition, the target delay is assumed to be the upper bound ($\mu + 3\sigma$) of the circuit delay, where each node has an ideal supply voltage. It is observed that the optimal pad assignment can significantly improve the timing yield. For example, the yield increases by 30.6% in the case of the s38417 circuit.

The runtime of the optimization method consists of the time spent on macromodel extraction and the time is required to solve the MINLP problem. As summarized in Table 6.4, the components of the macromodel are efficiently extracted by using the submatrices of the Cholesky factors [169]. However, for a large number of pad candidates, a few minutes is needed to find the optimal pad assignment. This can be improved by defining the priority order in which the non-integer variables are selected.

The magnitude of the yield enhancement depends on several parameters, including the number of pads, size of the power grid, and current sources. In addition to the pads, a larger yield is observed for the larger power grids. When the size of the power grid increases, the effective impedance of the grid is reduced such that the voltage drop decreases. However, as depicted in Fig. 6.10, the circuits demonstrate different yield sensitivity in respect to the total number of supply pads. But, for all the circuits, the effectiveness of the pad assignment decreases as the number of pads increases.

Table 6.4: Runtime of the pad assignment optimization and macromodel extraction.

Circuit			Runtime		
Name	Number of Candidate Pads	Number of Nodes	Macro Model Extraction (second)	Optimization (second)	Total (second)
s1196	16	4,762	1	4	5
s5378	64	16,642	1	14	15
s9234	64	16,642	1	16	17
s13207	64	16,642	1	22	23
s15850	64	82,370	1	19	20
s35932	64	82,370	1	20	22
s38417	100	434,282	16	136	152
s38584	100	434,282	16	138	154

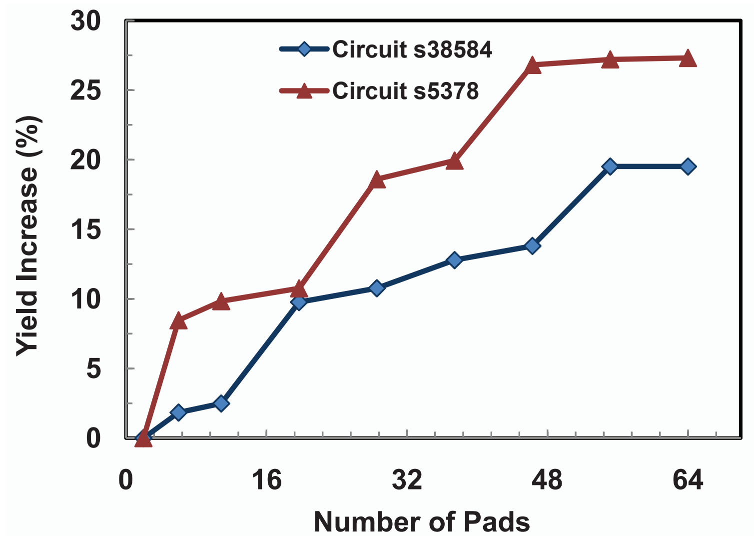


Figure 6.10: Timing yield sensitivity of two test circuits as a function of the number of supply pads.

6.4 Conclusions

Two solutions are proposed in this chapter that address the variability in VLSI systems. First, the correlation between the total power consumption and temperature variations, associated with a floorplan, is investigated. With this correlation, it is demonstrated that the architectural blocks of a floorplan can be placed such that total power consumption is

reduced and a more uniform thermal profile is achieved. However, due to the possibility of longer wire lengths, the most uniform thermal profile does not necessarily yield the highest total power reduction. Also, it is shown that for a small increase in the total power, a significant reduction in the temperature variations can be achieved. The proposed methodology can be used in the early phase of the design to reduce the total power and to address the thermal integrity issues.

As the next solution, the power supply pads assignment is presented to maximize the timing yield. Supply pad assignment significantly impacts the chip performance and reliability. Under process variations, a non-optimal pad selection can result in a design that does not meet its set objective. Here, the statistics of the currents, drawn off the power grid, are used to bound the voltage drops and supply currents across the die. A Mixed-Integer Non-Linear Programming (MINLP) problem is formed to maximize the timing yield while addressing the noise margin and electromigration concerns. The results demonstrate large yield improvements. Finally, by utilizing the macromodel technique, the optimal solution converges in a few minutes.

Chapter 7

Conclusions and Future Work

7.1 Summary of Contributions

In this thesis, yield optimization and the design of integrated circuits and systems under variability are studied. Initially, at the circuit level, a statistical methodology is proposed to achieve the circuit robustness, and find the best possible trade-off between power and performance. By using this design centering methodology, a pair of optimal V_{dd} and V_{th} is obtained for which the parametric yield is maximized. In addition, the scaling trend of variations in the parametric yield is presented. The maximum yield design centers for future nodes are also suggested. The impact of the switching activity and device sizing on the parametric yield is explored. It is demonstrated that by enhancing the performance, reducing the power consumption, and attaining an acceptable level of reliability beyond $45nm$ technology, calls for close attention to the trend of changes in the parametric yield.

At the system level, a power grid analysis is proposed to map the process variations to the upper bounds of the voltage drop across a chip, by considering variations in temperature-dependent power consumption. Here, the statistical thermal profile is generated across a power grid. Then, the close loop, between the temperature and leakage power consumption, is used to map the process variations to the voltage drop statistics. It is found that the IR drop is significantly impacted by the statistical thermal profile. This analysis, then, is utilized to propose a comprehensive method for the timing yield analysis under variations. Here, the statistical profiles of temperature and voltage drops are developed to estimate the circuit delay. By ignoring the variations of the temperature and their associated voltage drops, a significant yield loss results. In addition, an analysis is proposed to accurately estimate the power yield under process variations. This study integrates the thermal analysis, power estimation, and IR drop calculation to ensure the robustness and reliability of the system.

Lastly, two solutions are proposed to alleviate the impact of variations on the VLSI systems and enhance parametric yield. In the first solution, the correlation between the temperature variations, associated with a floorplan and the total power consumption, is examined. Then, this correlation is utilized to optimize the floorplan to reduce the power and address the thermal integrity of the architecture blocks. It is also shown that, due to the possibility of longer wire lengths, the most uniform thermal profile does not necessarily yield the highest total power reduction. Also, it is denoted that, in high performance applications, for a small increase in the total power, a significant reduction in the temperature variations can be achieved. As the second solution, supply pad assignment is proposed. Here, a Mixed-Integer Non-Linear Programming (MINLP) is formed not only to maximize the timing yield, but also to address the noise margin and electromigration concerns. The results demonstrate large timing yield improvements.

