Statistical Analysis and Optimization for VLSI: Timing and Power

Ashish Srivastava
Dennis Sylvester
David Blaauw
University of Michigan, Ann Arbor

Springer

Library of Congress Cataloging-in-Publication Data

A C.I.P. Catalogue record for this book is available
from the Library of Congress.

ISBN-10: 0-387-25738-1
ISBN-13: 9780387257389

ISBN-10: 0-387-26528-7 (e-book)
ISBN-13: 9780387265285

Printed on acid-free paper.

Printed in the United States of America.

9 8 7 6 5 4 3 2 1 SPIN 11380634

springeronline.com

Dedicated to our families

Preface

Traditional deterministic computer-aided-design (CAD) tools no longer serve the needs of the integrated circuit (IC) designer. These tools rely on the use of corner case models which assume worst-case values for process parameters such as channel length, threshold voltage, and metal linewidth. However, process technologies today are pushed closer to the theoretical limits of the process equipment than ever before (sub-wavelength lithography is a prime example) – this leads to growing levels of uncertainty in these key parameters. With larger process spreads, corner case models become highly pessimistic forcing designers to overdesign products, particularly in an application-specific integrated circuit (ASIC) environment. This growing degree of guardbanding erodes profits, increases time to market, and generall will make it more difficult to maintain Moore's Law in the near future.

The concept of statistical CAD tools, where performance (commonly gate delay) is modeled as a distribution rather than a deterministic quantity, has gained favor in the past five years as a result of the aforementioned growing process spreads. By propagating expected delay distributions through a circuit and not a pessimistic worst-case delay value, we can arrive at a much more accurate estimation of actual circuit performance. The major tradeoff in taking this approach is computational efficiency. Therefore, we can only afford to use statistical CAD tools when their performance benefit is compelling. In earlier technologies this was not the case. However, many companies now feel that the levels of variability, and the stakes, are high enough that the day of statistical CAD has arrived. An inspection of current CAD conference technical programs reflect a large amount of interest from both academia and industry; the current year's Design Automation Conference (DAC) has at least a dozen papers on this topic, nearly 10% of the conference program. While a large fraction of this work has been in extending traditional deterministic static timing analysis (STA) to the statistical regime, power is also critical due to the exponential dependencies of leakage current on process parameters.

As a result of the above trends, the pace of progress, in the past few years in statistical timing and power analysis has been rapid. This book attempts to

summarize recent research highlights in this evolving field. Due to the rapid pace of progress we have made every effort to include the very latest work in this book (e.g., at least five conference publications from the current year are included in the reference list). The goal is to provide a "snapshot" of the field circa mid-2005, allowing new researchers in the area to come up to speed quickly, as well as provide a handy reference for those already working in this field. Note that we do not discuss circuit techniques aimed at reducing the impact of variability or monitoring variability, although we feel these will play a key role in meeting timing, power, and yield constraints in future ICs. The focus here is on CAD approaches, algorithms, modeling techniques, etc.

On a final note, a key to the widespread adoption of statistical timing and power analysis/optimization tools is designer buy-in. This will only come about when there is open discussion of variability data, variation modeling approaches (e.g., Does a Quad-Tree model accurately capture the actual behavior of spatially correlated process parameters?), and related topics. We believe that the recent progress in algorithms for statistical analysis and optimization has brought us to the point where these practical issues, and not the underlying tool capabilities, are the limiting factor in commercial acceptance of the approaches described in this book.

This book is organized into six chapters. The first chapter provides an overview of process variability: types, sources, and trends. The second chapter sets the stage for the following four chapters by introducing different statistical modeling approaches, both generic (Monte Carlo, principal components) and specific to the topic of integrated circuit design (Quad-Tree). The third chapter summarizes recent work in statistical timing analysis, a ripe field of research in the past 4-5 years. Both block-based and path-based techniques are described in this chapter. Chapter 4 turns attention to power for the first time – both high-level and gate-level approaches to modeling variation in power are presented with emphasis on leakage variability. Chapter 5 combines ideas from the previous two chapters in examining parametric yield. This important performance metric may replace other more traditional metrics, such as delay or power, in future ICs as the primary objective function during the design phase. Finally, Chapter 6 describes current state-of-the-art in the statistical optimization area – the work to date is primarily aimed at timing yield optimization and ranges from sensitivity-based to dynamic programming and Lagrangian relaxation techniques.

The authors would like to thank Carl Harris of Springer Publishers for arranging for this book to be published and also for consistently pushing us to the finish line. We thank Sachin Sapatnekar for comments on the general content of the book and we also thank Amanda Brown and Paulette Ream for help in proofreading and generating figures.

Ann Arbor Michigan,
May 2005

Ashish Srivastava
Dennis Sylvester
David Blaauw

Contents

1

Introduction

The impact of process and environmental variations on performance has been increasing with each semiconductor technology generation. Traditional corner-model based analysis and design approaches provide guard-bands for parameter variations and are, therefore, prone to introducing pessimism in the design. Such pessimism can lead to increased design effort and a longer time to market, which ultimately may result in lost revenues. In some cases, a change in the original specifications might also be required while, unbeknownst to the designer performance is actually left on the table. Furthermore, traditional analysis is limited to verifying the functional correctness by simulating the design at a number of process corners. However, worst case conditions in a circuit may not always occur with all parameters at their worst or best process conditions. As an example, the worst case for a pipeline stage will occur when the wires within the logic are at their slowest process corner and the wires responsible for the clock delay or skew between the two stages is at the best case corner. However, a single corner file cannot simultaneously model best-case and worst-case process parameters for different interconnects in a single simulation. Hence, a traditional analysis requires that two parts of the design are simulated separately, resulting in a less unified, more cumbersome and less reliable analysis approach. The strength of statistical analysis is that the impact of parameter variation on all portions of a design are simultaneously captured in a single comprehensive analysis, allowing correlations and impact on yield to be properly understood.

As the magnitude of process variations have grown, there has been an increasing realization that traditional design methodologies (both for analysis and optimization) are no longer acceptable. The magnitude of variations in gate length, as an example, are predicted to increase from 35% in a 130 nm technology to almost 60% in a 70 nm technology. These variations are generally specified as the fraction $3\sigma/\mu$ (3σ is assumed to be the worst case shift in the parameter), where σ and μ are the standard deviation and mean of the process parameter, respectively. Thus a 60% variation in 70 nm technology implies that the standard deviation of the distribution of gate length across a

large number of samples is 14 nm. With variations as large as these, it becomes extremely important that the designers treat these variation in a statistical manner rather than using gaurd-bands in deterministic analysis.

