

Statistical Performance Analysis and Optimization of Digital Circuits

by

Kaviraj Chopra

A dissertation submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy
(Computer Science Engineering)
in the University of Michigan
2008

Doctoral Committee:

Associate Professor David Blaauw, Chair
Assistant Professor Scott Mahlke
Associate Professor Igor L. Markov
Associate Professor Dennis M. Sylvester

To my family and friends for their love and support

ACKNOWLEDGMENTS

The past four years of graduate school have been extremely educational for me, not only as a student of computer engineering, but as a person as well. For this, I would earnestly like to thank my adviser, David Blaauw. It has been his understanding and faith that have helped me grow over the past four years, encouraging my good habits and inspiring me to address my flaws. I would like to thank him, not only for his invaluable guidance but also for changing the way I approach and think about a problem, the result of which is the work in this dissertation. I can only hope that his modesty, humility and attitude towards helping people have somehow been passed onto me over time.

I would also like to thank my committee members, Professor Sylvester , Professor Mahlke and Professor Markov for taking time from their already busy schedules by agreeing to serve on my committee and providing valuable inputs to this work. I would also like to thank Professor Sylvester for advising me on numerous discussions and useful advice on most of the problems addressed in this work.

My earnest acknowledgments go to my lab-mates Aseem Agarwal, Sanjay pant, Visvesh Sathe and Ashish Srivastav for their invaluable help and inputs during several discussions and implementations of thsi work. Many thanks also go to the ACAL staff who often helped me with loads of administrative work. My years in Ann Arbor have been memorable mostly because of the people I encountered and the friends I have made. I would particularly like to thank Ashish, Pant, Sathe, Ravi, Shantanu, Carlos, Prashant and Vivek for their encouragement, understanding and support.

CONTENTS

DEDICATION	ii
ACKNOWLEDGMENTS	iii
LIST OF FIGURES	vii
LIST OF TABLES	x

CHAPTER

1 Introduction	1
1.1 Sources of Performance Variation	2
1.1.1 Process, Environmental and Model Uncertainties	2
1.1.2 Sources of Process Variation	5
1.2 Statistical Timing: Literature Review	10
1.2.1 Background: Static Timing	10
1.2.2 SSTA Problem Formulation	13
1.2.3 Challenges in SSTA	17
1.2.4 SSTA Solution Approaches	22
1.3 Block-Based SSTA	25
1.3.1 Distribution Propagation Approaches (Gate-Delay Space)	25
1.3.2 Dependence Propagation Approaches (Parameter Space)	30
1.4 Key Contribution and Dissertation Outline	33
2 Robust Extraction of Process Variation Model	37
2.1 Variogram Function	40
2.2 Process Variation Extraction	43

2.3	Simulation Results	45
3	A New Statistical Maximum Operation for Propagating Skewness in SSTA	49
3.1	Modeling Skewness	51
3.2	Skew-Normal Max Operation	55
3.2.1	Applying Skew-Normal Max to SSTA	59
3.3	Simulation Results	60
4	Parametric Yield Maximization using Gate Sizing	67
4.1	Review of Yield Analysis	69
4.2	Gradient Computation	72
4.2.1	Timing Perturbation Computation	72
4.2.2	Power Perturbation Computation	77
4.3	Yield Gradient	78
4.4	Simulation Results and Implementation Details	80
4.4.1	Runtime Comparisons	81
4.4.2	Yield Optimization	81
5	A Statistical Approach for Full-chip Gate-oxide Reliability Analysis	85
5.1	Modeling Thickness Variation	87
5.2	Problem Formulation	89
5.3	Computing Oxide Reliability	92
5.3.1	Reliability Function of One Chip	92
5.3.2	Overall Reliability Function	94
5.4	Simulation Results	97
6	Conclusion	102
6.1	Modeling: Extraction of Process Variation Model	102
6.2	Analysis: Skew-normal Maximum Operation	103
6.3	Optimization: Cut-set based Joint Timing and Power Yield Maxi- mization	103

6.4	Reliability: A Statistical Approach for Full-chip Gate-oxide Reliability Analysis	104
6.5	Future Directions	104
	BIBLIOGRAPHY	106

LIST OF FIGURES

Figure

1.1	Steps of the design process and their resulting timing uncertainties.	2
1.2	Physical parameter variations resulting in electrical parameter variations, which in turn, result in circuit delay variations.	6
1.3	Taxonomy of process variations.	10
1.4	An example circuit in (a) and its timing graph in (b).	14
1.5	The timing elements of a sequential circuit path (a) and its timing graph (b).	15
1.6	The probability distribution function and cumulative distribution function	16
1.7	Non-normal CD distribution due to nonlinear dependence of CD on DOF.	19
1.8	Skewness due to nonlinear maximum operation.	21
1.9	Shift with scaling and grouping techniques to perform convolution of input and gate-delay PDFs to compute the output delay PDF.	27
1.10	The upper bound of a delay CDF provides a conservative estimate of circuit delay for a given timing yield.	28
1.11	The principal components of two positively correlated random variables.	31
1.12	Modeling spatial correlation using Quad-Tree partitioning. The numbering of regions in different levels is as shown in the figure. A region (i, j) intersects the regions $(i+1, 4j-3) - (i+1, 4j)$	32
2.1	A typical Variogram in 1-dimension	41
2.2	A typical anisotropic Variogram in 2-dimension	42
2.3	Proposed algorithm for process variation extraction	44

2.4	The Variogram of measured gate length data	46
3.1	Examples for Asymmetric PDFs	50
3.2	The γ parameter of $f_\gamma(x)$ vs. Skewness Sk_γ	54
3.3	Comparison between skew-normal distribution and Normal distribution for a typical Monte Carlos based Arrival Time distribution.	54
3.4	Standard deviations of a bivariate skew-normal distribution and seven re- gions of integration for $\mu_x > \mu_y$	56
3.5	Example: Input X PDF.	62
3.6	Example: Input Y PDF.	62
3.7	Example: Result $\max(X, Y)$ PDF.	62
3.8	Comparison of standard deviation σ_{\max} error (%) as a function of input arrival time skewness Sk_x	62
3.9	Comparison of standard deviation σ_{\max} error (%) as a function of $\frac{\mu_{X\gamma} - \mu_{Y\gamma}}{a}$	63
3.10	Comparison of error in correlation coefficient calculation as a function of input arrival time skewness Sk_x	64
4.1	Transformation of the feasible region from (a) to (b) under the transforma- tion expressed in (21) for negative values of correlation.	71
4.2	A timing graph showing a linear topological ordering for the nodes and cut-sets for nodes 8 and 9	73
4.3	A timing graph showing the ATSet (nodes within the shaded ellipse) and cut-set source set (nodes within the dashed shape) for node 8	75
4.4	Pseudo-code for the CutSetSSTA routine	76
4.5	Pseudo-code for the computation of the yield gradient	79
5.1	Oxide-thickness distribution of one sample die	91
5.2	Compact representation of oxide-thickness variation of ensemble of all die	93
5.3	Oxide-thickness distribution of ensemble of all die	97

5.4	Comparison between Reliability function estimation result, monte carlo simulations, worst case oxide-thickness and best-case oxide-thickness	99
5.5	Oxide-thickness distribution generated using MC simulations for 8 samples of Design C (100K gates, Grid size 20×20 for different values of normalized ρ_{dist}	99
5.6	Comparison between monte carlo simulations of Reliability function estimation assuming actual spatial correlation, perfect correlation and zero correlation	101

LIST OF TABLES

Table

2.1	Accuracy and Robustness of proposed approach	47
3.1	Comparison of maximum and average error of standard deviation between Clark's max and skew-normal max over all the max operations	65
4.1	Comparison of Yield gradient computation using FASTYIELDGRADIENT AND BRUTE-FORCE APPROACH.	82
4.2	Yield Optimization Results	84
5.1	Accuracy comparison between proposed approach and MC Simulations for different correlation distance	100
5.2	Accuracy comparison between proposed approach and MC Simulations for different grid resolution for design B	100

CHAPTER 1

Introduction

Aggressive technology scaling has been the driving force that has enabled the design of CMOS integrated circuits in the nanometer regime. Advancements in CMOS technology have led to the development of complex information systems but a number of design and manufacturing issues are emerging that may limit technology scaling. For example, process control precision is worsening with continuous technology scaling due to smaller dimensions, smaller number of doping atoms and aggressive lithographic techniques. Consequently, loss of predictability in CMOS technology at nanometer integration scales has become a fundamental concern due to two main factors. First, *variability* in manufacturing process leads to uncertainty in electrical characteristics of circuit components resulting in chip performance variation. Traditionally corner based analysis has been used to guard against yield loss resulting from these variations; however, with increasing number of sources of variation, corner based methods are becoming overly pessimistic and computationally expensive. Second, device scaling accelerates the wear-out mechanisms, such as oxide breakdown and negative bias temperature instability (NBTI), which occur over a parts lifetime. Furthermore, the existence of multiple sources of uncertainty causes a loss in predictability of device parameters that degrades the reliability of increasingly complex VLSI circuits. It has therefore become imperative to model, analyze and optimize the effects of uncertainty for reliable operation to avoid the large costs associated with guardbanding VLSI circuits. In this chapter, we give a brief overview of the sources, taxonomy and impact of variability on performance analysis with emphasis on timing. We also review

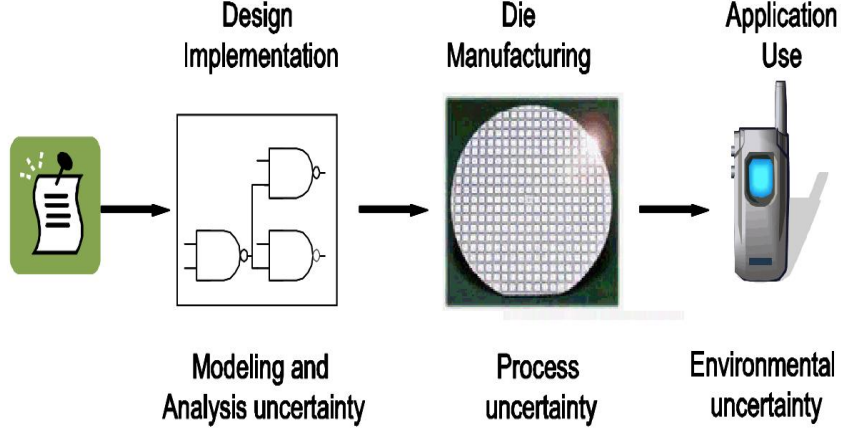


Figure 1.1. Steps of the design process and their resulting timing uncertainties.

in detail the start of the art and discuss open challenges that serves as a motivation for the research work proposed in this manuscript.

1.1 Sources of Performance Variation

In this section, we discuss the key sources of variation in timing prediction that make timing analysis a challenging task for nano-scale digital circuits. We first discuss different types of uncertainties that arise as a design moves from specification to implementation and final operation in the field. We then focus on process variations in more detail and discuss the distinction between die-to-die and within-die variations and the source of so-called spatial correlations. Finally, we discuss the impact of different types of process variations on the timing of a circuit.

1.1.1 Process, Environmental and Model Uncertainties

The uncertainty in the timing estimate of a design can be classified into three main categories.

- Modeling and analysis errors - inaccuracy in device models, in the extraction and

reduction of interconnect parasitics and in the timing analysis algorithms.

- Manufacturing variations - uncertainty in the parameters of a fabricated devices and interconnects from die-to-die and within a particular die.
- Operating context variations - uncertainty in the operating environment of a particular device during its lifetime, such as temperature, supply voltage, mode of operation and lifetime wear-out.

To illustrate each of these uncertainties, consider the stages of design, from initial specification to final operation, as illustrated in Figure 1.1. The design process starts with a broad specification of the design and then goes through several implementation steps, such as logic synthesis, buffer insertion, and place-and-route (P&R). At each step, timing analysis is used to guide the design process. However, timing analysis is subject to a host of inaccuracies, such as undetected false paths, cell delay error, error in interconnect parasitics, SPICE models, etc. These modeling and analysis errors result in a deviation between the expected performance of the design and its actual performance characteristics. For instance, the STA tool might utilize a conservative delay noise algorithm resulting in certain paths operating faster than expected.

In the next stage, the design is fabricated and each individual die incurs additional manufacturing related variations due to equipment imprecisions and process limitations. Finally, a manufactured die is used in an application such as a cell-phone or a laptop. Each particular die then sees different environmental conditions, depending on its usage and location. Since environmental factors such as temperature, supply voltage and work load affect the performance of a die, they give rise to the third class of uncertainty. To achieve the required timing specification for all used die throughout their entire lifetime, the designer must consider all three sources of uncertainty. However, a key difference between the three classes of uncertainty is that each has a sample space that lies along a different dimension. Hence, each class of uncertainty calls for a different analysis approach.

First, we recall that the sample space of an experiment or a random trial is the set of all possible outcomes. The timing uncertainty caused by modeling and analysis errors has as its sample space the set of design implementations resulting from multiple design attempts. Each design attempt results in an implementation that triggers particular inaccuracies in the models and tools resulting in a timing distribution across this sample space. However, a design is typically implemented only once and there needs to be a high level of confidence that the constraints will be met in the first attempt. Hence, the designer is interested in the worst-case timing across this sample space. Thus, margins are typically added to the models to create sufficient confidence that they are conservative, and will result in a successful implementation. Although a statistical analysis of model and analysis uncertainty is uncommon, it could aid in a more accurate computation of the delay with a specified confidence level.

In the case of process variations, the sample space is the set of manufactured die. In this case, a small portion of the sample space is allowed to fail the timing requirements since those die can be discarded after manufacturing. This relaxes the timing constraints on the design considerably and allows designers to significantly improve other performance metrics, such as power dissipation. In microprocessor design, it is common to perform so-called *binning* where die are targeted to different applications based on their performance level. This lessens the requirement that all or a very high percentage of the sample space meets the fastest timing constraint. Instead, each performance level in the sample space represents a different profit margin and the total profit must be maximized.

The sample space of environmental uncertainty is across the operational life of a part and includes variations in temperature, modes of operation, executed instructions, supply voltage, lifetime wear out, etc. Similar to model and analysis uncertainty, the chip is expected to function properly throughout its operational lifetime in all specified operating environments. Even if a design fails only under a highly unusual environmental condition, the percentage of parts that will fail at some point during their operational life can still be

very high. Therefore, a pessimistic analysis is required to ensure a high confidence of correct operation through out the entire life time of the part. Naturally, this approach results in a design that operates faster than necessary for much of its operational life, leading to a loss in efficiency. For instance, when a part is operating at a typical ambient temperature, the device sizing or supply voltage could be relaxed reducing power consumption. One approach to address this inefficiency is to use runtime adaptivity of the design[1][2].

Since each of the three discussed variabilities represent orthogonal sample spaces, it is difficult to perform a combined analysis in a meaningful manner. Environmental uncertainty and uncertainty due to modeling and analysis errors are typically modeled using worst-case margins, whereas uncertainty in process is generally treated statistically. Hence, most SSTA research, as well as this chapter, focus only on modeling process variations. However, the accuracy gained by moving from DSTA to SSTA methods must be considered in light of the errors that continue to exist due to the other sources of timing error, such as analysis and modeling error, uncertainty in operating conditions and lifetime wear-out phenomena. Below, we discuss the sources of process variation in more detail.

1.1.2 Sources of Process Variation

Physical parameters, electrical parameters and delay variation

The semiconductor manufacturing process has become more complex while process control precision is struggling to maintain relative accuracy with continued process scaling. As a result, a number of steps throughout the manufacturing process are prone to fluctuations. These include effects due to chemical mechanical polishing (CMP), which is used to planarize the oxides surrounding metal lines, optical proximity effects (OPE), which are a consequence of patterning features smaller than the wavelength of light [3, 4, 5], and lens imperfections in the optical system. These, as well as numerous other effects, cause variation of device and interconnect physical parameters such as gate length (or Critical Dimension - CD), gate-oxide thickness, channel doping concentration, interconnect thickness and

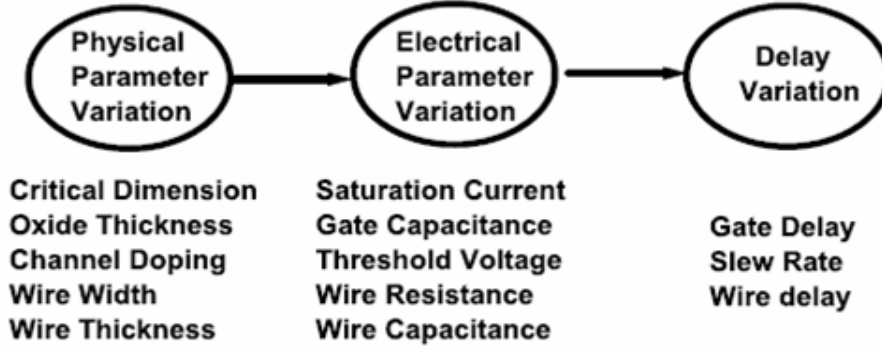


Figure 1.2. Physical parameter variations resulting in electrical parameter variations, which in turn, result in circuit delay variations.

height, etc., as shown in Figure 1.2. Among these, CD variation and channel doping fluctuations have typically been considered dominant factors. However, many SSTA methods model a much wider range of physical parameters. Variations in these physical parameters, in turn, result in variations in electrical device characteristics, such as the threshold voltage, drive strength of transistors, and the resistance and capacitance of interconnects. Finally, the variations in electrical characteristics of circuit components result in delay variations of the circuit.

It is important to note that more than one electrical parameter may have a dependence on a particular physical parameter. For example, both resistance and capacitance of an interconnect are affected by variation in wire width. An increase in interconnect width reduces the separation between wires resulting in an increased coupling capacitance, while decreasing the resistance of the wire. Similarly, perturbations in the gate oxide thickness influence the drive current, the threshold voltage and the gate capacitance of the transistors. Dependence of two or more electrical parameters on a common physical parameter gives rise to correlation of these electrical parameters and ignoring this correlation can result in pessimistic results. For instance, if we ignore the negative correlation between capacitance and resistance, there is a non-zero probability that both resistance and capacitance are at their worst cases values. However, this is physically impossible and leads to unrealistic RC delay estimates. In [6], the authors present a method to determine the process parameter

values that result in a more realistic worst-case delay estimate.

Along similar lines, a particular equipment variation can impact multiple physical parameter values, resulting in a correlation of the physical parameters themselves. For instance, consider the physical parameter variations due to lens aberration. If multiple masks are illuminated with the same lens, the variation of all metal layers and even polysilicon will be correlated.¹ In Section 1.3, we will discuss methods for modeling correlated parameters using a smaller number of independent parameters, such as principal component analysis.

It would be ideal to model each process step in the manufacturing process to determine the variations and correlations in the physical parameters. However, such an analysis is complex and impractical due to the number of equipment related parameters in each fabrication step and the total number of steps. Hence, most SSTA approaches have taken the physical parameters themselves (such as CD, doping concentration and oxide thickness) to be the basic random variables (RVs). These variables are either assumed to be independent or to have well understood correlations.

Classification of Physical Parameter Variation

Physical parameter variations can be classified based on whether they are deterministic or statistical and based on the spatial scale over which they operate, as illustrated in Figure 1.3.

1 - Systematic variations are components of physical parameters variation that follow a well understood behavior and can be predicted upfront by analyzing the designed layout. Systematic variations arise in large part from optimal proximity effects, CMP and its associated metal fill. These layout dependent variations can be modeled pre-manufacturing by performing a detailed analysis of the layout. Therefore, the impact of such variations can be accounted for using deterministic analysis at later stages of the design process [7, 8] and particularly at timing sign-off. However, since we do not have layout information early in

¹Multiple scanners may be used to manufacture a particular part. This can reduce the here discussed correlation, but will not eliminate it.

the design process, it is common to treat these variations statistically. In addition, the models required for analysis of these systematic variations are often not available to a designer, which makes it advantageous to treat them statistically, particularly when it is unlikely that all effects will assume their worst case values.

2 - *Non-Systematic or Random* variations represent the truly uncertain component of physical parameter variations. They result from processes that are orthogonal to the design implementation. For these parameters, only the statistical characteristics are known at design time and hence, they must be modeled using RVs throughout the design process.

Line-edge roughness (LER) and random dopant fluctuations (RDF) are examples of non-systematic, random sources of variation.

It is common that earlier in the design flow both systematic and non-systematic variations are modeled statistically. As we move through the design process and more detailed information is obtained, the systematic components can be modeled deterministically, if sufficient analysis capabilities are in place, thereby reducing the overall variability of the design.

Spatial Reach of Variations

Non-systematic variations can be further analyzed by observing that different sources of variations act on different spatial scales. Some parameters shift when the equipment is loaded with a new wafer or between processing one lot of wafers to the next - this can be due to small, unavoidable changes in the alignment of the wafers in the equipment, changes in the calibration of the equipment between wafer lot processing, etc. On the other hand, some shift can occur between the exposure of different reticles on a wafer, resulting in reticle-to-reticle variations. A reticle is the area of a wafer that is simultaneously exposed to the mask pattern by a scanner. The reticle is approximately $20\text{mm} \times 30\text{mm}$ and will typically contain multiple copies of a the same chip layout or multiple different chip layouts. At each exposure, the scanner is aligned to the previously completed process steps, giving rise to a variation in the physical parameters from one reticle to the next. Finally,

some shift can occur during the reticle exposure. For instance, a shift in a parameter, such as laser intensity, may occur while a particular reticle is scanned, leading to within-reticle variations. Another example is non-uniform etch concentration across the reticle leading to variation in the CD.

These different spatial scales of variation give rise to a further classification of non-systematic variations into two categories.

- Die-to-die variations (also referred to as global or inter-die variations) affect all the devices on the same die in the same way. For instance, they cause the CD of all devices on the same chip to be larger or smaller than nominal. We can see that die-to-die variations are the result of shifts in the process that occur from lot-to-lot, wafer-to-wafer, reticle-to-reticle, and across a reticle, if the reticle contains more than one copy of a chip layout.
- Within-die variations (also referred to as local or intra-die variations) affect each device on the same die differently. In other words, some devices on a die have a smaller CD while others devices on the same die have a larger CD than nominal. Within-die variations are only caused by across-reticle variations within the confines of a single chip layout.

Finally, within-die variations can be categorized into spatially correlated and independent variations, as discussed below.

- *Spatially correlated variations.* Many of the underlying processes that give rise to within-die variation change gradually from one location to the next. Hence, these processes tend to affect closely spaced devices in a similar manner, making them more likely to have similar characteristics than those placed far apart. The component of variation that exhibits such spatial dependence is known as spatially correlated variation. We discuss the modeling of spatial correlated device parameters in more detail in Section 1.3.2.
- *Independent variations.* The residual variability of a device that is statistically inde-

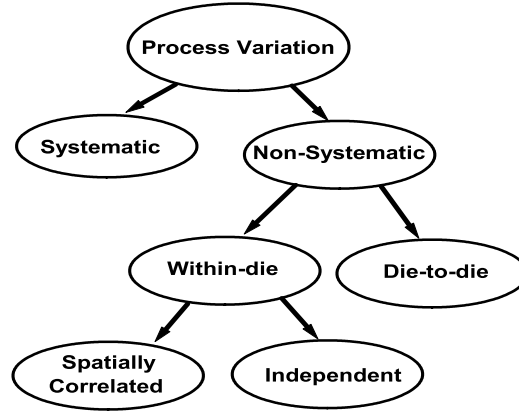


Figure 1.3. Taxonomy of process variations.

pendent from all other devices and does not exhibit spatially dependent correlations is referred to as independent variation.² These variations include effects such as random dopant fluctuations (RDF) and line edge roughness (LER). It has been observed that with continued process scaling the contribution of independent, within-die variation is increasing. Models such as those of Pelgrom [9], which express the amount of independent variation as a function of nominal device parameters, are gaining increased importance.

1.2 Statistical Timing: Literature Review

1.2.1 Background: Static Timing

Since the early 1990s, static timing analysis has been a widely adopted tool for all facets of VLSI chip design. Static timing analysis (STA) is not only the universal timing sign-off tool, but also lies at the heart of numerous timing optimization tools. The main advantage of STA over vector based timing simulation is that it does not rely on input vectors, which can be difficult to construct and can easily miss an obscure performance-limiting path in the circuit. The wide spread use of STA can be attributed to several factors: 1) The basic STA

²In the SSTA literature this independent component of non-systematic process variation is often referred to as the *random* component. However, this is an unfortunate misnomer since all non-systematic variations are random.

algorithm is linear in runtime with circuit size, allowing analysis of designs in excess of 10 million instances. As discussed in Section 1.2, the propagation of arrival times through the combinational portion of a circuit using the CPM algorithm has a run time that is linear with circuit size. However, industrial STA tools often include methods for common path removal in the clocking network and for false path elimination. These methods have a higher run time complexity than the simple CPM algorithm. 2) The basic STA analysis is conservative in the sense that it will over-estimate the delay of long paths in the circuit and under-estimate the delay of short paths in the circuit. This makes the analysis "safe", guaranteeing that the design will function at least as fast as predicted and will not suffer from hold-time violations. 3) The STA algorithms have become fairly mature, addressing critical timing issues such as interconnect analysis, accurate delay modeling, false or multi-cycle paths, etc. 4) Delay characterization for cell libraries is clearly defined, forms an effective interface between the foundry and the design team, and is readily available.

Traditional STA tools are deterministic and compute the circuit delay for a specific process condition. Hence, all parameters that impact the delay of a circuit, such as device gate length and oxide thickness, as well as the operating voltage and temperature are assumed to be fixed and are uniformly applied to all the devices in the design. In this work, we refer to traditional, deterministic STA as DSTA. In DSTA, process variation is modeled by running the analysis multiple times, each at a different process condition. For each process condition, a so-called *corner file* is created that specifies the delay of the gates at that process condition. By analyzing a sufficient number of process conditions, the delay of the circuit under process variation can be bounded.

The fundamental weakness of DSTA is that while global shifts in the process (referred to as die-to-die variations) can be approximated by creating multiple corner files, there is no statistically rigorous method for modeling variations across a die (referred to as within-die variations).³ However, with process scaling progressing well into the nano-meter regime,

³While the deterministic model of gate-delay as used in DSTA excludes a statistical treatment of across

process variations have become significantly more pronounced and within-die variations have become a non-negligible component of the total variation. We will show later in this chapter that the inability of DSTA to model within-die variation can result in either an over- or under-estimate of the circuit delay, depending on the circuit topology. Hence, DSTA's desirable property of being conservative may no longer hold for certain circuit topologies while at the same time, DSTA may be overly pessimistic for other circuit topologies. The accuracy of DSTA in an advanced processes is therefore a serious concern.

In addition to the growing importance of within-die process variations, the total number of process parameters that exhibit significant variation has also increased [10]. Hence, even the modeling of only die-to-die variations in DSTA now requires an untenable number of corner files. For instance, in addition to device parameters, interconnect parameters must be considered, and which combination of interconnect and device parameters results in the worst-case (or best-case) delay often depends on the circuit structure. In an attempt to capture the worst-case die-to-die variation for all cases, the number of corner files used in industry has risen sharply. It is now common to use more than a dozen corner files [11], while the number can even exceed over one hundred, thereby increasing the effective runtime of DSTA by one order of magnitude or more. Recently, a linear-time approach for STA which computes timing upper bounds across all process corners in a single pass was proposed in [12].