The analyses and optimization methodologies proposed in this thesis can be applied to estimate and improve the parametric yield. Voltage scaling, low power floorplanning, and designing robust power grids, are few examples of such applications.

7.2 Future Research Directions

The proposals in the thesis can be extended in various levels of abstraction. At the circuit level, the dual-supply voltage (V_{dd}) and dual-threshold voltage (V_{th}) framework can be covered by introducing new design variables. Since many designs have already adopted these power reduction techniques, this extension should also be beneficial.

In addition, one of the application that can benefit from the circuit level study is subthreshold circuit design. The operation of such applications greatly depends on the leakage current, and, thus, variations in V_{th} significantly impact the parametric yield. Therefore, the robustness of the subthreshold design can be enhanced by carefully constructing the feasible region and optimizing the yield.

At the system level, the study can be extended to include a transient analysis for the power grid and parametric yield. Here, in addition to the IR drop, the statistical moments of the voltage drop in an AC analysis can be extracted. Then the moments, are used to compute the timing and power yield. Also, the inductive effect of the power grid can be significant in high-performance applications. Therefore, adding Ld_i/d_t component to the voltage drop makes the analyses more accurate for such applications.

Moreover, the timing and power yield depend on the same underlying parameters. Therefore, it is very useful to simultaneously analyze the timing and power yield. As a result, the parametric yield can be estimated and analyzed under temperature and supply voltage variations.

The aforementioned extensions are also applicable to the proposed solutions. The supply voltage variations can be added so that the floorplanning methodology is more effective. The interdependency of the voltage and temperature variations can be addressed at the system level. For this, the cost function needs to be redefined and the total power should be minimized more effectively.

Furthermore, the pad assignment can address applications for maximizing the power yield in a power limited design. This is achievable by changing the constraints and objective function.

There are other solutions that can exploit the analyses for managing the variability. Placement and optimization of on-chip decoupling capacitors are two examples that can utilize the results and methodologies presented in this thesis.

Appendix A

Publications from this Research

The following is the related list of publications:

J1 K. Haghdad and M. Anis, "Design-specific optimization considering supply and threshold voltage variations," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, vol. 27, no. 10, pp. 1891-1901, Oct. 2008.

J2 K Haghdad, M. Anis, and Y. Ismail, "Floorplanning for low power IC design considering temperature variations," *Microelectronics Journal*, vol. 42, no. 1, pp. 89-95, January 2011.

J3 K. Haghdad and M. Anis, "Power Yield Analysis under Process and Temperature Variations," submitted revision to *IEEE Transactions on Very Large Scale Integration Systems (TVLSI)*.

J4 K. Haghdad and M. Anis, "Power Supply Pads Assignment for Maximum Timing Yield," submitted revision to *IEEE Transactions on. Circuits and Systems II (TCAS II)*.

J5 K. Haghdad and M. Anis, "Timing Yield Analysis Considering Process-Induced Temperature and Supply Voltage Variations," submitted to *ACM Transactions on Design Automation of Electronic Systems*.

J6 K. Haghdad, J. Jaffari, and M. Anis, "Power Grid Analysis and Verification Considering Temperature Variations," submitted to *Microelectronics Journal*.

C1 K. Haghdad and M. Anis, "Design-Specific Supply and Threshold Voltage Optimization in Nanometer Era," in *Proc. IEEE International Midwest Symposium on Circuits and Systems/IEEE International Northeast Workshop on Circuits and Systems (MWSCAS/NEWCAS)*, pp. 1054-1057, Montreal, Canada 2007.

C2 K. Haghdad and M. Anis, "Scaling analysis of yield optimization considering supply and threshold voltage variations," in *Proc. IEEE International Symposium on Circuits and Systems (ISCAS)*, Paris, France 2010, pp. 3665-3668.

Appendix B

Acronyms

ABB	Adaptive body bias
AC	Alternating current
ACL	Asynchronous level converter
AR	Aspect ratio
ASV	Adaptive supply voltage
AVS	Adaptive voltage scaling
CAD	Computer-aided design
CDF	Cumulative distribution function
CMP	Chemical-mechanical polishing
CVS	Clustered voltage scaling
D2D	Die-to-die
DB-PDF	Double-bounded probability density function
DC	Direct current
DIBL	Drain-induced barrier lowering
DPM	Dynamic power management
DTM	Dynamic thermal management
DVS	Dynamic voltage scaling
RBB	Reverse body bias
ECVS	Extended clustered voltage scaling
EDP	Energy delay product
FBB	Forward body bias
HiK+MG	High-k dielectric and metal gate device
HPWL	Half perimeter wire length
IC	Integrated circuit
ICA	Independent component analysis
ILD	Inter-layer dielectric

ISCAS	International symposium on circuits and systems
ITRS	International technology roadmap for semiconductors
LER	Line edge roughness
MCNC	Microelectronics center of North Carolina
MINLP	Mixed-integer nonlinear programming
MNA	Modified nodal analysis
MTTF	Mean time to failure
NBTI	Negative bias temperature instability
OPC	Optical proximity correction
OPE	Optical proximity effect
PC	Principal component
PCA	Principal component analysis
PDF	Probability density function
PDP	Power delay product
PERT	Program evaluation and review technique
PI	Primary input
PO	Primary output
PTM	Predictive technology models
PVT	Process, voltage, and temperature
RDF	Random dopant fluctuation
RV	Random variable
SCE	Short channel effect
SOC	System-on-chip
SPA	Statistical power analysis
SQP	Sequential quadratic programming
SSTA	Statistical static timing analysis
STA	Static timing analysis
TDDB	Time dependant dielectric breakdown
TDDS	Temperature dependant deactivation scheme
VDDH	High supply voltage
VDDL	Low supply voltage
VLSI	Very large scale integrated
WID	Within-die
Wmin	Minimum width

Appendix C

Variables

ΔL	Variation in channel length
ΔT_{ox}	Variation in oxide thickness
$\hat{I}_{leak-i}$	Random variable for leakage current in grid i
μ_{delay}	Expected value of longest path delay
ν_{sat}	Saturation velocity
ν_{sat0}	Saturation velocity at ambient temperature
σ_{delay}	Standard deviation of longest path delay
$\sigma_{T_{ox}}$	Standard deviation of oxide thickness
σ_L	Standard deviation of channel length
σ_T	Standard deviation of temperature
θ_{ja}	Junction to ambient thermal resistance
ε_{ox}	Oxide permittivity
ε_{S_i}	Silicon permittivity
$A_{m\times m}$	Inverse of admittance matrix
a_{orig}	Original area of a block
A_{orig}	Original area of the floorplan
AR	Aspect ratio of the floorplan
C_{eff}	Effective capacitance
d_{max}	Random variable associated with maximum delay
d_{sum}	Random variable associated with sum of the delays
d_i	Discritized delay in grid i
d_t	Target delay
$f_{EDP_{min}}$	Frequency at minimum EDP point
f_{min}	Minimum frequency
G	Conductance matrix
I_D	Drain current