1.1 Sources of Variations

The traditional approach to ensuring acceptable yield is to estimate margins, while assuming worst-case process and environmental conditions. With increasing clock frequency and the growth of variations, these margins have become a larger fraction of the total clock cycle, making the traditional techniques hard to sustain. Part of this difficulty is that margins do not result from a single source of randomness. They are, in fact, used to capture a host of physical effects that are either truly statistical (and hence unknown at design time), or are hard to model while performing analysis.

The first step to consider the impact of variations during the design process is to understand the sources of variations and the impact they have on performance. We first characterize the variations based on their sources.

1.1.1 Process Variations

Process variations are fluctuations in the value of process parameters observed after fabrication. These variations result from a wide range of factors during the fabrication process which determine the ranges of variations. It is obvious that large variations in process parameters will lead to designs that deviate strongly from their specifications. These variations effect the performance characteristics of devices as well as interconnects. The resulting distribution for performance across a large set of fabricated samples leads to the definition of *parametric yield*, which is the fraction of manufactured samples that meet the performance constraints. Parametric yield should be contrasted to *manufacturing yield* that defines the fraction of samples manufactured without catastrophic manufacturing failures (such as wire shorts and opens) that render a given sample useless at any frequency.

For a given process technology, two different designs can have significantly different parametric yield. This results from the fact that the same variations in process parameters may influence two designs in very different manners. For example, we will see in Chap. 2 that designs with a large number of timing critical signals have an increased susceptibility to process variations. In this context, we define the so-called *timing yield* as the fraction of samples of a design that meet the timing constraint, and similarly we define the *power yield* as the fraction of samples that meet the power constraint.

1.1.2 Environmental Variations

These variations capture the variations in the surrounding environment in which a chip sits during its operation. This includes temperature variations,

variation in the power supply and variations in switching activity (defined by the input vectors). A reduced power supply lowers the drive strengths of the devices and hence degrades performance. Similarly, an increased temperature results in performance degradation for both devices and interconnects. It is important to understand that these variations depend on the work-load of the processor and are hence time-dependent. Thus, the set of input vector combinations that result in a worst-case voltage supply drop can occur on any possible sample of the design but will, in all likelihood, occur only intermittently during its operational life time. Thus, power supply and temperature variations are generally not treated statistically, since every shipped chip is required to operate without failures over its entire operational life-time. Power supply drops and high temperatures are, therefore, assumed during the verification of a design. However, identifying specific worst-case conditions for temperature and power supply variation is extremely difficult. Therefore, designers often focus on minimizing temperature and supply variations as much as possible, such as ensuring that the voltage drop on a power grid is always within 5%–10% of the nominal supply voltage.

A particularly interesting situation occurs when process variations increases the current demands on the power supply grids. In older technologies, leakage power dissipation was a concern only in designs that spent a large fraction of their time in stand-by. With leakage power becoming a significant contributor to total power dissipation, leakage currents flowing through the power grid can result in significant supply voltage drops. Moreover, assuming that all devices are operating at their highest leakage will be extremely pessimistic. In this situation, it becomes important to estimate the mean and variance of voltage drops and temperature hot-spots based on variation in process parameters [50], [51], since worst-case leakage induced power-supply drops and hot-spots cannot be expected to occur on each sample of a design.

Leakage currents themselves also increase strongly with an increase in temperature, just as increasing leakage currents may result in a higher temperature. In certain cases, this positive feedback can be strong enough to cause *thermal runaway*, where the currents and temperature in the design continue to increase until failure. Thus, it is important that chip level leakage and temperature analysis are performed in a self-consistent manner [156].

1.1.3 Modeling Variations

These variations result from the fact that the power and delay models used to perform design analysis and optimization are inaccurate and do not perfectly capture device characteristics. These models, if conservative, will make it harder to meet design specifications, whereas aggressive models will result in yield loss. The sample-space of these variations is over design iterations, with different modeling errors at different design points. The tradeoff, in using smaller margins to capture modeling variations, involves the likelihood of tuning particular paths post-fabrication or going through the entire design

process again. Thus, we typically want to be conservative while accounting for modeling variations, since it affects all fabricated samples of a design.

1.1.4 Other Sources of Variations

Though most variations are included within the previous three classes of variations, there are physical effects that result in a change in process parameter with time. These effects include phenomena such as hot electrons, negative bias temperature instability (NBTI) and electromigration. Hot electron and NBTI effects result in device degradation with time causing the threshold voltage of the device to rise. Electromigration may cause increased wire resistance due to a reduction in the width of a wire, which increases the resistance of the wire and increases propagation delay. In the worst case, it will result in wire opens and shorts causing functional failure. The impact of these variations depends strongly on process and environmental variations. A wire that has a smaller width to start-off (due to patterning) and is used to provide current to a hot section of the design that demands large currents is much more likely to fail due to electromigration. If these effects are not properly accounted during the design process, they may result in timing errors that become visible during operation or burn-in. The analysis of these variations is particularly difficult, since they become visible after a reasonable time of operation. Therefore, techniques such as burn-in, which are accelerated test techniques, are used. These testing techniques are used to stress the design to operate under worst-case conditions. However, these testing techniques are expensive and have a large application time.

1.2 Components of Variation

For the purpose of design analysis, it is beneficial to divide the variations into two categories: inter-die and intra-die variations. As we will see in later chapters, these components influence the performance of a design differently. Moreover, the influence of these components also depends on how well the design is optimized, which impacts the number of critical paths in a design.

