Forward Discrete Probability Propagation Method for Device Performance Characterization under Process Variations*

Rasit Onur Topaloglu and Alex Orailoglu
University of California San Diego
Computer Science and Engineering Department
La Jolla, CA, 92093
{rtopalog,alex}@cse.ucsd.edu

Process variations are becoming influential at the device level in deep sub-micron and sub-wavelength design regimes, whereas they used to be a few generations away only influential at circuit level. Process variations cause device performance parameters, such as current or output resistance, to acquire a probability distribution. Estimation of these distributions has been accomplished using Monte Carlo techniques so far. The large number of samples needed by Monte Carlo methods adversely affects the possibility of integrating probabilistic device performance at the circuit level due to run-time inefficiency. In this paper, we introduce a novel technique called Forward Discrete Probability Propagation (FDPP). This method discretizes the probability distributions and effectively propagates these probabilities across a device formula hierarchy, such as the one present in the SPICE3v3 model. Consequently, probability distributions for process parameters are propagated to the device level. It is shown in the paper that with far fewer number of samples, comparable accuracy to a Monte Carlo method is achieved.

I. Introduction

Estimation of the effects of process variations on device performance has long been a concern. The computational complexity of current simulators precludes incorporation of process variations to device performance. This can be attributed to the lack of accurate methods and models for process variations. Designers have been trying to cope with this absence through worst-case analysis, Monte Carlo techniques or through the invocation of Gaussian distribution assumptions. But these approaches can no longer be counted upon to provide sufficiently accurate and fast results, as deep sub-micron silicon technologies rapidly push manufacturers to device parameter characterizations of increased accuracy in order to obviate the increasing number of design iterations.

The effects of process variations on device parameters further indicate that the relationships between factors causing process variations and device parameters are deviating from a linear approximation even for a small input domain. This implies that the Gaussian distribution assumption attributed to device performance parameters is no longer accurate. Therefore, a more accurate methodology is necessary to estimate the effects of mismatch on high-level parameters.

The paper presents a methodology to deterministically estimate the results of process variations on device parameters using connectivity graphs. The proposal consists of an algebraically tractable method, leading to the possibility of manual or simulator-guided implementations. In contrast to the Monte Carlo approach, there does not exist any non-determinism in the system. In contrast to Gaussian-based methods, the system is not restricted to Gaussian distributions, thus providing accurate device characterization. The method can also outperform the accuracy and run-time of a Monte Carlo-based approach in certain applications or conditions as indicated in the paper. With no reliance on a random method, it can nonetheless take advantage of analytical fabrication and device models already present in the literature and present the probabilistic device parameters formed as a result of process variations.

The paper proceeds by presenting the motivation and the previous work. The discussion is followed up by the introduction of a mathematical basis for discretization of probability distribution functions, introducing formalism through new operators and domains, followed by experimental data comparing Monte Carlo methods and FDPP.

II. Motivation

Monte Carlo methods are frequently used in engineering applications [1] [2], though they exhibit a number of shortcomings. A foremost one consists of the dependence of Monte Carlo techniques on a random number generator, signifying Monte Carlo as a non-deterministic method. Most computational packages only provide random number generators for a limited set of well-known distributions such as Gaussian or uniform. As a result, the users of Monte Carlo methods are limited in assigning distributions to low level parameters, such as process parameters in device characterization. Though a number of remedies have been suggested, such as importance sampling [18], these modifications usually necessitate an increased number of samples for sufficient accuracy. An approach for accurate consideration of arbitrary distributions may be of utmost importance for certain engineering applications where Gaussian or other distribution assumptions for low level parameters may cause large error build-up during the computation of the distributions of high level parameters.