I_s	Zero-threshold leakage current
K_B	Boltzmann's constant
L	Channel length
L_{eff}	Effective channel length
L_{ov}	Junction and channel overlap
L_d	Logic depth
$M_{P_{n\times1}}$	Matrix of expected values of power
$M_{T_{n\times1}}$	Matrix of expected values of temperature
N_a	Doping concentration
$P_{leak-j}^{(k)}$	Leakage power at time step k in grid j
P_{dyn}	Dynamic power
P_{leak}	Leakage power
$P_{m\times1}$	Chip to ambient removing power
P_{tot}	Total power
P_b	Power budget
pc_j	j^{th} Principal component
q	Elementary charge
R_θ	Thermal resistance
R_{con}	Heat sink to air heat resistance
$S_{P_{n\times n}}$	Matrix of covariance of power
$S_{T_{n\times n}}$	Matrix of covariance of temperature
$T_i^{(k+1)}$	Temperature at time step k+1 in grid i
T_{amb}	Ambient temperature
T_{avg}	Average temperature
t_{ILD}	Inter-layer dielectric thickness
$t_{m\times1}$	Vector of grid temperature
T_{max}	Maximum temperature
T_{ox}	Oxide thickness
T_{ref}	Reference temperature
T_g	Gate delay
T_j	Junction temperature
V_{dd}	Supply voltage
V_{dd}^l, V_{dd}^u	Lower and upper values of supply voltage
V_{th}	Threshold voltage
V_{th}^l, V_{th}^u	Lower and upper values of threshold voltage
V_{th0}	Threshold voltage at ambient temperature
W	Transistor width
W_{eff}	Effective width
WS	White space
WS_{max}	Maximum white space

Y_j	Junction depth

References

[1] W. Huang, S. Ghosh, S. Velusamy, K. Sankaranarayanan, K. Skadron, and M. R. Stan, "Hotspot: A compact thermal modeling methodology for early-stage vlsi design," *IEEE Trans. on VLSI*, vol. 14, no. 5, pp. 501–513, 2006. xiv, 69, 70

[2] "Electronic engineering times," 2007. [Online]. Available: http://www.eetimes.com 1

[3] "International technology roadmap for semiconductors (*ITRS*)." [Online]. Available: http://public.itrs.net/. 2, 39, 49, 60

[4] K. Banerjee, M. Pedram, and A. H. Ajami, "Analysis and optimization of thermal issues in high-performance vlsi," in *Proc. ISPD*, April 2001, pp. 230–237. 2

[5] H. Su, F. Liu, A. Devgan, E. Acar, and S. Nassif, "Full chip leakage estimation considering power supply and temperature variations," in *Proc. ISLPED*, 2003, pp. 78–83. 2, 34, 67

[6] A. Srivastava, D. Sylvester, and D. Blaauw, *Statistical Analysis and Optimization for VLSI: Timing and Power.* Springer, 2005. 4

[7] P. Yu, S. X. Shi, and D. Z. Pan, "Process variation aware opc with variational lithography modeling," in *Proc. DAC.* ACM, 2006, pp. 785–790. 6

[8] K. Agarwal and S. Nassif, "Characterizing process variation in nanometer cmos," in *Proc. DAC*, 2007, pp. 396–397. 7

[9] J. Watts, N. Lu, C. Bittner, S. Grundon, and J. Oppold, "Modeling fet variation within a chip as a function of circuit design and layout choices," in *Proc. Nanotech Workshop on Compact Modeling*, 2005, pp. 87–92. 7

[10] S. R. Nassif, "Modeling and analysis of manufacturing variations," in *Proc. IEEE Conference on Custom Integrated Circuits*, 2001, pp. 223–228. 7

[11] M. Hane, T. Ikezawa, and T. Ezaki, “Coupled atomistic 3d process/device simulation considering both line-edge roughness and random-discrete-dopant effects,” in *Proc. SISPAD*, 2003, pp. 99–102. 7

[12] K. Kuhn, C. Kenyon, A. Kornfeld, M. Liu, A. Maheshwari, W. kai Shih, S. Sivakumar, G. Taylor, P. VanDerVoorn, and K. Zawadzki, “Managing process variation in intels 45nm cmos technology,” *Intel Technology Journal*, 2008. 8, 10, 59, 60

[13] T. Mizuno, J. ichi Okamura, and A. Toriumi, “Experimental study of threshold voltage fluctuation due to statistical variation of channel dopant number in mosfets,” *IEEE Trans. on Electron Devices*, vol. 41, no. 11, pp. 2216–2221, 1994. 7

[14] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De, “Parameter variations and impact on circuits and microarchitecture,” in *Proc. DAC*, 2003, pp. 338–342. 8, 9, 12, 17, 18, 34, 106

[15] H. F. Dadgour, S.-C. Lin, and K. Banerjee, “A statistical framework for estimation of full-chip leakage-power distribution under parameter variations,” *IEEE Trans. on Electron Devices*, vol. 54, no. 11, pp. 2930–2945, 2007. 9

[16] S. Narendra, D. Antoniadis, and V. De, “Impact of using adaptive body bias to compensate die-to-die vt variation on within-die vt variation,” in *Proc. ISLPED*, 1999, pp. 229–232. 8

[17] C. Diaz, H.-J. Tao, Y.-C. Ku, A. Yen, and K. Young, “An experimentally validated analytical model for gate line-edge roughness (ler) effects on technology scaling,” *IEEE Trans. on Electron Devices*, vol. 22, no. 6, pp. 287–289, 2001. 10

[18] S. Dimitrijev, *Principles of semiconductor devices.* Oxford University Press Inc., 2005. 10, 16

[19] P. R. Kinget, “Device mismatch and tradeoffs in the design of analog circuits,” *IEEE Trans. on Solid-State Circuits*, vol. 40, no. 6, pp. 1212–1224, 2005. 10

[20] M. Choi and L. S. Milor, “Impact on circuit performance of deterministic within-die variation in nanoscale semiconductor manufacturing,” *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems*, vol. 25, no. 7, pp. 1350–1367, 2006. 10