The need for effective modeling of process variations in timing analysis has led to extensive research in statistical static timing analysis (SSTA). Some of the initial research dates back to the very introduction of timing analysis in the 1960s [13] as well as the early 1990s [14, 15]. However, the vast majority of research on SSTA dates from the last five years, with well over a hundred papers published in this research field since 2001.

die variation, industry tools have over time developed a number of methods to approximate the impact of such variations. A common method is to use a predetermined delay scaling factor for all circuit elements (delay is increased for long path analysis and decreased for short path analysis). However, if the scaling factor is set to the worst-case within-die variation, the analysis becomes exceedingly pessimistic. On the other hand, lesser values cannot be proved to be conservative, negating one of the major advantages of DSTA.

In this chapter, we give a brief review of the different issues and approaches to SSTA. In section 1.1.2, we examine the different sources of uncertainty and their impact on circuit performance. In section 1.2, we present the formulation of the SSTA problem and discuss its key challenges and approaches. In Section 1.3, we review the so-called "block based" approaches in more detail and present their strengths and weaknesses.

1.2.2 SSTA Problem Formulation

The traditional DSTA procedure abstracts a timing graph from a combinational circuit. The nodes of the timing graph represent primary inputs/outputs of the circuit and gate input/output pins. Its edges represent the timing elements of the circuit, namely, the gate input-pin to output-pin delay and wire delay from a driver to a receiver, as shown in Figure 1.4. The weight on these edges represents the delay of the corresponding timing element. For a combinational circuit, it is convenient to connect all primary inputs to a virtual source node with virtual edges having weight equal to the input arrival-times. Similarly, all the primary outputs are connected to a virtual sink node through virtual edges with weights representing the required arrival-times. The resulting timing graph therefore has a single source and sink node.

A similar timing graph can be constructed for sequential circuits. Figure 1.5 illustrates the additional timing elements pertaining to a clock network (i.e the launch and capture paths of the clock tree) and the sequential elements. In the corresponding timing graph the virtual source node corresponds to the input driver of the on-chip clock network. The clock driver delays and interconnect delays on the launch path, the clock-to-q delay and set-up times of the sequential elements are again modeled using weights on their corresponding graph edges. Similarly, the virtual sink node also corresponds to the clock input driver and the capture path is represented with nodes and edges in the timing graph. In this case, however, the weights of edges corresponding to the capture path are assigned negative delay values, as opposed to the positive values for the launch path. Apart from this distinction,

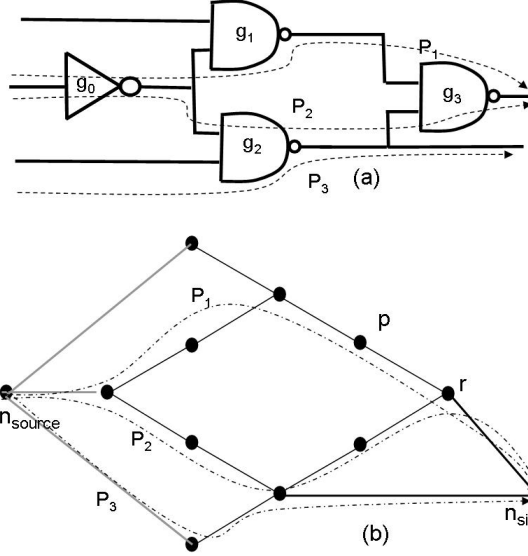


Figure 1.4. An example circuit in (a) and its timing graph in (b).

the timing graphs for flip-flop based sequential circuits are a direct extension of those for the combination circuits and can be analyzed with the same timing algorithms. However, significant complications arise when transparent latches are used in place of flip-flops or when the launch and capture paths of the clock tree share the same drivers, as is common.

As discussed in Section 1.1.2, device parameters such as gate length, doping concentration and metal thickness must be treated as RVs due to process variation. The delay of each edge, being a function of extend the concept of the traditional timing graph to a statistical timing graph defined as follows:

Definition: A timing graph $G = \{N, E, n_s, n_f\}$ is a directed graph having exactly one source node n_s and one sink node n_f , where N is a set of nodes and E is a set of edges. The weight associated with an edge corresponds to either the gate delay or interconnect delay. The timing graph is said to be a statistical timing graph if i^{th} edge weight d_i is a random variable.

The arrival time at the source node of the timing graph typically has a deterministic zero value. This reflects the fact that in combination timing graphs clock tree skew is not represented, while in sequential circuits the source node is pulled back to a common point

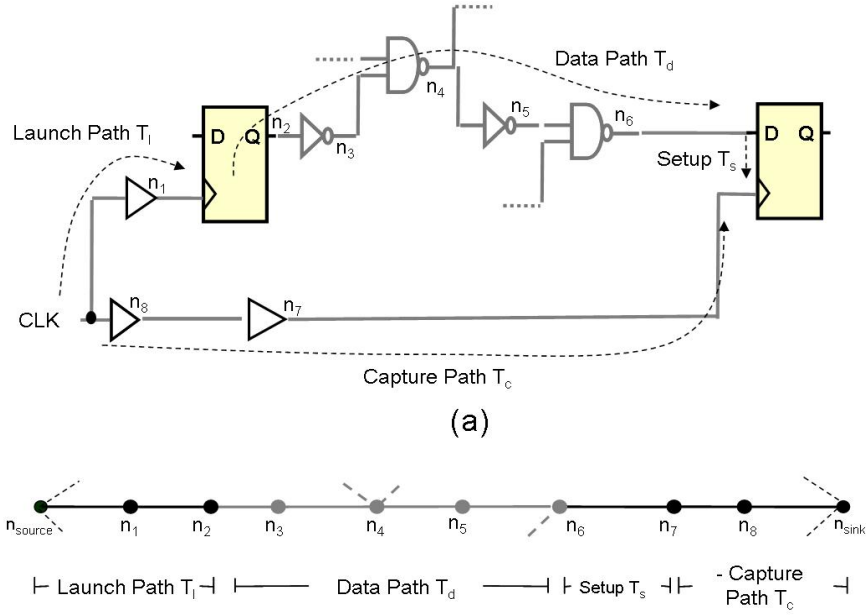


Figure 1.5. The timing elements of a sequential circuit path (a) and its timing graph (b).

on the launching and capturing clock paths⁴ In traditional DSTA, the most basic goal of the analysis is to find the maximum delay between the source node and the sink node of a timing graph, which is the delay of the longest path in the circuit. When modeling process induced delay variations, the sample space is the set of all manufactured dies. In this case, the device parameters will have different values across this sample space and, hence the critical path and its delay will change from one die to the next. Therefore, the delay of the circuit is also a RV and the first task of SSTA is to compute the characteristics of this random variable. This is performed by computing its probability distribution function (PDF) or cumulative distribution function (CDF) (see Figure 1.6). Alternatively, only specific statistical characteristics of the distribution, such as its mean and standard deviation can be computed. Note that the CDF and the PDF can be derived from one another through differentiation and integration. Given the CDF of circuit delay of a design and the required performance constraint, the anticipated yield can be determined from the CDF. Conversely,

⁴Note that a deterministic value at the source node of a sequential timing graph does not account for jitter from the PLL or other sources.

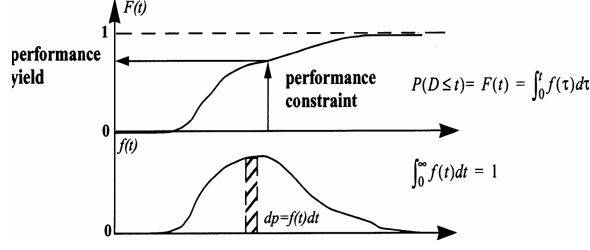


Figure 1.6. The probability distribution function and cumulative distribution function

given the CDF of the circuit delay and the required yield, the maximum frequency at which the set of yielding chips can be operated at can be found.

Definition: Let a path p_i be a set of ordered edges from the source node to the sink node in G and let D_i be the path length distribution of p_i , computed as the sum of the weights d for all edges k on the path. Finding the distribution of $D_{max} = \text{maximum}(D_1, \dots, D_i, \dots, D_{npaths})$ among all paths (indexed from 1 to n paths) in the graph G is referred to as the problem of statistical static timing analysis (SSTA) of a circuit.

Similar to traditional DSTA, we can formulate the statistical timing analysis problem as that of finding the latest arrival-time distribution at the sink node in the timing graph [16, 17]. The latest arrival-time distribution at the sink node can be found by propagating the arrival-time from the source node through the timing edges while computing the latest arrival-time at every node in topological order. Subsequently, the latest arrival-time distribution at the sink node is the circuit delay distribution. It is worth noting that the basic DSTA algorithm is based on the project planning technique known as the critical path method (CPM) and involves a simple breadth first traversal [18]. Likewise, the basic SSTA formulation for circuit designs was first motivated from the project evaluation and review technique (PERT) literature [19, 13]. In a typical PERT problem, a project is specified by precedence relations among tasks and task durations are independent random variables with discrete, finite ranges. However, in contrast to DSTA, PERT is shown to be an NP-complete problem [20].

In addition to the problem of finding the delay of the circuit, which we have posed as

the basic SSTA problem, it is also key to improve this delay when the timing requirements are not met. Hence, DSTA methods typically report the slack at each node in the circuit, in addition to the circuit delay and critical paths. The slack of a node is the difference between the latest time a signal can arrive at that node such that the timing constraints of the circuit are satisfied (referred to as the *required time*) and the actual latest arrival time of the signal at that node [21]. Similar to the circuit delay, the slack of a node is a random variable in the SSTA formulation. We also do not discuss latch based sequential timing analysis, which involves multiple phase clocks, cycle stealing, clock schedule verification, etc. Methods for statistical sequential timing analysis using latches and clock skew analysis can be found in [22, 23, 24, 25, 26].

1.2.3 Challenges in SSTA

The statistical formulation of timing analysis introduces several new modeling and algorithmic issues that make SSTA a complex and enduring research topic [27]. In this section, we introduce some of these issues as well as the relevant SSTA terminology.

Topological Correlation

So called *re-convergent paths* in a circuit are paths that start with one or more common edges after which the paths separate and join again at a later node, called the re-convergent node. For instance, in Figure 1.4, the two paths P_1 and P_2 share the same first edge (corresponding to gate g_1) and re-converge at the output of gate g_0 (node r). In such a case, the input arrival-times at the re-convergent node become dependent on each other because of the shared edge delay. This dependence leads to so-called *topological correlation* between the arrival-times and complicates the maximum operation at the re-convergent node. To perform accurate analysis, the SSTA algorithm must capture and propagate this correlation so that it is correctly accounted for during the computation of the maximum function.

Spatial Correlation

As discussed in Section 1.1.2, within-die variation of the physical device parameters often exhibits spatial correlation, giving rise to correlation between the gate-delays. Hence, if the gates that comprise two paths have spatially correlated device parameters, they will have correlated path delays. In this way, correlation can be introduced between paths that do not share any common timing edges. For instance, in Figure 1.4, the paths P_1 and P_3 do not share any common delay edges, but if gates g_1 and g_2 are within close proximity on the die, their spatially correlated delays can give rise to correlation between the two path delays. Hence, spatial correlation of the arrival-times must be captured and propagated during SSTA so that it is correctly accounted for during the maximum operation. Spatial correlation also impacts the sum operation. For example, if in Figure 1.4, gates g_1 and g_3 have spatially correlated delays then the arrival-time at node p will be correlated with the delay of gate g_3 .

While topological correlation only affects the maximum operation, spatial correlation affects both the sum operation as well as the maximum operation. This raises two fundamental challenges for SSTA: (i) How to model gate-delays and arrival-times such that the spatial correlation of the underlying device parameters can be expressed. (ii) Given a model of the spatial correlation, how to propagate and preserve the correlation information while performing the sum and maximum operations in SSTA. A common approach to this problem has been to represent delay in terms of the device parameter-space basis, which is common to all gate-delays. This approach is discussed in more detail in Section 1.3.

Non-normal Process Parameters and Non-linear Delay Models

Normal or Gaussian distributions are found to be the most commonly observed distributions for RVs and a number of elegant analytical results exist for them in the statistics literature. Hence, most of the initial work in SSTA assumed normal distributions for physical device parameters, electrical device parameter, gate-delays and arrival-times. However,

some physical device parameters may have significantly non-normal distributions. In this subsection, we discuss the source and impact of such non-normal distributions.

An example of a non-normal device parameter is CD (or gate length), due to variation in depth of focus (DOF). As a result of equipment limitations and non-planarity of the die, the focus point of the exposed image on the die exhibits some amount of variation. This impacts the development of the photo resist layer and consequently impacts the CD of the device. However, both large and small values of DOF result in an underdevelopment of the photo resist and the dependence of CD on DOF is non-linear. Even if the variation of DOF is normal, the CD variation will be decidedly non-normal. As illustrated in Figure 1.7, the PDF of CD is clearly (negatively) skewed and non-normal.⁵

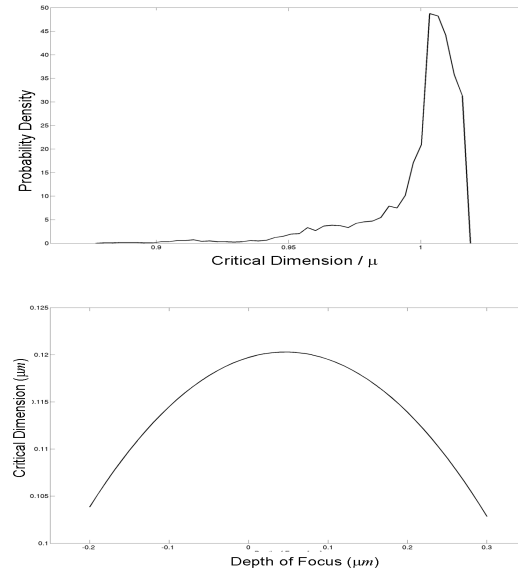


Figure 1.7. Non-normal CD distribution due to nonlinear dependence of CD on DOF.

Even if the physical device parameters are indeed normally distributed (e.g., doping concentration has a normal distribution) the dependence of the electrical device parameters and gate-delay on these physical parameters may not be linear, giving rise to non-normal gate-delays. Initial work in modeling spatial correlations [28, 29, 30] used a first-order

⁵A probability distribution is said to have negative skewness if it has a long tail in the negative direction of the RV, such as the CD distribution shown in Figure 1.7. Conversely, a positive skewness indicates a long tail in the positive direction.

delay model which assumed a linear dependence of the gate-delay on physical device parameters. If the variations are small, this linear approximation is justified as the error introduced by ignoring higher order terms is negligible. However, with reduction of geometries, process variation is becoming more pronounced and the linear approximation may not be accurate for some parameters.

Non-normal delay and arrival-time distributions introduce significant challenges for efficient SSTA. While this is a relatively new area of research, several researchers have proposed approaches to address this issue [31, 32, 33, 34, 35, 36]. Finally, it should be noted that apart from the difficulty of modeling the non-normality of an individual random variable, the dependence between two non-normal RVs is no longer expressed by a simple correlation factor. This further complicates the correct treatment of topological and spatial correlations.

Skewness due to Maximum Operation

Even if gate-delays are assumed to be normal, SSTA has to cope with the fact that the maximum operation is an inherently non-linear function. The maximum of two normal arrival times will result in a non-normal arrival time that is typically positively skewed.⁶ Also, the non-normal arrival-time distribution produced at one node is the input to the maximum computation at downstream nodes. Hence, a maximum operation that can operate on non-normal arrival times is required.

Most of the existing approaches ignore the skewness introduced by the maximum operation and approximate the arrival-times with normal distributions. The error of this normal approximation is larger if the input arrival-times have similar means and dissimilar variances [37]. In other words, the error is most pronounced when two converging paths have nominally balanced path delays, but one path has a tighter delay distribution than the other.

⁶It is possible to obtain much more complex distributions, such as bimodal distributions, even when the input parameters remain normal. While such occurrence are rare, they introduce significant modeling difficulties.

This can occur in a circuit when two paths with equal nominal delay have a different number of gates or the correlation among their gates differ. Another example is when one path is dominated by interconnect delay while the other is dominated by gate-delay.

An example of two such delay distributions is shown in Figure 1.8(a). Intuitively, we can see that RV B will dominate the maximum delay for values greater than its mean, since B has significantly higher probabilities in this range. For delay values below the mean, RV A will dominate. Since A and B have different variance, skew is introduced in their maximum. For two input distributions that have identical means and variances, the resulting maximum exhibits smaller skewness (Figure 1.8(b)). Finally, Figure 1.8(c) shows that if the means of the input distributions are significantly different, the resulting maximum is entirely dominated by one distribution and skew is negligible.

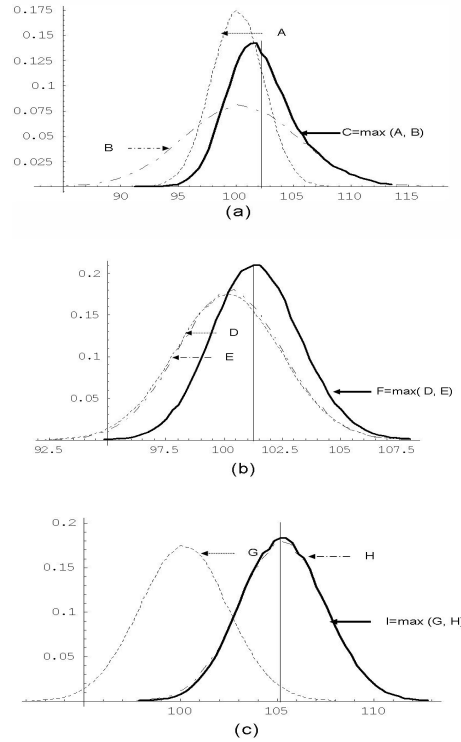


Figure 1.8. Skewness due to nonlinear maximum operation.

The above issues address four basic challenges in SSTA that have received significant attention in the literature. However, many other critical challenges to developing a mature

SSTA tool remain. For instance, availability of statistical data remains difficult.

1.2.4 SSTA Solution Approaches

We now give a brief overview of the principle approaches to SSTA, moving from traditional methods to more recent approaches.

Numerical Integration Method

The simplest SSTA approach follows directly from the problem definition given in Section 1.2. A numerical integration over the process parameter space is performed to compute the yield of the circuit for a particular delay. Typically, the delay of a set of critical paths is expressed as a linear function of the physical device parameters and a feasible region in this parameter space is defined by the desired circuit delay. This region is then numerically integrated, exploring all possible permutations of physical device parameter values that lie in the feasible region. Efficient numerical integration methods were proposed in [38]. The advantage of this method is that it is completely general and process variation with any type of distribution and correlation can be modeled. However, it can be quite expensive in runtime, in particular for balanced circuits with a large number of critical paths.

Monte-Carlo Methods

The second general approach performs a statistical sampling of the sample space using Monte-Carlo simulation, based on the Metropolis sampling algorithm [39]. Instead of enumerating the entire sample space explicitly, the key idea is to identify the regions of significant probability and to sufficiently sample these regions. Using the PDF of the physical device parameters, a number of samples is drawn. For each sample, the circuit delay is computed using traditional DSTA methods. Thereafter, by evaluating the fraction of samples that meet the timing constraint, an estimate of timing yield is found. If a sufficient number of samples is drawn, the estimation error is small. By sweeping the timing constraint and finding the yield for each value, the entire circuit delay distribution can be found.

As with numerical integration, the Monte-Carlo approach has the advantage of being completely general. Furthermore, it is based on existing mature DSTA methods and performs significantly faster than the numerical integration-based methods. However, since DSTA is in the inner-loop of the Monte-Carlo simulation, the runtime can still be significant, especially if a fully featured, industrial DSTA tool is used. Using Monte-Carlo simulation, it is also difficult to perform incremental analysis after a designer makes a small change to the circuit. It has been shown that the performance of Monte Carlos techniques can be improved using methods such as importance sampling [11, 40, 41, 42]. However, more research is required to examine if fast sampling techniques can be effective for SSTA.

Probabilistic Analysis Methods

Both previous approaches are based on sample space enumeration. In contrast, probabilistic methods explicitly model gate-delay and arrival-times with RVs. These methods typically propagate arrival-times through the timing graph by performing statistical sum and maximum operations.⁷ They can be classified into two broad classes: (i) path-based approaches and (ii) block-based approaches. The key difference between the two approaches is where in the algorithm the maximum function is invoked.

Path-based approaches: In path-based SSTA algorithms, a set of paths which are likely to become critical is identified and a statistical analysis is performed over these paths to approximate the circuit delay distribution. First, the delay distribution of each path is found by summing the delay of all its edges. Assuming normal gate-delays, the path delay distribution is normal and can be analytically computed [43, 44, 45]. The overall circuit delay distribution is then found by performing a statistical maximum operation over all the path delays. (discussed in more detail in Section 1.3).

The basic advantage of this approach is that the analysis is clearly split into two parts

⁷

The minimum operation is also needed for the computation of the shortest path, clock skew, and slack computations. However, it can be derived from the maximum operation.

- the computation of path delays followed by the statistical maximum operation over these path delays. Hence, much of the initial research in SSTA was focussed on path based approaches [15, 43, 44, 46, 47, 48, 49]. Clearly, a difficulty with the approach is how to rigorously find a subset of candidate paths such that no path that has significant probability of being critical in the parameter space is excluded. Also, for balanced circuits, the number of paths that must be considered can be very high. Therefore, most of the later research has focussed on the block-based approaches.

One of the methods that fall in the path based category approximates the statistical delay of a circuit early in the design process when the exact gate-level implementation is not yet known [50, 51]. In this work, the circuit is modeled using a set of generic paths whose specifications are provided by the designer. The method also determines a setting of the transistor level electrical parameters, give a specific yield goal. These settings can then be used in a traditional deterministic timing verification flow. The usefulness of applying SSTA methods early in the design process, when exact gate level implementations are not yet available, depends on the relative magnitude of the delay uncertainty introduced by process variations versus the uncertainty due to the undetermined circuit implementation.

Block-based approaches: The block-based methods follow the DSTA algorithm more closely and traverses the circuit graph in a breadth first manner. The arrival-time at each node is computed using two basic operations: (i) For all fan-in edges of a particular node, the edge delay is added to the arrival-time at the source node of the edge using the sum operation and (ii) given these resulting arrival times, the final arrival-time at the node is computed using the maximum operation. Hence, block-based SSTA methods propagate exactly two arrival-times (a rise and a fall arrival-time) at each circuit node resulting in a runtime that is linear with circuit size. The computation of the sum function is typically not difficult; however, finding the statistical maximum of two correlated arrival-times is non-trivial.

Due to its runtime advantage, many current research and commercial efforts have taken

the block-based approach. Furthermore, unlike other approaches, the block-based approach lends itself to incremental analysis which is advantageous for diagnostic/optimization applications. In block-based SSTA methods, the result of the maximum operation performed at one node is the input to the maximum operation which is performed at downstream nodes. It is therefore essential that the sum and maximum operation preserves the correlation information of the arrival-times so that this information is available at later operations. Furthermore, the skewness introduced by the maximum operation must be considered.

1.3 Block-Based SSTA

In this section, we discuss block-based SSTA methods in more detail. The different methods are presented in the order of increasing complexity. We start with simpler early methods that were based on a normal, independent approximation of the arrival-times. We then discuss methods that model topological correlation due to re-convergence of arrival-times. This is followed by a number of methods that account for spatial within-die variations. Finally, we briefly survey more recently proposed non-linear and non-normal block-based methods.

1.3.1 Distribution Propagation Approaches (Gate-Delay Space)

Initial efforts in block-based SSTA approaches focussed on directly representing gate-delays with RVs characterized by their distribution or statistical characteristics. The common technique employed by all these approaches is to explicitly propagate the arrival-time distributions through the timing graph. This is achieved by employing a statistical sum operator to compute the sum of the timing arc delay and the source node arrival-time distribution. In the case of multi-fan-in nodes, a statistical maximum operator is also applied to the arrival-times corresponding to different fan-ins edges of a node.

A basic block-based SSTA algorithm based on a PERT like traversal was first given in [13]. Later, the authors in [52] presented a linear runtime algorithm for propagating mean

and variance of timing variables. In this approach, both gate-delays and latest arrival-time distributions are assumed to be independent normal RVs. Based on these simplifying assumptions, the sum and maximum of arrival-time RVs are computed using analytical results for normal RVs.

In [53], the authors extend this analytical approach to handle topological correlation due to re-convergent paths and correlation between edge delays that correspond to the same gate, at the cost of increased complexity. The approach uses the statistical sum operation to reduce series edges in the timing graph to a single edge. At each step of the reduction the correlation of the reduced edge with the edges with which it has non-zero correlation is recomputed. A similar reduction procedure is then performed for parallel edges using the statistical maximum operation under the normal assumption using the analytical results given in [37]. This maximum operation is explained in more detail in the following subsection. The proposed approach limits the number of edges whose correlation information is stored in memory by identifying those nodes whose correlation information is no longer required. This approach was extended in [54],[55] by assigning a level to each node in the DAG using a depth-first search. The level is used to identify the nodes whose correlation information can be discarded at each stage of the arrival-time propagation.

In [56, 57, 58], the authors propose an alternative, discrete representation for relaxing the normal distribution assumption. The gate-delays are now modeled as discrete delay distributions that are generated by sampling a continuous distribution. Note that the discrete PDFs are re-normalized after sampling to ensure that the sum of the probabilities for the discrete events is equal to 1.

The approach then utilizes discrete sum and maximum operations for arrival-time propagation. In the case of a degenerate or deterministic input delay distribution, the sum operation is simple and the output delay PDF is obtained by simply *shifting* the gate-delay distribution by the input delay. However, in the case where the input delay PDF is non-degenerate, a set of shifted output delay distributions is generated as shown in Figure 1.9.

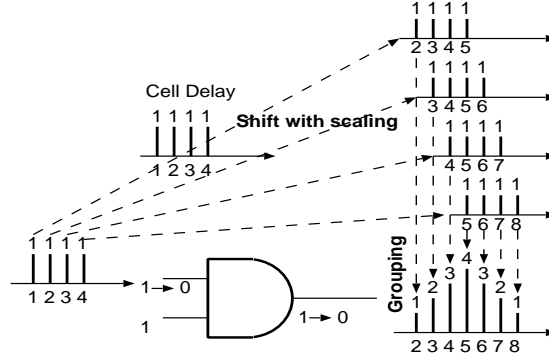


Figure 1.9. Shift with scaling and grouping techniques to perform convolution of input and gate-delay PDFs to compute the output delay PDF.

