国外电子信息精品著作(影印版)

VLSI的统计分析和优化：时序和功耗

Statistical Analysis and Optimization for VLSI: Timing and Power

Srivastava Ashish
Sylvester Dennis
Blaauw David

科学出版社
www.sciencep.com

VLSI的统计分析和优化：时序和功耗

Statistical Analysis and Optimization for VLSI: Timing and Power

Ashish Srivastava
Dennis Sylvester
David Blaauw

科学出版社

国外电子信息精品著作（影印版）

Statistical Analysis and Optimization for VLSI: Timing and Power

VLSI 的统计分析和优化：时序和功耗

Srivastava Ashish
Sylvester Dennis
Blaauw David

科学出版社
北 京

图字：01-2007-3876

内 容 简 介

该书介绍了集成电路的统计CAD工具的相关知识。主要面向CAD工具开发人员、集成电路工艺技术人员，以及相关学科的学生和研究人员。书中介绍了统计时序和功耗分析技术中的最新研究成果，并结合参数化的产量作为设计过程中的主要目标函数。该书强调算法、过程变量的建模方法，以及统计方法。既可作为刚涉足CAD工具开发领域的人员的入门书籍，也可作为该领域工程师的参考手册。

图书在版编目(CIP)数据

VLSI的统计分析和优化：时序和功耗= Statistical Analysis and Optimization for VLSI: Timing and Power：英文/（美）安歇斯（Ashish, S.）等著. —北京：科学出版社，2007
（国外电子信息精品著作）
ISBN 978-7-03-018850-2

Ⅰ.V… Ⅱ.安… Ⅲ.超大规模集成电路-电路设计-英文 Ⅳ.TN470.2

中国版本图书馆CIP数据核字(2007)第071686号

责任编辑：余 丁/责任印制：刘士平/封面设计：陈 敬

科学出版社 出版
北京东黄城根北街16号
邮政编码：100717
http://www.sciencep.com

源海印刷有限责任公司 印刷

科学出版社发行 各地新华书店经销

*

2007年8月第 一 版 开本：B5(720×1000)
2007年8月第一次印刷 印张：18.1/4
印数：1—2 500 字数：457 000

定价：42.00元

（如有印装质量问题，我社负责调换〈新欣〉）

《国外电子信息精品著作》序

20世纪90年代以来，信息科学技术成为世界经济的中坚力量。随着经济全球化的进一步发展，以微电子、计算机、通信和网络技术为代表的信息技术，成为人类社会进步过程中发展最快、渗透性最强、应用面最广的关键技术。信息技术的发展带动了微电子、计算机、通信、网络、超导等产业的发展，促进了生命科学、新材料、能源、航空航天等高新技术产业的成长。信息产业的发展水平不仅是社会物质生产、文化进步的基本要素和必备条件，也是衡量一个国家的综合国力、国际竞争力和发展水平的重要标志。在中国，信息产业在国民经济发展中占有举足轻重的地位，成为国民经济重要支柱产业。然而，中国的信息科学支持技术发展的力度不够，信息技术还处于比较落后的水平，因此，快速发展信息科学技术成为我国迫在眉睫的大事。

要使我国的信息技术更好地发展起来，需要科学工作者和工程技术人员付出艰辛的努力。此外，我们要从客观上为科学工作者和工程技术人员创造更有利于发展的环境，加强对信息技术的支持与投资力度，其中也包括与信息技术相关的图书出版工作。

从出版的角度考虑，除了较好较快地出版具有自主知识产权的成果外，引进国外的优秀出版物是大有裨益的。洋为中用，将国外的优秀著作引进到国内，促进最新的科技成就迅速转化为我们自己的智力成果，无疑是值得高度重视的。科学出版社引进一批国外知名出版社的优秀著作，使我国从事信息技术的广大科学工作者和工程技术人员能以较低的价格购买，对于推动我国信息技术领域的科研与教学是十分有益的事。

此次科学出版社在广泛征求专家意见的基础上，经过反复论证、仔细遴选，共引进了接近30本外版书，大体上可以分为两类，第一类是基础理论著作，第二类是工程应用方面的著作。所有的著作都涉及信息领域的最新成果，大多数是2005年后出版的，力求“层次高、内

容新、参考性强”。在内容和形式上都体现了科学出版社一贯奉行的严谨作风。

当然，这批书只能涵盖信息科学技术的一部分，所以这项工作还应该继续下去。对于一些读者面较广、观点新颖、国内缺乏的好书还应该翻译成中文出版，这有利于知识更好更快地传播。同时，我也希望广大读者提出好的建议，以改进和完善丛书的出版工作。

总之，我对科学出版社引进外版书这一举措表示热烈的支持，并盼望这一工作取得更大的成绩。

中国科学院院士
中国工程院院士
2006 年 12 月

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