[21] T. H. Park, “Characterization and modeling of pattern dependencies in copper interconnects for integrated circuits,” Ph.D. dissertation, Massachusetts Institute Of Technology, 2002. 10, 11

[22] R. Chang, “Integrated cmp metrology and modeling with respect to circuit performance,” Ph.D. dissertation, University Of California, Berkeley, 2004. 10, 11

[23] S. Zhang, V. Wason, and K. Banerjee, "A probabilistic framework to estimate full-chip subthreshold leakage power distribution considering within-die and die-to-die p-t-v variations," in *Proc. ISLPED*, 2004, pp. 156–161. 12, 14, 34, 48, 69

[24] S. G. Narendra, "Challenges and design choices in nanoscale cmos," *ACM Journal on Emerging Technologies in Computing Systems*, vol. 1, no. 1, pp. 7–49, 2005. 12, 60

[25] K. Skadron and M. Stan, "A quick thermal tutorial," University of Virginia, Charlottesville, Virginia, 2005. 13, 14

[26] R. Kumar and V. Kursun, "Impact of temperature fluctuations on circuit characteristics in 180nm and 65nm cmos technologies," in *Proc. ISCAS*, 2006. 13

[27] S. Im, N. Srivastava, K. Banerjee, and K. Goodson, "Scaling analysis of multilevel interconnect temperatures for high-performance ics," in *Proc. IEEE Conference on Electron Devices*, 2005, pp. 2710–2719. 13

[28] I. A. Ferzli and F. N. Najm, "Statistical estimation of leakage-induced power grid voltage drop considering within-die process variations," in *Proc. DAC*, 2003, pp. 856–859. 13, 32

[29] I. A. Ferzli and F. N. Najm., "Statistical verification of power grids considering process-induced leakage current variations," in *Proc. ICCAD*, 2003, p. 770. 13

[30] O. Semenov, A. Vassighi, M. Sachdev, A. Keshavarzi, and C. Hawkins, "Effect of cmos technology scaling on thermal management during burn-in," *IEEE Trans. on Semiconductor Manufacturing*, vol. 16, no. 4, pp. 686–695, Nov. 2003. 15

[31] K. A. Bowman, S. G. Duvall, and J. D. Meindl, "Impact of die-to-die and within-die parameter fluctuations on the maximum clock frequency distribution for gigascale integration," *IEEE Trans. on Solid-State Circuits*, vol. 37, no. 2, pp. 183–190, 2002. 15

[32] O. S. Unsal, J. W. Tschanz, K. Bowman, V. De, X. Vera, A. Gonzalez, and O. Ergin, "Impact of parameter variations on circuits and microarchitecture," *IEEE Micro*, p. 3039, Nov. 2006. 16

[33] S. M. Burns, M. Ketkar, N. Menezes, K. A. Bowman, J. W. Tschanz, and V. De, "Comparative analysis of conventional and statistical design techniques," in *Proc. DAC*, 2007, pp. 238–243. 16

[34] A. Srivastava, S. Shah, K. Agarwal, D. Sylvester, D. Blaauw, and S. Director, "Accurate and efficient gate-level parametric yield estimation considering correlated variations in leakage power and performance," in *Proc DAC*, 2005, pp. 535–540. 18, 19

[35] M. Mani, A. Devgan, and M. Orshansky, "An efficient algorithm for statistical minimization of total power under timing yield constraints," in *Proc. DAC*, 2005, pp. 309–314. 18, 21, 22, 86

[36] I. Ferzli and F. Najm, "Analysis and verification of power grids considering process induced leakage current variations," *IEEE Trans. on CAD*, pp. 126–143, Jan 2006. 18, 31, 67, 77, 89, 121

[37] T. Chen and S. Naffziger, "Comparison of adaptive body bias (abb) and adaptive supply voltage (asv) for improving delay and leakage under the presence of process variation," *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 11, no. 5, 2003. 18, 24

[38] M. Alioto and G. Palumbo, "Impact of supply voltage variations on full adder delay: Analysis and comparison," *IEEE Trans. VLSI Syst.*, vol. 14, no. 12, pp. 1322–1335, 2006. 18, 27

[39] L. Yan, J. Luo, and N. Jha, "Joint dynamic voltage scaling and adaptive body biasing for heterogeneous distributed real-time embedded systems," *IEEE Trans. on CAD*, vol. 24, no. 7, pp. 1030–1041, 2005. 18, 27, 28

[40] K. Agarwal and K. Nowka, "Dynamic power management by combination of dual static supply voltages," in *Proc. ISQED*, 2007, pp. 85–92. 18, 22, 23, 24

[41] A. Agarwal, K. Kang, S. Bhunia, J. Gallagher, and K. Roy, "Device-aware yield-centric dual-v_t design under parameter variations in nanoscale technologies," *IEEE Trans. VLSI Syst.*, vol. 15, no. 6, pp. 660–671, 2007. 18, 21

[42] N. Azizi, M. Khellah, V. De, and F. Najm, "Variations-aware low-power design with voltage scaling," in *Proc. DAC*, 2005, pp. 529–534. 18, 26, 27

[43] M. Elgebaly and M. Sachdev, "Variation-aware adaptive voltage scaling system," *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 15, no. 5, pp. 560–571, 2007. 18, 23

[44] S. Kulkarni, A. Srivastava, and D. Sylvester, "A new algorithm for improved vdd assignment in low power dual vdd systems," in *Proc. ISLPED*, 2004, pp. 200–205. 18, 24

[45] B. Lasbouygues, R. Wilson, N. Azémard, and P. Maurine, "Temperature and voltage-aware timing analysis," *IEEE Trans. on CAD*, vol. 26, no. 4, pp. 801–815, 2007. 18, 32, 97

[46] D. Sengupta and R. Saleh, "Generalized power-delay metrics in deep submicron cmos designs," *IEEE Trans. on CAD*, pp. 183–189, Jan. 2007. 20, 40, 55, 61

[47] R. Rao, K. Agarwal, A. Devgan, K. Nowka, D. Sylvester, and R. Brown, "Parametric yield analysis and constrained-based supply voltage optimization," in *Proc. ISQED*, 2005, pp. 284–290. 20, 40

[48] R. Gonzalez, B. M. Gordon, and M. A. Horowitz, "Supply and threshold voltage scaling for low power cmos," *IEEE J. Solid-State Circuits*, vol. 32, pp. 1210–1216, 1997. 20, 40, 41, 42

