Novel Algorithms for Fast Statistical Analysis of Scaled Circuits

Lecture Notes in Electrical Engineering
Volume 46

For other titles published in this series, go to
www.springer.com/series/7818

Amith Singhee • Rob A. Rutenbar

Novel Algorithms for Fast Statistical Analysis of Scaled Circuits

Springer

Dr. Amith Singhee
IBM Corporation
T. J. Watson Research Center
1101 Kitchawan Road
Route 134
PO Box 218
Yorktown Heights, NY 10598
USA
asinghee@us.ibm.com

Rob A. Rutenbar
Carnegie Mellon University
Dept. Electrical & Computer Engineering
5000 Forbes Ave.
Pittsburg, PA 15213-3890
USA
rutenbar@ece.cmu.edu

ISSN 1876-1100 Lecture Notes in Electrical Engineering
ISBN 978-90-481-3099-3 e-ISBN 978-90-481-3100-6
DOI 10.1007/978-90-481-3100-6
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2009931791

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To my parents
– Amith

Introduction

I.1 Background and Motivation

Very Large Scale Integration (VLSI) technology is moving deep into the nanometer regime, with transistor feature sizes of 45 nm already in widespread production. Computer-aided design (CAD) tools have traditionally kept up with the difficult requirements for handling complex physical effects and multi-million-transistor designs, under the assumption of fixed or deterministic circuit parameters. However, at such small feature sizes, even small variations due to inaccuracies in the manufacturing process can cause large relative variations in the behavior of the circuit. Such variations may be classified into two broad categories, based on the source of variation: (1) *systematic* variation, and (2) *random* variation. Systematic variation constitutes the deterministic part of these variations; e.g., proximity-based lithography effects, nonlinear etching effects, etc. [GH04]. These are typically pattern dependent and can potentially be completely explained by using more accurate models of the process. Random variations constitute the unexplained part of the manufacturing variations, and show stochastic behavior; e.g., gate oxide thickness (t_{ox})variations, poly-Si random crystal orientation (RCO) and random dopant fluctuation (RDF) [HIE03]. These random variations cannot simply be accounted for by more accurate models of the physics of the process because of their inherent random nature (until we understand and model the physics well enough to accurately predict the behavior of each ion implanted into the wafer).

As a result, integrated circuit (IC) designers and manufacturers are facing difficult challenges in producing reliable high-performance circuits. Apart from the sheer size and complexity of the design problems, a relatively new and particularly difficult problem is that of these *para-*

metric variations (threshold voltage (V_t), gate oxide thickness, etc.) in circuits, due to nonsystematic variations in the manufacturing process. For older technologies, designers could afford to either ignore the problem, or simplify it and do a worst-case corner based conservative design. At worst, they might have to do a re-spin to bring up the circuit yield. With large variations, this strategy is no longer efficient since the number of re-spins required for convergence can be prohibitively large. *Per-transistor* effects like RDF and line edge roughness (LER) [HIE03] are becoming dominant as the transistor size is shrinking. As a result, the relevant statistical process parameters are no longer a few inter-wafer or even inter-die parameters, but a huge number of inter-device (intra-die) parameters. Hence, the dimensionality with which we must contend is also very large, easily 100s for custom circuits and millions for chip-level designs. Furthermore, all of these inter-die and intra-die parameters can have complex correlation amongst each other. Doing a simplistic conservative design will, in the best case, be extremely expensive, and in the worst case, impossible. These variations must be modeled accurately and their impact on the circuit must be predicted reliably in most, if not all, stages of the design cycle. These problems and needs have been widely acknowledged even amongst the non-research community, as evidenced by this extensive article [Ren03].

Many of the electronic design automation (EDA) tools for modeling and simulating circuit behavior are unable to accurately model and predict the large impact of process-induced variations on circuit behavior. Most attempts at addressing this issue are either too simplistic, fraught with no-longer-realistic assumptions (like linear [CYMSC85] or quadratic behavior [YKHT87][LLPS05], or small variations), or focus on just one specific problem (e.g., Statistical Static Timing Analysis or SSTA [CS05][VRK$^+$04a]). This philosophy of doing "as little as needed", which used to work for old technology nodes, will start to fail for tomorrow's scaled circuits. There is a dire need for tools that efficiently model and predict circuit behavior in the presence of large process variations, to enable reliable and efficient design exploration. In the cases where there are robust tools available (e.g., Monte Carlo simulation [Gla04]), they have not kept up with the speed and accuracy requirements of today's, and tomorrow's, IC variation related problems.