Another shortcoming is that Monte Carlo, due to its random sampling mechanism, may require an increasingly high number of samples to reduce the error for regions of the probability distribution that have a reduced occurrence probability. This last bottleneck may cause certain regions in a distribution to be missed by the method altogether when computational effectiveness issues limit the number of samples, causing an under-estimation of the probability. Similarly, fewer than the adequate number of

*This project is partially supported by the Semiconductor Research Corporation, under task ID 906.001.

samples may cause an over-estimation at certain regions. This over-estimation may be the result of choosing a point at a low probability region and not being able to normalize with an adequate number of drawn samples. Increasing the number of samples to prevent these bottlenecks, on the other hand, may result in an unmanageable run-time complexity. In most engineering applications, reducing this complexity without an accuracy compromise is nevertheless of utmost interest.

Finally, sampling from correlated parameters may bring forth a large inaccuracy. In a directed acyclic graph, if a node has more than one outgoing edge, assigning the same sample to each of the outgoing edges will over-estimate the variance of a high level node that is a function of these nodes. A user directed sampling mechanism can avoid this error, yet its implementation can be quite cumbersome and error-prone for complex trees. Instead, a probabilistic approach on a tree, on the other hand, may take advantage of Bayes rule [17]. Hence, ancestral nodes can be treated as being conditionally independent while calculating the posterior distribution of the descendant node.

III. Previous Work

Monte Carlo based methods are predominantly used in device parameter characterization [3] [4]. In [3], a Monte Carlo based method has been used for the simulation of impact ionization while in [4], a Schottky barrier is simulated with a Monte Carlo method. Monte Carlo methods are used for the newest technologies as well [5]. [6] and [7] have pointed out the inaccuracy of Monte Carlo methods and formulated it as a variance representing the deviation from estimated values.

Process variations can be attributed to physical parameters as suggested in [8]. In [9], a technique is presented to estimate the device characteristics using the sensitivities of device parameters to physical parameters. Means and variances of device parameters can be approximated in this method. This technique though falls short of being sufficiently accurate in deep sub-micron and sub-wavelength technologies due to the Gaussian distribution assumption attributed to device parameters, as device parameters are sharply deviating from Gaussian distributions with newer technologies, as can be seen in [10], [11] and [12]. Inaccurate information regarding the distribution of device parameters provided to the designers may cause a major bottleneck in the design cycle increasing or elongating iterations. The importance of avoiding such worst-case approximations in deep sub-micron designs has been identified in [13].

The effects of various steps in a semiconductor fabrication on device parameters has been analytically modeled in a number of papers in the literature [14] [15] [16]. However, a continuous time probabilistic analysis is usually not provided when process variations need to be accounted for. Powerful models have been so far presented in the literature. These models should be incorporated into the design in an accurate manner as we progress to newer technologies.

IV. Probability Discretization Theory

Accurate simulation of devices has exceeded computational practicality thresholds. The computational cost of simulating process variations introduces an additional exponential increase to this already inordinately high computational time requirements. The necessity to accurately estimate device parameters has become quite significant as a result of this. To close this gap, we propose a methodology that provides a way for the estimation of device parameters. This methodology is both manually tractable and can be incorporated into a simulator.

In order to introduce the proposed technique, FDPP, a number of definitions will be useful. Let X be a random variable. We will denote the probability distribution of X as $pdf(X)$. $pdf(X)$ is assumed to be continuous. We propose to attain an approximation of this *pdf* by sampling the *pdf* at equidistant points of the random variable X.

In reality, X may extend to positive or negative infinity for certain distributions. In these situations, the tails of the *pdf* will be terminated after a certain value of X, which corresponds to band-pass filtering the *pdf*. This will define a boundary of the form $[m, n]$ for X, where m and n are practical lower and upper limits. The probability that X will *not* fall within this region is given by:

$$\int_{-\infty}^{+\infty} pdf(X) - \int_{m}^{n} pdf(X) = 1 - \int_{m}^{n} pdf(X) \qquad (1)$$

This difference should be chosen as small as possible to reduce the filtering error.