[49] A. Basu, S.-C. Lin, V. Wason, A. Mehrotra, and K. Banerjee, "Simultaneous optimization of supply and threshold voltages for low-power and high-performance circuits in the leakage dominant era," in *Proc. DAC*, 2004, pp. 884–887. 20, 40, 42, 44, 45

[50] D. Sengupta and R. A. Saleh, "Power-delay metrics revisited for 90nm cmos technology," in *Proc. ISQED*, 2005, pp. 291–296. 20, 40

[51] T. Karnik, Y. Ye, J. Tschanz, L. Wei, S. Burns, V. Govindarajulu, V. De, and S. Borkar, "Total power optimization by simultaneous dual-vt allocation and device sizing in high performance microprocessors," in *Proc. DAC*, 2002, pp. 486–491. 21

[52] J. W. Tschanz, J. T. Kao, S. G. Narendra, R. Nair, D. A. Antoniadis, A. P. Chandrakasan, and V. De, "Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage," *IEEE Trans. on Solid-State Circuits*, vol. 37, no. 11, pp. 1396–1402, 2002. 23, 25

[53] J. W. Tschanz, S. Narendra, R. Nair, and V. De, "Effectiveness of adaptive supply voltage and body bias for reducing impact of parameter variations in low power and high performance microprocessors," *IEEE Trans. on Solid-State Circuits*, vol. 38, no. 5, pp. 826–829, 2003. 25

[54] Y. HAN, "Temperature aware techniques for design, simulation and measurement in microprocessors," Ph.D. dissertation, University of Massachusetts Amherst, 2007. 29

[55] M. B. Healy, H. S. Lee, G. H. Loh, and S. K. Lim, "Thermal optimization in multi-granularity multi-core floorplanning," in *ASP-DAC*, 2009, pp. 43–48. 29, 30, 105

[56] M. Pedram and S. Nazarian, “Thermal modeling, analysis, and management in vlsi circuits: Principles and methods,” in *Proc. of the IEEE*, vol. 94, no. 8, August 2006, pp. 1487–1501. 28

[57] A. H. Ajami, K. Banerjee, and M. Pedram, “Modeling and analysis of nonuniform substrate temperature effects on global ulsi interconnects,” *IEEE Trans. on CAD*, vol. 24, no. 6, June 2005. 28, 104, 115

[58] A. H. Ajami, M. Pedram, and K. Banerjee, “Effects of non-uniform substrate temperature on the clock signal integrity in high performance designs,” in *Proc. CICC*, May 2001, pp. 233–236. 28

[59] K. Banerjee, S.-C. Lin, and N. Srivastava, “Electrothermal engineering in the nanometer era: from devices and interconnects to circuits and systems,” in *Proc. DAC*, 2006, p. 8. 29

[60] K. Sankaranarayanan, S. Velusamy, M. Stan, and K. Skadron, “A case for thermal-aware floorplanning at the microarchitectural level,” *The Journal of Instruction-Level Parallelism*, vol. 7, p. 816, 2005. 29, 30, 105

[61] D. Brooks and M. Martonosi, “Dynamic thermal management for high-performance microprocessors,” in *Proc. Seventh International Symposium on High-Performance Computer Architecture (HPCA-7)*, Jan. 2001, pp. 171–182. 29, 47

[62] M. Sarrafzadeh and C. K. Wong, *An Introduction to VLSI Physical Design.* McGraw-Hill Higher Education, 1996. 29, 105

[63] Y.-W. Wu, C.-L. Yang, P.-H. Yuh, and Y.-W. Chang, “Joint exploration of architectural and physical design spaces with thermal consideration,” in *Proc. ISLPED*, 2005, pp. 123–126. 29, 105

[64] P. Zhou, Y. Ma, Z. Li, R. P. Dick, L. Shang, H. Zhou, X. Hong, and Q. Zhou, “3d-staf: scalable temperature and leakage aware floorplanning for three-dimensional integrated circuits,” in *Proc. ICCAD*, 2007, pp. 590–597. 30

[65] A. Gupta, N. D. Dutt, F. J. Kurdahi, K. S. Khouri, and M. S. Abadir, “Leaf: A system level leakage-aware floorplanner for socs,” in *Proc. ASP-DAC*, Jan. 2007, pp. 274–279. 30, 105

[66] H. Mogal and K. Bazargan, “Microarchitecture floorplanning for sub-threshold leakage reduction,” in *Proc. DATE*, 2007, pp. 1238–1243. 30, 105

[67] D. Kouroussis, R. Ahmadi, and F. N. Najm, “Voltage-aware static timing analysis,” *IEEE Trans. on CAD*, vol. 25, no. 10, pp. 2156–2169, Jan 2006. 32, 33

[68] F.Najm, N. Menezes, and I. A. Ferzli, "A yield model for integrated circuits and its application to statistical timing analysis," *IEEE Trans. on CAD*, vol. 26, no. 3, pp. 574–591, 2007. 33

[69] T. D. Burd, T. A. Pering, A. J. Stratakos, and R. W. Broderson, "A dynamic voltage scaled microprocessor system," *IEEE J. Solid-State Circuits*, vol. 35, no. 11, pp. 1571–1580, Nov. 2000. 39

[70] T. Kuroda, T. Fujita, S. Mita, T. Nagamatsu, S. Yoshioka, K. Suzuki, F. Sano, M. Norishima, M. Murota, M. Kako, M. Kinugawa, M. Kakumu, and T. Sakurai, "A 0.9-v 150-mhz 10-mw, 4 mm 2-d discrete cosine transform core processor with variable-threshold-voltage (vt) scheme," *IEEE J. Solid-State Circuits*, vol. 31, no. 11, pp. 1770–1779, 1996. 39

[71] S. Lee and T. Sakurai, "Run-time voltage hopping for low-power real-time systems," in *Proc. DAC*, 2000, pp. 806–809. 39, 58

[72] C. Kim and K. Roy, "Dynamic vth scaling scheme for active leakage power reduction," in *Proc. DATE*, March 2002, pp. 163–167. 39

[73] A. Keshavarzi, S. Ma, S. Narendra, B. Bloechel, K. Mistry, T. Ghani, S. Borkari, and V. De, "Effectiveness of reverse body bias for leakage control in scaled dual-vt cmos ics," in *Proc. ISLPED*, 2001, pp. 207–212. 39

[74] K. Nose, M. Hirabayashi, H. Kawaguchi, S. Lee, and T. Sakurai, "V-hopping scheme for power saving in low-voltage processors," in *Proc. Custom Integrated Circuits Conference*, 2001, pp. 93–96. 39

