Forward Discrete Probability Propagation Method as a Deterministic and Non-parametric Means for Device Performance Characterization

Connectivity graphs are powerful tools for a number of operations on circuits and devices. With connectivity graphs, prediction of probability distributions of device parameters is also possible. Such estimates are useful in analyzing the effects of process variations on device performance. Such information will be necessary in deep sub-micron and sub-wavelength design regimes to guide designers and simulators in taking these effects into consideration. This paper presents how such a powerful estimation technique is realized through the use of connectivity graphs and a non-parametric probabilistic method. The method can be considered as an alternative to Monte Carlo and parametric methods for accurate device performance estimation.

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 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 push manufacturers to more accurately characterize device parameters rapidly to cope with 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. As opposed to the Monte Carlo approach, there does not exist any non-determinism in the system. In contrast to Gaussian-based methods, the system is non-parametric, 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 throughout the paper. Without employing a random method, it can 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 an introduction of a mathematical basis for discretization of probability distribution functions, bringing formalism by introducing new operators and domains. With this mathematical background, practical TCAD algorithms for implementing forward discrete probability propagation, FDPP, using these operators are presented, followed by an error analysis and experimental data comparing Monte Carlo methods and FDPP.

II. MOTIVATION

Monte Carlo methods are frequently used in engineering applications [1] [2], though they lack two major constituents. A foremost shortcoming consists of the non-determinism of Monte Carlo resulting in the absence of any algebraic framework. The randomness stems from the fact that Monte Carlo employs sampling using a random number generator. At consequent runs, the same random seed is usually employed. A non-ideal random number generator will inject a bias into the system, causing the simulations to perturb from the expected nominal. Additionally, most computational packages only provide random number generators for a couple of well-known distributions such as Gaussian or uniform. As a matter of fact, the users of Monte Carlo methods are limited in assigning distributions to low level parameters. An approach for 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 computation of the distributions of high level parameters.

The other shortcoming is that Monte Carlo, due to its random sampling mechanism, may miss points that are less likely to occur. This last bottleneck may cause certain regions in a distribution to be missed by the method altogether, or similarly it may lead to an overly optimistic result by not being able to sample the end-tails of distributions, missing the worst-case points. Increasing the number of samples to prevent the latter bottleneck does not guarantee a fix to this problem, while it leads at the same time to unmanageable run-time complexity. As Monte Carlo is a random sampling method, increasing the number of samples may lead to sampling the same points more than once, as opposed to discovering the least likely points.
III. PREVIOUS WORK

Monte Carlo based methods are predominantly used in device parameter characterization [3] [4]. In the former one, a Monte Carlo based method has been used for simulation of impact ionization while in the latter one, 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 identified 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 is not highly accurate in deep sub-micron and sub-wavelength technologies due to the Gaussian distribution assumption attributed to device parameters, as device parameters are deviating from Gaussian distributions with newer technologies. In [10], [11] and [12], it can be observed that device parameters deviate from a Gaussian approximation. 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. The necessity for accurately employing powerful models so far presented in the literature is unavoidably present and becomes predominant as we progress towards newer technologies.

IV. PROBABILITY DISCRETIZATION THEORY

Accurate simulation of devices has exceeded practical time limits. The computational cost of simulating process variations introduces exponential increase to this time. The necessity to 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 estimation of device parameters. This methodology is both manually tractable and can be incorporated into a simulator.

In order to introduce our methodology, a number of definitions are necessary; we introduce them herein. 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:

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

This difference should be chosen as small as possible to reduce the filtering error. The sampling is accomplished by binning the band-pass filtered distribution. The band-pass filtering and binning operations are illustrated in Figures 1 and 2, respectively.

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 midpoint 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 $\Delta$ is the step-size defined by $\frac{n - m}{N}$. The $N$’th bin, however, is bounded by $[m + (N - 1)\Delta, m + N\Delta]$, which is equal to $[n - \Delta, n]$. We denote the sampled $pdf(X)$ as $\phi(X)$ or equivalently $spdf(X)$, 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 $Q^N_{\phi}$ operator:

$$\phi(X) = Q^N_{\phi}(pdf(X))$$ \hspace{1cm} (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:

$$\phi(X) = \sum_{i=1}^{N} p_i \delta(x - u_i)$$ \hspace{1cm} (3)

where

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

$$u_i = m + (i - 1)\frac{\Delta}{2}$$ \hspace{1cm} (5)

In these equations, $p_i$ corresponds to the probability that a sample of the random variable falls within the $i$’th bin $b_i$ and $v_i$ denotes the approximation of values of samples of $X$ within $b_i$, $u_i$ is the mid-point in the particular bin.