Each of these shifted PDFs corresponds to a discrete event from the input delay PDF. This set of shifted PDFs is then combined using *Bayes' Theorem* - the shifted PDFs are first *scaled*, where the scaling factor is the probability of the associated discrete input event. The scaled events are then *grouped* by summing the probability at each of the discrete time points. The actual probability of an event can be obtained by dividing the total value for each discrete point of the PDF by the sum of the numbers corresponding to all the events in each discrete PDF. The overall computation can be expressed succinctly as

$$f_s(t) = \sum_{i=-\infty}^{\infty} f_x(i)f_y(i-t) = f_x(t) \star f_y(t) \quad (1.1)$$

where $s = x + y$, and implies that the PDF of the sum of two RVs can be expressed as a convolution of their PDFs.

The statistical maximum is computed using the relation

$$f_z(t) = F_x(t)f_x(t) + F_y(t)f_y(t) \quad (1.2)$$

where $z = \text{maximum}(x, y)$, and f and F represent the PDF and CDF of the RV respectively and x and y are assumed to be independent. The above equation expresses mathematically that the probability that the maximum of two discrete RVs has a value t_0 is equal to the probability that one of the RVs has a value equal to t_0 and the other has a value less than or equal to t_0 .

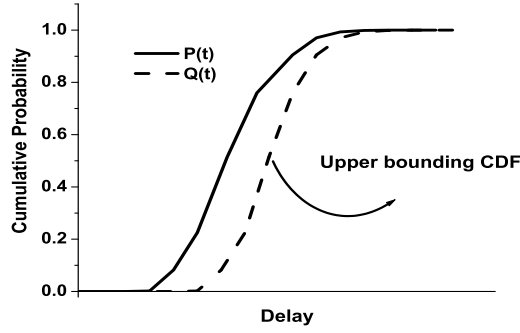


Figure 1.10. The upper bound of a delay CDF provides a conservative estimate of circuit delay for a given timing yield.

For handling topological correlation due to re-convergent paths, a partitioning based approach is used to decompose the circuit into so-called *super-gates*. Each super-gate is a sub-circuits with a single fan-out and one or more inputs, all of which are statistically independent. The discrete events at the inputs of the super-gates are propagated separately and the resulting distributions are finally combined at the output of the super-gate using Bayes' Theorem. The process of propagating each of the discrete events of the PDFs separately is referred to as *enumeration*. Special care has to taken in the case where a multiple fan-out node lies in the fan-out cone of another multiple fan-out node. Unfortunately, the runtime complexity of this algorithm depends on the circuit structure, and is exponential in the worst case.

The authors in [16, 59, 60] extend the work on handling topological correlation while using the same discrete framework for representing PDFs. The authors present an approach to determine the minimum set of nodes that need to be enumerated to handle re-convergence exactly. As expected, the worst case computational complexity of enumeration remains exponential. Nevertheless, the authors show the useful property that ignoring topological correlation results in a distribution that is an upper bound on the exact distribution of the circuit delay. The stochastic upper bound of a delay distribution with CDF $P(t)$ is a distribution whose CDF $Q(t)$ has a value which is always smaller than or equal to $P(t)$ for all values of t , as shown in Figure 1.10. Such an upper bound results in a pessimistic

estimate of the timing yield of the circuit at a given performance target.

Based on this result, the authors developed a linear runtime method for computing both a lower and an upper bound on the exact delay distribution of the circuit. These bounds are then used to obtain an estimate of the circuit delay at a desired confidence point as well as the accuracy of the bounds. In the case when the bounds are not sufficiently close to each other, a heuristic method is used to iteratively improve the bounds using selective enumeration of subset of the nodes. The results presented in [16] showed that performing enumeration at a small set of carefully selected nodes leads to a significant improvement in the quality of the bounds. This is due to the fact that correlation between two arrival-times only impacts the maximum if the two arrival-times have comparable means. Hence, the correlation between two arrival times that are substantially shifted can be ignored without incurring significant error.

In a related work [61], a Bayesian Network based approach for representing the statistical timing graph is presented for handling topological correlations. The Bayesian Network formulation prescribes an efficient method to factorize the joint distribution over the entire timing graph into an optimal set of factors. Though the worst case runtime complexity of such an approach remains exponential, the complexity grows exponentially with the size of the largest clique in the circuit, which, in practice, is found to grow much more slowly than the circuit size.

In [62], the authors model arrival times as CDFs and delays as PDFs. Using a piecewise linear model for CDFs and PDFs, they present a computationally simple expression for performing the sum and maximum operations. Furthermore, they also presented a method for handling re-convergent fanouts using a dependency list associated with every arrival time, which are propagated through the circuit and pruned for increased efficiency. Using error budgeting, an approach to optimize the runtime of this method was presented in [63]. A method to generate device level discrete delay distributions from the underlying process parameter distribution was presented in [64].

1.3.2 Dependence Propagation Approaches (Parameter Space)

In the previous subsection we discussed techniques that consider topological correlations. The next crucial step in the development of block-based SSTA was to account for spatial correlation of the underlying physical device parameters. The basic difference between the two cases is that the correlation among arrival-times now originates from the correlation of the device parameters. Also, an arrival-time at the input of a gate can be correlated with the delay of the gate itself, impacting the sum operation in addition to the maximum operation.

In the distribution propagation approaches, the gate-delays are the basic random variables in the formulation. However, to model the correlation in the physical device parameters, it is necessary to model these device parameters themselves as the basic RVs. The delay of the gates is therefore expressed as a function (linear or non-linear) of the physical or electrical device parameters. It is this functional form, that expresses the dependence of the gate-delays on the device parameters, that is propagated through the circuit. This concept of representing delay dependencies with a parametric delay model was first introduced by [65]. To enable such techniques, it is necessary to develop a model of the spatial correlation of the device parameters. Therefore, we first discuss some of the models for expressing the correlation of device parameters and then show how these can be used to compute the final circuit delay.

Correlation Models

To exactly model spatial correlation between the physical parameters of two devices a separate RV is required for each device. However, the correlation between two device is generally a slow monotonically decreasing function of their separation, decaying over distances of hundreds of microns. Therefore, simplified correlation structures using a grid model [30] or Quad-Tree model [29] have been proposed. These models allow the correlation among gates of the die to be expressed using a smaller set of RVs.

In a grid model the overall die area is divided using a square grid. It is assumed that

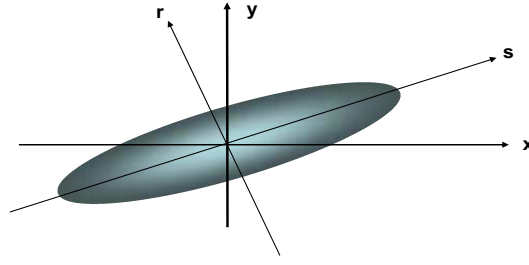


Figure 1.11. The principal components of two positively correlated random variables.

the grid size is chosen such that all gates within a single square on the grid can be assumed to have perfectly correlated spatial variations. Let us now consider the RVs required to model variations in a given process parameter. Each square in the grid corresponds to a RV of a device parameter which has correlations with all other RVs corresponding to the other squares. To simplify the correlation structure of the RVs this set of correlated RVs is mapped to another set of mutually independent RVs with zero mean and unit variance using the *principal components* of the original set of correlated RVs. The original RVs are then expressed as a linear combination of the principal components. These principal components can be obtained by performing an eigenvalue decomposition of the correlation matrix as explained in more detail in [66]

Intuitively, this is illustrated in Figure 1.11 where the distribution of two correlated, jointly normal random variables A and B is shown. In the scatter plot, the X-axis is the value of A while the Y-axis is the value of B . If A and B were independent, the scatter plot would form a perfect circle or a horizontal or vertical oval. The diagonal distribution shown indicates positive correlation between A and B since large values of A tend to correspond to large values of B . The principle component analysis (PCA) method expresses A and B using two new RVs C and D , using the rotated axes r and s . RVs A and B can be expressed using a linear combination of C and D . Furthermore, the rotation of r and s ensures that C and D are independent.

It is important to note that constructing the correlation matrix directly from a distance

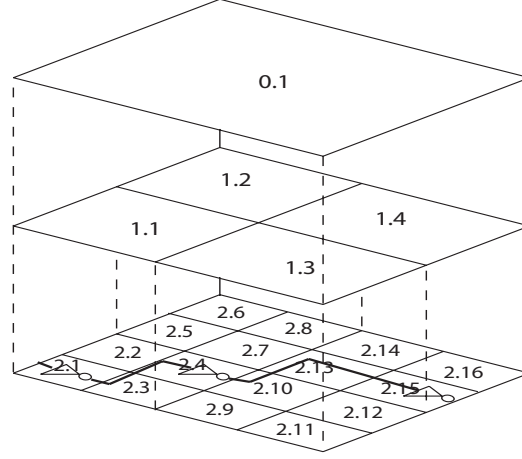


Figure 1.12. Modeling spatial correlation using Quad-Tree partitioning. The numbering of regions in different levels is as shown in the figure. A region (i, j) intersects the regions $(i+1, 4j-3) - (i+1, 4j)$.

based correlation function may result in a non-positive-definite matrix. Furthermore, the correlation matrix must be positive definite and that this condition may be violated if the matrix is constructed from an arbitrary distance based function or from measured data (especially if noisy). Hence, some techniques, such as replacing all negative eigenvalues by zero, may need to be used. This problem has been investigated in [67, 68, 69, 70, 71, 72].

The Quad-Tree model proposed in [44, 29] also partitions the overall die area into a number of regions. However, instead of using principle component analysis to express the correlated components of variations, it uses an additive approach to consider the spatial dependence of process parameters. This is achieved by recursively dividing the area of the die into four equal parts, which is known as *Quad-Tree Partitioning*. As the regions of the die are recursively divided into parts, the number of parts at each level increase by a factor of four as shown in Figure 1.12. Each partition, at all levels of the quad tree, is assigned an independent RV. The spatially correlated variation associated with a gate is then defined to be the sum of the RV associated with the lowest level partition that contains the gate and the RVs at each of the higher partitioning levels that intersects the position of the gate. The correlation among gates arises from the sharing of some of the RVs at higher levels of the Quad-Tree. The number of shared RVs depends on the distance between the two gates. Moreover, a larger fraction of the variance can be allocated at higher levels of the

Quad-Tree if the parameter is known to be more strongly correlated over large distances. In [73], a method for determining the values of the RVs associated with each partition is presented.

An alternative grid based model was proposed in [34] where only four RVs are used to model the correlation structure. The four RVs are assumed to be associated with the four corners of a die and the RVs associated with the gates inside the design are represented as a weighted sum of these four RVs, where the weighting coefficients are functions of the distance between the position of a gate and each of the four corners of the die.

In [74], the authors use the *Karhunen-Loeve expansion* (KLE) to express the distance based correlation function in terms of a set of RVs and an orthonormal set of deterministic functions related to the position of a gate on the die. This allows the correlation to be expressed as a continuous function of the location of a gate. In addition, the authors show that KLE provides much greater accuracy as compared to PCA of a grid model [30], or equivalently, it provides similar accuracy with a reduction in the number of RVs.

1.4 Key Contribution and Dissertation Outline

In this research, we have developed four techniques focusing on analysis, modeling and optimization while addressing variability. A detailed background and previous work relevant to each of the following four solutions is given in the subsequent four chapters:

- **Robust Extraction of Process Variation Model.** Aggressive device scaling has made it imperative to account for process variations in the design flow. A robust model of process variations is an essential requirement for any meaningful variation aware design analysis and optimization. Unfortunately, the previous approaches on extracting spatial correlation function assume ergodicity and isotropy while estimating the inter-die(global) component of variation and spatial correlation function, respectively. We find that making such simplifying assumptions may result in significant estimation error. In Chapter 2, addressing these issues, we propose an alternative approach to

extract spatial variation models based on the theory of spatial statistics. The proposed approach uses the concept of variogram function that represents how parameters can co-vary as a function of spatial distance. The variogram function provides us with a representation that is independent of the global component of variation. This allows us to directly estimate the within die component of variations and thus circumvents the need for making ergodicity assumption. We further show that using two dimensional variogram functions allows us to model geometrically anisotropic process variation data. Additionally, for extracting process variation models in the presence of significant measurement noise, we employ weighted least squares regression technique, which is known to be statistically more robust technique than the previously used ordinary least square technique.

- **A New Statistical Max Operation for Propagating Skewness in SSTA.** Statistical static timing analysis (SSTA) is emerging as a solution for predicting the timing characteristics of digital circuits under process variability. For computing the statistical max of two arrival time probability distributions, existing analytical SSTA approaches use the results given by Clark in [37]. These analytical results are exact when the two operand arrival time distributions have jointly Gaussian distributions. Due to the nonlinear max operation, arrival time distributions are typically skewed. Furthermore, nonlinear dependence of gate delays and non-gaussian process parameters also make the arrival time distributions asymmetric. Therefore, for computing the max accurately, a new approach is required that accounts for the inherent skewness in arrival time distributions. In Chapter 3, we present analytical solution for computing the statistical max operation. First, the skewness in arrival time distribution is modeled by matching its first three moments to a so-called skewed normal distribution. Then by extending Clarks work to handle skewed normal distributions we derive analytical expressions for computing the moments of the max. We then show using initial simulations results that using a skewness based max operation has

a significant potential to improve the accuracy of the statistical max operation in SSTA while retaining its computational efficiency.

- **Parametric Yield Maximization using Gate Sizing.** With the increased significance of leakage power and performance variability, the yield of a design is becoming constrained both by power and performance limits, thereby significantly complicating circuit optimization. In Chapter 4, we propose a new optimization method for yield optimization under simultaneous leakage power and performance limits. The optimization approach uses a novel leakage power and performance analysis that is statistical in nature and considers the correlation between leakage power and performance to enable accurate computation of circuit yield under power and delay limits. We then propose a new heuristic approach to incrementally compute the gradient of yield with respect to gate sizes in the circuit with high efficiency and accuracy. We then show how this gradient information can be effectively used by a non-linear optimizer to perform yield optimization. We consider both inter-die and intra-die variations with correlated and random components. The proposed approach is implemented and tested and we demonstrate up to 40% yield improvement compared to a deterministically optimized circuit.
- **Statistical Full-chip Gate-oxide Reliability Analysis.** Gate oxide breakdown is a key factor limiting the useful lifetime of an integrated circuit. Unfortunately, the conventional approach for full chip oxide reliability analysis assumes a uniform oxide thickness for all devices. In practice, however, gate-oxide thickness varies from die-to-die and within die due to temperature and pressure non-uniformity of the gate-oxidation process. As the precision of process control worsens with aggressive oxide thickness scaling, an alternative reliability analysis approach is needed for modeling oxide thickness variations. In Chapter 5, we propose a statistical framework for chip level gate oxide reliability analysis while considering both die-to-die and within-die components of thickness variation. The thickness of each device is modeled as a

distinct random variable and thus the full chip reliability estimation problem is defined on a huge sample space of several million devices. We observe that the full chip oxide reliability is independent of the relative location of the individual devices. This enables us to transform the problem such that the resulting representation can be expressed in terms of only two distinct random variables. Using this transformation we present a computationally efficient and accurate approach for estimating the full chip reliability while considering oxide thickness variation that captures their spatial correlations. Simulation results validate the accuracy of the proposed approach.

The main contribution of this thesis is the development of a collection of tools and techniques for the statistical analysis, modeling and optimization of circuits in order to support a variation aware methodology. While the techniques presented in this work have made some progress in advancing the current state of variation aware design methodology, several other issues are needed to be addressed for adoption of a complete statistical design methodology. In the concluding chapter, Chapter 6, we list some of these issues and suggest directions for future work.

CHAPTER 2

Robust Extraction of Process Variation Model

The semiconductor manufacturing process has become more complex, while process control precision is struggling to maintain its accuracy with continuous process scaling. As a result, a variety of steps throughout the manufacturing process are prone to fluctuations. These include effects due to chemical mechanical polishing which is used to planarize oxides surrounding metal lines and optical proximity effects which are a consequence of patterning features smaller than the wavelength of light [3, 4].

These in combination with numerous other effects cause an increase in variation of device and interconnect physical parameters such as gate length (or Critical Dimension - CD), gate-oxide thickness, channel doping concentration, interconnect thickness and height, etc., which in turn affects their electrical characteristics. These variations in electrical characteristics of circuit components have led to increased variability in circuit performance, resulting in yield degradation. Unfortunately, as listed in [75], no manufacturable solutions are available for controlling the variability of several parameters (for e.g., CD) for sub-45 nm technology nodes. Thus, for achieving reasonable yield it has become important to model variability during design. Acknowledging these issues, statistical performance analysis and optimization approaches [30, 28, 76] have emerged as a possible solution for statistically quantifying the variability in performance. A model of parameter variation is an essential input for both statistical circuit analysis and optimization, therefore it is important to understand and characterize parameter variations.

Numerous research efforts [30, 28, 77], have focused on developing statistical analysis

and optimization techniques assuming that of a model of parameter variation is known a priori. However not much attention has been paid on extracting such a model from experimental measurements. In [78] the authors focused on extracting the deterministic component of CD variation for $0.18\mu\text{m}$ CMOS technology. As random variations were insignificant for the $0.18\mu\text{m}$ CMOS technology, the random component of spatial variations was ignored. A simple computation of the spatial correlation coefficient as a function of distance from wafer-scale measurement data was presented in [79]. From their measurement results, it was found that spatial correlation was significantly different in the horizontal and vertical direction. In the field of spatial statistics, this difference in spatial dependence across all die with respect to the direction is referred to as *anisotropy*.

Recently the authors in [69, 80] introduced a formal approach to model and extract correlation functions. They noted an important result from random field theory that a correlation function is valid only if any correlation matrix derived from it is a positive semi-definite and therefore, any arbitrary function such as those derived from polynomial curve fitting cannot be a valid spatial correlation function. They also presented an algorithm to extract such a valid correlation function from measurement data in the presence of random noise. In cases when measured variation data is found not to be stationary, they further proposed a technique to extract a valid spatial correlation matrix by employing a modified alternative projection algorithm. A study of different correlation models and their associated timing methods using measured critical dimension (CD) measurement data is given in [73]. A heuristic algorithm to fit quad-tree based correlation models was also presented in [73]. Recently, we found out independent of our work another approach [81] was simultaneously developed based on the spatial statistics concepts [82]. In [81], the authors proposed a geo-statistics based mathematical method to extract a model of spatial correlation trend of a single die.

In this chapter, we appeal to spatial statistics methods to develop a method for robust extraction of an anisotropic process variation model. Our approach is based on the con-

cept of variogram functions, which is an alternative representation of covariance function. The variogram function can be understood as the variance of the difference of two random variables at different locations in a die. By construction itself, the variogram representation becomes independent of the global component of variation. This approach also side-steps the need to compute the global variance prior to computing the correlation model. Apart from this key motivating property, it has been shown that the variogram function can be estimated more reliably from measurement data than the covariance function [82]. A further study of spatial statistics revealed that two dimensional valid variogram functions can be used to model anisotropic process variation data [83]. Using such a two dimensional anisotropic valid variogram function we develop a method for extracting the process variation model. First, we compute an empirical variogram from the given measurement data set. Then, we extract an estimate of a valid anisotropic variogram function from the empirical variogram using nonlinear regression. For nonlinear regression, the weighted least square is chosen as the objective, which is known to be more robust to outliers than the previously used least square objective. Finally, the global component is computed by subtracting the within-die component of variation, found from the variogram function, from the overall variation.

The rest of the chapter is organized as follows: in Section 2.1 we explain the concept of variogram. Section 2.2 explains the proposed algorithm for extracting the process variation. Experimental results that show the accuracy of the proposed approach are given in Section 2.3.

Our goal is to extract a process variation model for some parameter of interest, which can be either a process parameter such as depth of focus and dose, a device parameter such as gate length and oxide thickness, or an electrical parameter like drive current, gate delay and threshold voltage. Due to global variations, manufactured values of a parameter vary from die-to-die whereas due to within-die variation, they also vary with its location within a die. Therefore, for modeling within-die variations, a distinct random variable

has to be associated corresponding to each location in the die. Thus, all components of process variation can be modeled using a two dimensional random function that associates a random variable for each location within the die. To set notation, a model of process variation at location $\mathbf{x} = (x, y)$ can be given as follows:

$$F(\mathbf{x}) = f_0 + X + Y(\mathbf{x}) + Z(\mathbf{x}), \quad (2.1)$$

where (i) f_0 is the deterministic mean of the random function, (ii) X represents the global component of variation that is same for all locations in the die, (iii) $Y(\cdot)$ is a zero-mean random function defined over each location $\mathbf{x} = (x, y)$ in the die, representing the parameter's spatial dependence (after mean is removed) (iv) $Z(\cdot)$ is also a zero-mean random function, but it represents the residual independent component of variation. For any location \mathbf{x} , the corresponding random variables $X, Y(\mathbf{x})$ and $Z(\mathbf{x})$ are independent of each other. Recall, that the mean f_0 is also dependent on location due to design dependent systematic variations. In [78], the authors have presented a detailed method for extracting systematic variation, using which mean trends resulting from systematic variations can be removed from the measured characterization data. Therefore, in this work, we focus on extracting the non-systematic components of variation. In particular, we wish to find the global variance σ_X^2 , the spatial variance σ_Y^2 , the variance of independent component σ_Z^2 and the correlation function.

2.1 Variogram Function

In this section, we present the concept of variogram function for modeling spatial variations. As mentioned earlier, the variogram function models the spatial dependence by quantifying the amount by which the parameters of two devices can vary from each other as a function of their spatial distance. The definition of variogram function is given as follows:

Definition 1 *For a random function $F(\mathbf{s})$, the **variogram function** between any two points*

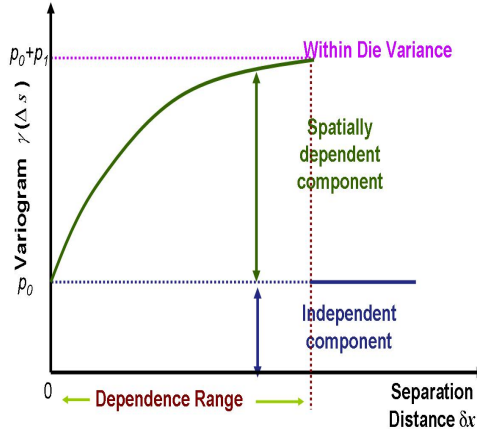


Figure 2.1. A typical Variogram in 1-dimension

\mathbf{s}_1 and \mathbf{s}_2 is defined as

$$2\gamma(\mathbf{s}_1, \mathbf{s}_2) = \text{Variance}(F(\mathbf{s}_1) - F(\mathbf{s}_2)),$$

where $\gamma(\cdot)$ is referred to as the semi-variogram.

Similar to the second order stationary assumption required for extracting correlation function [84], a more general intrinsic stationary assumption is required for characterizing random functions using variogram functions.

Definition 2 *If the mean of the random function $F(\mathbf{s})$ is constant and its variogram function $2\gamma(\mathbf{s}_1, \mathbf{s}_2)$ between any two points \mathbf{s}_1 and \mathbf{s}_2 depends only on the difference vector $\mathbf{v} = \mathbf{s}_1 - \mathbf{s}_2$,*

$$2\gamma(\mathbf{s}_1, \mathbf{s}_2) = 2\gamma(\mathbf{s}_1 - \mathbf{s}_2) = 2\gamma(\mathbf{v}),$$

*then it is called an **intrinsically stationary random process**.*

To understand the variogram function, consider an example semi-variogram plot shown in Figure 2.1. In the presence of spatial dependence, two random variables (device parameters) located close to each other tend to vary in a similar manner. Therefore, the variogram function, which represents the variance of their difference, monotonically increases with the separation distance. Due to the device scale independent component of variation, the y intercept of the variogram function can be greater than zero.

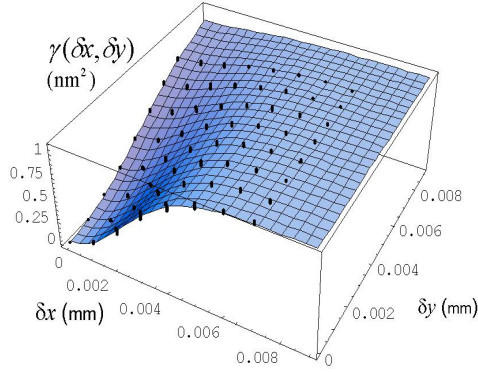


Figure 2.2. A typical anisotropic Variogram in 2-dimension

All monotonically decreasing functions cannot be considered as valid candidates for a covariance function. Similarly, not all monotonically increasing functions can qualify for a valid variogram function. A spectral theory also exists for valid variogram functions that defines negative definiteness as the necessary condition for any arbitrary function to be a valid variogram function. A detailed description of the spectral theory of variogram functions can be found in [82]. For our purpose, a valid class of functions, known as Matern-class, can be used to extract a valid variogram function. A parametric family of valid variogram functions based on the Matern-class can be given as

$$2\gamma(\mathbf{v}) = p_0 + p_1 \left(1 - \frac{1}{2^{p_3-1}\Gamma(p_3)} \left(\frac{\mathbf{v}}{p_2} \right)^{p_3} K_{p_3} \left(\frac{\mathbf{v}}{p_2} \right) \right),$$

where p_0 represents the variance of the independent component (i.e., σ_Z^2), p_1 represents the variance of the spatially dependent component (i.e., σ_Y^2), p_2 is the spatial scale parameter, p_3 is the shape parameter, $\Gamma(\cdot)$ is the gamma function and $K_{p_3}(\cdot)$ is the modified Bessel function of the second kind and order p_3 . The flexibility of the above mentioned Matern-class has been demonstrated on a variety of spatial data sets [84].

If the variogram depends only on the Euclidean distance (i.e., $\|s_1 - s_2\|$, where $\|\cdot\|$ represents the l^2 norm) between locations, then the variogram is said to be isotropic; otherwise, it is considered to be anisotropic. Anisotropy is caused by the underlying physical sources of variation evolving differentially in space. A classification of types of anisotropy

arising in spatial analysis and methods for modeling them using appropriate variogram functions can be found in [83]. A common theme in most of the methods is to generalize the isotropic function to model the asymmetries. The simplest and the most common form of anisotropy is the *geometric anisotropy*, where the random function is not isotropic in the original space, but it is in some linearly transformed space. For modeling geometric anisotropy, a valid variogram function can be achieved from the isotropic variogram function by linearly transforming the distance vector \mathbf{v} . Based on this concept, the geometrically anisotropic Matern-class of variogram functions can be given as

$$2\gamma(\mathbf{v}) = p_0 + p_1 \left(1 - \frac{1}{2^{p_3-1}\Gamma(p_3)} \left(\frac{\|\mathbb{P}\mathbf{v}\|}{p_2} \right)^{p_3} \left(\frac{\|\mathbb{P}\mathbf{v}\|}{p_2} \right) \right), \quad (2.2)$$

where \mathbb{P} is 2×2 -matrix of parameters that models the axes and scale of anisotropy. For simplicity, we denote the set of all parameters with $\mathcal{P} = \{p_0, p_1, p_2, p_3, \mathbb{P}\}$. Figure 2.2 shows a surface plot of a possible anisotropic semi-variogram as a function of distance vector $\mathbf{v} = (\delta x, \delta y)$. If \mathbb{P} is an identity matrix then the above expression reduces to the isotropic case.