[75] K. Nose and T. Sakurai, "Optimization of vdd and vth for low power and high speed applications," in *Proc. ASP-DAC*, 2000, pp. 469–474. 41

[76] V. De and S. Borkar, "Technology and design challenges for low power and high performance," in *Proc. ISLPED*, 1999, pp. 163–168. 41

[77] G. Nabaa and F. N. Najm, "Minimization of delay sensitivity to process induced voltage threshold variations," in *Proc. NEWCAS*, June 2005, pp. 171–174. 42

[78] K. Banergee and A. Mehrotra, "A power-optimal repeater insertion methodology for global interconnects in nanometer designs," *IEEE Trans. on Electron Devices*, vol. 49, pp. 2001–2007, Nov. 2002. 42

[79] J. Rabaey, *Digital Integrated Circuits: A Design Perspective.* Upper Saddle River, NJ: Prentice Hall, 1996. 42

[80] B. Nikolic, "Design in the power-limited scaling regime," *IEEE Trans. on Electron Devices*, vol. 55, pp. 71–83, Jan. 2008. 42, 69

[81] "Intel's transistor technology breakthrough." [Online]. Available: http://www.intel.com/pressroom/archive/releases/2007/20070128comp.htm 42

[82] Y. Cheng, K. Imai, M. Jeng, Z. Liu, K. Chen, and C. Hu, "Modeling temperature effects of quarter micrometer mosfets in bsim3v3 for circuit simulation," *Semicond. Sci. Tech*, pp. 1349–1354, 1997. 43, 91

[83] B. Chatterjee, M. Sachdev, S. Hsu, R. Krishnamurthy, and S. Borkar, "Effectiveness and scaling trends of leakage control techniques for sub-130nm cmos technologies," in *Proc. ISLPED*, 2003. 45

[84] K. Choi, R. Soma, and M. Pedram, "Off-chip latency-driven dynamic voltage and frequency scaling for an mpeg decoding," in *Proc. DAC*, 2004, pp. 544–549. 46

[85] W. Lee, K. Patel, and M. Pedram, "Dynamic thermal management for mpeg-2 decoding," in *ISLPED*, 2006, pp. 316–321. 46

[86] N. Weste and D. Harris, *A Circuits and Systems Perspective, Third Edition.* Upper Saddle River, NJ: Prentice Hall, 2004. 47

[87] S. Im and K.Banerjee, "Full chip thermal analysis of planar (2-d) and vertically integrated (3-d) high performance ics," in *Proc. Tech. Digest IEDM*, 2000, pp. 727–730. 47

[88] A. Devgan, S. Narendra, D. Blaauw, and F. Najm, "Leakage issues in ic design: Trends, estimation and avoidance," in *Proc. ICCAD*, 2003. 48, 107

[89] P. Kumaraswamy, "A generalized probability density function for double-bounded random processes," *J. of Hydrology*, vol. 46, pp. 79–88, 1980. 48

[90] K. Ponnambalam, A. Seifi, and J. Vlach, "Probabilistic design of systems with general distributions of parameters," *Int'l J. of Circuit Theory and Appl*, vol. 29, no. 6, pp. 527–536, 2001. 49

[91] X. Xi, K. M. Cao, H. Wan, M. Chan, and C. Hu, *BSIM4 MOSFET model-user manual*, Available: http://www.eas.asu.edu/~ptm/latest.html. 51

[92] Y.-C. Ban, S.-H. Choi, K.-H. Lee, D.-H. Kim, J. ong, Y.-H. Kim, Moon-Hyun, and J.-T. Kong, "A fast lithography verification framework for litho-friendly layout design," in *Proc. ISQED*, March 2005, pp. 169–174. 52, 63

[93] "Calibre lfd: Litho-friendly design." [Online]. Available: http://www.mentor.com/products/ic_nanometer_design/bl_phy_design/calibre_lfd 52, 63

[94] Y. C. Huifang, Y. Cao, H. Qin, R. Wang, P. Friedberg, A. Vladimirescu, and J. Rabaey, "Yield optimization with energy-delay constraints in low power digital circuits," in *Proc. IEEE Conference on Electron Devices and Solid-State Circuits*, December 2003, pp. 285–288. 56, 63

[95] J. Pouwelse, K. Langendoen, and H. Sips, "Dynamic voltage scaling on a low-power microprocessor," in *Proc. International Conference on Mobile Computing and Networking*, 2001, pp. 251–259. 58

[96] H. Soeleman, K. Roy, and B. Paul, "Robust ultra-low power sub-threshold dtmos logic," in *Proc. ISLPED*, 2000, pp. 25–30. 59

[97] D. Sylvester and A. Srivastava, "Computer-aided design for low-power robust computing in nanoscale cmos," *Proc. of the IEEE*, vol. 95, no. 3, pp. 507–529, 2007. 60

[98] K. Itoh and R. Takemura, "Low-voltage limitations of nano-scale cmos lsis: current status and future trends," in *Proc. IEEE Conference on Electron Devices and Solid-State Circuits*, 2007, pp. 83–86. 60

[99] S. Mukhopadhyay, K. Kim, K. A. Jenkins, C. T. Chuang, and K. Roy, "Statistical characterization and on-chip measurement methods for local random variability of a process using sense-amplifier-based test structure," in *Proc. IEEE International Solid-State Circuits Conference*, 2007, pp. 400–401. 60

[100] "Predictive technology model (ptm)." [Online]. Available: http://www.eas.asu.edu/~ptm/ 61

[101] H. Chang and S. Sapatnekar, "Prediction of leakage power under process uncertainties," *ACM Trans. Des. Autom. Electron. Syst.*, vol. 12, no. 2, p. 127, 2007. 67, 72

[102] N. A. Ghani and F. Najm, "Fast vectorless power grid verification using an approximate inverse technique," in *Proc. DAC*, 2009, pp. 184–189. 67, 79, 84

[103] N. Mi, J. Fan, S. X. d. Tan, Y. Cai, and X. Hong, "Statistical analysis of on-chip power delivery networks considering lognormal leakage current variations with spatial correlation," *Trans. on Circuits and Systems (TCAS-1)*, vol. 55, no. 7, pp. 2064–2075, 2008. 67

[104] H. Qian, S. R. Nassif, and S. S. Sapatnekar, "Early-stage power grid analysis for uncertain working modes," *IEEE Trans. on CAD*, vol. 24, no. 5, pp. 676–682, 2005. 67, 118

[105] Y. Zhong and M. D. F. Wong, "Thermal-aware ir drop analysis in large power grid," in *Proc. ISQED*, 2008, pp. 194–199. 67