1.2.1 Inter-die Variations

Inter-die variations refer to a parameter variation that has the same value across a single die, and hence captures variations that occur from die-to-die, wafer-to-wafer and lot-to-lot. Since these variations are independent, they are all represented using a single variational term for ease of analysis. These variations are thus represented by a single value for each die and represent a shift in the mean or expected value of the parameter distribution from the nominal value. These variations include gate-length variations due to fluctuations in the time of exposure during fabrication and metal thickness variations

between different metal layers. Thus, considering inter-die variations for a process parameter, we can write the value of a parameter for a device as a random variable (RV).

$$P = P_{\text{nom}} + \Delta P_{\text{inter}} \tag{1.1}$$

where P_{nom} is the nominal value of the process parameter and P_{inter} is a zero mean RV that captures the inter-die variation. The RV P_{inter} has a single value for all components on the die. The inter-die variations are generally assumed to have a simple distribution, such as Gaussian, with a given variance. These variations may have systematic trends across dies that can be captured if the specific orientation and location of a die on the wafer is known. However, the designer typically has no control where his chip will be placed on a wafer. Moreover, this information is not available at design time and hence the impact of these factors on process parameters must be captured using a random variable.

Inter-die variations in a single process parameter are easily captured by corner models, which assume that all devices and interconnects on a given sample of the design have a value that is shifted away from the mean by a fixed value that degrades (improves) performance, for slow (fast) path analysis. However, when a number of process parameters are considered simultaneously it is important to consider the correlation between these process parameters. As discussed above, thickness of metal layers that are negatively correlated can result in timing failures when the logic is slower than nominal and clock is faster than nominal. The number of process corners at which a design needs to be simulated for functional correctness thus increase exponentially with the increase in process parameters.

1.2.2 Intra-die Variations

Intra-die variation is the component of variation that causes device parameters to vary across different locations within a single die. Thus, each device on a die requires a separate RV to represent its intra-die variation. Depending on the source of variations, intra-die variations may be spatially correlated or spatially uncorrelated. Though all variations are random, the accepted terminology is to use the term random variations specifically to refer to the uncorrelated component of intra-die variations.

It is obvious that intra-die variations result in a huge increase in the dimensionality of the problem by requiring an extra RV for each device. In addition, these RVs are correlated due to proximity-effects. Since, it is computationally very expensive to generate samples of correlated RVs of high dimensionality, traditional statistical analysis methodologies such as Monte Carlo become unsuitable in scenarios where intra-die variations are significant, whereas deterministic approaches fail to capture the effect of intra-die variations completely. Spatially correlated random variations can be handled

by dividing the chip into regions that can be assumed to be perfectly correlated and using a correlation matrix to capture the correlation among these RVs. If the number of these perfectly correlated regions are small, they can be handled easily.

Now, considering both intra-die and inter-die variations for a process parameter, we can write the value of a process parameter as

$$\begin{aligned} P &= P_{\text{nom}} + \Delta P_{\text{inter}} + \Delta P_{\text{intra}}(x_i, y_i) \\ &= P_{\text{nom}} + \Delta P_{\text{inter}} + \Delta P_{\text{spatial}}(x_i, y_i) + \Delta P_{\text{random, i}} \end{aligned} \tag{1.2}$$

where $\Delta P_{\text{intra}}(x_i, y_i)$ represents intra-die variation that consists of a spatially correlated component $\Delta P_{\text{spatial}}$, which is a function of the location on the die and an independent or so-called random component $\Delta P_{\text{random, i}}$ that has no correlation with other devices and is represented as a separate RV for each device.

Intra-die variations can also be classified based on their origin as: wafer-level trends, layout dependent variations and statistical variations.

Wafer-level Variations

Wafer-level variation originate due to effects such as *lens aberrations* and result in *bowl-shaped* or other known distributions over the entire reticle, which results in a *slanted* profile of the process parameter across a single die. Again, the direction of slant varies depending on the orientation of the die on the wafer and cannot be ascertained a priori.

Layout Dependent Variations

Layout dependent variations result in different geometric dimensions due to lithographic and etching techniques that are used during fabrication. These include fabrication steps such as chemical mechanical polishing (CMP) and optical proximity correction (OPC). CMP results in variations in dimensions due to *dishing* (shown in Fig. 1.1) and *erosion*. Dishing arises from the fact that all excess copper must be removed from the wafer – to accomplish this goal, a wafer is typically over-polished, removing some of the copper that is supposed to remain. As copper etches much faster than the surrounding dielectric, the wire ends up being shorter than the oxide. Dishing is the vertical distance between the final oxide level and the lowest point in the copper wire. A substantial amount of dishing leads to increased resistance, worsened planarity, and overall process non-uniformity. Constraints are set on the processing equipment (including slurries and pads) to limit the amount of dishing in the widest wire expected in a given process. Oxide erosion is another problem – normally in this case CMP is applied to an array of dense lines. The oxide between wires in a dense array tends to be over-polished compared to nearby areas of wider insulators (that is, oxide between sparse features will be thicker

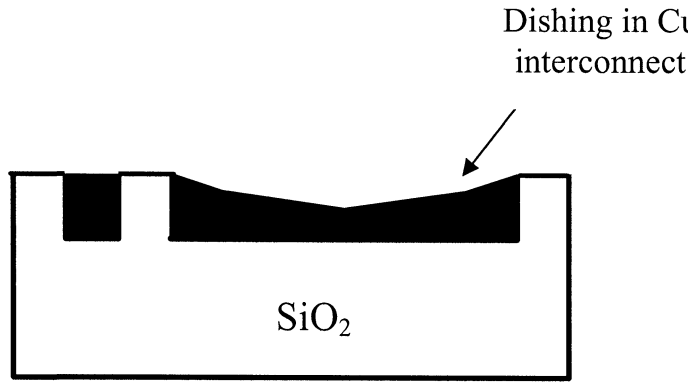

Fig. 1.1. Dishing results in smaller height of copper interconnects resulting in higher resistance, with wider wires having the largest impact.

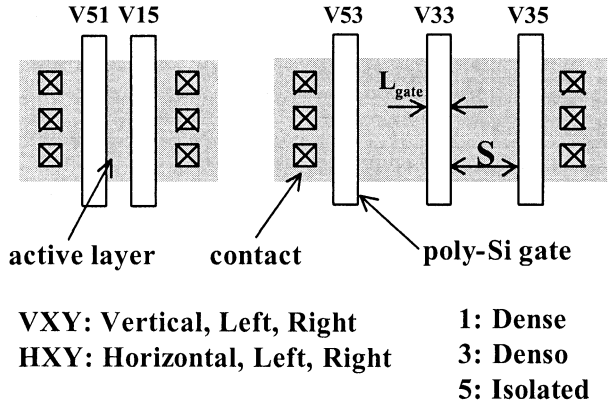

Fig. 1.2. Characterization of polysilicon lines based on their orientation and distance to nearby polysilicon lines [104]. (©2005 IEEE)

than that between dense features). Both dishing and oxide erosion are problematic in wide lines and dense arrays, respectively, and are therefore layout dependent. They lead to higher resistances and more surface non-uniformity.