In this thesis we propose a set of novel algorithms that discard simplifications and assumptions as much as possible and yet achieve the necessary accuracy at very reasonable computational costs. We recognize that these variations follow complex statistics and use *statistical* approaches based on accurate statistical models. Apart from being flexible and scalable enough to work for the expected large variations in

future VLSI technologies, these techniques also have the virtue of being independent of the problem domain: they can be applied to any engineering or scientific problem of a similar nature. In the next section we briefly review the specific problems targeted in this thesis and the solutions proposed.

I.2 Major Contributions

In this thesis, we have taken a wide-angle view of the issues mentioned in the previous section, addressing a variety of problems that are related, yet complementary. Three such problems have been identified, given their high relevance in the nanometer regime; these are as follows.

I.2.0.1 SiLVR: Nonlinear Response Surface Modeling and Dimensionality Reduction

In certain situations, SPICE-level circuit simulation may not be desired or required, for example while computing approximate yield estimates inside a circuit optimization loop [YKHT87][LGXP04]: circuit simulation is too slow in this case and we might be willing to sacrifice some accuracy to gain speed. In such cases, a common approach is to build a model of the relationship between the statistical circuit parameters and the circuit performances. This model is, by requirement, much faster to evaluate than running a SPICE-level simulation. The common term employed for such models is *response surface models* (RSMs). In certain other cases, we may be interested in building an RSM to extract specific information regarding the circuit behavior, for example, sensitivities of the circuit performance to the different circuit parameters. Typical RSM methods have often made simplifying assumptions regarding the characteristics of the relationship being modeled (e.g., linear behavior [CYMSC85]), and have been sufficiently accurate in the past. However, in scaled technologies, the large extent and number of variations make these assumptions invalid.

In this thesis, we propose a new RSM method called *SiLVR* that discards many of these assumptions and is able to handle the problems posed by highly scaled circuits. SiLVR employs the basic philosophy of *latent variable regression*, that has been widely used for building linear models in chemometrics [BVM96], but extends it to flexible nonlinear models. This model construction philosophy is also known as *projection pursuit*, primarily in the statistics community [Hub85]. We show how SiLVR can be used not only for performance modeling, but also for extracting sensitivities in a nonlinear sense and for output-driven dimensionality reduction from 10–100 dimensions to 1–2. The ability to extract insight regarding the circuit behavior in terms of numerical quantities,

even in the presence of strong nonlinearity and large dimensionality, is the real strength of SiLVR. We test SiLVR on different analog and digital circuits and show how it is much more flexible than state-of-the-art quadratic models, and succeeds even in cases where the latter completely breaks down. These initial results have been published in [SR07a].

I.2.0.2 Fast Monte Carlo Simulation Using Quasi-Monte Carlo

Monte Carlo simulation has been widely used for simulating the statistical behavior of circuit performances and verifying circuit yield and failure probability [HLT83], in particular for custom-designed circuits like analog circuits and memory cells. In the nanometer regime, it will remain a vital tool in the hands of designers for accurately predicting the statistics of manufactured ICs: it is extremely flexible, robust and scalable to a large number of statistical parameters, and it allows arbitrary accuracy, of course at the cost of simulation time. In spite of the technique having found widespread use in the design community, it has not received the amount of research effort from the EDA community that it deserves. Recent developments in number theory and algebraic geometry [Nie88][Nie98] have brought forth new techniques in the form of *quasi-Monte Carlo*, which have found wide application in computational finance [Gla04][ABG98][NT96a]. In this thesis, we show how we can significantly speed up Monte Carlo simulation-based statistical analysis of circuits using quasi-Monte Carlo. We see speedups of 2× to 50× over standard Monte Carlo simulation across a variety of transistor-level circuits. We also see that quasi-Monte Carlo scales better in terms of accuracy: the speedups are bigger for higher accuracy requirements. These initial results were published in [SR07b].