2.2 Process Variation Extraction

In this section, we present the overall algorithm for extracting the process variation model (see Figure 2.3). As mentioned earlier, by construction the semi-variogram function is independent of the global component of variation. Therefore, by estimating the variogram representation, we can directly find the spatially dependent within-die component of variation. The measurement data of N die at M locations per die represents a set of samples of the random function $F(\cdot)$. For distinguishing between the actual function and measurement data, we use the over-line to denote a sampling, for example, the set of samples of random variable at location \mathbf{s} is denoted by $\overline{F}(\mathbf{s})$. The k^{th} measured sample of the random variable $F(\mathbf{s})$ at location \mathbf{s} is given by $f_k(\mathbf{s})$. From the measurement data, first, we find the unbiased

1. Compute the experimental variogram using (2.3)
2. Using nonlinear regression solve (2.2), to find all parameters \mathcal{P}
3. From (2.5), find $\bar{\rho}(\mathbf{v})$ by substituting all parameters \mathcal{P}
4. Evaluate σ_X^2 from (2.6)

Figure 2.3. Proposed algorithm for process variation extraction

estimate of the sample semi-variogram, using the following expression:

$$\begin{aligned}
\bar{\gamma}(\bar{F}(\mathbf{s}_i), \bar{F}(\mathbf{s}_j)) &= \frac{1}{2} \text{Variance}(\bar{F}(\mathbf{s}_i) - \bar{F}(\mathbf{s}_j)) \\
&= \frac{1}{2(N-1)} \sum_{k \in N} (f_k(\mathbf{s}_i) - f_k(\mathbf{s}_j))^2 \\
&\quad - \frac{(\sum_{k \in N} (f_k(\mathbf{s}_i) - f_k(\mathbf{s}_j)))^2}{2N(N-1)}.
\end{aligned} \tag{2.3}$$

The sample semi-variogram $\bar{\gamma}(\bar{F}(\mathbf{s}_i), \bar{F}(\mathbf{s}_j))$ is computed for every location pair $(\mathbf{s}_i, \mathbf{s}_j)$ separated by the distance vector $\mathbf{v} = \mathbf{s}_i - \mathbf{s}_j$. This gives us the so-called *empirical semi-variogram* as a finite set of ordered pairs $(\mathbf{v}, \bar{\gamma}(\mathbf{v}))$.

Now, for estimating a valid variogram function, we wish to find the set of parameters \mathcal{P} of the anisotropic Matern-class given in (2.1). This can be formulated as a nonlinear estimation problem. It is known from nonlinear estimation theory [85] that the method of least squares is not statistical; it is purely a numerical criterion. When the response variable (i.e., the experimental variogram $\bar{\gamma}(\mathbf{v})$) are random variables with a non-scalar variance matrix, the method of least squares is found to be an inefficient estimator [85]. Due to the spatial dependence among each of the measured locations, the variance-matrix of the empirical variogram data set is not a diagonal matrix. An alternative objective, known as generalized least squares provides statistically robust estimates, however it is found to be computationally more expensive. Nevertheless, an intermediate approach, the method of weighted least squares, provides reasonable compromise between robustness and computational efficiency. For extracting valid variogram function, a nonlinear estimation formulation based on the weighted least square objective can be given as follows:

$$\min_{\mathcal{P}} \left\| \frac{(\bar{\gamma}(\mathbf{v}) - \gamma(\mathbf{v}))}{\text{Variance}(\bar{\gamma}(\mathbf{v}))} \right\|, \tag{2.4}$$

where $\bar{\gamma}(\mathbf{v})$ is the empirical semi-variogram, $\gamma(\mathbf{v})$ is the family of variogram functions given in (2.1), and $Variance(\bar{\gamma}(\mathbf{v}))$ is the weight representing the diagonal of the covariance matrix of estimates in the empirical variogram.

Solving the above optimization problem gives statistically robust estimates of parameters \mathcal{P} . The intra-die variation components $\bar{\sigma}_Y^2 \approx p_1$ and $\bar{\sigma}_Z^2 \approx p_0$ are obtained directly by estimating the variogram function. The estimate of spatial correlation function can be obtained from the variogram function and $\bar{\sigma}_Y^2$ and $\bar{\sigma}_Z^2$ using the following relation:

$$\bar{\rho}(\mathbf{v}) = \frac{\sigma_Z^2 + \sigma_Y^2 - \bar{\gamma}(\mathbf{v})}{\sigma_Y^2}. \quad (2.5)$$

Finally, the only component remaining to be estimated is the global variance σ_X^2 . Now at any location s , an unbiased estimate of overall sample variance of random variable $F(\mathbf{s})$ can be found as follows:

$$Variance(\bar{F}(\mathbf{s})) = \frac{1}{N-1} \left(\sum_{k \in N} f_k(\mathbf{s}_i)^2 - \frac{1}{N} \sum_{k \in N} f_k(\mathbf{s}_i) \right)$$

As all components of variation X , Y and Z are independent of each other, we know that the total variance of every random variable $F(\mathbf{s})$ is the sum of all three components out of which we already know $\bar{\sigma}_Y^2$ and $\bar{\sigma}_Z^2$. Thus for each random variable the global variance be estimated by subtracting the $\bar{\sigma}_Y^2$ and $\bar{\sigma}_Z^2$ from its overall variance. An estimate of the overall global variance can be found by taking the average over all M measurement locations on the die,

$$\bar{\sigma}_X^2 \approx \frac{1}{M} \sum_{s \in \text{Die}} Variance(\bar{F}(\mathbf{s}) - \sigma_Y^2 - \sigma_Z^2). \quad (2.6)$$

2.3 Simulation Results

In our first experiment, we applied our algorithm to extract a process variation model from 130nm Electrical Line-width Measurement (ELM) data taken from horizontal polysilicon lines that have optical proximity corrections included [86]. The data-set included measurements from 5 different wafers, each wafer contained 23 fields, and each field included

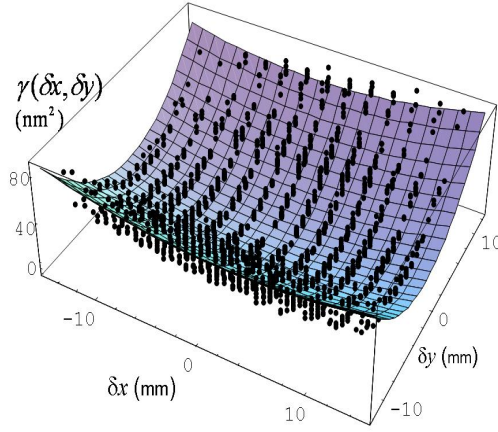


Figure 2.4. The Variogram of measured gate length data

308 measurement points: 14 points in the horizontal direction and 22 points in the vertical direction. Individual measurement points were spaced horizontally by 2.19mm and vertically by 1.14mm. For these 5 wafers, we divided the reticles into various die sizes in order to investigate the effects of die size on the gate-length variation. We diced a reticle into 4 die, a 2-die by 2-die configuration where each die is 15mm x 12mm. A surface plot of anisotropic variogram function extracted from the measurement data and the scatter plot of exact empirical variogram (dots) numerically computed from the ELM data are shown in Figure 2.4. It can be seen that the proposed algorithm can extract good fits of valid anisotropic variogram functions from measured ELM data. Similar results were observed when the entire reticle was considered and when the reticle was divided into 16 die.

In the second experiment, we validate the accuracy and robustness of our algorithm using a Monte Carlo model of process variation. For the sake of comparison, we closely follow the experimental setup used in [69]. A valid correlation function $\rho(\mathbf{s})$ and known variance of all variation components are used by the Monte Carlo model. To model the measurement errors inherent in real data, different amounts of gaussian noise is added to the Monte Carlo model. A representative measurement data-set is generated using Monte Carlo simulations for N number of sample die and M number of measurement locations on each die. By applying the proposed algorithm, all components of process variation are

Table 2.1. Accuracy and Robustness of proposed approach

M	N	%Noise	% $err(\sigma_X^2)$	% $err(\sigma_Y^2)$	% $err(\rho(\mathbf{s}))$
64	2000	10	0.35	2.20	2.95
		50	2.61	-0.24	2.91
		100	-2.66	2.17	3.02
64	1500	10	-1.41	-0.76	3.82
		50	5.11	-0.01	3.54
		100	-1.30	2.13	3.57
64	1000	10	-1.53	2.01	4.01
		50	-0.11	-0.94	4.20
		100	2.08	1.33	4.31
64	500	10	-8.91	2.26	6.13
		50	0.37	3.27	5.84
		100	11.76	-2.08	6.40
49	1000	10	-5.98	1.94	4.21
		50	-5.22	3.85	4.30
		100	8.82	-1.12	4.34
36	1000	10	5.72	-3.39	4.06
		50	-6.44	-2.02	4.18
		100	-9.38	4.33	4.82

extracted from the measurement data. A comparison of extracted values of σ_X^2 , σ_Y^2 , and $\rho(\mathbf{s})$ with the known variation components used in the Monte Carlo model, will determine the robustness and accuracy of the proposed extraction algorithm.

Similar to [69], we report the percentage error for the global variation and spatial variation relative to the Monte Carlo model values, but not for the random variation as it is indistinguishable from the added measurement noise. The error in spatial correlation function, was measured using the same metric as [69], given by $err(\rho\mathbf{s}) = \frac{\bar{\rho}(\mathbf{s}) - \rho(\mathbf{s})}{\rho(\mathbf{s})}$. The results of our experiment are given in Table 2.1. The number of sample die N, the number of measurement location M and the amount of random noise added into the Monte Carlo model relative to the total variation ($\sigma_X^2 + \sigma_Y^2 + \sigma_Z^2$) are listed in the first 3 columns. Similar to [69], by varying M and N different data-sets were generated and for each data-set we tested our algorithm for 10 %, 50 % and 100% noise. It is evident from Table 2.1 that the proposed algorithm gives accurate and robust estimates of all variation components.

We also implemented the algorithm given in [69]. As previous approach was proposed

assuming isotropy, we used an isotropic correlation function in the Monte Carlo model. To our surprise, even for a very similar setup, we were not able to reproduce the previously published results given in [69]. We found that even when no measurement noise is added, our implementation of the previous approach resulted in unreasonable errors (for e.g., $err(\sigma_X^2) = 35\%$, $err(\sigma_Y^2) = 65\%$). After analyzing the reason for the error at each step, we found that this error is due to an incorrect estimation of the overall within-die variation. The ergodicity assumption made while estimating the overall within-die variance seems to be the most plausible reason for such high errors.

A comparison between the previously published results [69] on a similar experimental setup favors the proposed algorithm in terms of overall accuracy. For example, the worst case error of $err(\sigma_X^2)$, $err(\sigma_Y^2)$ and $err(\rho(s))$ observed in the previous approach were 18%, 11% and 9% but it should be noted that different data-sets were used in both the cases. For the second experiment, as neither the implementation nor the data-set used for the previous approach [69] were not publicly available, we were not able to make a direct comparison between the two approaches.

To summarize, we have presented a new approach to extract spatial variation models based on the concept of variogram function. The key advantage of the variogram function is that it provides us with a representation that is independent of the global component of variation. It allows us to directly estimate the within-die component of variations and thus circumvents the need for making ergodicity assumption. In this work, we further showed that using two dimensional variogram functions allows us to model geometrically anisotropic process variation data. Additionally, for extracting process variation models in the presence of significant measurement noise, we employ *weighted least squares* regression technique, a technique which is known to be statistically more robust than the previously used ordinary least square technique. The experiment results on both Monte-Carlo models and ELM measurement data confirm the validity of the proposed approach.

CHAPTER 3

A New Statistical Maximum Operation for Propagating Skewness in SSTA

A path-based SSTA requires enumeration of an exponential number of paths, therefore, block-based SSTA is considered to be a more efficient technique. As discussed in Chapter 1, the analytical methods presented in [52, 30, 28] appeal to be the more promising approaches for a computationally efficient implementation of SSTA. In [52], the author introduced a linear time analytical SSTA algorithm assuming uncorrelated normal random variables for delay distribution. Using a first order parametric delay model, a method for handling correlations in global sources of variation due to both spatial correlation and path re-convergence was presented in [30, 28]. Recall that the basic block based SSTA algorithm included a PERT-like topological traversal of a circuit graph, where at each node the maximum arrival time distribution is computed in terms of the parametric delay model. For propagating arrival time distributions, one needs to compute the sum and the maximum of two arrival times at each node in the circuit graph. The computation of the sum function is relatively simple; however, the statistical max of two correlated arrival time variables is non-trivial.

The max operation in existing SSTA approaches is invariably based on analytical results given in [37]. Clark derived analytical expressions for finding the moments of the max of two correlated normal random variables and an expression for computing the correlation of the resulting max with any other jointly normal variable. The Clark's max results are exact when the two operand random variables have a bivariate normal distribution. However, the result of the max of two normal variables is typically a positively skewed non-normal dis-

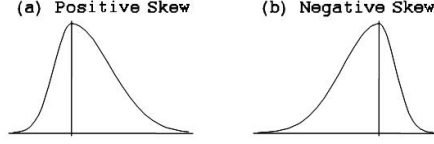


Figure 3.1. Examples for Asymmetric PDFs

tribution. Skewness is a statistical parameter used to describe asymmetry in a probability distribution. A probability distribution is said to have positive(negative) skewness if it has a long tail in the positive(negative) direction (see Figure 3.1). Both the above mentioned analytical approaches [30, 28] use these expressions for computing the moments of statistical max of two arrival time random variables. Unfortunately in SSTA, the asymmetric non-normal arrival time distributions resulting from the max operation performed at one node are inputs to the max operation which is needed to be performed at a downstream node. Additionally, variations in interconnect and few process parameters also have asymmetric non-normal distributions [33]. However, existing analytical SSTA approaches have to approximate the non-normal arrival time distribution with a normal distribution for applying Clark's max. The error of this approximation increases when the difference of the mean relative to the standard deviation decreases and it becomes maximum when two means are equal [37]. For a typical design, there can be several thousand critical paths and the means of their output arrival time distributions and arrival time distributions at common internal nodes will be closely aligned with each other. Therefore, in such a case Clark's max based SSTA methods may result in inadequate accuracy, in particular, for power-optimized designs having a large number of nodes with zero or small slack.

Recently, SSTA algorithms using higher order nonlinear parametric delay models with non-normal distributions were proposed in [31, 34, 32, 33]. However, for computing the max operation, these approaches either use numerical techniques and/or employ the Clark's max requiring normal approximation. A conditional max based heuristic analytical method was presented in [33] where the max operations is postponed until the mean of the two arrival time distributions are far not close to each other. In this work, we extend Clark's

max approach and give analytical results for computing the approximate maximum of a set of asymmetric random variables. The problem of computing the max of a finite set of random variables has been well studied. Several approaches derived Clark's results using different methods [87, 88]. In our method, given the first three moments of any asymmetric distribution, we give analytical expressions to map it to a skew-normal (explained later) representation having same moments. We then derive analytical results for computing the moments of the max of two correlated skew-normal distributions assuming a bivariate skew-normal distribution. The derivation is similar in spirit to Clark's approach, although it is more general since we can compute the moments for a bivariate skew-normal random variables.

The rest of the chapter is organized as follows. In Section 3.1 we explain the skew-normal distribution and give analytical expressions for computing the parameters of a skew-normal distribution from the mean, variance and skewness of arrival time distribution. A bivariate skew-normal distribution and the derivation for the proposed max operation are given in Section 3.2. In Section 3.3, we give numerical results illustrating the efficacy of the proposed max operation.

3.1 Modeling Skewness

Arrival time distributions and circuit delay distributions are typically skewed, due to the nonlinear max operation and nonlinear dependence of delay on process parameters. We need an analytical representation that is flexible enough to capture the skewness in asymmetric arrival time distributions and at the same time be of the functional form which allows analytical derivation of the maximum operation. After studying several skewed representations, in [89], we found a general method for introducing skewness into any unimodal symmetric distribution. Their basic idea is to simply introduce inverse scale factors in the left and the right half real lines around the mean. Let $f(x)$ be the normal distribution with mean μ and variance σ given by

$$f(x) = \frac{1}{\sigma} \phi\left(\frac{x-\mu}{\sigma}\right), \text{ where } \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

Using the method presented in [89], a skew-normal distribution $f_\gamma(x)$ can be computed from the normal distribution $f(x)$, by scaling its left half and right half by factors γ and its inverse $1/\gamma$, respectively. This gives us the skew-normal distribution,

$$f_\gamma(x) = \frac{2}{\sigma(\gamma + 1/\gamma)} \left\{ \phi\left(\frac{(x-\mu)\gamma}{\sigma}\right) I_{(-\infty, \mu]}(x) + \phi\left(\frac{x-\mu}{\gamma\sigma}\right) I_{(\mu, \infty)}(x) \right\},$$

where, $I_{(-\infty, \mu]}(x)$ and $I_{(\mu, \infty)}(x)$ are the Indicator functions:

$$I_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise.} \end{cases}$$

For a skew-normal distribution, we can observe that scaling variable x corresponds to an inverse scaling of the standard deviation σ around its mean. Therefore, $f_\gamma(x)$ can be alternatively written as

$$f_\gamma(x) = \frac{2}{\sigma_l + \sigma_r} \left\{ \phi\left(\frac{(x-\mu)}{\sigma_l}\right) I_{(-\infty, \mu]}(x) + \phi\left(\frac{x-\mu}{\sigma_r}\right) I_{(\mu, \infty)}(x) \right\},$$

where,

$$\sigma_l = \frac{\sigma}{\gamma} \quad \text{and} \quad \sigma_r = \sigma\gamma.$$

Note that the resulting skewed distribution $f_\gamma(x)$ has a functional form similar to the original non-skewed distribution $f(x)$. If the skewness parameter γ is greater(less) than unity then $f_\gamma(x)$ is positively(negatively) skewed. For $\gamma = 1$ we get back the original symmetric normal distribution. Furthermore, $f_\gamma(x)$ is both continuous and differentiable and

is completely defined by only three parameters μ , σ and γ . These were the key appealing properties that motivated us to use this representation for deriving the proposed max operation.

Existing SSTA approaches model and propagate only the mean and variance of the arrival time distribution. For improving the accuracy of SSTA algorithm, in addition to the mean and variance, we wish to propagate the skewness in asymmetric arrival time distributions. In such an SSTA framework, the input parameters of the max operation will include mean, variance and skewness of the two input arrival time distributions and their correlation. We first want to map the arrival time distribution characterized by its mean, variance and skewness to a skew-normal distribution $f_\gamma(x)$. Let μ_γ , σ_γ and Sk_γ be the given mean, standard deviation and skewness of a skewed arrival time distribution and μ , σ and γ are the three parameters that define the desired skew-normal distribution $f_\gamma(x)$. For finding $f_\gamma(x)$, we express the mean, variance and skewness of the skew-normal distribution as function of its parameters μ , σ and γ and then match these to the μ_γ , σ_γ and Sk_γ of a skewed arrival time distribution to solve for μ , σ and γ . The analytical expressions for mean μ_γ , variance σ_γ^2 and skewness Sk_γ of $f_\gamma(x)$ derived in terms of its parameters (μ , σ and γ) are given as follows:

$$\mu_\gamma = \mu + \sqrt{\frac{2}{\pi}} \left(\gamma - \frac{1}{\gamma} \right) \sigma \quad (3.1)$$

$$\sigma_\gamma^2 = \frac{(\pi \gamma^4 - 2 \gamma^4 - \pi \gamma^2 + 4 \gamma^2 + \pi - 2) \sigma^2}{\pi \gamma^2} \quad (3.2)$$

The skewness of distribution defined by the ratio of the third centered moment and cubed standard deviation is given by

$$Sk_\gamma = \frac{\sqrt{2} (1 - \gamma^2) \left(\pi (\gamma^4 - 3 \gamma^2 + 1) - 4 (\gamma^2 - 1)^2 \right)}{\left(\pi (\gamma^4 - \gamma^2 + 1) - 2 (\gamma^2 - 1)^2 \right)^{\frac{3}{2}}}. \quad (3.3)$$

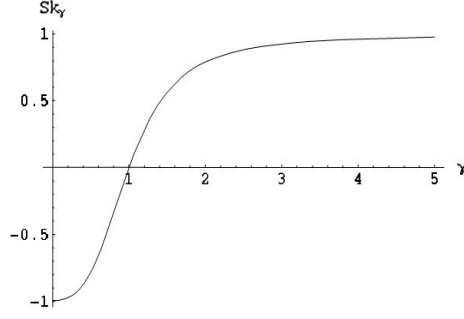


Figure 3.2. The γ parameter of $f_\gamma(x)$ vs. Skewness Sk_γ

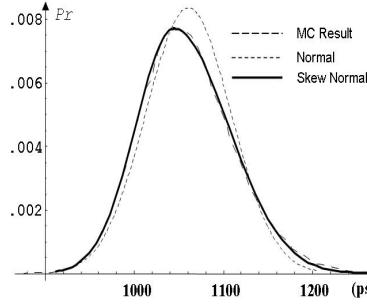


Figure 3.3. Comparison between skew-normal distribution and Normal distribution for a typical Monte Carlos based Arrival Time distribution.

Fortunately, the skewness Sk_γ (Eq. 3.3) is only a function of γ and is independent of the other two parameters μ and σ . A plot of this function is given in Figure 3.2, where it can be seen that skewness Sk_γ is a well behaved function and it monotonically increases with γ . Therefore, for a given Sk_γ , one can efficiently compute γ either using pre-computed look-up tables or using numerical methods with very fast convergence. Using γ , σ_γ and μ_γ we can analytically solve equations 3.2 and 3.1 for parameters σ and μ , respectively. Thus given mean, variance and skewness of an arrival time distribution we can easily map it to a skew-normal distribution. In Figure 3.3, we show plots of a typical skewed arrival time distribution approximated by a skew-normal distribution and normal distribution. It is evident that compared to existing normal approximations, skew-normal is a much better representation that can accurately capture the inherent skewness in arrival time distributions.

3.2 Skew-Normal Max Operation

Based on the skew-normal representation explained in the previous section, we now present the skew-normal max operation. For analytically expressing the max function of two correlated arrival time random variables X and Y , we need to know their joint probability distribution function. In [37], the author uses the following bivariate normal distribution for the two operand random variables.

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y}\phi\left(\frac{x-\mu_x}{\sigma_x}, \frac{y-\mu_y}{\sigma_y}\right)$$

$$\text{where, } \phi(x,y) = \frac{1}{\sqrt{1-\rho^2}} e^{-\frac{x^2-2\rho xy+y^2}{2(1-\rho^2)}}$$

Recall that bivariate normal representation being symmetric will introduce errors in the computation of the recursive max operation for SSTA purposes. Therefore, similar to the univariate skew-normal presented in the previous section, we add two inverse scale parameters γ_x and γ_y for random variables X and Y around their respective means μ_x and μ_y for introducing skewness in the bivariate distribution.

$$f_Y(x,y) = \frac{1}{\Gamma\sigma_x\sigma_y} \left(\begin{aligned} &\phi\left(\frac{x-\mu_x}{\sigma_{x_l}}, \frac{y-\mu_y}{\sigma_{y_l}}\right) I_{(-\infty, \mu_x)}(x) I_{(-\infty, \mu_y)}(y) \\ &+ \phi\left(\frac{x-\mu_x}{\sigma_{x_l}}, \frac{y-\mu_y}{\sigma_{y_r}}\right) I_{(-\infty, \mu_x)}(x) I_{[\mu_y, \infty)}(y) \\ &+ \phi\left(\frac{x-\mu_x}{\sigma_{x_r}}, \frac{y-\mu_y}{\sigma_{y_l}}\right) I_{[\mu_x, \infty)}(x) I_{(-\infty, \mu_y)}(y) \\ &+ \phi\left(\frac{x-\mu_x}{\sigma_{x_r}}, \frac{y-\mu_y}{\sigma_{y_r}}\right) I_{[\mu_x, \infty)}(x) I_{[\mu_y, \infty)}(y) \end{aligned} \right)$$

$$\text{where } \Gamma = \frac{\pi}{2} \left(\gamma_x + \frac{1}{\gamma_x} \right) \left(\gamma_y + \frac{1}{\gamma_y} \right)$$

$$+ \left(\gamma_x - \frac{1}{\gamma_x} \right) \left(\gamma_y - \frac{1}{\gamma_y} \right) \tan^{-1} \left(\frac{\rho}{\sqrt{1-\rho^2}} \right);$$

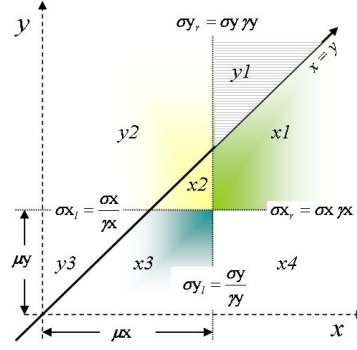


Figure 3.4. Standard deviations of a bivariate skew-normal distribution and seven regions of integration for $\mu_x > \mu_y$

$$\sigma_{x_l} = \frac{\sigma_x}{\gamma_x}; \quad \sigma_{y_l} = \frac{\sigma_y}{\gamma_y}; \quad \sigma_{x_r} = \sigma_x \gamma_x; \quad \text{and} \quad \sigma_{y_r} = \sigma_y \gamma_y.$$

Due to the correlation ρ , the normalizing constant term Γ differs from the univariate case. Figure 3.4 graphically illustrates how the two indicator functions partition the X, Y plane into 4 quadrants having different standard deviations around the mean vector (μ_x, μ_y) . Strictly speaking the arrival time distributions may not necessarily have a bivariate skew-normals; however, introducing additional skewness parameters allows us more degrees of freedom in comparison with [37]. Furthermore, in the absence of skewness the bivariate skew-normal representation reduces to the bivariate normal representation and therefore in this case we get the exact same results as [37]. For this bivariate skew-normal distribution we now derive results for computing the moments of the max of X and Y based on the original derivation given in [37]. Let $v(i)$ be the i^{th} moment of $\max(X, Y)$ given by

$$\begin{aligned} v(i) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\max(x, y))^i f_{\gamma}(x, y) dy dx \\ &= \oint_{(x, y) \in X > Y} x^i f_{\gamma}(x, y) d(x, y) \\ &+ \oint_{(x, y) \in X \leq Y} y^i f_{\gamma}(x, y) d(x, y) \end{aligned}$$

As shown in Figure 3.4 the region $X > Y$ gets further partitioned into 4 sub-regions x_1 ,

x_2 , x_3 and x_4 , where the sub-script denotes the standard deviation quadrant and likewise region $X \leq Y$ gets partitioned into sub-regions y_1 , y_2 and y_3 . Therefore, we can write the i^{th} moment of $\max(X, Y)$ as

$$v(i) = \sum_{j=1}^3 v_{y,j}(i) + \sum_{j=1}^4 v_{x,j}(i)$$

where, $v_{x,j}(i)$ and $v_{y,j}(i)$ are the i^{th} moment of $\max(X, y)$ in the j^{th} quadrant. The complete derivation of $v(i)$ over all the seven regions is repetitive and tedious. Therefore, in this chapter we will present the key steps encountered while deriving the expression for moments of sub-region y_1 . The i^{th} moment of max, for $Y > X$ in the 1st quadrant, is given as follows:

$$v_{y,1}(i) = \frac{1}{\Gamma\sigma_x\sigma_y} \int_{\mu_x}^{\infty} \int_x^{\infty} y^i \phi\left(\frac{x-\mu_x}{\sigma_{x_r}}, \frac{y-\mu_y}{\sigma_{y_r}}\right) dy dx$$