The sampling can be done by dividing the band-pass filtered $pdf(X)$ to bins and approximating the values that fall in any bin by the value at the mid-point of the bin. Let b_i be an enumeration over the bins where $1 \leq i \leq N$ and N is the total number of bins. b_i will be defined to be bounded by $[m + (i-1)\Delta, m + i\Delta)$, where Δ is the step-size defined by $\frac{n-m}{N}$. We denote the sampled $pdf(X)$ as $\phi(X)$ or $spdf(X)$, and we introduce two domains such that $pdf(X)$ is in the *p-domain* and $\phi(X)$ is in the *r-domain*.[1]

The procedure of converting a *pdf* to an *spdf* will be represented with the $\mathcal{Q}_{\mathcal{N}}$ operator:

$$\phi(X) = \mathcal{Q}_N(pdf(X)) \qquad (2)$$

The domain of this operator is a band-pass filtered *pdf*, and the range of this operator is an *spdf*. The result of this operator on the *pdf* of a random variable X, $\phi(X)$, is essentially a Riemann sum of impulses and is given by:

$$\phi(X) = \sum_{i \epsilon 1..N} p_i \delta(x - w_i) \qquad (3)$$

where

$$p_i = \int_{m+(i-1)\Delta}^{m+i\Delta} pdf(X)dx \qquad (4)$$

$$v_i = m + (i-1)\frac{\Delta}{2} \qquad (5)$$

In these equations, p_i corresponds to the probability that a sample of the random variable X falls within the i'th bin b_i and v_i denotes the approximation of values of samples of X within b_i.

[1] This nomenclature has been motivated by the similarity of these domains to the *s-domain* (Laplace domain), as some operations such as filtering and band-passing can be depicted and formulated easily in the *r-domain* than in the *p-domain*, just as some operations are easily applied in the *s-domain* than in the time domain.

v_i is the mid-point in the particular bin. Hence it is given as $m + (i-1)\frac{\Delta}{2}$.

Assume that we have a number of random variables given as X_1,..,X_r, whose sampled *pdf*'s are respectively given by $\phi(X_1)$,..,$\phi(X_r)$. Let Y be another random variable that is given by a deterministic function f of the given random variables: $Y = f(X_1, .., X_r)$. Then $\phi(Y)$ is given by the $\mathcal{F}$ operator as:

$$\phi(Y) = \mathcal{F}(\phi(X_1), .., \phi(X_r)) \tag{6}$$

which is defined as being equivalent to:

$$\phi(Y) = \sum_{\forall s_1,..,s_r} p_{s_1}^{X_1} .. p_{s_r}^{X_r} \delta(y - f(w_{s_1}^{X_1}, .., w_{s_r}^{X_r})) \tag{7}$$

The domain of the $\mathcal{F}$ operator is at least one *spdf* and the range is a single *spdf*. Hence this operator presents a many-to-one relationship. The $\mathcal{F}$ operator essentially provides another sampled *pdf*, where the multiplication term $p_{s_1}..p_{s_m}$ denotes the probability at the point $f(w_{s_1}^{X_1}, .., w_{s_r}^{X_r})$. Here, s_i corresponds to the set of all samples belonging to the random variable X_i. Since f is any function, the samples may no longer be situated at fixed distances. Elimination of the points that are practically infeasible can be attained through a subsequent band-pass filtering. This step should be followed by another binning process on the sampled probability distribution. Samples that fall in any particular bin will be approximated by a single impulse at the center of the corresponding bin, with the height of this impulse being the sum of the impulses that fall in this bin. The re-binning operation precludes the number of samples from reaching computationally intractable numbers for each new random variable. Notably, each new random variable is a function of random variables all of which have had their sampled probability distributions computed. The band-pass and re-bin operators are applied once for each new random variable that is acquired through the forward operator. Also, binning of impulses makes interpolation of the impulses possible; without the binning process, neighboring impulses with largely differing sample probabilities would cause a great amount of noise. The band-pass filtering, on the other hand, restrains samples of impractically low occurrence probability, which would otherwise have caused Δ to increase for the same number of bins, thus resulting in an inaccurate binning of the values of a random variable. The band-pass and re-bin operators are defined respectively as:

$$\phi'(X) = \mathcal{B}_e(\phi(X)) \tag{8}$$

$$\phi''(X) = \mathcal{R}_{\mathcal{N}}(\phi(\mathcal{X})) \tag{9}$$

The domain and range of the band-pass function can either be a *pdf* or an *spdf*. The domain and range of the re-bin function is an *spdf*. The subscript e in the $\mathcal{B}_e$ operator denotes the *error-rate*, a parameter defined to eliminate least likely samples to ensure computational time efficiency. The band-pass operation can also be defined as $\mathcal{B}_{[m,n]}$, directly providing lower and upper limits [m,n]. A band-pass operation with margins over an *spdf* equals:

$$\phi'(X) = \sum_{i:(w_i \epsilon [m,n]) \wedge (w_i \epsilon (\phi(X)))} p_i \delta(x - w_i) \tag{10}$$

A band-pass operation with *error-rate* over an *spdf* equals:

$$\phi'(X) = \sum_{i:(p_i > \frac{max_i(p_i)}{e}) \wedge (p_i \epsilon (\phi(X)))} p_i \delta(x - w_i) \tag{11}$$

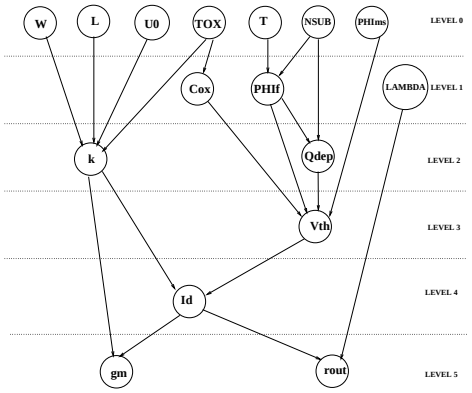

Fig. 1. The connectivity graph of a device.

The $\mathcal{R}_{\mathcal{N}}$ is defined as:

$$\phi''(X) = \sum_i p_i \delta(x - w_i) \tag{12}$$

where $p_i = \sum_{s_j} p_j$ such that $w_j \epsilon b_i$. N in the $\mathcal{R}_{\mathcal{N}}$ operator corresponds to the new number of bins, as the re-bin operator can bin the samples into a different number of bins than the number of bins present in the spdf which this operation is applied to.

V. Experimental Results

In order to verify the proposed techniques, the connectivity graph shown in Figure 1 is used. This connectivity graph is structured according to the relationships of formulas that model a transistor. A connectivity graph, originally presented in [10] to model mismatch between transistors, essentially defines a relationship between parameters according to underlying formulas. Nodes that have incoming edges are functions of nodes from which these edges originate.[2] The values used for the parameters are taken from a 0.13μm TSMC process.

To verify the forward probability propagation method, Gaussian distributions with 10% standard deviation have been assigned to the lowest level physical parameters.[3]

The validity of these sampled probability distributions has been confirmed through a Monte Carlo sampling approach. Such a comparison is given in Figure 2 for the current. Depending on the band-pass filtering and re-binning processes and their related parameters, such as error rate for the former and M and N for the latter, a variation for FDPP results from the real distribution may be observable. Since the calculation of real distributions is not feasible or very complex in most cases, comparisons have been made with the Monte Carlo method. It is observable that the results are quite close. The minor deviation can be attributed to the fact that in a connectivity graph, whenever a node has more than one outgoing edge, the Monte Carlo based method assigns the same sample values to all of these nodes. As a result of this correlation error, Monte Carlo methods result in an estimation error for the distribution of final parameters. This can be important depending on the technology and its accompanying model.

[2]The formulas used in this connectivity graph have been chosen to be well known design functions instead of the more complex BSIM 3v3 formulas. This choice has been made so as to obtain a better intuition of the proposed methodologies, as qualitative relationships between these formulas are known to the designers.

[3]Notice that the lowest level parameters are assumed Gaussian in both FDPP and Monte Carlo in this paper for a fair comparison.