The patterning of features smaller than the wavelength of light used in optical lithography results in distortions due to the diffraction of light referred to as optical proximity effects (OPE). Shorter wavelength lithography technology is too costly and unstable to be used in current technologies. Changes made to the mask layout to account for these distortions are known as optimal proximity corrections (OPC). Another technique that is used to improve

the performance of sub-wavelength lithography is phase-shift masks (PSM), which exploits the phenomenon of interference to enable patterning of features with higher resolution. OPEs are also layout dependent and result in different CD variations depending on their environment (presence of neighboring lines) and orientation (vertical or horizontal). Figure 1.2 shows the classification of polysilicon lines based on their orientation and distance to the neighboring lines from the left and right edges. The edge is characterized as being *dense* if the next line is at the minimum possible distance, *denso* if the next line is at some intermediate distance, and *isolated* if the next line is further apart. Based on test-chip measurements, the work in [104] found that proximity CD variation is a strong function of both the orientation and the nearby environment. Controlling these variations has become extremely critical in current technologies and has resulted in an explosion in the number of design rules. Polysilicon routing in two orthogonal directions may no longer be allowed in certain technologies, so that better control can be achieved in one single direction. Since these variations are layout dependent, they are generally treated as spatially correlated intra-die variations.

Statistical Variations

Statistical quantization effects, such as random dopant variations, have also grown with scaling of process dimensions. The number of dopant atoms in the channel region of a device decreases as the critical dimension is scaled down. As the number of dopant atoms becomes less, small variation in their number result in a large variation in device performance. Moreover, the actual location of these atoms also plays a role in determining the threshold voltage of a device, further increasing the variability. These variations are true random variations with no correlation across devices and represent one source of intra-die random variations. Such random variations can result from a host of other sources as well, such as lithography, etching, CMP etc. Although their impact in current technologies is small, it is expected to grow as process parameters scale. Their impact on performance has been manageable since random intra-die variations have the well known *averaging effect*, and their impact on path delay decreases with increasing logic depth. However, they result in an increase in mean circuit delay. In addition, the trend to increase clock frequency of a design using aggressive pipelining has resulted in smaller logic depths, which increases the effect of these random intra-die variations.

These variations have a strong influence on leakage power as well, which has become a big cause for concern even in current technologies. As an example, increased V_{th} variability and lower V_{th} values (which result in a much higher leakage) can result in functional failures in dynamic logic designs. To counter worst-case leakage scenarios, a stronger *keeper device* is required which has a negative impact on both power and performance. Adaptive post-fabrication techniques such as [74], which turn on a subset of parallel keeper devices depending on the variations will become useful in these scenarios.

We have classified variations as being inter- and intra-die variations with intra-die variations having spatially correlated and random components. Another equivalent view is to divide variations as being spatially uncorrelated and correlated with the correlated variation further divided as being intra- or inter-die variations depending on their correlation distance [158]. However, we will work with the previous definition of variations throughout the remainder of this book.

1.3 Impact on Performance

In this section, we will discuss the impact of variation on performance parameters. However, first we need to establish the components of variations that dominate each of the device and interconnect parameters. Variation in gate-length is perhaps the most critical device variation and has significant components of both inter-die variation (resulting from variation in duration of exposure) and intra-die variation (resulting from lens aberration and other lithography effects) [158], [124]. The intra-die variations in gate length are also expected to have significant components of spatially correlated variation with a small amount of random variations.

Device threshold voltage presents an interesting picture, since it is dependent on a number of process parameters such as channel doping concentration and gate length. Variations in gate length result in a change in the Drain Induced Barrier Lowering (DIBL) coefficient which results in a change in the threshold voltage. Thus, it is beneficial to separate the variation of threshold voltage between gate length independent variation, resulting from channel doping variations which are random intra-die variations, and gate length dependent variation (which has equal components of inter-die and spatially correlated intra-die variations). In current technologies, most of the variation in threshold voltage is due to variation in gate length and is thus spatially correlated. However, in future technologies random dopant variations are expected to increase raising the level of random variations significantly. In terms of interconnect parameters variations, most of the variations are spatially correlated intra-die variations and inter-die variations.

The trends in the magnitude of process variations is shown in Fig. 1.3 based on the National Technology Roadmap of Semiconductors [99]. The figure shows the increase in the variability of interconnect parameters such as wire width W, wire thickness T, wire height H and resistivity ρ, along with device parameters such as gate-oxide thickness T_{ox} and threshold voltage V_T and environmental factors such as power supply voltage V_{dd}. It shows that variations in gate-length are expected to increase significantly as compared to other process parameters, with variability increasing in all parameters.

The impact of the variations on power and performance was highlighted in [20], which showed measured data over 1000 samples of a design manufactured in an 180 nm technology. The results showed a 20X variation in leakage current

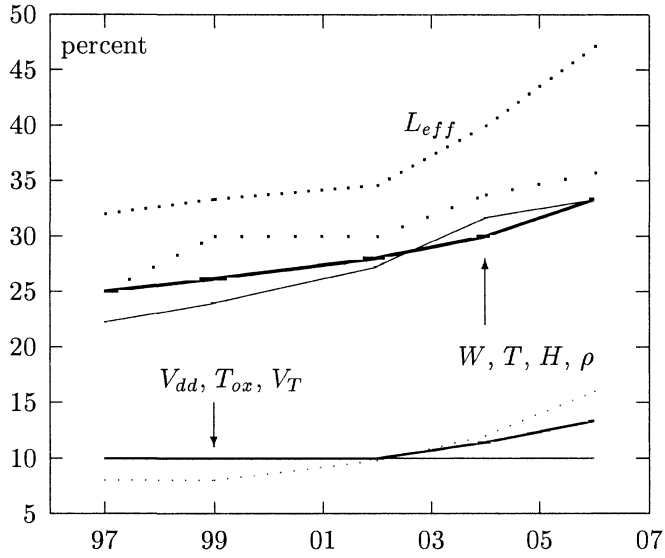

Fig. 1.3. Variability trends in key process parameters with scaling process technology. The x-axis is time with numbers representing the last two digits of the year and the y-axis represents variability in process parameters [99]. (©2005 IEEE)