Using the Lebnitz rule, we compute the partial derivative of $v_{y,1}(i)$ with respect to μ_x :

$$\begin{aligned} \frac{\partial v_{y,1}(i)}{\partial \mu_x} &= \frac{1}{\Gamma\sigma_x\sigma_y} \int_{\mu_x}^{\infty} \int_x^{\infty} y^i \frac{\partial}{\partial \mu_x} \phi\left(\frac{x-\mu_x}{\sigma_{x_r}}, \frac{y-\mu_y}{\sigma_{y_r}}\right) dy dx \\ &\quad - \frac{1}{\Gamma\sigma_x\sigma_y\sqrt{1-\rho^2}} \int_{\mu_x}^{\infty} y^i e^{-\frac{\left(\frac{y-\mu_y}{\sigma_{y_r}}\right)^2}{2(1-\rho^2)}} dy \end{aligned}$$

We first change order of integration variables in the inner integral,

$$\begin{aligned} \frac{\partial v_{y,1}(i)}{\partial \mu_x} &= \frac{1}{\Gamma\sigma_x\sigma_y\sqrt{1-\rho^2}} \int_{\mu_x}^{\infty} y^i e^{-\frac{(y-\mu_y)^2}{2\sigma_{y_r}^2}} \int_{\mu_x}^y \frac{\partial e^{-\frac{\left(\frac{x-\mu_x}{\sigma_{x_r}} - \frac{\rho(y-\mu_y)}{\sigma_{y_r}}\right)^2}{2(1-\rho^2)}}}{\partial \mu_x} dx dy \\ &\quad - \frac{\gamma_1}{\Gamma\sigma_x\sigma_y\sqrt{1-\rho^2}} \int_{\mu_x}^{\infty} y^i e^{-\frac{\left(\frac{y-\mu_y}{\sigma_{y_r}}\right)^2}{2(1-\rho^2)}} dy \end{aligned}$$

Now, the inner integral of the first term in the above expression is in an integrable form. We evaluate this integral and an additional term due to the integration cancels out the second term and gives us the following simplified result.

$$\frac{\partial v_{y,1}(i)}{\partial \mu x} = -\frac{1}{\Gamma \sigma x \sigma y \sqrt{1-\rho^2}} \int_{\mu x}^{\infty} y^i e^{-\frac{(y-\mu y)^2}{2\sigma y_r^2}} e^{-\frac{\left(\frac{y-\mu x}{\sigma x_r} - \frac{\rho(y-\mu y)}{\sigma y_r}\right)^2}{2(1-\rho^2)}} dy$$

Similar to [37], we first make the substitution $y = \frac{(\sigma x_r \sigma y_r \sqrt{1-\rho^2})n}{a} + \mu y + \frac{(\mu x - \mu y) \sigma y_r (\sigma y_r - \sigma x_r \rho)}{a^2}$ and then $\mu x = \mu y + am$.

$$\begin{aligned} \frac{\partial v_{y,1}(i)}{\partial m} &= -\frac{1}{\Gamma} e^{-\frac{m^2}{2}} \int_{\frac{m(\sigma x_r - \sigma y_r \rho)}{\sigma y_r \sqrt{1-\rho^2}}}^{\infty} e^{-\frac{n^2}{2}} \\ &\quad \left(\mu y + \frac{n \sigma x_r \sigma y_r \sqrt{1-\rho^2}}{a} + \frac{m \sigma y_r (\sigma y_r - \sigma x_r \rho)}{a} \right)^i dn \end{aligned}$$

where,

$$a^2 = \sigma x_r^2 + \sigma y_r - 2\rho \sigma x_r \sigma y_r$$

Now note that for $m = \infty$, the random variable $X \gg Y$ and therefore, at $m = \infty$ all moments $v_{y,1}(i) = 0$. Using this observation one can express $v_{y,1}(i)$ as follows:

$$v_{y,1}(i) = \frac{1}{\Gamma} \int_{\alpha}^{\infty} e^{-\frac{m^2}{2}} \int_{k_3 m}^{\infty} (\mu y + k_1 n + k_2 m)^i e^{-\frac{n^2}{2}} dn dm \quad (3.4)$$

where,

$$\begin{aligned} k_1 &= \frac{\sigma x_r \sigma y_r \sqrt{1-\rho^2}}{a}, & k_2 &= \frac{\sigma y_r (\sigma y_r - \sigma x_r \rho)}{a}, \\ k_3 &= \frac{\sigma x_r - \sigma y_r \rho}{\sigma y_r \sqrt{1-\rho^2}} \quad \text{and} \quad \alpha = \frac{(\mu x - \mu y)}{a}. \end{aligned}$$

For a given positive integer value of i , the above integral can be expressed in terms of well known special functions. For example the first moment can be written as

$$\begin{aligned} v_{y,1}(1) &= \frac{\sqrt{\pi}}{\Gamma \sqrt{2}} \left(\frac{(k_1 - k_2 k_3)}{\sqrt{k_3^2 + 1}} \operatorname{erfc} \left(\frac{\alpha \sqrt{k_3^2 + 1}}{\sqrt{2}} \right) \right. \\ &\quad \left. + e^{-\frac{\alpha^2}{2}} k_2 \operatorname{erfc} \left(\frac{k_3 \alpha}{\sqrt{2}} \right) \right) + \frac{\mu y}{\Gamma} T(\alpha, k_3 \alpha) \end{aligned}$$

where, $\text{erfc}(x) = 1 - \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$ is the complementary error function and $T(x,y)$ is the Owen's T-function [90], given by

$$T(x,y) = \frac{1}{\pi} \int_0^y \frac{e^{-\frac{1}{2}x^2(1+t^2)}}{1+t^2} dt.$$

The special functions $\text{erfc}(x)$ and $T(x,y)$ are commonly encountered while integrating univariate and bivariate normal distributions, respectively. Precise numerical tables or accurate closed form analytical approximations exist for both $\text{erfc}(x)$ and $T(x,y)$. Efficient analytical solution for computing $T(x,y)$ is given in [91]. Thus similar to [37], the moments of the max can be found in a computationally efficiently manner. Likewise, higher moments can also be found by evaluating the integral given in Equation 3.4 at higher values of i . Using similar manipulations, the moments of $\max(X,Y)$ in all seven regions can be computed. Note that the case when $\mu_x = \mu_y$ we will only have 6 regions instead of 7 as $(\mu_x = \mu_y)$ will lie on the line $x = y$. For such a case, $T(x,y)$ can be analytically integrated by changing the variables to polar coordinates.

3.2.1 Applying Skew-Normal Max to SSTA

In this sub-section, we will briefly discuss how the proposed skew-normal max analytical results can be applied to block based SSTA. First we review how Clark's max results are currently being used for computing $\max(X,Y)$ of two arrival times X and Y . As mentioned earlier, X and Y are expressed in terms of a canonical form [30, 28]. First, the variances and the covariance of X and Y are computed from the canonical form. Then by substituting the mean, variance and correlation values into the analytical results given in [37], the statistics of the max distribution, namely, tightness probability (i.e., probability of one input being greater than the other [28]), the mean and the variance of their maximum are computed. Thereafter, the approximate distribution of the $\max(X,Y)$ in canonical form is determined by taking a linear combination of the two input arrival times weighted by their tightness probability [30, 31, 33]. Finally, the mean and variance of the resulting canonical form distribution are matched with the analytically computed mean and variance of the max

computed from Clark’s max results.

The focus of the proposed skew-normal max operation is to take into account skewness of X and Y in addition to mean and variance of the arrival time distribution. nonlinear canonical timing models (for example, the quadratic timing model) proposed in previous approaches [31, 34, 32, 33], are inherently skewed, and if X and Y are expressed in terms of such nonlinear canonical forms, then the skewness of X and Y can also be analytically found in addition to the variance and covariance. Recall from the discussion in Section 3.1, that the parameters of skew-normal approximations of X_γ and Y_γ can be efficiently found from their respective mean, variance and skewness. Assuming that X_γ and Y_γ have a joint skew-normal distribution given in the previous section, the approximate moments of the $\max(X, Y)$ can be computed from the analytical results derived in the previous section. The tightness probability, mean, variance and skewness of the maximum can be found from these moments of the max. Thus, given the statistics of two correlated arrival time distribution we can compute the statistics of their maximum and their respective tightness probability needed for the statistical maximum algorithm in SSTA. Therefore, these results can be applied to existing SSTA approaches that rely on [37] for computing the approximate $\max(X, Y)$ in canonical form.

3.3 Simulation Results

In this section, we will present a comparison between the proposed skew-normal max results and Clark’s max results. Our goal is to show the usefulness of the proposed max operation in an SSTA framework. Therefore, to emulate the actual use of these results in true statistical max algorithm, we generated a test suite consisting of skewed arrival time distributions by running 50,000 Monte Carlo(MC) simulations on a toy circuit that mimics the behavior of a real circuit. In this setup, the arrival times at the primary inputs were assumed to have a correlated multivariate normal distribution. The relative mean alignment, the ratio of the variance and the correlation were swept within a reasonable range for gen-

erating skewed distributions at the internal nodes. For each max operation performed at an internal node, the two input operand arrival time distributions and their resulting maximum arrival time distributions were logged during MC simulations. The statistics of the operand arrival times were used as an input to the proposed max function implemented in C++. We implemented the analytical results presented in [91] for evaluating $T(X,Y)$. For comparison purposes, we also implemented the 5-parameter Clark's max results. The error in the result of both the proposed max operation and Clark's max operation was computed relative to the MC simulation results of the output arrival distribution for each test case.

Now for every test case, we computed the statistical parameters of the two input arrival time distributions, namely, μ_{x_γ} , σ_{x_γ} , Sk_{x_γ} , μ_{y_γ} , σ_{y_γ} , Sk_{y_γ} and ρ . These 7 statistical parameters were the input to the proposed skew-normal function. Using the moment matching method presented in Section 3.1, we first find the parameters of skew-normal distribution and then using the analytical max results derived in Section 3.2, we compute the moments of $\max(X,Y)$. An example illustrating the efficacy of the max operation is given in Figure 3.5, 3.6 and 3.7. Given the statistics of $X(1060.55, 58.56, 0.56)$, $Y(1045.53, 66.73, 0.80)$ and their correlation, the parameters of skew-normal probability distribution function are first computed. It can be seen from these figures that the skew-normal distribution accurately represents the MC generated skewed arrival time distribution as compared to the symmetric normal distribution for both the inputs. Consequently, as shown in Figure 3.7, a skewness based treatment of the input arrival time distribution gives a $\max(X,Y)$ distribution that accurately matches the MC simulation results.

We found that the error in the standard deviation of the max operation based on a normal assumption increases significantly with increase in skewness of the two input arrival time distributions. This is illustrated in Figure 3.8 where we show a plot of percentage error in computing the standard deviation of the $\max(X,Y)$ as a function of the skewness in $X(Sk_x)$. It is evident from this plot that the proposed skew-normal max operation can significantly improve the accuracy of existing SSTA approaches.

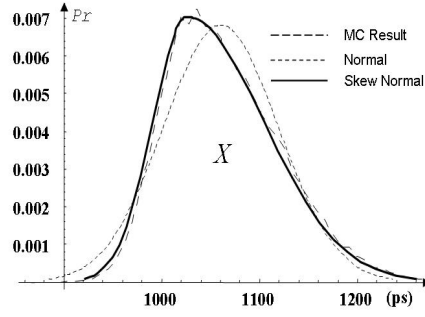


Figure 3.5. Example: Input X PDF.

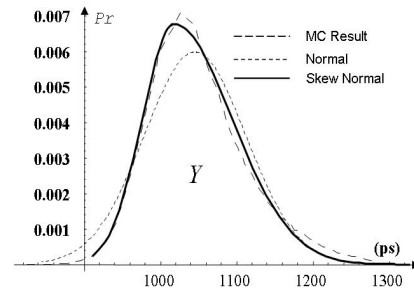


Figure 3.6. Example: Input Y PDF.

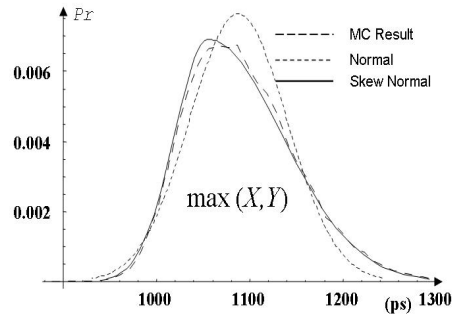


Figure 3.7. Example: Result $\max(X,Y)$ PDF.

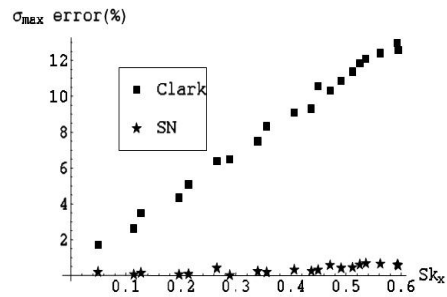


Figure 3.8. Comparison of standard deviation σ_{\max} error (%) as a function of input arrival time skewness Sk_x

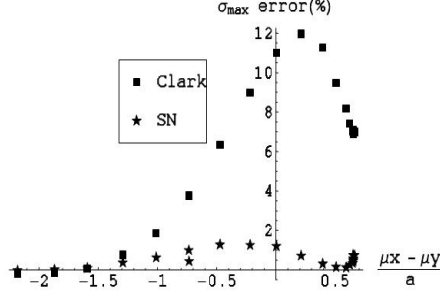


Figure 3.9. Comparison of standard deviation σ_{max} error (%) as a function of $\frac{\mu X_\gamma - \mu Y_\gamma}{a}$

Furthermore, as mentioned in [37], the error of the max operation also increases when the difference between μX_γ and μY_γ decreases. We observed a similar trend in our simulation results. In Figure 3.9 we present error plots of percentage error in standard deviation of output arrival time as a function of $\frac{\mu X_\gamma - \mu Y_\gamma}{a}$. It is clear from this plot that the proposed method exhibits much better robustness to difference in the mean of input arrival time distribution.

Recall from the discussion in Section 3.2.1, that while computing the correlated coefficients of $Z = \max(X, Y)$, a linear sum of the two operand coefficients weighted by their respective tightness probability is computed. In Figure 3.10, we show a comparison of error for this step between the skew-normal max and Clark’s max operation. It is evident from this plot that the error in computing the correlation coefficient using both the max operations is very similar. This result illustrates the fact that correlated propagation of canonical forms can be achieved using tightness probabilities computed from skew-normal max results.

In addition to the above results, for evaluating the impact of the proposed skew-normal max operation on the benchmark circuits we also compared the proposed max operation with Clark’s max by comparing them with MC simulations. The benchmark circuits were synthesized using an industrial 0.13μ technology and placed using Cadence Silicon Ensemble. The $3\sigma/\text{mean}$ of 20% was considered for channel length and gate-length independent threshold-voltage variations. All variation in threshold-voltage was assumed to be random

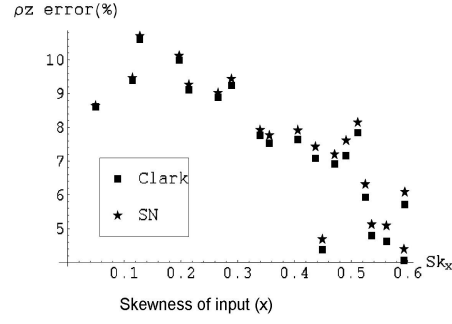


Figure 3.10. Comparison of error in correlation coefficient calculation as a function of input arrival time skewness Sk_x

(due to random dopant effects), whereas half the variation in channel length was considered to be correlated. The gates in the library were characterized for delay using SPICE simulations for different values of channel length and threshold-voltage, which were fit to a linear canonical form. A grid based spatial correlation model similar to the one proposed in [30] was used. In the absence of real correlation data, the correlation coefficient among different squares on the grid was assumed to be inversely proportional to the distance between the centers of their grids. MC simulations were performed by generating correlated normal random variables for the process parameters and arrival time distributions were logged for each max operation computed in the circuit. Both the max results were tested against these MC generated statistics for each max operation computed in the circuit.

Similar to the previous research [30, 28], the maximum error in mean computed from both Clark and skew-normal max was found to be less than 1%; however, as shown in Table 3.1, for standard deviation the accuracy of skew-normal max was found to be better than the Clark's max. The maximum and average error in standard deviation computed using both Clark and the proposed skew-normal max over all max operations performed in the circuit is listed in Table 3.1. Note that in most of the cases the proposed skew-normal max operation has better accuracy in standard deviation than the Clark's max approach. Interestingly, although the maximum error in Clark's standard deviation is on the high side but the average error is comparable to the skew-normal max. This result suggests that for an efficient implementation in an SSTA algorithm one can selectively use the proposed skew-

normal max operation where necessary. On the other hand, due to the linear canonical form and normal distributions of process parameters, the current results only model the skewness introduced due to the nonlinear max operation. Therefore, we believe for more realistic nonlinear delay models and non-normal process parameters, due to additional inherent skewness, further improvement over Clark's max can be achieved.

Additional investigation of the sources of error in the proposed max operation revealed that most of the maximum error cases for the skew-normal max occurred when the correlation was relatively high (typically > 0.9). Because of the systematic nature of this error we believe it will be possible to reduce it further. Nevertheless, to the best of our knowledge, this is the first work that analytically addresses the problem of computing the maximum of non-normal distributions in SSTA. We believe that the proposed work appeals to be a promising new direction for improving the accuracy of max operation in SSTA algorithms.

Table 3.1. Comparison of maximum and average error of standard deviation between Clark's max and skew-normal max over all the max operations

Circuit	#Gates	max. % Error in SD		Avg. Error in SD	
		Clark	skew-normal	Clark	skew-normal
c432	257	2.800	1.153	0.054	0.051
c499	545	1.334	1.685	0.067	0.066
c880	501	2.345	-0.780	0.054	0.066
c1908	604	2.588	2.090	0.078	0.070
c2670	781	1.209	0.688	0.048	0.047
c3540	1164	3.489	2.222	0.064	0.068
c5315	1693	6.736	2.589	0.064	0.063
c7552	2153	3.694	3.063	0.085	0.078
i2	193	0.776	0.561	0.044	0.041
i4	265	1.072	-0.773	0.156	0.137
i5	424	1.834	0.912	0.145	0.091
i6	462	0.977	-0.651	0.090	0.066
i7	770	1.086	0.879	0.099	0.087
i8	1014	1.795	1.151	0.072	0.059
i10	2483	2.583	1.486	0.050	0.049

In this work we present an analytical approach that extends Clark's max results to skew-normal distributions for computing the statistical maximum of two skewed arrival time dis-

tributions. An efficient method is presented to approximate the arrival time distribution using skew-normal representation. This done by matching the mean, the variance and the skewness of arrival time distributions to that of the skew-normal approximation. Using this method we then derived analytical results for computing the approximate moments of the maximum of the arrival time distribution assuming their joint PDF as a bivariate skew-normal distribution. From these moments the tightness probability, mean, variance and skewness of the maximum can be computed and therefore the presented results can be applied in existing SSTA algorithms that work on Clark's results. Our numerical results show that the proposed max operation can improve the accuracy of existing SSTA approaches. Furthermore, the skewness based proposed max function can be used to augment existing SSTA framework to propagate three moments.

CHAPTER 4

Parametric Yield Maximization using Gate Sizing

A number of analysis techniques have been developed to consider the impact of variability on timing [30, 28, 60, 62] and power [77, 92]. However, these approaches do not estimate the true parametric yield of a design considering both power and performance correlation. Reference [93] was one of the first works to consider the impact of variability on circuit optimization. The authors described the formation of a timing wall due to deterministic power optimization which increases the susceptibility of the design to process variations, and proposed a heuristic approach to prevent the build-up of a large number of paths near the critical delay of the circuit. This was achieved by adding a penalty function which had a negative impact on the objective function value whenever a path had a delay which was near critical. However, the approach was deterministic in nature and did not use any statistical information during optimization. Following this work, several statistical timing [94, 95, 96] and power optimization [76, 92, 97] approaches were also proposed. However, all these approaches neglect the correlation between power and performance. Therefore, timing yield improvements inevitably result in degradation in power yield and vice-versa. Moreover, most of these approaches suffer from large computational complexity and runtime. Thus, there is a critical need to develop approaches that perform true and efficient parametric yield optimization, where yield is defined using both power and timing constraints. The authors in [98], proposed an approach to perform gate-level yield analysis in a computationally efficient manner while considering the correlation between power and performance. The approach was based on a principal-component based process

variation modeling technique to perform timing and power analysis using the same set of underlying random variables (RVs), allowing the correlation in power and performance to be captured. Additionally the approach considers all components of process variations: inter-die and intra-die (spatially correlated and random) variability and can therefore serve as a framework to enable true parametric yield optimization.

In this work, we propose an alternative approach to perform yield optimization using gate sizing. The yield optimization is formulated as an unconstrained optimization problem, where the objective is to maximize the parametric yield of a design. The optimization is performed using a gradient-based non-linear optimizer. A brute-force gradient computation approach based on iterative yield analysis, however, leads to large computational overheads. Therefore, we propose an efficient heuristic technique to perform the computation of the yield gradient. This is achieved by perturbing the size of a gate in the circuit and heuristically recalculating the delay and power probability distribution functions (pdfs) of the perturbed circuit. The timing pdf of the perturbed circuit is calculated based on a novel cut-set approach that analyzes only a subset of the nodes in the circuit to estimate the complete delay pdf. The power pdf of the perturbed circuit is calculated with an incremental power analysis. This involves subtracting the power dissipation of the perturbed gate from the pdf of total power for the complete circuit and then adding the power dissipation of the perturbed gate while accounting for the gate size change. These perturbed pdfs are then used to compute the yield of the perturbed circuit by integrating the perturbed bivariate Gaussian distribution over the region defined by the leakage power and timing constraint. This gradient computation technique is then integrated with LANCELOT [99], a large-scale non-linear optimizer, to improve the parametric yield of the design, and is found to provide an 8X improvement in runtime with an average error of 0.1% with respect to the brute force approach.

The remainder of the chapter is organized as follows. Section 4.1 briefly reviews the principal component based approach to perform yield analysis. Section 4.2 presents the

incremental timing and power analysis techniques, which are then used in section 4.3 to compute the gradient of yield. In Section 4.4, we provide details regarding the implementation of our yield optimization approach and present results including a comparison of our approach to deterministic optimization.

4.1 Review of Yield Analysis

In this section, we briefly discuss our modeling assumptions and yield analysis approach. We also define the yield optimization problem and describe a brute-force approach to perform yield optimization. The computational complexity of this brute-force approach motivates the need for more efficient gradient computation techniques.

Our yield analysis framework is based on the approach in [98], and we express the delay and a leakage of a gate as

$$\begin{aligned} Delay &= d_{nom} + \sum_{i=1}^p \alpha_i \Delta P_i, \\ Leakage &= \exp \left(V_{nom} + \sum_{i=1}^p \beta_i \Delta P_i \right). \end{aligned}$$

where d_{nom} and $\exp(V_{nom})$ are the nominal gate delay and leakage, respectively, and α_i and β_i captures the dependence of gate delay and the logarithm of gate leakage on the p process parameters of interest. The RVs ΔP_i in the above equation refers to the variations in these process parameters that are assumed to be Gaussian. The variation in process parameter is then partitioned into a correlated and random component. The correlated component is handled by partitioning the design as shown in Figure 1. For each parameter, each square in the grid is assigned to a Gaussian RV that captures the correlated variation in that process parameter, which is defined using a correlation matrix. This gives a total of N_g RVs for each parameter, which are assumed to have a joint multivariate-normal distribution. Using principal components analysis [66], the correlated component is expressed as a linear combination of N_g independent Gaussian RVs (z_i), and the random variation in all the process parameters is lumped into a single RV - η_d for delay and η_l for leakage, and the

coefficient of the random component is calculated by matching the variance of the random contribution. This approach gives us canonical expressions for gate delay and leakage power which are expressed as:

$$\begin{aligned} Delay &= d_{nom} + \sum_{i=1}^p (\alpha_p \sum_{j=1}^n \gamma_j z_{j,i}) + \eta_d R, \\ Leakage &= \exp(V_{nom} + \sum_{i=1}^p \beta_i \sum_{j=1}^n \gamma_j z_{j,i} + \eta_l R). \end{aligned}$$

Timing analysis is then performed in the spirit of [30, 28], and the delay is propagated through the circuit, while maintaining the node delays in the same canonical form with different coefficients. The sum operation is performed by simply adding the coefficients for each of the RVs, other than the random component whose coefficient is obtained as the square root of the sum of the squared coefficients of the random components of the summed pdfs. The max operation is performed by matching the mean, variance and the correlation of the max of two RV (which are obtained using [37]) and the canonical expression of the max. Leakage power analysis is based on summing log-normal RVs using Wilkinsons method [100] as proposed in [77]. The leakage of each gate is iteratively added to the sum which is maintained in canonical form. For each addition, the coefficients of the canonical expression for the sum are calculated by matching the mean, variance and correlation with the principal components of the sum (obtained using [77]) and the canonical expression for the sum. At the end of timing and power analysis (that provides the delay and leakage power canonical form) the correlation between delay and leakage power is estimated using:

$$Cov(Delay, Leakage) = \sum_{i=1}^n \alpha_i \beta_i. \quad (4.1)$$

The five parameters (mean and variance of delay and leakage power and their correlation) are used to define a bivariate Gaussian distribution for delay and log of leakage power as shown in Figure 2. The parametric yield which is defined as:

$$Y = P(D \leq D_0, P \leq P_0), \quad (4.2)$$

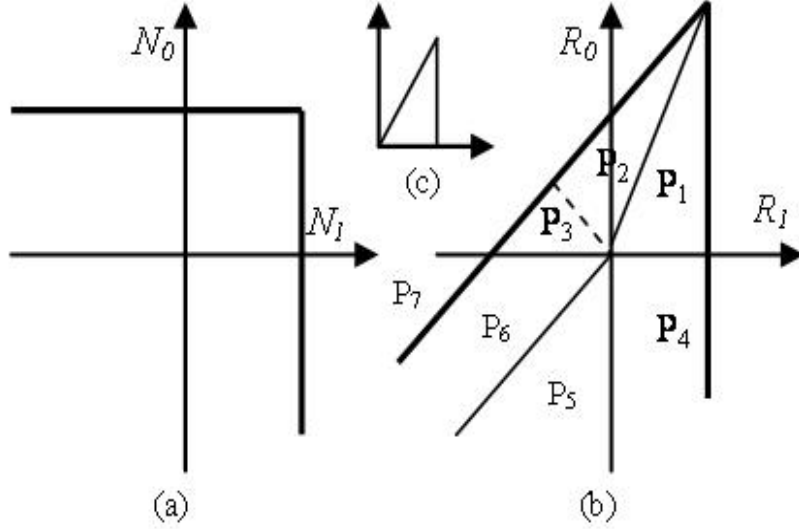


Figure 4.1. Transformation of the feasible region from (a) to (b) under the transformation expressed in (21) for negative values of correlation.

where D is delay of the circuit constrained to be less than D_0 and P is the power of the design constrained to be less than P_0 . In this work, we assume that variations in power are dominated by variations in leakage power and the dynamic power dissipation is assumed to be a fixed number and subtracted out of the total power budget of the design to define the leakage power constraint. Based on this assumption, we can rewrite (4) as

$$Y = P(D \leq D_0, \log P_L \leq \log(P_0 - P_D)), \quad (4.3)$$

where P_L and P_D are the leakage and dynamic power of the design. The above yield expression is now equivalent to the integral of a bivariate Gaussian RV over a rectangular region, and can be evaluated using expression developed in [90]. Both the timing and power computation steps require $O(nN_g)$ steps, where n is the number of gates in the design and N_g is the number of regions into which the design is partitioned to capture the correlation structure of correlated process variations. However, the final yield computation step (5) itself runs in constant time since the yield computation is always performed using the set of five parameters, independent of the size of the problem.