[106] A. H. Ajami, K. Banerjee, and M. Pedram, "Scaling analysis of on-chip power grid voltage variations in nanometer scale ulsi," *Journal of Analog Integrated Circuits and Signal Processing*, vol. 42, no. 3, pp. 277–290, 2005. 67, 81, 122

[107] J. Jaffari and M. Anis, "Statistical thermal profile considering process variations: analysis and applications," *IEEE Trans. on CAD*, vol. 27, pp. 1027–1040, 2008. 68, 69, 73, 74

[108] A. Agarwal, D. Blaauw, V. Zolotov, S. Sundareswaran, M. Zhao, K. Gala, and R. Panda, "Statistical delay computation considering spatial correlations," in *Proc. ASPDAC*, 2003, pp. 271–276. 69

[109] M. G. Kendall, A. Stuart, and J. K. Ord, *Kendall's advanced theory of statistics*. New York, NY: Oxford University Press Inc., 1987. 77

[110] N. C. Beaulieu, A. A. Abu-Dayya, and P. J. McLane, "Estimating the distribution of a sum of independent lognormal random variables," *IEEE Trans. on Communication*, vol. 43, no. 22, pp. 2869–2873, 1995. 78

[111] S. Ross, *Introduction to Probability Models*. New York, NY: New York, NY, Academic, 2003. 78

[112] M. J. Grote and T. Huckle, "Parallel preconditioning with sparse approximate inverses," *SIAM Journal on Scientific Computing*, vol. 18, no. 3, pp. 838–853, 1997. 79

[113] "Hotspot 3.1 temperature modeling tool." [Online]. Available: http://lava.cs.virginia.edu/HotSpot/ 79

[114] G. M. Link and N. Vijaykrishnan, "Thermal trends in emerging technologies," in *Proc. IEEE Int. Symp. Quality Electron. Des.*, 2006, pp. 625–632. 79

[115] J. Xiong, V. Zolotov, and L. He, "Robust extraction of spatial correlation," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 26, no. 4, pp. 619–631, 2007. 80, 94, 123

[116] S. R. Nassif, "Power grid analysis benchmarks," in *Proc. ASPDAC*, 2008, pp. 376–381. 80, 94, 100, 123

[117] M. Nizam, F. N. Najm, and A. Devgan, "Power grid voltage integrity verification," in *Proc. IEEE International Symposium on Low Power Electronics and Design*, 2005, pp. 239–244. 84

[118] R. Raoand, A. Srivastava, D. Blaauw, and D. Sylvester, "Statistical analysis of subthreshold leakage current for vlsi circuits," *IEEE Trans. on VLSI*, vol. 12, no. 2, pp. 131–139, 2004. 86

[119] F. Wang, G. Sun, and Y. Xie, "A variation aware high level synthesis framework," in *Proc. DATE*, 2008, pp. 1063–1068. 86

[120] D. Sylvester, K. Agarwal, and S. Shah, "Variability in nanometer cmos: Impact, analysis, and minimization," *Integration, the VLSI Journal*, vol. 41, no. 3, pp. 319–339, 2008. 86

[121] N. Banerjee, S. Chandra, S. Ghosh, S. Dey, A. Raghunathan, and K. Roy, "Coping with variations through system-level design," in *Proc. International Conference on VLSI Design*, 2009, pp. 581–586. 86

[122] K. Agarwal, R. Rao, D. Sylvester, and R. Brown, "Parametric yield analysis and optimization in leakage dominated technologies," *IEEE Trans. on VLSI*, vol. 15, no. 6, pp. 613–623, 2007. 86, 87

[123] R. Rao, A. Devgan, D. Blaauw, and D. Sylvester, "Parametric yield estimation considering leakage variability," in *Proc. DAC*, 2004, pp. 442–447. 86

[124] K. Heloue, N.Azizi, and F.N.Najm, "Modeling and estimation of full-chip leakage current considering within-die correlation," in *Proc. DAC*, 2007, p. 9398. 86

[125] T. Enami, S. Ninomiya, and M. Hashimoto, "Statistical timing analysis considering spatially and temporally correlated dynamic power supply noise," *IEEE Trans. on CAD*, vol. 28, no. 4, pp. 541–553, 2009. 86

[126] K. R. Heloue and F. N. Najm, "Early analysis and budgeting of margins and corners using two-sided analytical yield models," *IEEE Trans. on CAD*, vol. 27, no. 10, pp. 1826–1839, 2008. 86

[127] A. Devgan and C. Kashyap, "Block-based static timing analysis with uncertainty," in *Proc. ICCAD*, 2003, pp. 607–614. 86

[128] C. Visweswariah, K. Ravindran, K. Kalafala, S. G. Walker, and S. Narayan, "First-order incremental block-based statistical timing analysis," in *Proc. DAC*, 2004, pp. 331–336. 86

[129] H. Chang and S. S. Sapatnekar, "Statistical timing analysis under spatial correlations," *IEEE Trans. on CAD*, vol. 24, no. 9, pp. 1467–1482, 2005. 86, 97, 100

[130] J. Singh and S. S. Sapatnekar, "A scalable statistical static timing analyzer incorporating correlated non-gaussian and gaussian parameter variations," *IEEE Trans. on CAD*, vol. 27, no. 1, pp. 160–173, 2008. 86, 90

[131] S. Pant and D. Blaauw, "Static timing analysis considering power supply variations," in *Proc. ICCAD*, 2005, pp. 365–371. 87, 118, 124

[132] K. Haghdad and M. Anis, "Design-specific optimization considering supply and threshold voltage variations," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 27, no. 10, pp. 1891–1901, 2008. 87

[133] B. Li, L. Peh, and P. Patra, "Impact of process and temperature variations on network-on-chip design exploration," in *Proc. IEEE International Symposium on Networks-on-Chip (NOCS)*, 2008. 87

[134] T. Kirkpatrick and N. Clark, "Pert as an aid to logic design," *IBM Journal of Research and Development*, vol. 10, no. 2, pp. 135–141, 1966. 90

[135] E. Chiprout, "Fast flip-chip power grid analysis via locality and grid shells," in *Proc. DAC*, 2004, pp. 485–488. 91, 121

[136] C. E. Clark, "The greatest of a finite set of random variables," *Oper. Res.*, vol. 9, no. 2, pp. 145–162, 1961. 92

[137] A. Papoulis and S. U. Pillai, *Probability in Random Variables and Stochastic Processes*. New York, NY: McGraw-Hill, 2001. 92, 99