I.2.0.3 Statistical Blockade: Estimating Rare Event Statistics, with Application to High Replication Circuits

Certain small circuits have millions of identical instances on the same chip, for example, the SRAM (Static Random Access Memory) cell. We term this class of circuits as *high-replication circuits*. For these circuits, typical acceptable failure probabilities are extremely small: orders of magnitude less than even 1 part-per-million. Here we are restricting ourselves to failures due to parametric manufacturing variations. Estimating the statistics of failures for such a design can be prohibitively slow, since only one out of a million Monte Carlo points might fail: we might need to run millions to billions of simulations to be able to estimate the statistics of these very rare failure events. Memory designers have often avoided this problem by using analytical models, where available, or by making

"educated guesses" for the yield, using large safety margins, worst-case corner analysis, or small Monte Carlo runs. Inaccurate estimation of the circuit yield can result in significant numbers of re-spins if the margins are not sufficient, or unnecessary and expensive (in terms of power or chip area) over-design if the margins are too conservative. In this thesis, we propose a new framework that allows fast sampling of these rare failure events and generates analytical probability distribution models for the statistics of these rare events. This framework is termed *statistical blockade*, inspired by its mechanics. Statistical blockade brings down the number of required Monte Carlo simulations from millions to very manageable thousands. It combines concepts from machine learning [HTF01] and extreme value theory [EKM97] to provide a novel and useful solution for this under-addressed, but important problem. These initial results have been published in [SR07c][WSRC07][SWCR08].

I.3 Preliminaries

A few conventions that will be followed throughout the thesis are worth mentioning at this stage. Each statistical parameter will be modeled as having a probability distribution that has been extracted and is ready for use by the algorithms proposed in this thesis. The parameters considered are SPICE model parameters, including threshold voltage (V_t) variation, gate oxide thickness (t_{ox}) variation, resistor value variation, capacitor value variation, etc. It will be assumed for experimental setup, that the statistics of any variation at a more physical level, e.g., random dopant fluctuation, can be modeled by these probability distributions of the SPICE-level device parameters.

Some other conventions that will be followed are as follows.

- All vector-valued variables will be denoted by bold small letters, for example $\mathbf{x} = \{x_1, \ldots, x_s\}$ is a vector in s-dimensional space with s coordinate values, also called an s-vector. Rare deviations from this rule will be specifically noted. Scalar-valued variables will be denoted with regular (not bold) letters, and matrices with bold capital letters; for example, $\mathbf{X}$ is a matrix, where the i-th row of the matrix is a vector $\mathbf{x}_i$. All vectors will be assumed to be column vectors, unless transposed. $\mathbf{I}_s$ will be the $s \times s$ identity matrix.

- We will use s to denote the dimensionality of the statistical parameter space that any proposed algorithm will work in.

- Following standard notation, $\mathbb{R}$ denotes the set of all real numbers, $\mathbb{Z}$ denotes the set of all integers, $\mathbb{Z}_+$ denotes the set of all nonneg-

ative integers, and $\mathbb{R}^s$ is the s-dimensional space of all real-valued s-vectors.

I.4 Organization

The ideas proposed in this thesis are born out of a large body of knowledge from several different fields. Hence, there is no practical limit to the amount of background material that could be considered relevant. It is out of the practical scope of any single volume to cover all such "relevant" material in detail. However, to make these ideas accessible to the general reader, a reasonably comprehensive discussion of the background is needed. In its attempt to achieve a balance between conciseness and completeness, this thesis reviews relevant background material that is required for a clear understanding of the proposed ideas, and avoids lengthy expositions of background material on related or competing ideas. The latter can easily be found in referenced literature in, or related to, electronic design automation, and is not immediately required for a clear understanding of the proposed ideas. In certain cases, small diversions are made to review interesting concepts from some field outside of electrical and computer engineering, to enable a more expansive understanding of the underlying concepts. An example is the brief review of Asian option pricing in Sect. 2.2.1.1.

This thesis is organized into three nearly independent chapters, each presenting one of the three contributions of this work. Chapter 1 introduces SiLVR, the proposed nonlinear RSM method. For this purpose, it first reviews typical RSM techniques and relevant background relating to latent variable regression, projection pursuit, and the specific techniques employed by SiLVR. The chapter ends with a section comparing the modeling results of SiLVR against simulation and an optimal quadratic RSM (PROBE from [LLPS05]). Chapter 2 provides the necessary application and theoretical background for Monte Carlo simulation and the proposed quasi-Monte Carlo (QMC) simulation technique. It then details the proposed QMC flow and present experimental results validating its gains over standard Monte Carlo. Chapter 3 introduces the problem of yield estimation for high-replication circuits and reviews relevant background from machine learning and extreme value theory. It then explains the proposed statistical blockade flow in detail and present validation using different relevant circuit examples. Chapter 4 provides concluding remarks. Suggestions for future research directions are provided at the end of each of Chaps. 1, 2 and 3.

Contents