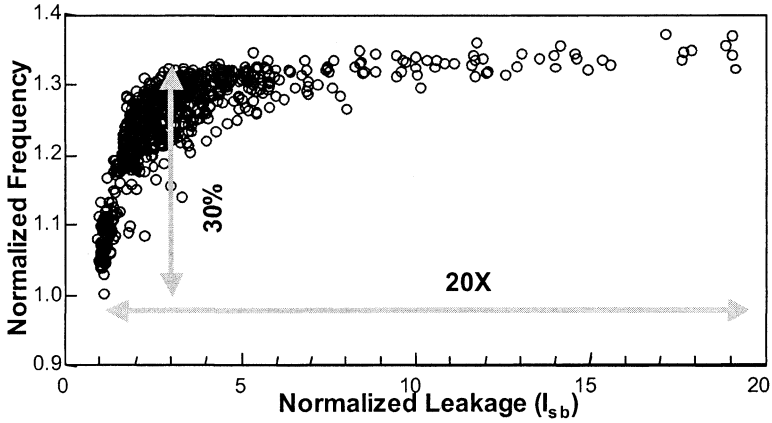

Fig. 1.4. Large variations in leakage power and performance are attributed to process variations [20]. (©2005 IEEE)

for a 1.3X variation in performance. The large variations in leakage result in a large fraction of samples that fail to meet the power constraint. Moreover, these samples are the high performance samples of a design and hence result in a two sided constraint on the region that represents samples that meet both the timing and power constraint.

Though the problem of variations seems to be growing tremendously, [124] recently showed that spatial correlated variations have been kept within manageable limits due to better polysilicon CD control. It was argued that the impact of inter-die variation can be kept within limits through better analysis and design techniques.

2

Statistical Models and Techniques

Traditionally, circuit performance has been modeled in the industry using worst-case models which are used to predict the performance of a design under worst-case process, temperature, and voltage conditions. However, with scaling process dimensions, the impact of process variations has grown, making traditional worst-case models extremely pessimistic. This results in the reduction of feasible regions for the design and increases design effort. Additionally, most of this effort is aimed at accounting for worst-case situations that will most likely not occur in actual designs. This has resulted in significant interest in statistical modeling techniques that can be used to enable statistical analysis and optimization.

Although the need for statistical modeling has been acknowledged to be critical, industry has been reluctant in adopting modeling techniques that can be used to replace traditional worst-case models. This stems from the fact that statistical models are expensive and difficult to construct, and unless analysis and optimizations tools are built on top of these modeling techniques, the utility and validity of these models will be questionable.

In this chapter, we will discuss key statistical techniques, such as principal component analysis, that have been extensively used in developing techniques for process variation modeling and analysis to simplify the problem of simultaneously considering different components of variations. We will also look at specialized modeling techniques to account for sources of variations as discussed in Chap. 1. Having developed the basic infrastructure to model process variation, we will then discuss performance modeling techniques using response surfaces. Then we will discuss statistical gate-delay models and interconnect-delay models that have seen substantial research activity in the past few years.

Before we discuss modeling techniques, let us spend some time understanding the basics of a crucial statistical technique known as Monte Carlo. This will serve as a benchmark against which all modeling and analysis techniques will be tested for accuracy. The need for techniques such as Monte Carlo becomes obvious as soon as we look at the scale of the problem at hand. We

will show that the error in Monte Carlo techniques reduces with the number of samples n as $O(n^{-1/2})$. Hence, obtaining an accuracy improvement of two orders of magnitude requires that the number of samples be increased by four orders of magnitude. Thus, the number of simulations required to obtain reasonable accuracy using Monte Carlo is generally extremely large and using a Monte Carlo based analysis or optimization engine will be prohibitive. Even though this seems to be computationally demanding, this dependence is much better than non-statistical techniques where the error reduces as $O(n^{-1/d})$, where d is the dimensionality of the problem.

Therefore, Monte Carlo methods are used in almost all cases to evaluate the results obtained using newly developed analysis techniques. These techniques, which are, in general, orders of magnitudes faster than performing Monte Carlo simulations, lay the framework for the development of optimization engines that provide improvements in a reasonable amount of time. However, it is important to understand the basics of Monte Carlo simulations, so that they are used reasonably as golden models to test the accuracy of new techniques.

2.1 Monte Carlo Techniques

Numerical methods that make use of random numbers are known as *Monte Carlo* methods. One of the most important applications of Monte Carlo methods is in the evaluation of multi-dimensional integrals, and hence finds extensive application in areas such as yield estimation [154].

Non-statistical numerical techniques to estimate one dimensional definite integrals proceed by dividing the region, over which the integration needs to be performed, into a number of identical parts. Let us apply the technique to estimate the definite integral as shown in Fig. 2.1

$$I = \int_a^b f(x)\mathrm{d}x. \tag{2.1}$$

The interval $[a, b]$ is divided into n equal subintervals such that $a = x_0 < x_1 < x_2 < \cdots < x_n = b$. The integral (2.1) can then be approximated by

$$I = \int_a^b f(x)\mathrm{d}x \approx \sum_{i=0}^{i=n-1} f\left(\frac{x_i + x_{i+1}}{2}\right) h \tag{2.2}$$

where $h = (b - a)/n$. This method is known as the *midpoint* method, since it approximates the area under the curve $f(x)$ in a subinterval using the value of the function at the midpoint of the subinterval. If the function varies linearly within the subinterval, then the value estimated using the midpoint method is exact. Hence, in the general case, midpoint method incurs an $O(h^2)$ error in each subinterval of the integral. Since the total number of subintervals is