[138] "Capo: A large-scale fixed-die placer," UCLA. [Online]. Available: http://vlsicad.eecs.umich.edu/BK/PDtools/Capo/ 94, 100, 123

[139] R. Kumar and V. Kursun, "Reversed temperature-dependent propagation delay characteristics in nanometer cmos circuits," *IEEE Trans. on Circuits and Systems*, vol. 53, pp. 1078–1082, 2006. 104, 115

[140] D. F. Wong and C. L. Liu, "A new algorithm for floorplan design," in *Proc. DAC*, 1986, pp. 101–107. 105

[141] P. Guo, C. Cheng, and T. Yoshimura, "An o-tree representation of non-slicing floorplan and its applications," in *Proc. DAC*, 1999, pp. 268–273. 105

[142] L. Chuan, Z. Hai, and C. Chris, "A revisit to floorplan optimization by lagrangian relaxation," in *Proc. ICCAD*, 2006, pp. 164–171. 105

[143] M. B. Healy, M. Vittes, M. Ekpanyapong, C. Ballapuram, S. Lim, H. Lee, and G. H. Loh, "Microarchitectural floorplanning under performance and thermal tradeoff," in *Proc. DATE*, 2006, pp. 1288–1293. 105

[144] J. Cong, J. Wei, and Y. Zhang, "A thermal-driven floorplanning algorithm for 3d ics," in *Proc. ICCAD*, 2004, pp. 306–313. 105

[145] W. Hung, Y. Xie, N. Vijaykrishnan, C. Addo-Quaye, T. Theocharides, and M. J. Irwin, "Thermal-aware floorplanning using genetic algorithms," in *Proc. ISQED*, 2005, pp. 634–639. 105

[146] H. D. Mogal and K. Bazargan, "Thermal-aware floorplanning for task migration enabled active sub-threshold leakage reduction," in *Proc. ICCAD*, 2008, pp. 302–305. 105, 106

[147] S. N. Adya and I. L. Markov, "Fixed-outlined floorplanning through better local research," in *Proc. IEEE/ACM Int. Conf. on Computer Aided Design*, 2001, pp. 328–334. 106

[148] K. Skadrona, M. R. Stan, K. Sankaranarayanan, W. Huang, S. Velusamy, and D. Tarjan, "Temperature-aware microarchitecture," in *Proc. IEEE International Symposium on Circuits and Systems*, 2003. 106

[149] S. Mukhopadhyay, C. Neau, R. T. Cakici, A. Agarwal, C. H. Kim, and K. Roy, "Gate leakage reduction for scaled devices using transistor stacking," *IEEE Trans. on VLSI*, vol. 11, no. 4, pp. 716–730, 2003. 107

[150] Y. Han, I. Koren, and C. Moritz, "Temperature aware floorplanning," June 2005. 111, 117

[151] D. Brooks, V. Tiwari, and M. Martonosi, "Wattch: A framework for architectural-level power analysis and optimizations," in *Proc. ACM/IEEE International Symposium on Computer Architecture*, June 2000, pp. 83–94. 112

[152] Y. Zhang, D. Parikh, K. Sankaranarayanan, K. Skadron, and M. R. Stan, "Hotleakage: An architectural, temperature-aware model of subthreshold and gate leakage," Mar. 2003. 112

[153] W. Navidi, *Statistics for Engineers and Scientists.* McGraw-Hill, New York, NY, 2006. 112

[154] K. Skadron, "Hybrid architectural dynamic thermal management," in *Proc. DATE*, 2004, pp. 10–15. 115

[155] T. Wang, J. Tsai, and C. Chen, "Power-delivery networks optimization with thermal reliability integrity," in *Proc. ISPD*, 2004, pp. 124–131. 115

[156] V. Nookala, Y. Chen, D. J. Lilja, and S. S. Sapatnekar, "Microarchitecture-aware floorplanning using a statistical design of experiments approach," in *Proc. DAC*, 2005, pp. 579–584. 117

[157] I. A. Ferzli, F. N. Najm, and L. Kruse, "A geometric approach for early power grid verification using current constraints," in *Proc. ICCAD*, 2007, pp. 40–47. 118

[158] Z. Zeng and P. Li, "Locality-driven parallel power grid optimization," *IEEE Trans. on CAD*, vol. 28, no. 8, pp. 1190–1200, 2009. 118

[159] C. Liu and Y. Chang, "Power/ground network and floorplan cosynthesis for fast design convergence," *IEEE Trans. on CAD*, vol. 26, no. 4, pp. 693–704, 2007. 118

[160] J. Singh and S. S. Sapatnekar, "Congestion-aware topology optimization of structured power/ground networks," *IEEE Trans. on CAD*, vol. 24, no. 5, pp. 683–695, 2005. 118

[161] J. Singh and S. Sapatnekar, "A partition-based algorithm for power grid design using locality," *IEEE Trans. on CAD*, vol. 24, no. 5, pp. 664–677, 2006. 118

[162] R. Jakushokas and E. Friedman, "Multi-layer interdigitated power distribution networks," *IEEE Trans. on VLSI*, vol. pp, no. 99, pp. 1–13, 2010. 118

[163] H. Chen, C. Cheng, A. B. Kahng, Q. Wang, and M. Mori, "Optimal planning for mesh-based power distribution," in *Proc. ASPDAC*, 2004, pp. 444–449. 118

[164] M. Zhao, Y. Fu, V. Zolotov, S. Sundareswaran, and R. Panda, "Optimal placement of power supply pads and pins," *IEEE Trans. on CAD*, vol. 25, no. 1, pp. 144–154, 2006. 118

[165] T. Sato, H. Onodera, and M. Hashimoto, "Successive pad assignment algorithm to optimize number and location of power supply pad using incremental matrix inversion," in *Proc. ASPDAC*, 2005, pp. 723–728. 118

[166] Y. Zhong and M. Wong, "Fast placement optimization of power supply pads," in *Proc. ASPDAC*, 2007, pp. 763–767. 118

[167] K. Wang and M. Marek-Sadowska, "On-chip power supply network optimization using multigrid-based technique," in *Proc. DAC*, 2003, pp. 113–118. 118

[168] D. Kouroussis, R. Ahmadi, and F. N. Najm, "Worst-case circuit delay taking into account power supply variations," in *Proc. DAC*, 2004, pp. 652–657. 118

[169] M. Zhao, R. V. Panda, S. S. Sapatnekar, and D. T. Blaauw, "Hierarchical analysis of power distribution networks," *IEEE Trans. on CAD*, vol. 21, no. 2, pp. 159–168, 2002. 119, 120, 121, 125