Based on this yield analysis engine, a brute-force approach to perform yield optimization using gate sizing can be developed. This involves computing the gradient of yield to the size of each gate, which can be estimated by resizing each gate and performing yield analysis and setting the gate back to its original size. After computing the gradient, we use a large scale non-linear optimizer to improve the yield of the circuit. We now consider the computational complexity of one iteration of this approach. Each gradient computation requires $n + 1$ yield analysis runs and thus has an overall complexity of $O(n^2 N_g)$. As the size of each partitions is fixed, the number of partitions N_g can also be expected to increase with the size of the design. Thus, the overall computational requirements soon become untenable for large designs. Also, note that the brute-force approach spends most of the time recalculating the same information for most of the circuit and motivates the need for an efficient gradient computation approach.

4.2 Gradient Computation

In this section, we will discuss our new gradient computation approach that calculates the updated timing and delay pdfs based on a change in gate size. Both the timing and power perturbation analysis techniques update the coefficients of the delay and leakage pdf expression based on the change in gate size. These updated delay and leakage power pdfs are then used to compute the yield of the perturbed design. The change in yield is used to estimate the gradient of yield to the size of each gate in the design.

4.2.1 Timing Perturbation Computation

We will explain our timing perturbation computation approach based on cut-sets using the following graph representation for our circuits.

Definition 3 *Definition 1: A timing graph is a directed acyclic graph having exactly one source and one sink: $G = N, E, ns, nf$, where $N = n_1, n_2, \dots, n_k$ is a set of nodes, $E = e_1, e_2, \dots, e_l$ is a set of edges, $ns \in N$ is the source node and $nf \in N$ is the sink node and*

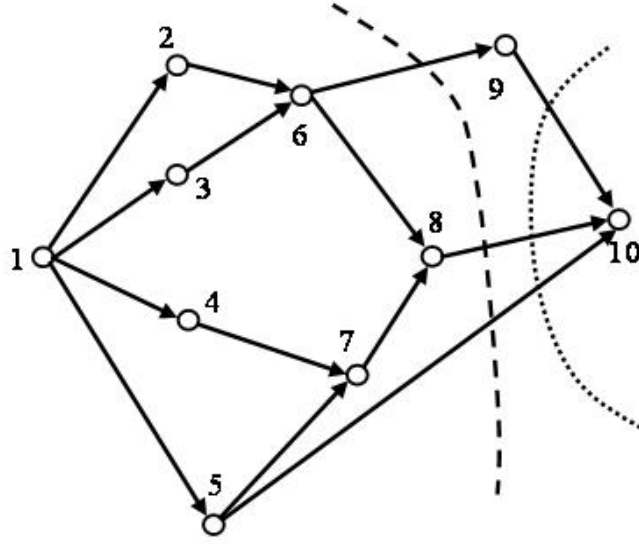


Figure 4.2. A timing graph showing a linear topological ordering for the nodes and cut-sets for nodes 8 and 9

each edge is an ordered pair of nodes $e = (n_i, n_j)$ and each node is associated with a delay for each fanin edge, which depends on the characteristics of the fanout nodes.

The nodes in the timing graph correspond to gates and the edges correspond to nets in a circuit. A probabilistic timing graph is defined as a timing graph where each node is associated with a RV for the delay for each fanin edge. Figure 3 shows an example timing graph with ten nodes, eight of which represent actual gates and nodes 1 and 10 represent the source and sink nodes, respectively. The latest arrival time (AT) and required arrival time (RAT) probability distribution functions (pdf) for each node in the timing graph are now defined as:

Definition 4 The latest arrival time (AT) at an edge e in the probabilistic timing graph is a RV whose CDF $A_e(t)$ gives the probability that a deterministic sample of this timing graph has an arrival time less than t .

Definition 5 The earliest required arrival time (RAT) at an edge e in the probabilistic timing graph is a RV whose CDF $R_e(t)$ gives the probability that the deterministic sample meets the timing constraint T_{crit} if the deterministic arrival time at the node is less than t .

Note that the sum of the AT and RAT at a node represents the partial pdf of delay since it does not take into account the influence of the edges that are not present in either the fanin or the fanout cone of the node on the pdf of circuit delay. To express the dependence of circuit delay on the delay of one of the nodes let us define the following terms.

Definition 6 *A linear topological ordering (LTO) of the nodes in a timing graph is a total order based on the relationship that the order of any node x that lies in the fanout cone of a node n is strictly larger than the order of node n , and that no two nodes in a timing graph have the same order.*

An LTO of a timing graph can be easily determined by performing a breadth-first traversal of the timing graph. Though a given timing graph can have many LTOs, finding the optimal LTO is not the focus of this chapter. Figure 3 illustrates a timing graph with nodes labeled according to a LTO of the timing graph. Note that swapping nodes 8 and 9 will still maintain a valid LTO of the nodes.

Definition 7 *A cut-set of a timing graph with a given LTO of a node n is defined to be the set of edges (n_i, n_j) of the timing graph which satisfy $LTO(n_i) \leq LTO(n)$ and $LTO(n_j) > LTO(n)$.*

Definition 8 *A node x of the timing graph belongs to the cut-set source of node n if there exists an edge $(x, *)$ which belongs to the cut-set of node n .*

Definition 9 *The fanin-set of a node n of a timing graph is the set of immediate predecessor nodes of node n .*

Definition 10 *The arrival time set or ATSet of a node n is the union of the fanin-set of node n and the nodes in the fanout cone of the fanin-set of node n that have order less than or equal to the order of node n .*

The convolution-set or ConvSet of a node n is the intersection of the ATSet and cut-set source of node n . Any cut-set of the timing graph divides the timing graph into two disconnected components and the statistical maximum of the sum of the AT and RAT of all edges in the cut-set gives the complete pdf of circuit delay. Now, if we perturb the delay

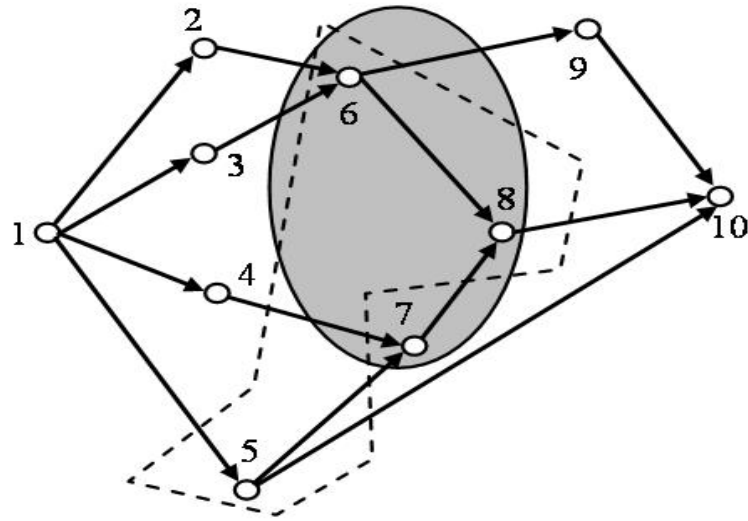


Figure 4.3. A timing graph showing the ATSet (nodes within the shaded ellipse) and cut-set source set (nodes within the dashed shape) for node 8

characteristics of a node n (by gate sizing) we also change the capacitive loading of the fanin gates, affecting their delay characteristics as well. To compute the new circuit delay we note that the RAT of the edges in the cut-set does not change, since all the gates in their fanout cone includes gates that have order strictly greater than the order of node n , and have unchanged delay characteristics.

The AT for all edges that are in the fanout cone of the fanin-set of node n changes. However, we are only interested in AT changes for edges that are driven by nodes that have order less than the order of gate n , since we need to compute the AT for the edges in the cut-set only. This is exactly the set of nodes defined by the ATSet of node n . If the AT of an edge in the cut-set changes we need to recompute the convolution of the AT and RAT at that edge. These edges are driven by the nodes in the intersection of the ATSet and the cut-set which is defined as the ConvSet of node n .

Let us revisit the example timing graph in Figure 3 and consider node 8. The cut-set for this node is the set of edges (6,9), (8,10) and (5,10) as shown by the dashed line. The ATSet for the node can be identified as the set of nodes 6, 7 and 8 as shown in Figure 4. The intersection of the cut-set source and ATSet defines the ConvSet and is the set of nodes

```

CUTSETSTA (n)
  for each node ( $x \in \text{ATSET}(n)$ )
    Compute  $AT(x)$ ;
  for each node ( $x \in \text{CONVSET}(n)$ )
     $CT(x) \leftarrow \text{convolution}(AT(x), RT(x))$ 
  for each edge ( $x \in \text{CUTSET}(n)$ )
     $T_n \leftarrow \text{maximum}(T_n, CT(x))$ 
  return  $T_n$ 

```

Figure 4.4. Pseudo-code for the CutSetSTA routine

6 and 8. Note that the ConvSet identifies that the AT and RAT has not changed on the edge (5,10) and we do not need to recompute the convolution of the AT and RAT for this edge. However, if we consider node 9 the cut-set is defined by the edges from nodes 5, 8 and 9 to node 10, as shown as the dotted line in Figure 3. The pseudo-code to calculate the delay pdf of the perturbed circuit is shown below, where we refer to the edge by the name of the driving node. The pseudo-code involves the computation of the AT for all nodes in the ATSet, convolution of the AT and RAT for all nodes in the ConvSet and the statistical maximum of the convolution for all edges in the cut-set.

Note that all the computations in CutSetSTA are performed using the same canonical expression for the delay pdf. Thus, the final delay pdf of the perturbed circuit is also expressed in the same form. Although, the approach as described seems exact, it is heuristic. This results from the fact that the computation of the max function of delay pdfs is not exact and forward and backward traversals of the graph result in timing delays that are not exactly same. However, this error is very small as will be shown later in the results section.

4.2.2 Power Perturbation Computation

The statistical power computation is performed by summing the power dissipation of each gate in a circuit to compute the complete pdf of leakage power. To perform power analysis of a circuit with perturbations in the size of a gate, we first perform statistical power analysis of the unperturbed circuit as described in Section 2. Now the leakage power after the size of gate i has been perturbed is expressed as

$$\begin{aligned} P_{circ}^{pert} &= P_{circ}^{unpert} - P_{gate,i}^{unpert} + P_{gate,i}^{pert} \\ &= P_{circ\ i}^{unpert} + P_{gate,i}^{pert}, \end{aligned}$$

where P_{pert} and P_{unpert} refer to the perturbed and unperturbed power, respectively and the subscript indicates whether the power refers to the circuit or to the gate. Since the leakage power is expressed a log-normal (exponential of a Gaussian) RV, we can approximate their sum using another log-normal. In general, if we sum P_{leak}^b and P_{leak}^c to obtain P_{leak}^a , which is mathematically expressed as,

$$\begin{aligned} P_{leak}^a &= \exp\left(a_0 + \sum_i^n a_i z_i + a_{n+1}\right) \\ &= \exp\left(b_0 + \sum_i^n b_i z_i + b_{n+1}\right) + \exp\left(c_0 + \sum_i^n c_i z_i + c_{n+1}\right) \\ &= P_{leak}^b + P_{leak}^c, \end{aligned}$$

the coefficients in the expression for P_{leak}^a can be obtained by matching the mean, variance and the correlation coefficient with the exponential of the principal components (z_i s). This gives us a set of $n+2$ equations in $n+2$ variables which can be analytically solved to obtain the following expression for the coefficients associated with the principal components,

$$\begin{aligned} a_i &= \log\left(\frac{E(P_{leak}^a e^{z_i})}{E(P_{leak}^a)E(e^{z_i})}\right) \\ &= \log\left(\frac{E(P_{leak}^b e^{z_i}) + E(P_{leak}^c e^{z_i})}{(E(P_{leak}^b) + E(P_{leak}^c))E(e^{z_i})}\right). \end{aligned}$$

Using the expressions developed in [98], the remaining two coefficients in the expression for P_{leak}^a can be expressed as

$$a_0 = \frac{1}{2} \log \left(\frac{(E(P_{leak}^b) + E(P_{leak}^c))^4}{(E(P_{leak}^b) + E(P_{leak}^c))^2 + \text{Var}(P_b) + \text{Var}(P_c) + 2\text{Cov}(P_b P_c)} \right),$$

$$a_{n+1} = \left(\log \left(1 + \frac{\text{Var}(P_b) + \text{Var}(P_c) + 2\text{Cov}(P_b P_c)}{(E(P_{leak}^b) + E(P_{leak}^c))^2} \right) - \sum_{i=1}^n a_i^2 \right)^{\frac{1}{2}}.$$

Note that to compute the final perturbed leakage, we need to use the above expressions twice. However, when one of the log-normals is subtracted the signs associated with its expected value and covariance terms in the above expressions are reversed.

4.3 Yield Gradient

To this point we have developed efficient approaches to perform statistical timing and power perturbation computation. Now, we will use these techniques to perform the computation of the gradient of yield in an efficient manner. The computation of yield gradient involves the computation of the perturbation in yield for small changes in the size of gates in the design. The pseudo-code for the computation of yield is included in Figure 6. After each resizing move, the non-linear optimizer calls the yield computation routine FastYield-Gradient. The first step is to initialize the circuit so that all nodes are assigned the correct load capacitance based on the sizes of the gates in its immediate fanout, and the correct leakage power based on its own size. Based on the load capacitance of the node each input of a node is assigned to a delay pdf, which represents the delay of the timing arc from that particular input to the output of the gate.

After the initialization step, the next step involves the propagation of the AT from the source node to the sink node in the timing graph. This is represented as ForwardSSTA in the pseudo-code. The next step is to perform statistical power analysis and generate the leakage current pdf using StatPowerAnalysis. The Yield function is then used to compute the yield based on the timing and leakage power pdfs given a leakage power constraint P and a delay constraint D , as outlined in Section 2.

```

FASTYIELDGRADIENT (CIRCUIT, SIZE )
  for each gate ( $g \in \text{CIRCUIT}$ )
    update load cap and size of ( $g$ ) using SIZE;
  for each gate ( $g \in \text{CIRCUIT}$ )
    compute new gate delay & power of ( $g$ )
   $T \leftarrow \text{FORWARDSSSTA} (\text{CIRCUIT})$ 
   $P \leftarrow \text{STATPOWERANALYSIS} (\text{CIRCUIT})$ 
   $Y \leftarrow \text{YIELD} (P, T)$ 
  do REVERSESSSTA (CIRCUIT)
  for each gate( $g \in \text{CIRCUIT}$ )
    save current state of CIRCUIT
     $s^+ \leftarrow \text{SIZE}(g) + \Delta\text{SIZE}(g)$ 
    compute new gate delay & power of  $g$ 
    for each gate( $i \in \text{FANIN}(g)$ )
      update load cap and delay of  $i$ 
     $T^+ \leftarrow \text{CUTSETSSSTA} (\text{CIRCUIT})$ 
     $P^+ \leftarrow \text{INCREMSPA} (\text{CIRCUIT}, P)$ 
     $Y^+ \leftarrow \text{YIELD} (P^+, T^+)$ 
     $\nabla Y(g) \leftarrow (Y^+ - Y) / \Delta\text{SIZE}(g)$ 
    restore the original state of the CIRCUIT
  return  $\nabla Y$ 

```

Figure 4.5. Pseudo-code for the computation of the yield gradient

To perform the computation of yield gradient, we first propagate the RAT from the sink node to the source node using ReverseSSTA. Then we go through each node in the circuit iteratively and perturb the size of each gate by a small amount. The load capacitance of the nodes in the fanin-set of the node and the delay pdf assigned to each timing arc of this node and the nodes in the fanin-set are updated. Then using the statistical timing and power perturbation computation techniques discussed in Sections 3.1-3.2 we compute the delay and leakage power pdfs of this perturbed circuit. The yield corresponding to the perturbed circuit is then calculated and the change in yield is used to define the particular component of the yield gradient.

Let us consider the computational complexity of our proposed approach and compare it to the brute-force approach, where each iteration had a complexity of $O(n^2 N_g)$. Each iteration, in our proposed approach, involves a single run of the complete yield analysis approach, as discussed above, which has a complexity of $O(n N_g)$. The timing and power perturbation computation is repeated $O(n)$ times. The complexity of the incremental timing analysis is $O(N_g w)$ where w is the width of the cut-set and that of power analysis is $O(N_g)$ since we require only two sum operations of the lognormal RVs. Thus the overall complexity of the approach is $O(n N_g w)$. In a typical combinational circuit $w \ll n$, therefore we achieve significant runtime improvement compared to the brute-force approach.

4.4 Simulation Results and Implementation Details

We implemented the proposed approach in C and compared our yield improvements to a deterministic circuit optimization technique. Our proposed approach for the computation of the yield gradient is written as a subroutine which the optimizer uses to calculate the gradient of the objective function. The yield analysis engine serves as the subroutine to calculate the objective function itself. Following, we present the accuracy and runtime results for the proposed approach.

4.4.1 Runtime Comparisons

If we assume that reverse and forward SSTA give exact timing distributions at each node, then the procedure would be exact as well. However, as noted before, due to the Gaussian approximation considered while computing the maximum introduces a small inaccuracy while performing forward and reverse SSTA. Now, since this inaccuracy is a function of circuit topology and reconvergence structure the sensitivity of yield computed using only ForwardSSTA based brute-force is negligible. Table 4.1 shows the runtime comparison and accuracy results of the proposed gradient computation procedure as compared to the naive brute force approach. The circuit size in terms of the number of gates and the average cut-width over all nodes in the circuit is also reported in the second and the third columns, respectively. Runtime per gradient vector computation using the brute approach and the proposed procedure are given in Columns 4 and 5, respectively. The speed up of the proposed method over the brute-force approach is given in Column 6, and ranges between 3X to 20X and is found to be larger for bigger circuits. The maximum error, over all gates, found using gradient computation normalized with respect to the brute-force method is given in Column 7, and is found to be small in most cases with an average of 2.4%. The error averaged error over all gates in the circuit is given in the last columns of Table 1 and is found to be extremely small.

4.4.2 Yield Optimization

The gates in our standard cell library are characterized for a set of sizes in the range from minimum size to maximum size and the delay and leakage power for intermediate gate sizes is obtained using linear interpolation. All designs are then deterministically optimized for power under delay constraints using either design compiler or LANCELOT [99]. We use our statistical yield maximization approach to improve the yield of this optimized design for a set of different power and timing constraints. Our results indicate that performing statistical optimization can significantly improve the timing yield of the design. We compare

Table 4.1. Comparison of Yield gradient computation using FASTYIELDGRADIENT AND BRUTE-FORCE APPROACH.

Circuit	No. Gates	Average Cut-Size	Brute-Force (s)	Fast-Grad. (s)	Speed-up	Max. Error (%)	Avg. Error (%)
c432	257	46.8	0.6	0.1	7.0	1.2	6.7E-03
c499	545	96.0	5.2	0.7	7.1	0.3	3.4E-03
c880	501	103.8	4.8	0.6	7.9	0.2	1.2E-02
c1908	604	84.3	6.5	0.7	9.8	2.6	1.7E-02
c2670	781	248.2	5.7	1.1	5.3	5.2	1.0E-03
c3540	1164	140.6	41.9	3.5	12.1	1.0	1.5E-03
c5315	1693	295.1	130.3	13.1	9.9	7.4	2.4E-03
c6288	3835	297.9	991.3	51.0	19.5	1.8	1.4E-03
c7552	2153	317.8	979.4	51.1	19.2	1.8	1.4E-03
i2	193	105.0	214.8	20.3	10.6	2.0	1.5E-03
i4	265	105.8	0.4	0.1	2.9	1.3	1.8E-02
i5	424	137.5	0.6	0.2	2.8	8.0	9.4E-02
i6	462	177.8	3.2	0.7	4.5	0.1	2.3E-02
i7	770	252.3	2.0	0.7	3.0	1.8	6.0E-02
i8	1014	243.7	10.4	2.2	4.7	1.0	4.0E-02
i10	2483	413.0	19.6	3.8	5.2	2.3	2.3E-03

our results based on the ISCAS85 [101] benchmarks which were synthesized in a 130 nm technology.

The yield optimization results are given in Table 4.2. The first sub-section including columns 2, 3, 4 and 5 report the initial timing and power statistics resulting from a deterministically optimized circuit. The deterministic optimization was performed using nominal delay and power models. We present yield optimization results for two sets of constraints. The first set includes yield optimization for aggressive nominal value constraints. As a deterministic optimizer is unaware of the variation in power and timing and their correlation, the initial yield at nominal constraints is extremely small. However, the proposed variability aware yield optimization significantly improves the yield. For example, the yield for benchmark circuits c432, c499 and c880 dramatically improves from close to 0% to up to 40%. Column 6, 7, 8 and 9 report the post optimization timing and power statistics of the circuit. The initial yield subject to nominal value constraints and the yield after performing optimization are given in Columns 10 and 11, respectively. The second set

of results report the performance of yield optimization under pessimistic constraints. For this case we use the nominal value offset by one standard deviation as the constraint for both power and timing while defining the objective function for optimization. Again columns 12, 13, 14 and 15 list the post optimization statistics of the circuit whereas columns 16 and 17 report the results achieved after performing the proposed yield optimization. As the constraints are relaxed in this case the initial yield of the circuit improves and for the same reason the improvements achieved are relatively smaller as compared to the previous case. The maximum improvement in this case is found to be greater than 20% for the benchmark circuit c1908.

Table 4.2. Yield Optimization Results

Circuits	Initial solution				D ; Dm, P ; Pm						D ; Dm + Ps, P ; Pm +Ps					
	Delay(ps)		Power (uW)		Delay(ps)		Power (uW)		Yield(%)		Delay(ps)		Power (uW)		Yield(%)	
	Dm	Ds	Pm	Ps	Dm	Ds	Pm	Ps	Init.	Opt.	Dm	Ds	Pm	Ps	Init.	Opt.
c432	670	23	10.35	3.78	627	22	10.2	3.72	0	49	651	22	9.98	3.62	42	49
c499	755	26	30.66	10.56	716	23	31.01	10.1	0	47	739	25	28.81	9.87	42	47
c880	702	23	25.48	8.83	669	21	23.1	7.99	0	45	695	22	23.2	8	42	45
c1908	925	27	16.09	5.47	928	27	15.4	5.2	1	19	938	31	4.8	1.6	43	66
c2670	685	23	6.3	2.16	690	24	6.1	2.09	1	19	657	22	6.15	2.15	42	49
c3540	1134	35	48.73	16.14	1126	35	46.1	15.2	0	29	1126	35	46.1	15.25	42	45
c5315	982	31	74.97	24.4	995	31	69.7	22.6	0	16	985	31	72.6	23.6	42	42
c6288	2639	72	98.37	30.66	2651	72	91.2	28	1	19	2591	71	96.5	30	42	48
c7552	1181	34	72.39	23.54	1186	35	67.4	21.8	0	19	1150	30	67.8	21	42	48

CHAPTER 5

A Statistical Approach for Full-chip Gate-oxide Reliability Analysis

Semiconductor reliability and manufacturing variability have become key challenging issues as device critical dimensions shrink and integration complexity continues to grow at a rapid pace. For assessing product reliability, it is important to quantify the reliability of oxide which is its ability to retain its dielectric properties while being subjected to high electric fields. Aggressive oxide-thickness scaling has led to huge vertical electric fields in metal oxide semiconductor devices that result in high direct tunneling gate oxide leakage current. The gate oxide leakage current creates defects such as electron traps, interface states, holes traps, etc. in the gate-dielectric. These defects gradually build up in the oxide until a critical defect density is reached when the oxide destructively breaks down leading to a large increase in gate oxide conductance. Such a break down eventually results in a functional failure of the product.

Over the last 30 years, numerous publications have focused on understanding and modeling the mechanisms leading to defect generation and breakdown in individual devices [102]. Some researchers have initiated an effort to understand the oxide breakdown mechanisms of simple circuits [103]. Recently, a product level approach performing oxide breakdown analysis on full-chip was proposed in [104]. In most of the existing approaches, simple test structures such as discrete devices or capacitors are used to characterize the oxide breakdown mechanism for a specific manufacturing process. These discrete device characterization results are then extrapolated to deduce a model for the full-chip oxide reliability which is later calibrated using lifetime tests of sample product.

However, a major concern with prior approaches is that they assume a uniform oxide-thickness for all devices on every chip. In practice, the non-uniformity in temperature and pressure during the gate-oxidation process leads to within-die and die-to-die variations in oxide-thickness. As the precision of oxidation temperature and pressure control is not scaling commensurately with oxide-thickness, the control over oxide-thickness is worsening with process scaling. For a given supply voltage and operating temperature, the reliability of oxide is an exponential function of its thickness and its sensitivity to thickness variations increases for thinner oxides [105]. In previous approaches, it is therefore imperative to consider a uniform minimum oxide-thickness across all devices on a chip and across all chips for a conservative worst-case analysis. This may lead to significantly pessimistic estimates of the overall oxide breakdown reliability of the product. Furthermore, oxide reliability is one of the key factors that sets constraints on the operating supply voltage and temperature of the chip. Any pessimism in oxide reliability analysis limits the maximum operating voltage and thus the maximum achievable chip-performance. In order to find consistent supply voltages and operating temperature limits, it is therefore critical to quantify the product oxide breakdown strength.

The goal of this work is to develop a methodology for chip level gate oxide breakdown analysis while considering the impact of thickness variations. Such an analysis is useful for finding accurate statistical estimates of the reliable lifetime of the chip. We present a statistical framework for chip level gate oxide reliability analysis while considering both die-to-die and within-die components of thickness variation while also considering spatial correlations. The thickness of each device is modeled as a distinct random variable and thus the full chip reliability estimation problem is defined on a huge sample space of several million devices. We observe that the full chip oxide reliability is independent of the relative location of the individual devices. This enables us to transform the problem such that the resulting representation can be expressed in terms of only two random variables. Using this transformation, we present a computationally efficient and accurate

approach for estimating the full chip reliability while considering oxide-thickness variation and correlations. To our knowledge, this is the first attempt to perform the chip-level oxide reliability analysis while modeling all components of oxide-thickness and correlations. Apart from the manufacturing variations in oxide-thickness, the operational life of a part depends on variations in temperature, modes of operation, executed instructions, supply voltage, lifetime wear out, etc. However, the uncertainty of the operational life of a part can not be modeled statistically as the chip is expected to function reliably throughout its operational lifetime even in a worst-case operating environments for all specified applications. Therefore, unlike manufacturing variability, in our approach we consider the worst-case operating temperature and supply voltage to ensure a correct operation throughout the entire life time of the part for any application profile. Simulation results validate the accuracy of the proposed approach and justify the need for a statistical framework for modeling manufacturing variability.