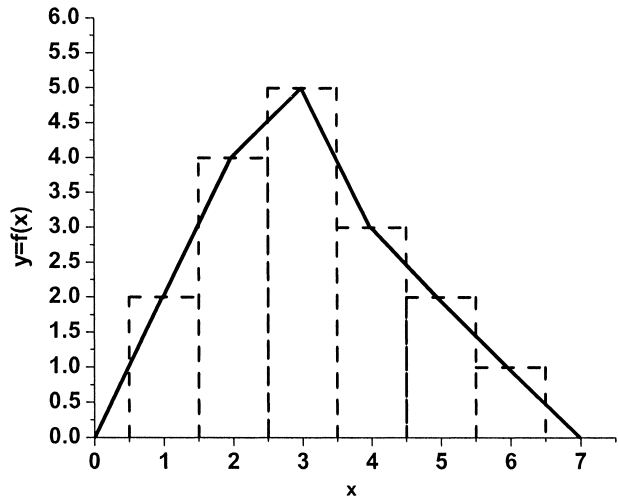

Fig. 2.1. Midpoint method to approximate the integral of $f(x)$, or the area under a curve.

inversely proportional to h, the overall error incurred in estimating the integral is $O(h)$. Thus, we can finally write

$$I = \int_a^b f(x)\mathrm{d}x = \sum_{i=0}^{i=n-1} f\left(\frac{x_i + x_{i+1}}{2}\right) h + O(h)$$
$$= \sum_{i=0}^{i=n-1} f\left(\frac{x_i + x_{i+1}}{2}\right) h + O(n^{-1}). \tag{2.3}$$

The approach can be easily extended to two dimensional integrals. We now consider the case where the area enclosed by a curve is estimated as shown in Fig. 2.2. Using the ideas from the one dimensional case, the two dimensional surface is divided into a set of n equal sized squares with dimensions (h, h). If the midpoint of the square is enclosed by the curve, then the square contributes to the integral, otherwise not. Note that the square either contributes fully to the area or contributes nothing. The error in estimating the area of the square that actually contributes to the area of the curve is therefore $O(h^2)$. Since the number of squares that intersect the curve is $O(h)$, the overall error in estimating the area is again $O(h)$. However, the number of function evaluations required to estimate the area is now proportional to $1/h^2$, which results in an overall error in the integral of $O(n^{-1/2})$. Note that if this idea is extended to the evaluation of multi-dimensional integrals of dimension d, the

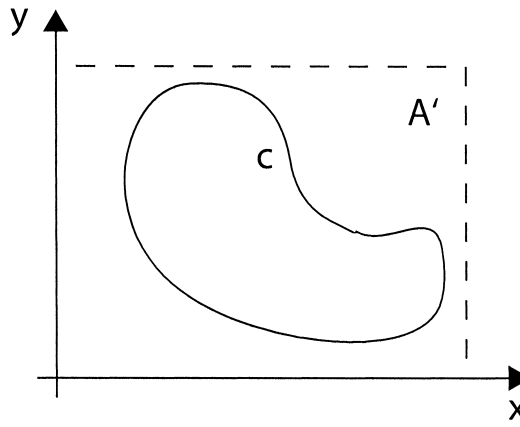

Fig. 2.2. Estimating the area enclosed by the curve C enclosed by a rectangular bounding box A'.

error falls off at a very slow rate of $O(n^{-1/d})$ as the number of samples in the d-dimensional space are increased. Thus we see that to maintain a reasonable accuracy, the number of function evaluations required by the midpoint method grows rapidly with the dimensionality of the integral.

Let us again estimate the area enclosed by a curve as shown in Fig. 2.2, now using a statistical technique. Instead of partitioning the entire region A', we generate n random points independently and assume that n_0 of these points lie within the region enclosed by the curve. Now we can approximate the area enclosed by the curve as

$$A_C \approx \widehat{A_C} = A_{A'} \frac{n_0}{n} \tag{2.4}$$

where $A_{A'}$ is the area of the region A' and A_C is the area enclosed by the curve C as shown in the figure. What is the advantage of this method compared to the midpoint method? To answer this question we need to estimate the error incurred in using approximation (2.4). The probability that a randomly generated point lies within the area enclosed by the curve is simply $A_C/A_{A'}$. If we generate n such samples, then the number of points found to be within C can be expressed as

$$n_0 = \sum_{i=1}^{n} x_i \tag{2.5}$$

where x_i is the result of the i^{th} measurement of x, which is 1 if the randomly generated i^{th} point lies within C and 0 otherwise. The expected value of n_0 can then be expressed as

$$E[n_0] = E\left[\sum_{i=1}^{n} x_i\right] = \sum_{i=1}^{n} E[x] \tag{2.6}$$

where $E[x]$ is the expected value of x, which has a binomial distribution with n samples and a probability of success $A_C/A_{A'}$. The expected value of x can then expressed as

$$E[x] = 0 * \left(1 - \frac{A_C}{A_{A'}}\right) + 1 * \frac{A_C}{A_{A'}} = \frac{A_C}{A_{A'}}. \tag{2.7}$$

Substituting (2.6) and (2.7) into (2.4) and taking expectations we get

$$E[\widehat{A_C}] = A_{A'}\frac{E[n_0]}{n} = A_{A'}\frac{nA_C}{nA_{A'}} = A_C \tag{2.8}$$

and we find that on average the measurement of n_0 will result in an accurate estimate of the area enclosed by C. The class of estimators whose expected value of error is zero are known as *unbiased estimators*, therefore Monte Carlo provides an unbiased estimate of the area.

Let us now consider the variance of the estimate provided by Monte Carlo. We know from *Chebyshev's inequality* [109] that for a RV x

$$\mathcal{P}\left(|x - \eta| \geq \epsilon\right) \leq \frac{\sigma^2}{\epsilon^2} \tag{2.9}$$

where η and σ are the expected value and the standard deviation of x, respectively. Setting $\delta = \sigma^2/\epsilon^2$ we can rewrite (2.9) as

$$\mathcal{P}\left(|x - \eta| \geq \frac{\sigma}{\sqrt{\delta}}\right) \leq \delta. \tag{2.10}$$

Since the expected value of n_0 gives the exact value of A_C, using (2.10) allows us to estimate the error in the value of n_0 in terms of the number of samples for a fixed desired level of accuracy. First, let us calculate the variance of n_0:

$$\begin{aligned}
Var[n_0] &= E\left[\left(n_0 - E[n_0]\right)^2\right] \\
&= E\left[\left(\sum_{i=1}^{n} x_i - E\left[\sum_{i=1}^{n} x_i\right]\right)^2\right] \\
&= E\left[\left(\sum_{i=1}^{n} \left(x_i - E\left[x\right]\right)\right)^2\right] \\
&= E\left[\sum_{i=1}^{n}(x_i - E[x])^2 + 2\sum_{i,j=1}^{n}(x_i - E[x])(x_j - E[x])\right].
\end{aligned} \tag{2.11}$$