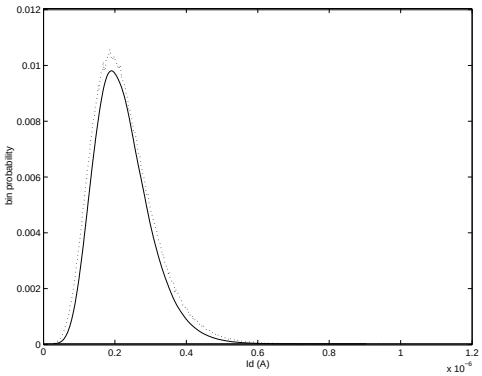

Fig. 2. Comparison of *pdf* of I_D: FDPP (solid line); Monte Carlo method (dotted line)

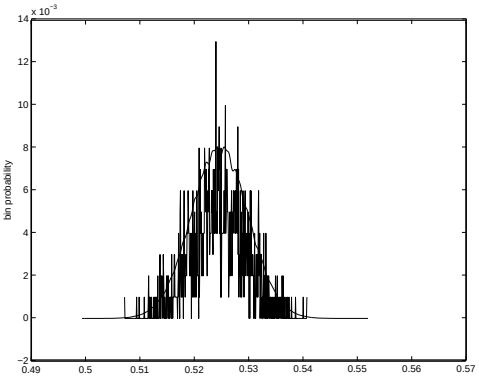

Fig. 3. Comparison of FDPP (solid line) and Monte Carlo (noisy line)

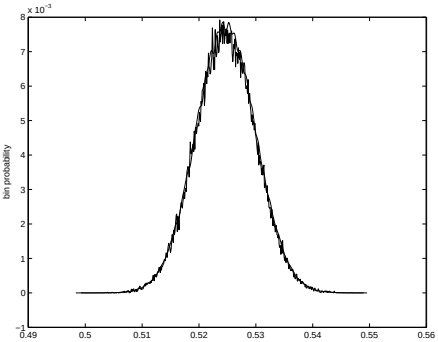

Fig. 4. Comparison of FDPP (solid line) and Monte Carlo (noisy line)

When only a moderate number of samples are to be used, Monte Carlo will generate highly inaccurate results. The relevant experimental results are presented in Figures 3-4. In both figures, FDPP has been applied with 100 samples each for T and N_{sub} to calculate the *spdf* of ϕ_F. On the other hand, 1000 samples of each have been used for the result shown in Figure 3 and 100000 samples of each have been used for the result in Figure 4 for the Monte Carlo Method. It can be observed that in the former case, even though more samples are employed, the result is highly inaccurate for the Monte Carlo method. In the latter case, a large number of samples produce a better but still much noisier result. Due to its deterministic sampling, the FDPP method provides an acceptable approximation with even a small number of samples. This improvement can be of use in either of two ways. First, a manual calculation can be used for estimation with FDPP due to it necessitating a number of orders of magnitude fewer samples in comparison to a Monte Carlo method. Secondly, when an iteration is needed that necessitates the fast estimation of device parameters, FDPP would result in more accurate results than Monte Carlo if a low number of samples are to be used.

It can also be easily observed that device parameters such as current deviate from a Gaussian distribution. For example, both methods indicate a visual deviation from Gaussian distribution in Figure 2.

VI. Conclusions

A technique called *forward discrete probability propagation* has been presented. Distributions of device parameters as a result of process variations can be accurately estimated through this method. Mathematical foundations for probability discretization have been presented along with experimental comparisons with Monte Carlo methods. FDPP can be considered to be an efficient alternative to the Monte Carlo approach. The methodology we propose is preferable to Monte Carlo methods when an algebraic intuition is needed, the importance of low probability samples is accentuated due to a limited number of samples, one-to-many relationships in the formula hierarchy are present or arbitrary distributions are needed for low level process parameters. It can be envisioned that the proposed method will be prominent in deep sub-micron and sub-wavelength technologies where effects of process variations need to be integrated into the design cycle in a fast and accurate manner.