The rest of the chapter is organized as follows: in Section 5.1 we describe the modeling of thickness variation. In Section 5.2, we discuss the oxide break down model and formulate the oxide reliability analysis problem. Section 5.3 explains the proposed methodology for estimating the full-chip oxide breakdown reliability. Simulation results illustrating the efficacy of the proposed approach are given in Section 5.4.

5.1 Modeling Thickness Variation

As discussed earlier like any other process parameter the oxide-thickness variation can be classified based on the spatial scale over which it manifests. Several factors gradually affect temperature from one location to another within a die (e.g., the emissivity variations resulting from location of die on the wafer). For accurate statistical analysis, it is necessary to capture the dependence between the global and spatial component of thickness variations. It is also important to model the residual *independent variation* resulting from certain local device-scaling effects such as different surface orientations, stress conditions

as well as poly-Si intrusion from the gate electrodes.

To exactly model spatial correlation between the oxide-thickness of two devices, a separate random variable is required for each device. However, the correlation between two devices is generally a slow monotonically decreasing function of their separation. Therefore, simplified correlation structures using a grid model [30] or quad-tree model [44] have been proposed in the literature. In this work, we discuss the proposed approach using the grid based model. In this model, the spatial component of oxide-thickness variation is modeled using n random variables, each representing the spatial component of variation in one of the p grids, and a covariance matrix of size $n \times n$ representing the spatial correlations among the grids. The covariance matrix could be determined from measurement data extracted from manufactured wafers using the method given in [69]. To simplify the correlation structure, this set of correlated random variables is mapped to another set of mutually independent random variables with zero mean and unit variance using the principal components of the original set. The original random variables are then expressed as a linear combination of the principal components. These principal components can be obtained by performing an eigenvalue decomposition of the correlation matrix. This representation of the correlation is expressed in a so-called canonical form [28, 30], where oxide-thickness x_i of any device in i^{th} grid is given by

$$x_i = \lambda_{i,0} + \sum_{j=1}^n \lambda_{i,j} z_j + \lambda_r \epsilon, \quad (5.1)$$

where $\lambda_{i,0}$ is the mean or nominal value of oxide-thickness in i^{th} grid, z_j represents the n independent random variables used to express the spatially correlated device parameter variations, ϵ is a distinct random variable for each device that represents the residual independent variation, and the coefficients $\lambda_{i,j}$ s represent the sensitivity of thickness variation in i^{th} to each of the j^{th} random variables.

5.2 Problem Formulation

The gate oxide degradation depends on the oxide-thickness, voltage, and temperature. There are many oxide breakdown models in the literature that attempt to explain the dependence on these factors. A widely accepted model is the anode hole injection model [106]. According to this model, injected electrons generate holes at the anode that can tunnel back into the oxide. Intrinsic breakdown occurs when a critical hole fluence is reached. A second model known as electron trap density model has been suggested, which claims that a critical density of electron traps generated during stress is required to trigger oxide breakdown [107]. Both models of oxide breakdown mechanisms note that the defect generation is a non-deterministic mechanism. As a result the oxide breakdown time is inherently a statistically distributed quantity. Thus the oxide breakdown time is modeled as a random variable typically characterized by a Weibull probability distribution function, given by [108, 105]

$$F(t) = 1 - e^{-a(\frac{t}{\alpha})^\beta}, \quad (5.2)$$

where F is the cumulative distribution function (CDF) of time-to-breakdown t , a is the device area normalized with respect to the minimum device area, α and β are the scale and shape parameters of the Weibull distribution. The scale parameter α represents the characteristic life which is the time where 63.2% of samples fail, whereas the shape parameter β is a function of critical defect density. The critical defect density depends on device oxide-thickness, the oxide field and temperature. For a given temperature and stress voltage, it has been shown that the slope parameter of the Weibull distribution varies linearly with oxide-thickness [109]. Thus if x denotes the oxide-thickness, we have

$$F(t) = 1 - e^{-a(\frac{t}{\alpha})^{bx}}, \quad (5.3)$$

where b is a constant for given worst-case temperature and supply voltage. Another major factor that affects the oxide lifetime is the oxide breakdown fail criterion. A commonly used fail criterion is soft-breakdown(SBD) which is characterized by a small increase in

gate leakage. In practice, however, after SBD the gate leakage current monotonically increases with time eventually leading to a hard breakdown [110]. The time between oxide SBD and HBD is a function of the gate area, oxide quality, and the bias conditions. For the purpose of this work, we limit our analysis assuming the initiation of SBD as the failure criterion since it is typically followed rapidly by HBD.

According to the SBD criterion the chip is considered to have failed as soon as the soft break down occurs for any device on the chip. We are interested in finding the reliable lifetime of the chip for which none of the devices fail. For such weakest link problems, it is more convenient to use an alternate representation known as reliability function $R(t)$ or survivor function, given by

$$R(t) = 1 - F(t) = P(T > t) = \int_t^{\infty} f(s)ds, \quad (5.4)$$

where $f(s)$ is the probability distribution function of oxide breakdown. The reliability function is complimentary to CDF $F(t)$, taking the value 1 at $t = 0$ and tending to 0 as t tends to infinity. Simply stated, a reliability function is the probability that a device(chip) does not fail by time t . Due to manufacturing variations the thickness of gate oxide is also a non-deterministic parameter at design time. Thus the reliability function of a device can be interpreted as its conditional reliability function for the given oxide-thickness. For i^{th} device having x_i oxide-thickness the conditional reliability function can be given as

$$R_i(t|x_i) = P(t > \mathbf{t}|x_i) = \int_t^{\infty} f(s|x_i)ds. \quad (5.5)$$

Due to the spatial component of oxide-thickness variation, the oxide-thicknesses of any two devices on a chip are correlated with each other. Therefore, in general, their respective reliability functions being functions of oxide-thickness are also correlated with each other. However, if the oxide-thicknesses are known apriori then the defect generation mechanism in one device is independent of any other device on the chip for constant worst-case voltage and temperature. Thus for a particular chip, if the thicknesses of all devices are known then any device fails independently of all other devices. Furthermore the reliability function

of the chip $R_c(t)$ requires that all devices on the chip are functioning reliably, therefore, from elementary statistics $R_c(t)$ can be given by the product of reliability functions of all individual devices:

$$R_c(t|x) = \prod_{i=1}^m R_i(t|x_i) \quad (5.6)$$

where x represent the vector of oxide-thickness (x_1, \dots, x_m) and m is the total number of devices on the chip.

As discussed in Section 5.1, the thickness of each device is modeled as a random variable that is correlated with the oxide-thickness of other devices. If the oxide-thickness of all devices is characterized by their joint PDF $f(x_1, \dots, x_m)$ then the overall reliability function of the entire ensemble of all manufactured chips can be given by

$$R_c(t) = \int_0^\infty \dots \int_0^\infty \prod_{i=1}^m R_i(t|x_i) f(x_1, \dots, x_n) dx_1 \dots dx_m. \quad (5.7)$$

Due to the huge dimensionality of the above integral, a straight forward numerical evaluation of the above integral is computationally impractical for full chip analysis. Using judicious engineering approximations we develop a computationally efficient approach to address this problem in the next Section.

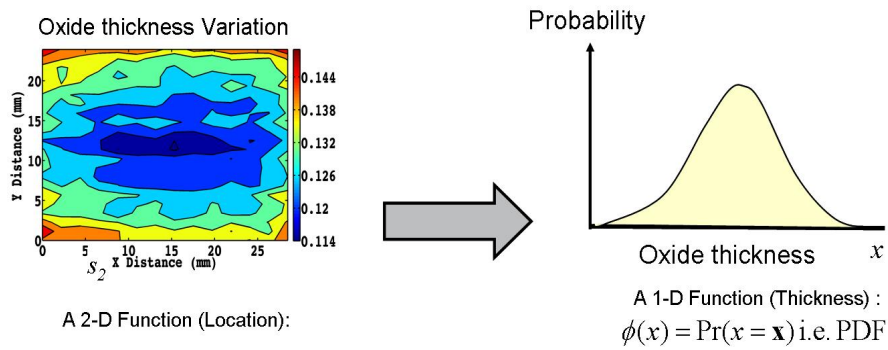


Figure 5.1. Oxide-thickness distribution of one sample die

5.3 Computing Oxide Reliability

The proposed approach for efficiently estimating the overall reliability function $R_c(t)$ is discussed in a bottom up manner. We first present expressions for finding the conditional reliability function of a device. Using this expression, the conditional reliability function of a particular chip can be found given the oxide-thickness of all devices on it. Finally we present how the overall reliability function is found for the entire ensemble of all manufactured chips. The key to our approach is the compact representation of the oxide-thickness variation that enables us to efficiently perform the enumeration across the thickness variation sample space while preserving spatial correlation.

5.3.1 Reliability Function of One Chip

Using the definition of the reliability function and the oxide breakdown time model of an individual device (Equation 5.3), the conditional reliability function of every i^{th} device having oxide-thickness x_i can be given by

$$R_i(t|x_i) = e^{-a_i t^{\frac{bx_i}{\alpha}}}, \quad (5.8)$$

For each device on the chip $x = (x_1, x_2, \dots, x_m)$ and their respective area a_i , the conditional reliability of the chip is given by

$$R_c(t|\mathbf{x}) = \prod_{i=1}^m R_i(t|x_i) = e^{-\sum_{i=1}^m a_i t^{\frac{bx_i}{\alpha}}}. \quad (5.9)$$

There can be several million devices on a chip, therefore it is impractical to evaluate the above exponent. In order to efficiently evaluate the overall reliability across all chips, we need to reduce the dimensionality of the above exponent. To achieve this, we represent the set of devices and their individual oxide-thicknesses by the distribution of the oxide-thickness for a particular chip. For the sake of understanding, we discretize this oxide-thickness distribution into k discrete intervals assuming a truncated distribution. It can be seen that when we make this transformation the area of the devices with identical thickness can be summed together. Let \bar{x}_i denote the midpoint of the i^{th} discrete interval and \bar{a}_i be

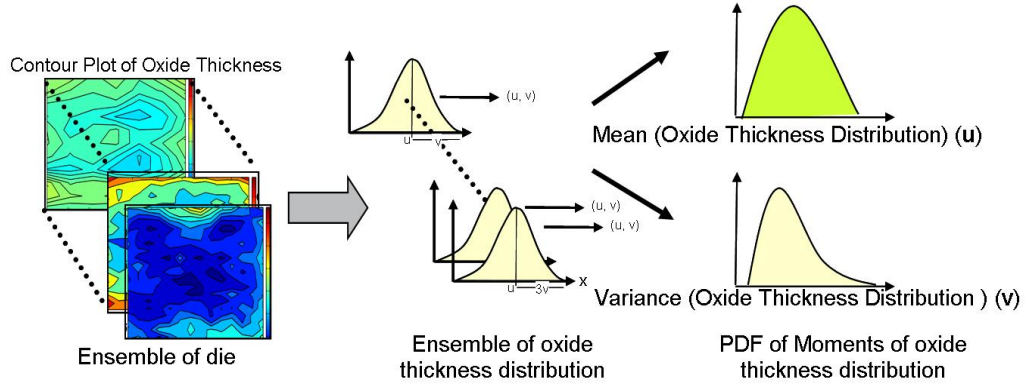


Figure 5.2. Compact representation of oxide-thickness variation of ensemble of all die

the total area of all devices having thickness \bar{x}_i . By applying this transformation the above expression for $R_c(t|x)$ can be rewritten as

$$R_c(t|x) = e^{-\sum_{i=1}^k \bar{a}_i t \frac{b\bar{x}_i}{\alpha}}. \quad (5.10)$$

By making such a transformation the dimensionality of $R_c(t|x)$ can be significantly reduced from number of devices n to the number of discrete intervals k . If we normalize the exponent with total area the above expression gives

$$R_c(t|x) = e^{-A \sum_{i=1}^k p_i t \frac{b\bar{x}_i}{\alpha}}, \quad (5.11)$$

where $p_i = \bar{a}_i/A$ represents the probability of observing an oxide-thickness x_i on a particular sample chip. Thus the oxide-thickness of all devices on a particular sample chip can be compactly characterized by a distribution function. As discussed in Section 5.1, the oxide-thickness variation of each device across the ensemble of chips is modeled as a gaussian random variable. If there is no spatial correlation in the within-die component then the thickness distribution across a sample of particular chip will also be a gaussian distribution. However, strictly speaking this may not be the case in general. Nevertheless as there are several million devices on a chip, we find that assuming gaussian distribution of the oxide-thickness across a particular sample chip is a reasonable approximation to its

exact distribution. Thus we have

$$R_c(t|u, v) = e^{-A \int_{-\infty}^{\infty} \phi(\frac{x-u}{v}) t^{\frac{bx}{\alpha}} dx}. \quad (5.12)$$

The multidimensional exponent in Equation 5.9 can be compactly represented using a function of two parameters u and v which represent the sample mean and variance of the oxide distribution function.

5.3.2 Overall Reliability Function

As shown in Figure 5.2, each sample die results in a different oxide distribution, therefore, the entire ensemble of all die can be represented with an ensemble of oxide distribution characterized by two random variables \mathbf{u} and \mathbf{v} . Let $f_{\mathbf{uv}}(u, v)$ denote the joint probability distribution function of \mathbf{u} and \mathbf{v} . For computing the overall reliability function, we need to integrate the above expression of reliability function of each chip over the joint probability distribution function $f_{\mathbf{uv}}(u, v)$ of random variables \mathbf{u} and \mathbf{v} .

$$R_c(t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} R_c(t|u, v) f_{\mathbf{uv}}(u, v) du dv \quad (5.13)$$

Now for a particular chip, the mean u and variance v of the oxide distribution can be estimated by calculating the unbiased statistical sample mean and sample variance of the oxide-thickness values observed across the die. Likewise, the random variables \mathbf{u} and \mathbf{v} can be found in terms of the thickness variation model discussed in Equation 5.1. Using the unbiased sample mean estimator, \mathbf{u} can be expressed as

$$\mathbf{u} = \frac{1}{n} \sum_{i=1}^n x_i = u_0 + \sum_{i=1}^n u_i z_i + u_{p+1} \mathbf{E}, \quad (5.14)$$

where

$$\begin{aligned} u_j &= \frac{1}{n} \sum_{i=1}^n \lambda_{i,j} \quad \forall j = 0 \dots n \\ u_{p+1} &= \frac{1}{n} \sqrt{\sum_{i=1}^n \lambda_r^2} = \frac{\lambda_r}{\sqrt{n}} \end{aligned}$$

The coefficient u_0 is the nominal value of \mathbf{u} , whereas coefficients u_i is the sensitivity of i^{th} to each of the principal component random variables. It is evident that u_{p+1} tends to zero for a large number of devices and thus can be safely neglected for a typical industrial design.

Similarly the expression for \mathbf{v} , the variance of the oxide distribution across the ensemble of all die, in terms of oxide variation model (Equation 5.1) can be given as follows:

$$\mathbf{v} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \mathbf{u})^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i^2 - \mathbf{u}^2) \quad (5.15)$$

Again the above expression can be expressed in terms of principal components as follows:

$$\mathbf{v} = v_0 + \sum_{i=1}^n \sum_{j=1}^n v_{i,j} z_i z_j, \quad (5.16)$$

where

$$v_0 = \lambda_r \quad \text{and} \quad v_{i,j} = \frac{1}{n-1} \sum_{k=1}^n \lambda_{k,i} \lambda_{k,j} \quad (5.17)$$

In this manner the oxide-thickness variation of all chips can be compactly represented by two random variables \mathbf{u} and \mathbf{v} . Furthermore, it can be shown as follows that \mathbf{u} and \mathbf{v} are uncorrelated (i.e., $E[\mathbf{u}\mathbf{v}] = E[\mathbf{u}]E[\mathbf{v}]$, where $E[\cdot]$ denotes the expectation).

$$E[\mathbf{u}\mathbf{v}] = E\left[\left(u_0 + \sum_{i=1}^n u_i z_i + u_{p+1} \epsilon\right) \left(v_0 + \sum_{i=1}^n \sum_{j=1}^n v_{i,j} z_i z_j\right)\right].$$

By construction each principal component z_i is an independent standard normal random variable, therefore we have $E[z_i] = E[z_i^2 z_j] = E[z_i z_j^2] = E[z_i^3] = 0$ and $E[z_i^2] = 1$ for all $i, j = i = 1 \dots p$. Likewise $E[\epsilon] = E[\epsilon^2 z_j] = E[z_i \epsilon^2] = E[\epsilon^3] = 0$ and $E[\epsilon^2] = 1$. Thus the above expression can be simplified and can be given by

$$E[\mathbf{u}\mathbf{v}] = u_0 v_0 + \sum_{i=1}^n u_0 v_{i,i} + u_0 v_{p+1} = E[\mathbf{u}]E[\mathbf{v}].$$

For two normal random variables to be independent, it is sufficient to show that they are uncorrelated, but in general this is not the case for non-gaussian random variables. Unfortunately, the sample variance \mathbf{v} is not a normal random variable and as both \mathbf{u} and \mathbf{v} are functions of the same principal components they are dependent on each other. However,

assuming u and v as independent, allows us to express the JPDF in terms of their marginal distributions $f_{\mathbf{u}}(u)$ and $f_{\mathbf{v}}(v)$. This approximation enables us to enumerate the individual reliability distribution functions (see Figure 5.3) of each chip by simply integrating the marginal distributions $f_{\mathbf{u}}(u)$ and $f_{\mathbf{v}}(v)$.

$$R_c(t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} R_c(t|u, v) f_{\mathbf{u}}(u) f_{\mathbf{v}}(v) du dv. \quad (5.18)$$

Since \mathbf{u} and \mathbf{v} are uncorrelated, in practice, we find that the independence-approximation between \mathbf{u} and \mathbf{v} gives us a reasonably accurate estimate of oxide variation with a significantly simpler approach. Now the sample mean \mathbf{u} is a sum of normal random variables, therefore $f_{\mathbf{u}}(u)$ can be characterized by distribution of a normal random variable and can be analytically computed. However, sample variance v is a quadratic expression of normal random variables. Such an expression is commonly found in several multivariate statistics application and is referred to as quadratic normal form. Several techniques have been proposed to accurately estimate the distribution function of quadratic normal form. In this work, we implemented the method listed in [111] to estimate the distribution of $f_{\mathbf{v}}(v)$. First the coefficients of the quadratic terms are diagonalized to express the quadratic normal form as linear combination of chi-square random variables. Then the characteristic function of the linear combination of the chi-square random variables is found. Using the numerical algorithm presented in [112], the inverse of the characteristic function is found which gives the probability distribution function of $R_c(t)$. The overall reliability distribution function can be computed by evaluating the numerical integration in two dimensions using these marginal distributions. The overall runtime complexity of the proposed approach is limited by the computation of $f_{\mathbf{v}}(v)$ which is $O(n)$, where n is the number of principal components. Thus the computation complexity is independent of the total number of devices on the chip and is only a function of the number of grids in the spatial variation model.

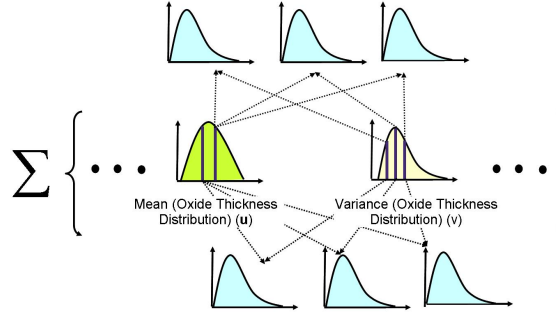


Figure 5.3. Oxide-thickness distribution of ensemble of all die

5.4 Simulation Results

A simple simulation methodology for estimating the critical defect density required for triggering a dielectric breakdown in an ultra thin oxide was originally developed by Degraeve in [109]. Using this methodology and the published defect generation relationships from an IBM technology node given in [108], the oxide breakdown reliability function was found for a set of characteristic devices differing in area and thickness. The technology dependent parameter of the oxide reliability function model given in Equation 5.3 was estimated by fitting it to the simulation results. In practice, such a model can also be characterized from real oxide breakdown distributions measured from test capacitors or discrete devices for the required process and technology. A prototype based on the proposed method was implemented in Mathematica.

In the first set of experiments, a set of 5 designs was considered to determine the accuracy of the proposed methodology. The chip area was divided into different grids sizes based upon the size of the circuit. The overall 3σ of oxide-thickness variation was assumed to be 4% of the nominal value and it was equally split amongst all three components of variation, namely, global, spatial and random component. As the real measurement data for thickness correlation was unavailable, the covariance matrix for thickness variations used in this work was derived from an exponential decaying function of the respective distance. The correlation distance of exponential correlation function is normalized with respect to the chip dimensions.

Given the post-layout design implementation and a process variation model of oxide-thickness, the proposed methodology can compute the overall reliability distribution function. To validate the results of the proposed method, the overall reliability distribution was also computed from 1000 samples of Monte Carlo simulations using the same oxide reliability model and thickness variation model. In Table 5.1, a comparison of lifetime estimation for 1 fault per million parts and 10 faults per million parts between the proposed approach and Monte Carlo simulations is shown for 5 design circuits. The size of the circuit under test in terms of number of devices is given in the second column. As the MC simulation time increases significantly for larger size circuits, it was necessary to limit the test cases to no more than 150 thousand devices. Unlike MC simulations the proposed approach was able to analyze circuits with several million devices. To verify the robustness of the proposed approach with respect to correlation distance we tested our approach for three different values of correlation distance normalized with respect to the die size. We also validated the approach by choosing 4 different resolutions of grid size for design C. The numerical results found for 4 different grid-size are given in Table 5.2. As the discretization error of the grid-based model decreases for larger grid size, it can be seen that the error in estimation of reliability function decreases.

Figure 5.4 shows the plot of overall $R_c(t)$ estimation computed using (1) Proposed Approach (2) MC Simulation (3) Minimum oxide-thickness (4) Maximum oxide-thickness. It can be seen that reliability distribution function result for the proposed approach is in good agreement with MC result whereas the curves corresponding to the pessimistic and optimistic analysis deviate significantly from the proposed approach. This clearly exemplifies the need for a statistical approach for reliability distribution function analysis.

For examining the gaussian-assumption made for the oxide-thickness distribution of a chip, we computed the oxide-thickness distribution for each sample using MC simulations. Figure 5.5 shows the plots of oxide-thickness distributions for 8 different samples of design C for different values of normalized correlation distance. For most samples, we found that

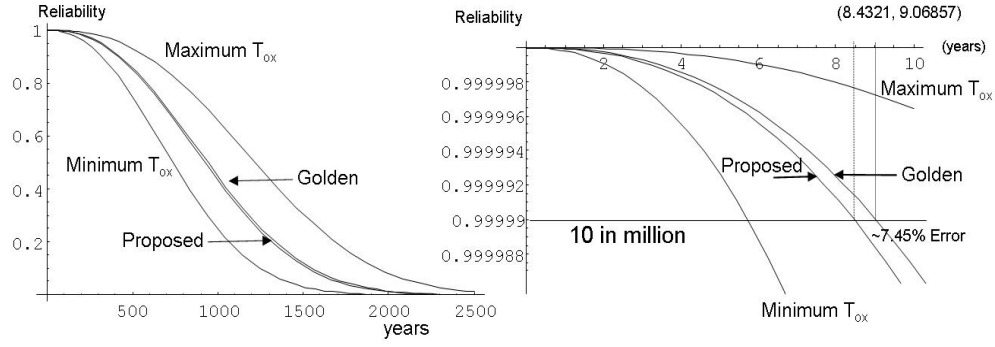


Figure 5.4. Comparison between Reliability function estimation result, monte carlo simulations, worst case oxide-thickness and best-case oxide-thickness

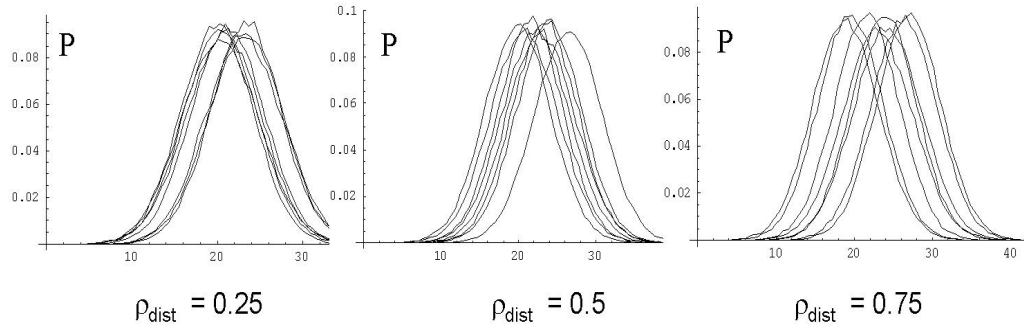


Figure 5.5. Oxide-thickness distribution generated using MC simulations for 8 samples of Design C (100K gates, Grid size 20×20 for different values of normalized ρ_{dist}

the gaussian distribution is a good fit for the oxide-thickness distribution. Likewise, we also varied the ratio of each component of thickness variation, the grid size and the chip size and similar results were observed.

To illustrate the significance of considering spatial correlations, we compare the difference between performing reliability analysis with and without actual spatial correlation. A set of MC simulations were done assuming accurate spatial correlation, no spatial-correlation and assuming perfect spatial correlation. Figure 5.6 compares the resulting reliability distribution functions with the result when spatial correlation is accurately modeled. It can be observed that the mean values of the uncorrelated case is much smaller than the correlated cases (MC) and the perfectly correlated case overestimates the variance of the reliability function. Therefore, statistical oxide reliability analysis without considering

Table 5.1. Accuracy comparison between proposed approach and MC Simulations for different correlation distance

		Lifetime Estimation Error w.r.t. MC simulations					
circuit		Rho-distance = 0.5		Rho-distance = 0.25		Rho-distance = 0.75	
name	No. devices	1/million	10/Million	1/million	10/Million	1/million	10/Million
A	50K	9.54%	8.79%	9.19%	8.74%	8.82%	8.38%
B	75K	8.87%	8.04%	8.17%	8.08%	7.84%	7.74%
C	100K	7.88%	7.14%	7.54%	6.94%	7.24%	6.65%
D	125K	7.67%	6.97%	7.24%	6.72%	6.95%	6.89%
E	150K	7.54%	7.02%	7.11%	6.93%	6.45%	7.04%

Table 5.2. Accuracy comparison between proposed approach and MC Simulations for different grid resolution for design B

B	Lifetime Estimation Error w.r.t. MC simulations					
(75K)	Rho-distance = 0.5		Rho-distance = 0.25		Rho-distance = 0.75	
Grid Size	1/million	10/Million	1/million	10/Million	1/million	10/Million
5x5	9.14%	8.28%	8.42%	8.32%	8.08%	7.86%
10x10	8.87%	8.04%	8.17%	8.08%	7.84%	7.63%
15x15	8.55%	7.74%	7.87%	7.78%	7.55%	7.34%
20x20	8.23%	7.45%	7.57%	7.49%	7.43%	7.06%

correlation may incorrectly predict the real lifetime of the circuit and could even underestimate the lifetime of the circuit.