Since different measurements of x are assumed to be independent, the second term on the right in (2.11) does not contribute to the expression and (2.11) can be simplified as

$$\begin{aligned} Var[n_0] &= E\left[\sum_{i=1}^{n}(x_i - E[x])^2\right] \\ &= nE[x^2 - 2x_iE[x] + E^2[x]] \\ &= n(E[x^2] - E^2[x]). \end{aligned} \tag{2.12}$$

Now

$$E[x^2] = 0^2 * \left(1 - \frac{A_C}{A_{A'}}\right) + 1^2 * \frac{A_C}{A_{A'}} = \frac{A_C}{A_{A'}} \tag{2.13}$$

therefore, the standard deviation σ of n_0 can be written as

$$\sigma_{n_0} = \sqrt{Var[n_0]} = \sqrt{n\left(\frac{A_C}{A_{A'}}\right)\left(1 - \frac{A_C}{A_{A'}}\right)}. \tag{2.14}$$

Since the estimate of the area enclosed by C is proportional to the ratio n_0/n, using (2.10) and (2.14), the error in the estimate is $O(n^{-1/2})$. Note that the estimation in error is independent of the dimensionality of the problem. This gives us the very interesting and important result that the *error incurred by Monte Carlo methods does not depend on the dimensionality of the problem.* Note that the error in Monte Carlo is fundamentally of a different nature. The error in the midpoint method was due to the inability of the linear approximation to fit the actual integrand, whereas in Monte Carlo methods, the error has a probabilistic origin. Additionally, for one dimensional integrals the midpoint method is more accurate since the error is $O(n^{-1})$ whereas Monte Carlo methods provide an accuracy which is $O(n^{-1/2})$. For two dimensional integrals both the methods provide similar accuracy, and for higher dimensions Monte Carlo methods are always more accurate. The disparity between the accuracy of both the methods increases with the dimensionality of the problem, since the inaccuracy of the midpoint method increases rapidly.

Note that to improve the accuracy of the integral by a factor of two while using Monte Carlo would always require an increase in the number of samples by a factor of four. On the other hand, analytical methods such as the midpoint method require an increase in the number of samples by a factor $2^{D/2}$, where D is the dimensionality of the integral. If $D > 4$, then Monte Carlo methods fare better in this respect as well as compared to analytical midpoint methods.

For our purposes, we will use Monte Carlo methods to estimate the moments of physical or performance parameters. The main goal will be to estimate the quantity

$$E[g(X)] = \int_{\Re} g(x)f(x)\mathrm{dx} \tag{2.15}$$

where X is a RV with probability density function $f(x)$, $g(x)$ is a function of the RV X, and $\Re$ is the region of interest. If we can generate samples of the RV X, then the integral can be estimated as an average of the values of $g(x)$ at these sample points. This approach shows better convergence properties and reduces the runtime of Monte Carlo based techniques.

2.1.1 Sampling Probability Distributions

Monte Carlo methods rely on sampling the space of interest using random samples by generating uniform statistically independent values in the region. As it turns out, it is very difficult to generate truly random numbers using computers. Specialized pieces of hardware are used in certain applications to generate random numbers that amplify the thermal noise of a resistor or a diode and then sample it using a *Schmitt trigger*. If these samples are taken at sufficient intervals of time, we obtain a series of random bits. However, in software, random numbers have to be modeled using *pseudo-random* number generators. Pseudo-random numbers, as the name suggests, are not truly random and are typically generated using a mathematical formula. Most computer languages use *linear congruential* generators. These generators are defined by three positive integers a (multiplier), b (increment), and m (modulus) and given an initial seed (the first pseudo-random number r_0), generates pseudo-random numbers in the following fashion:

$$r_{k+1} = ar_k + b(\mathrm{mod\,m}). \tag{2.16}$$

If desired, the random numbers generated can be mapped to a given range by dividing the numbers obtained using the above generator by m. Note that the r_k's can only take one of the m values. Hence, in all practical implementations m is a very large number (eg. 2^{32}). Also, the choice of a is critical to the randomness of the number generated. More details regarding pseudo-random generators can be found in [75].

We will now review some of the general techniques used to sample arbitrary probability distributions and algorithms to generate samples of some of the pertinent RVs that we will deal with throughout this book.

Inverse Transform Method

Let us assume that the probability distribution function (pdf) of a RV X that we want to sample is given by $f(x)$. The cumulative probability distribution (cdf) $F(x)$, which gives the probability that $X \leq x$, is then given by

$$F(x) = \int_{-\infty}^{x} f(x)\mathrm{d}x. \tag{2.17}$$

Let us take samples of X, which will have a probability density of $f(x)$. Now we will use these samples of X to obtain samples of F. Consider a small region $x < X < x + \mathrm{d}x$ on the x-axis of the cdf. The number of sample points in this region will be proportional to the integral of the pdf in this range. Note that this is equal to the change in the value of the cdf. Hence, the number of sampling points within a range is equal to the length of the region sampled as well. Therefore, these samples of $F(x)$ will be uniformly distributed in the range [0,1].

Using this idea we can write

$$\begin{aligned} u &= F(x) \\ x &= F^{-1}(u) \end{aligned} \tag{2.18}$$

where u represents samples of a uniformly distributed random variable, and F^{-1} is the inverse of F. Hence, if we can find the inverse of F we can use this technique to generate random numbers distributed according to the probability distribution $f(x)$.

Transformation Method

Now let us consider two RVs, X and Y, which are related such that $Y = f(X)$, where f is a monotonic function (inverse of f is well defined). Let the pdf of X and Y be $f_x(x)$ and $f_y(y)$, respectively. Then from the conservation of probability it follows that

$$|\mathcal{P}_x(x)\mathrm{d}x| = |\mathcal{P}_y(y)\mathrm{d}y| \tag{2.19}$$

which states that the probability of finding X between x and $x + \mathrm{d}x$ is the same as the probability of finding Y between $y = f(x)$ and $y + \mathrm{dy} = \mathrm{f(x + dx)}$ as illustrated in Fig. 2.3. From (2.19) it follows that

$$f_y(y) = \frac{f_x(x)}{|f'(x)|}. \tag{2.20}$$

When f is non-monotonic, the left hand side in (2.19) is replaced by a summation of the ranges of x that correspond to the given range of y on the right hand side in (2.19). An equivalent for (2.20) can then be immediately constructed [109]. Therefore, to generate samples of a RV Y we need to find a RV X whose samples can be easily obtained such that X and Y satisfy (2.20).