References

[1] J. S. Huang and A. S. Oates, "Monte-Carlo Simulation of Electromigration Failure Distributions of Submicron Contacts and Vias: A New Extrapolation Methodology for Reliability Estimate," *Interconnect Technology Conference*, pp.208-210, June 2000

[2] R. Burch, F. N. Najm, P. Yang and T. N. Trick, "A Monte Carlo Approach for Power Estimation," *IEEE Transactions on VLSI Systems*, Vol.1, Iss.1, pp.63-71, Mar.1993

[3] P. Palestri, A. Pacelli, M. Mastrapasqua and J. D. Bude, "Monte Carlo Simulation of Impact Ionization in SiGe HBTs," *IEEE Electron Device Letters*, Vol.22, Iss.11, pp.533-535, Nov.2001

[4] B. Winstead and U. Ravaioli, "Simulation of Schottky Barrier MOSFETs with a Coupled Quantum Injection/Monte Carlo Technique," *IEEE Tran. on Electron Devices*, Vol.47, Iss.6, pp.1241-1246, June 2000

[5] F. M. Bufer, Y. Asahi, H. Yoshimura, C. Zechner, A. Schenk and W. Fichtner, "Monte Carlo Simulation and Measurement of Nanoscale n-MOSFets," *IEEE Tran. on Electron Devices*, Vol.50, Iss.2, pp.418-424, Feb.2003

[6] C. Jungemann and B. Meinerzhagen, "Efficiency and Stochastic Error of Monte Carlo Device Simulations," *IEDM*, pp.109-112, Dec.1999

[7] N. Sano, "Sub-0.1μm Device Simulation Technology: Another Problem for Monte Carlo Simulations," *IEDM*, pp.167-170, Dec.2000

[8] P. G. Drennan and C. C. McAndrew, "A Comprehensive MOSFET Mismatch Model," *Int. Conf. on Simulation of Semiconductor Processes and Devices*, pp.23-26, Sept.1999

[9] N. Telang and J. M. Higman, "Statistical Modeling Techniques: FPV vs. BPV," *In Proceedings of IEEE Microelectronic Test Structures*, Vol.14, pp.71-75, Mar. 2001

[10] R. O. Topaloglu and A. Orailoglu, "On Mismatch in the Deep Sub-micron Era: From Physics to Circuits," *ASP-DAC*, pp.62-67, Jan.2004

[11] K. Aoyama, H. Kunitomo and K. Tsuneno, "Rigorous Statistical Process Variation Analysis for Quarter-μm CMOS with Advanced TCAD Metrology *Int. Workshop on Statistical Metrology*, pp.8-11, June 1997

[12] H. Sato, H. Kunitomo, K. Tsuneno, K. Mori and H. Masuda, "Accurate Statistical Process Variation Analysis for 0.25-μm CMOS with Advanced TCAD Methodology," *IEEE Transactions on Semiconductor Manufacturing*, Vol.11, Iss.4, pp.575-582, Nov.1998

[13] A. Nardi, A. Neviani, E. Zanoni and C. Guardiani, "Impact of Unrealistic Worst Case Modeling on the Performance of VLSI Circuits in Deep Submicron CMOS Technologies" *IEEE Transactions on Semiconductor Manufacturing*, Vol.12, Iss.4, pp.396-402, Nov.1999

[14] B. E. Stine, D. S. Boning, J. E. Chung, D. Ciplickas and J. K. Kibarian, "Simulating the Impact of Poly-CD Wafer-Level and Die-Level Variation on Circuit Performance" *IEEE Transactions on Semiconductor Manufacturing*, Vol.11, Iss.4, pp.552-556, Nov.1998

[15] S-D. Kim, H. Wada and J. C. S. Woo, "TCAD-Based Statistical Analysis and Modeling of Gate Line-Edge Roughness Effect on Nanoscale MOS Transistor Performance and Scaling," *IEEE Transactions on Semiconductor Manufacturing*, Vol.17, No.2, May 2004

[16] T. Mizuno, "Novel Statistical Fluctuation of Dopant Concentration and Its Influence on Scaled MOS Device Performance,' *Statistical Metrology*, pp.16-19, June 1997

[17] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann Publishers, Inc., 1988

[18] D. J. C. Mackay, *Introduction to Monte Carlo methods*, Kluwer Academic Press, 1997