Even for small circuits considered in the experimental setup and a Mathematica based prototype implementation the runtime of the proposed algorithm was 3 orders of magnitude faster than the brute-force MC simulation based approach. The runtime of the MC simulation based approach is prohibitively expensive for larger circuits as it scales super-linearly with the number of devices. However, the proposed approach is independent of the number of devices and scales linearly with grid size.

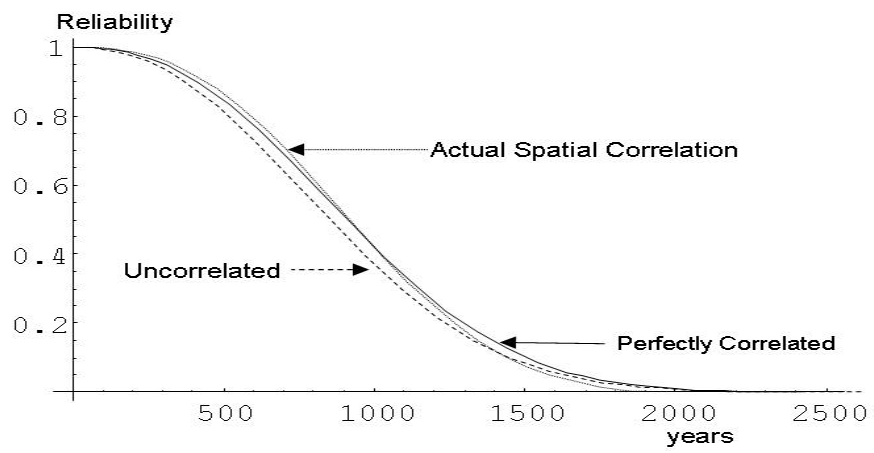


Figure 5.6. Comparison between monte carlo simulations of Reliability function estimation assuming actual spatial correlation, perfect correlation and zero correlation

CHAPTER 6

Conclusion

This chapter summarizes our work in statistical analysis and optimization and the contributions made in these areas. We discuss the merits and contributions of the techniques adopted in our work. We summarize the results of this dissertation research and also discuss open challenges for further research that will facilitate the adoption of statistical analysis in mainstream VLSI CAD and design.

6.1 Modeling: Extraction of Process Variation Model

Based on the key concept of variogram function, we have presented a new approach to extract spatial variation models. The key advantage of the variogram function is that it provides us with a representation that is independent of the global component of variation. It allows us to directly estimate the within die component of variations and thus circumvents the need for making ergodicity assumption. In this work, we further showed that using two dimensional variogram functions allows us to model geometrically anisotropic process variation data. Additionally, for extracting process variation models in the presence of significant measurement noise, we employ *weighted least squares* regression technique, a technique which is known to be statistically more robust than the previously used ordinary least square technique. The experiment results on both Monte-Carlo models and ELM measurement data confirm the validity of the proposed approach.

6.2 Analysis: Skew-normal Maximum Operation

We presented an analytical approach that extends Clark’s max results to skew normal distributions for computing the statistical maximum of two skewed arrival time distributions. An efficient method is presented to approximate the arrival time distribution using skew normal representation. This is done by matching the mean, the variance and the skewness of arrival time distributions to that of the skew normal approximation. Using this method, we then derived analytical results for computing the approximate moments of the maximum of the arrival time distribution assuming their joint PDF as a bi-variate skew normal distribution. From these moments, the tightness probability, mean, variance and skewness of the maximum can be computed and therefore the presented results can be applied in existing SSTA algorithms that work on Clark’s results. Our numerical results show that the proposed max operation can improve the accuracy of existing SSTA approaches. Furthermore, the skewness based proposed max function can be used to augment existing SSTA framework to propagate three moments.

6.3 Optimization: Cut-set based Joint Timing and Power Yield Maximization

To the best of our knowledge, we have presented the first approach to perform gate-level parametric yield optimization considering constraints on power and performance, along with their correlation. The approach for yield computation is shown to be computationally efficient and is shown to provide an 8X improvement in runtime, as compared to a brute-force gradient computation approach. The yield gradient is used to guide a large-scale non linear optimizer to improve the yield of a design that has been optimized deterministically, under varying power and delay constraints. The cut-set based method presented in this work for computing the gradient for circuit tuning was further improved by subsequent work published in [113] .

6.4 Reliability: A Statistical Approach for Full-chip Gate-oxide Reliability Analysis

A statistical methodology for performing full chip oxide reliability analysis has been proposed, considering all three components of oxide thickness variation. It is shown that worst-case oxide reliability analysis may not be adequate to predict chip lifetime accurately. The complexity analysis of the proposed methodology shows that the proposed approach is independent of the number of devices and is thus scalable to large industrial size circuits. Our simulation results exemplifies the accuracy and efficiency of the proposed method.

6.5 Future Directions

Over the past decade, statistical performance analysis has gained extensive interest and significant progress has been made towards developing the basic algorithm for statistical timing and leakage analysis approaches. However, the obstacles to wide-spread adoption of statistical in industry remain formidable. Much attention has been paid to the modeling and analysis of spatial correlations, non-normal physical device parameter distributions and non-linear delay dependencies. However, the current state-of-the-art statistical methods still do not address many of the issues that are taken for granted in deterministic approaches, such as interconnect analysis, coupling noise, clocking issues, and complex delay modeling. In addition, a major concern is the lack of silicon verification of the proposed methods and the underlying business model needed to communicate silicon data while abstracting the technology recipe of semiconductor foundries. Several possibilities still exist for future work in statistical modeling, analysis and optimization techniques developed in this manuscript.

The variogram based approach for extracting process variation model extracts the design independent non-systematic components of variation. In addition to the non-deterministic

component of manufacturing variation, one must also model and extract the design dependent systematic component. In [78], the authors have proposed a method to extract the systematic component for every standard cell, however, apart from the cell type several systematic variations such as optical proximity effects are a strong function of the proximity of the cell. Currently, this is modeled by performing a detailed lithography and resist process simulation on the entire design for optical proximity correction. A technique to model and characterize such proximity dependent variations can be very useful for modeling systematic variations in analysis.

The skew-normal maximum operation presented in this work improves the accuracy of the statistical maximum operation at the cost of additional computation its accuracy, however, while deriving the skew-normal parameters of individual input operands we assume a marginal skew normal distribution for each input operand. In practice, since the operands for the maximum operation are correlated deriving skew-normal parameters from the respective marginal distributions leads to some approximation error. Another promising alternative would be to develop a method to derive the parameters of a joint bi-variate skew-normal distribution from the bi-variate distribution of the operands. Such an approach may further increase the accuracy of the derivation developed in this work.

The current modeling framework for reliability analysis considers the variation in manufacturing parameters, however, it considers a worst-case temperature for the entire chip. This leads to unwarranted pessimism and thus limits the usefulness of developed framework for practical industrial designs. Augmenting the proposed reliability analysis framework to accurately model across chip temperature variations for different operating modes of a chip would be a very useful direction to consider for future work.

BIBLIOGRAPHY

BIBLIOGRAPHY

- [1] S. Das, S. Pant, D. Roberts, and S. Seokwoo, "A self-tuning dvs processor using delay-error detection and correction," in *IEEE Symp. on VLSI Circuits*, 2005.
- [2] J. Tschanz, J. Kao, S. Narendra, R. Nair, D. Antoniadis, A. Chandrakasan, and V. De, "Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage," *JSSC*, vol. 37, November 2002.
- [3] G. Nanz and L. Camilletti, "Modeling of chemicalmechanical polishing: A review," *IEEE Trans. on Semiconductor Manuf.*, vol. 8, no. 4, 1995.
- [4] C. Mack, "Understanding focus effects in submicrometer optical lithography: A review," *Optical Engineering*, vol. 32, no. 10, 1993.
- [5] L. Scheffer, "Physical cad changes to incorporate design for lithography and manufacturability," in *Proc. ASP-DAC*, 2004.
- [6] F. Huebbers, A. Dasdan, and Y. Ismail, "Computation of accurate interconnect process parameter values for performance corners under process variations," in *Proc. DAC*, 2006.
- [7] J. Yang, L. Capodieci, and D. Sylvester, "Advanced timing analysis based on post-opc extraction of critical dimensions," in *Proc. DAC*, 2005.
- [8] P. Gupta and F. Heng, "Toward a systematic-variation aware timing methodology," in *Proc. DAC*, 2004.
- [9] M. Pelgrom, A. Duinmaijer, and A. Welbers, "Matching properties of MOS transistors," *IEEE Journal of Solid-State Circuits*, vol. 24, October 1989.
- [10] S. Nassif, "Modeling and forecasting of manufacturing variations (embedded tuto-

- rial),” in *Proc. ASP-DAC*, 2001.
- [11] L. Scheffer, “The count of monte carlo,” in *Proc. TAU Int. Work. on Timing*, 2004.
 - [12] S. Onaissi and F. N. Najm, “A linear-time approach for static timing analysis covering all process corners,” in *ICCAD ’06: Proceedings of the 2006 IEEE/ACM international conference on Computer-aided design*, pp. 217–224, 2006.
 - [13] T. Kirkpatrick and N. Clark, “PERT as an aid to logic design,” *IBM Journal of Research and Development*, vol. 10, no. 2, 1966.
 - [14] H. Jyu, S. Malik, S. Devdas, and K. Keutzer, “Statistical timing analysis of combinational logic circuits,” *IEEE Trans. on VLSI*, vol. 1, June 1993.
 - [15] R. Brashear, N. Menezes, C. Oh, L. Pillage, and M. Mercer, “Predicting circuit performance using circuit-level statistical timing analysis,” in *Proc. DATE*, March 1994.
 - [16] A. Agarwal, D. Blaauw, and V. Zolotov, “Statistical timing analysis using bounds and selective enumeration,” *IEEE Trans. on CAD*, 2003.
 - [17] A. Agarwal, D. Blaauw, V. Zolotov, and S. Vrudhula, “Statistical timing analysis using bounds and selective enumeration,” in *TAU Int. Work. on Timing*, 2002.
 - [18] J. Kelley, “Critical-path planning and scheduling: Mathematical basis,” *Journal of Operations Research*, vol. 9, no. 3, 1961.
 - [19] D. Malcolm, J. Roseboom, C. Clark, and W. Fazar, “Application of a technique for research and development program evaluation,” *Journal of Operations Research*, vol. 7, no. 5, 1959.
 - [20] J. Hagstrom, “Computational complexity of PERT problems,” *Networks*, vol. 18, 1988.
 - [21] S. Sapatnekar, *Timing*. Springer-Verlag New York, Inc., 2004.
 - [22] R. Chen and H. Zhou, “Clock schedule verification under process variations,” in *Proc. ICCAD*, 2004.
 - [23] A. Agarwal, D. Blaauw, and V. Zolotov, “Statistical clock skew analysis considering

- intra-die process variations,” in *Proc. ICCAD*, 2003.
- [24] L. Zhang, Y. Hu, and C. Chen, “Statistical timing analysis in sequential circuit for on-chip global interconnect pipelining,” in *Proc. DAC*, 2004.
 - [25] R. Rutenbar, L. Wang, K. Cheng, and S. Kundu, “Static statistical timing analysis for latch-based pipeline designs,” in *Proc. ICCAD*, 2004.
 - [26] L. Zhang, J. Tsai, W. Chen, Y. Hu, and C. Chen, “Convergence-provable statistical timing analysis with level-sensitive latches and feedback loops,” in *Proc. ASP-DAC*, 2006.
 - [27] C. Visweswariah, “Death, taxes and failing chips,” in *Proc. DAC*, 2003.
 - [28] C. Visweswariah, K. Ravindran, K. Kalafala, S. Walker, and S. Narayan, “First-order incremental block-based statistical timing analysis,” in *Proc. DAC*, 2004.
 - [29] A. Agarwal, D. Blaauw, and V. Zolotov, “Statistical timing analysis for intra-die process variations with spatial correlations,” in *Proc. ICCAD*, 2003.
 - [30] H. Chang and S. Sapatnekar, “Statistical timing analysis considering spatial correlations using a single pert-like traversal,” in *Proc. ICCAD*, 2003.
 - [31] H. Chang, V. Zolotov, S. Narayan, and C. Visweswariah, “Parameterized block-based statistical timing analysis with non-gaussian parameters, nonlinear delay functions,” in *Proc. DAC*, 2005.
 - [32] Y. Zhan, A. Strojwas, X. Li, T. Pileggi, D. Newmark, and M. Sharma, “Correlation-aware statistical timing analysis with non-gaussian delay distributions,” in *Proc. DAC*, 2005.
 - [33] L. Zhang, W. Chen, Y. Hu, J. Gubner, and C. Chen, “Correlation-preserved non-gaussian statistical timing analysis with quadratic timing model,” in *Proc. DAC*, 2005.
 - [34] V. Khandelwal and A. Srivastava, “A general framework for accurate statistical timing analysis considering correlations,” in *Proc. DAC*, 2005.
 - [35] J. Singh and S. Sapatnekar, “Statistical timing analysis with correlated non-gaussian

- parameters using independent component analysis,” in *Proc. DAC*, 2006.
- [36] K. Chopra, B. Zhai, D. Blaauw, and D. Sylvester, “A new statistical max operation for propagating skewness in statistical timing analysis,” in *Proc. ICCAD*, 2006.
 - [37] C. Clark, “The greatest of a finite set of random variables,” *Journal of Operations Research*, vol. 9, 1961.
 - [38] J. Jess, K. Kalafala, S. Naidu, R. Otten, and C. Visweswariah, “Statistical timing for parametric yield prediction of digital integrated circuits,” in *Proc. DAC*, 2003.
 - [39] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, “Equation of state calculations by fast computing machines,” *Journal of Chemical Physics*, vol. 21, 1953.
 - [40] S. Tasiran and A. Demir, “Smart monte carlo for yield estimation,” in *Proc. TAU Int. Work. on Timing*, 2006.
 - [41] R. Kanj, R. Joshi, and S. Nassif, “Mixture importance sampling and its application to the analysis of sram designs in the presence of rare failure events,” in *Proc. DAC*, 2006.
 - [42] V. Veetil, D. Blaauw, and D. Sylvester, “Criticality aware latin hypercube sampling for efficient statistical timing analysis,” in *Proc. TAU Int. Work. on Timing*, 2007.
 - [43] A. Gattiker, S. Nassif, R. Dinakar, and C. Long, “Timing yield estimation from static timing analysis,” in *Proc. ISQED*, 2001.
 - [44] A. Agarwal, D. Blaauw, V. Zolotov, S. Sundareswaran, M. Zhou, K. Gala, and R. Panda, “Statistical delay computation considering spatial correlations,” in *Proc. ASP-DAC*, 2003.
 - [45] C. Amin, N. Menezes, K. Killpack, F. Dartu, U. Choudhury, N. Hakim, and Y. Ismail, “Statistical static timing analysis how simple can we get?,” in *Proc. DAC*, 2005.
 - [46] R. Lin and M. Wu, “A new statistical approach to timing analysis of VLSI circuits,” in *Proc. Int. Conf. on VLSI Design*, Jan 1998.

- [47] B. Choi and D. Walker, "Timing analysis of combinational circuits including capacitive coupling and statistical process variation," in *Proc. of Symp. on VLSI Test*, 2000.
- [48] M. Orshansky and K. Keutzer, "A general probabilistic framework for worst case timing analysis," in *Proc. DAC*, 2002.
- [49] H. Mangassarian and M. Anis, "On statistical timing analysis with inter- and intra-die variations," in *Proc. DATE*, 2005.
- [50] F. Najm and N. Menezes, "Statistical timing analysis based on a timing yield model," in *Proc. DAC*, 2004.
- [51] K. Heloue and F. Najm, "Statistical timing analysis with two-sided constraints," in *Proc. ICCAD*, 2005.
- [52] M. Berkelaar, "Statistical delay calculation, linear time method," in *Proc. TAU Int. Work. on Timing*, 1997.
- [53] S. Tsukiyama, M. Tanaka, and M. Fukui, "A statistical static timing analysis considering correlations between delays," in *Proc. ASP-DAC*, 2001.
- [54] J. Le, X. Li, and L. Pileggi, "STAC: Statistical timing analysis with correlation," in *Proc. DAC*, 2004.
- [55] K. Kang, B. Paul, and K. Roy, "Statistical timing analysis using levelized covariance propagation," in *Proc. DATE*, 2005.
- [56] J. Liou, K. Cheng, S. Kundu, and A. Krstic, "Fast statistical timing analysis by probabilistic event propagation," in *Proc. DAC*, 2001.
- [57] J. Liou, A. Krstic, L. Wang, and K. Cheng, "False-path-aware statistical timing analysis and efficient path selection for delay testing and timing validation," in *Proc. DAC*, 2002.
- [58] S. Naidu, "Timing yield calculation using an impulse-train approach," in *Proc. ASP-DAC*, 2002.
- [59] A. Agarwal, D. Blaauw, V. Zolotov, and S. Vrudhula, "Statistical timing analysis

- using bounds,” in *Proc. DATE*, 2003.
- [60] A. Agarwal, D. Blaauw, V. Zolotov, and S. Vrudhula, “Computation and refinement of statistical bounds on circuit delay,” in *Proc. DAC*, 2003.
 - [61] S. Bhardwaj, S. Vrudhula, and D. Blaauw, “ τ AU Timing Analysis under Uncertainty,” in *Proc. ICCAD*, 2003.
 - [62] A. Devgan and C. Kashyap, “Block-based static timing analysis with uncertainty,” in *Proc. ICCAD*, 2003.
 - [63] V. Khandelwal, A. Davoodi, and A. Srivastava, “Efficient statistical timing analysis through error budgeting,” in *Proc. ICCAD*, 2004.
 - [64] R. Topaloglu and A. Orailoglu, “Forward discrete probability propagation method for device performance characterization under process variations,” in *Proc. ASP-DAC*, 2005.
 - [65] L. Scheffer, “Explicit computation of performance as a function of process variation,” in *Proc. TAU Int. Work. on Timing*, 2002.
 - [66] I. Jolliffe, *Principal Component Analysis*. Springer-Verlag New York, Inc., 2002.
 - [67] D. Boning and S. Nassif, *Models of Process Variations in Device and Interconnect in Design of High-Performance Microprocessor Circuits*. A. Chandrakasan, 2000.
 - [68] Y. Cao and L. Clark, “Mapping statistical process variations toward circuit performance variability an analytical modeling approach,” in *Proc. DAC*, 2005.
 - [69] J. Xiong, V. Zolotov, and L. He, “Robust extraction of spatial correlation,” in *Proc. ISPD*, 2006.
 - [70] L. Zhang, J. Shao, and C. Chen, “Non-gaussian statistical parameter modeling for ssta with confidence interval analysis,” in *Proc. ISPD*, 2006.
 - [71] K. Chopra, N. Shenoy, and D. Blaauw, “Variogram based robust extraction of process variation,” in *Proc. TAU Int. Work. on Timing*, 2007.
 - [72] F. Liu, “How to construct spatial correlation models: A mathematical approach,” in *Proc. TAU Int. Work. on Timing*, 2007.

- [73] B. Cline, K. Chopra, and D. Blaauw, "Analysis and modeling of cd variation for statistical static timing," in *Proc. ICCAD*, 2006.
- [74] S. Bhardwaj, S. Vrudhula, P. Ghanta, and Y. Cao, "Modeling of intra-die process variations for accurate analysis and optimization of nano-scale circuits," in *Proc. DAC*, 2006.
- [75] "International technology roadmap for semiconductors 2005,"
- [76] M. Mani, A. Devgan, and M. Orshansky, "An efficient algorithm for statistical minimization of total power under timing yield constraints," in *Proc. DAC*, 2005.
- [77] R. R. Rao, A. Devgan, D. Blaauw, and D. Sylvester, "Parametric yield estimation considering leakage variability," in *DAC '04 Proceedings of the 41st annual conference on Design automation*, (New York, NY, USA), pp. 442–447, ACM Press, June 2004.
- [78] M. Orshansky, L. Milor, P. Chen, K. Keutzer, and C. Hu, "Impact of spatial intra-chip gate length variability on the performance of high-speed digital circuits," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, pp. 544–553, May 2002.
- [79] P. Friedberg, Y. Cao, J. Cain, R. Wang, J. Rabaey, and C. Spanos, "Modeling within-die spatial correlation effects for process-design co-optimization," in *Sixth Int. Symp. on Quality of Electronic Design*, pp. 516–521, 2005.
- [80] J. Xiong, V. Zolotov, and L. He, "Robust extraction of spatial correlation," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, April 2007.
- [81] F. Liu, "A general framework for spatial correlation modeling in vlsi design," in *TAU '07 international workshop on Timing issues in the specification and synthesis of digital systems*, 2007.
- [82] N. A. Cressie, *Statistics for Spatial Data*. Wiley Series in Probability and Statistics, 1993.

- [83] Z. D. L., “Another look at anisotropy in geostatistics,” *Mathematical geology (Math. geol.)*, pp. 453–470, 1993.
- [84] M. S. Handcock and J. R. Wallis, “An approach to statistical spatial-temporal modeling of meteorological fields,” *Journal of the American Statistical Association*, pp. 368–390, 1994.
- [85] D. M. Bates and D. G. Watts, *Nonlinear Regression and Its Applications*. New York: Wiley, 1988.
- [86] J. Cain and C. Spanos, “Electrical linewidth metrology for systematic cd variation characterization and causal analysis,” in *Proceedings of SPIE Int. Soc. Opt. Eng.*, pp. 516–521, 2003.
- [87] O. Kella, “On the distribution of the maximum of bivariate normal random variables with general means and variances,” *Communications in Statistics - Theory and Methods*, vol. 15(11), pp. 3265–3276, 1986.
- [88] M. Cain, “The moment-generating function of the minimum of bivariate normal random variables,” *The American Statistician*, vol. 48(2), pp. 124–125, 1994.
- [89] C. Fernandez and M. Steel, “On bayesian modelling of fat tails and skewness,” *Journal of the American Statistical Association*, pp. 359–371, 1998.
- [90] D. B. Owen, “Tables for computing bivariate normal probabilities,” *Ann. Math. Statist.*, vol. 27, pp. 1075–1090, 1956.
- [91] M. Patefield, “Fast and accurate calculation of Owen’s T function,” *Statistical Software*, vol. 5, no. 5, pp. 1–25, 2000.
- [92] A. Srivastava, D. Sylvester, and D. Blaauw, “Statistical optimization of leakage power considering process variations using dual-vth and sizing,” in *Proc. DAC*, 2004.
- [93] X. Bai, C. Visweswariah, and P. Strenski, “Uncertainty-aware circuit optimization,” in *Proc. DAC*, 2002.
- [94] S. Raj, S. Vrudhula, and J. Wang, “A methodology to improve timing yield in the

- presence of process variations,” in *Proc. DAC*, 2004.
- [95] S. Choi, B. Paul, and K. Roy, “Novel sizing algorithm for yield improvement under process variation in nanometer technology,” in *Proc. DAC*, 2004.
 - [96] A. Agarwal, K. Chopra, and D. Blaauw, “Statistical timing based optimization using gate sizing,” in *Proc. DATE*, 2005.
 - [97] A. Davoodi and A. Srivastava, “Probabilistic dual-V_{th} leakage optimization under variability,” in *Proc. ISLPED*, 2005.
 - [98] A. Srivastava, S. Shah, K. Agarwal, D. Sylvester, and D. Blaauw, “Accurate and efficient gate level parametric yield estimation considering correlated variations in leakage power and performance,” in *Proc. DAC*, 2005.
 - [99] A. Conn, N. Gould, and P. Toint, *LANCELOT: A Fortran package for large-scale non-linear optimizer (Release A)*. Springer-Verlag New York, Inc., 1992.
 - [100] S. Schwartz and Y. Yeh, “On the distribution function and moments of power sums with lognormal components,” *Bell Systems Technical Journal*, vol. 61, pp. 1441–1462, 1982.
 - [101] F. Brglez and H. Fujiwara, “A neutral netlist of 10 combinatorial benchmark circuits and a target translator in FORTRAN,” in *Symposium on Circuits and Systems, Special Session on ATPG and Fault Simulation*, 1985.
 - [102] C. Hu, “Gate oxide scaling limits and projection,” in *IEEE Int. Electron Devices Meeting*, p. 96, 1996.
 - [103] B. Kaczer, F. Crupi, R. Degraeve, P. Roussel, C. Ciofi, and G. Groeseneker, “Observation of hot-carrier-induced nfet gate-oxide breakdown in dynamically stressed cmos circuits,” in *IEEE Int. Electron Devices Meet.*, 2002.
 - [104] Y. Lee, N. Mielke, M. Agostinelli, S. Gupta, R. Lu, and W. McMahon, “Prediction of logic product failure due to thin-gate oxide breakdown,” in *IEEE Int. Reliability Physics Symp.*, 2006.
 - [105] J. Sune, “New physics-based analytic approach to the thin-oxide breakdown statis-

- tics,” *IEEE Electron Device Letter*, vol. 22, 2001.
- [106] Y. Lee, R. Nachman, S. Hu, N. Mielke, and J. Liu, “Implant damage and gate-oxide-edge effects on product reliability,” in *IEEE Int. Electron Devices Meeting*, p. 481, 2004.
 - [107] E. Avni, J. Shappir, and J. Appl., “New physics-based analytic approach to the thin-oxide breakdown statistics,” *Applied Physics*, vol. 64, p. 743, 1988.
 - [108] J. Stathis, “Physical and predictive models of ultra thin oxide reliability in cmos devices and circuits,” *IEEE Trans. on Devices and Materials Reliability*, 2001.
 - [109] R. Degraeve, “A consistent model for intrinsic breakdown in ultra-thin oxides,” in *IEEE Int. Electron Devices Meeting*, p. 866, 1995.
 - [110] J. Sune and E. Y.Wu, “Statistics of successive breakdown events in gate oxides,” in *IEEE Int. Electron Devices Meeting*, p. 481, 2004.
 - [111] J. P. Imhof, “Computing the distribution of quadratic forms in normal variables,” *Biometrika*, vol. 48, pp. 419–426, 1961.
 - [112] R. W. Farebrother, “Algorithm AS 256: The distribution of a quadratic form in normal variables,” *Applied Statistics*, vol. 39, pp. 294–309, 1990.
 - [113] J. Xiong, V. Zolotov, N. Venkateswaran, and C. Visweswariah, “Criticality computation in parameterized statistical timing,” in *Proc. DAC*, 2006.