Consider the case where we want to generate samples of a *Poisson distribution*. The pdf of the Poisson distribution is expressed as

$$f_y(y) = \begin{cases} \mathrm{e}^{-y} & \text{if } 0 \leq \mathrm{y} \leq \infty \\ 0 & \text{o.w.} \end{cases} \tag{2.21}$$

then choosing $y = -\ln x$ we get

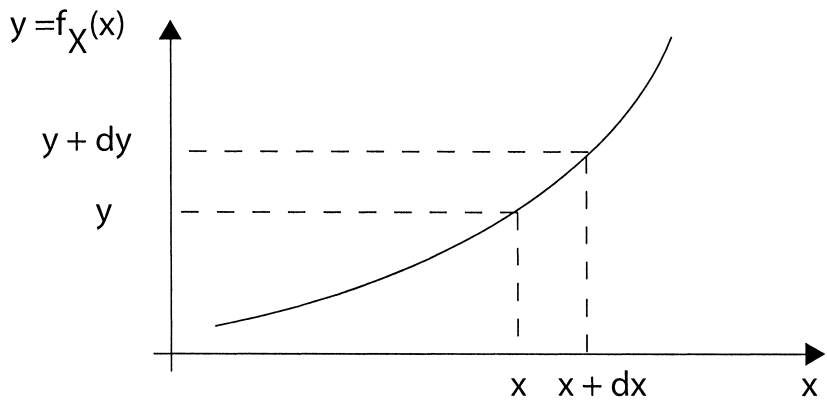

Fig. 2.3. The probability that $x \leq< X <\leq x + \mathrm{d}x$ is equal to the probability that $y \leq< Y <\leq y + \mathrm{d}y$ for the case when Y varies monotonically with X.

$$f_x(x) = \begin{cases} 1 & \text{if } 0 \leq \mathrm{x} \leq 1 \\ 0 & \text{o.w.} \end{cases} \tag{2.22}$$

hence the pdf of Y and X satisfy (2.20). Therefore, if we generate uniform samples in the range [0,1], then the negative natural log of these samples will have a Poisson distribution. This method requires a differentiable pdf, which is a restriction particularly when dealing with discrete RVs.

Acceptance-Rejection Method

If both the above methods are inapplicable due to the restrictions imposed on the pdf of the RV then the acceptance-rejection method may be used. Let us consider the case where we want to generate samples of a RV X whose pdf is as shown in Fig. 2.4. The acceptance-rejection method consists of the following steps. First, generate uniform samples in the range $[x_{min}, x_{max}]$. For each sample x_i evaluate the value of $f_x(x)$. Next, generate another random sample a in the range $[0, \max f_x(x)]$. If $x_i \geq a$, then accept the sample x_i, otherwise reject it. The accepted samples are then distributed according to the pdf f_x.

To generate samples of a Gaussian RV using this approach, we must truncate the pdf of the RV. Since most of the values of a Gaussian RV are concentrated around its mean, a $\pm 4\sigma$ range around the mean is sufficient to capture the behavior of the Gaussian RV. The steps outlined can then be applied to this *truncated* Gaussian RV to generate the desired random samples.

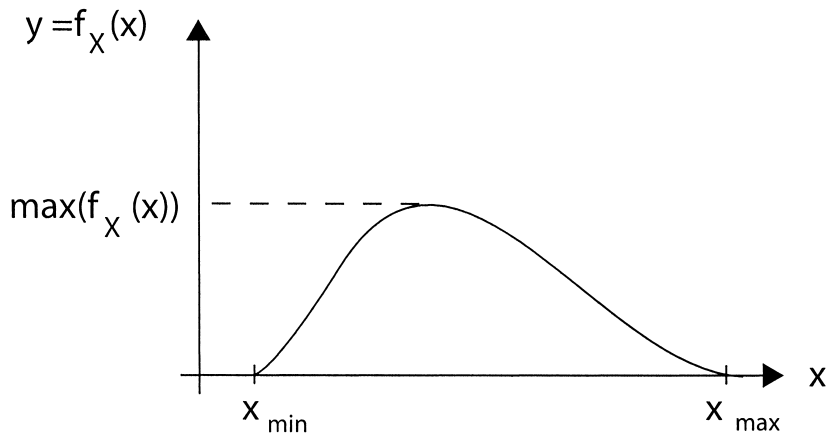

Fig. 2.4. The acceptance-rejection method to generate samples of a RV with a given distribution function.

Generating Multivariate Gaussian RVs

Now let us look at techniques that may be used to generate multivariate Gaussian RVs. We will use the transformation method to generate samples of a one dimensional Gaussian RV. If u_1 and u_2 are independent uniform RVs in the range [0,1], then

$$\begin{aligned} y_1 &= \sin 2\pi u_1 \sqrt{-2 \ln u_2} \\ y_2 &= \cos 2\pi u_1 \sqrt{-2 \ln u_2} \end{aligned} \tag{2.23}$$

are two independent Gaussian RVs with zero mean and unit variance. The Gaussian random numbers generated using the above transformation, also known as the *Box-Muller* transformation, can then be used to generate samples of a Gaussian RV with an arbitrary mean and variance. To obtain the desired mean and variance for the Gaussian RV, we use the fact that given two Gaussian RVs that are related as $Y = aX + b$

$$\begin{aligned} E[Y] &= aE[X] + b \\ Var[Y] &= E[Y^2] - E^2[Y] = a^2 Var[X]. \end{aligned} \tag{2.24}$$

To generate an n-dimensional multivariate random variable with a covariance matrix $\mathbf{\Sigma}$ and mean $\mathbf{\Delta}$, the first step is to generate n independent random variables with zero mean and unit variance. Then, take a sample of these RVs ($\mathbf{X}$), and generate a new sample $\mathbf{X'}$ from $\mathbf{X}$ such that