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 $F$ operator as:

$$\phi(Y) = F(\phi(X_1),...,\phi(X_r))$$ \hspace{1cm} (6)

which is defined as being equivalent to:

$$\phi(Y) = \sum_{x_{s_1},..,x_{s_r}} p_{x_{s_1}}...p_{x_{s_r}} \delta(y - f(x_{s_1},..,x_{s_r}))$$ \hspace{1cm} (7)

The domain of the $F$ operator is at least one $spdf$ and the range is a single $spdf$. Hence this operator presents a many-to-one relationship. The $F$ operator essentially provides another sampled $pdf$, where the multiplication term $p_{x_{s_1}}...p_{x_{s_r}}$ denotes the probability at the point $f(x_{s_1},..,x_{s_r})$. Here, the $p_{x_{s_j}}$ notation

---

$^1$This nomenclature has been chosen noting the similarity of these domains to the $s$-domain (Laplace domain), as some operations such as filtering can be depicted and formulated easily in the $p$-domain and $r$-domain just like some operations are easily applied in the $s$-domain.
impulses in a small neighborhood with largely differing sample filtering, on the other hand, avoids less probable samples which probabilities would cause a great amount of noise. Band-pass filtering will eliminate points that are practically impossible. Without the binning process, interpolations of the impulses possible. Without the binning process, distributions computed. Also, binning of impulses makes inter-
pertaining to connectivity graphs. It can be recalled that connectivity graphs are structured according to functional inference relationships. When the distributions of the lowest level parameters are given, the algorithm can then find the next lowest level parameters all of whose ancestors have their sampled probability distributions computed. The repetition of this procedure leads to the computation of the highest level parameters. This procedure is defined as the forward discrete probability propagation, FDPP. In other words, FDPP is the calculation of \( spdf \) of a random variable through the application of the algorithm implementing the \( \mathcal{F} \) operator.

Algorithm Implementing the \( \mathcal{F} \) Operator:

1. While each random variable has its \( spdf \) computed:
2. For each random variable which has all ancestor \( spdf \)'s computed:
3. For each sample in \( X_1 \)
   .
   .
4. For each sample in \( X_r \)
5. Place an impulse with height \( p_1 \ldots p_r \) at \( x=f(v_1, \ldots, v_r) \)
6. Apply the \( B \) and \( R \) algorithms to this random variable

For practical purposes, filtering and re-binning operators can be combined under a single algorithm:

Algorithm Implementing the \( B \) and \( R \) Operators:

1. Find maximum and minimum values \( w_i \) within impulses
2. Divide this range into \( M \) bins
3. For each bin
4. Place a quantizing impulse at the center of the bin with a height \( p_i \) equal to the sum of all impulses within the bin
5. Find maximum probability, \( p_{i=\text{max}} \), of quantized impulses within all bins
6. Eliminate impulses within bins which have a quantized impulse with smaller probability than error-rate* \( p_{i=\text{max}} \)
7. Find new maximum and minimum values \( v_i \) within impulses
8. Divide this range into \( N \) bins
9. For each bin
10. Place an impulse at the center of the bin with a height equal to the sum of all impulses within the bin

This algorithm employs the error-rate based filtering, as this type of filtering does not need a detailed inspection of the \( pdf \). Notice that two different total bin numbers, \( M \) and \( N \), have been used. The choice of \( M \) constitutes an initial binning, providing a discrete approximation. This enables a finer resolution for the bin widths, which provides the elimination of a selective section of the \( pdf \). As a result, these values should be chosen such that \( N \leq M \). In line 5, the bin with the highest probability is determined. This is used with the error-rate to eliminate low probability bins. Hence possible peaks and lobes of the \( pdf \) are retained while impractically low probability samples are discarded.\(^2\) As \( N \) will be chosen smaller for computational efficiency, the pruning of low probability bins precludes the possibility of the whole \( pdf \) coalescing into a single impulse.

\(^2\)As opposed to Monte Carlo, elimination of these low probability samples is a choice given to the user, enabling exploitation of this benefit in particular applications where low probabilities might still be considered important.
VI. Error Analysis

If the quantizer is uniform, the samples \( q \) of the quantization error random variable \( Q \) will be bounded by \(-\Delta < q \leq \Delta\). If \( \Delta \) is sufficiently small, then the quantization error is a uniformly distributed random variable. The probability distribution of the quantization error random variable \( Q \) is given by \( \frac{1}{\Delta} \) when \(-\Delta < q \leq \Delta\) and 0 otherwise. Approximating the mean of the quantization error as 0, the variance can be written as:

\[
\sigma_Q^2 = E[Q^2] = \int_{-\Delta}^{\Delta} q^2 \, pdf(Q) \, dq
\]  

which can be evaluated to be equal to \( \frac{\Delta^2}{12} \).

In the re-binning process, if we wish to evaluate the distortion caused by the representation of all impulses in a bin by a single impulse located at the center of the bin \( i, m_i \), we can use a distortion cost function, \( d \), such as the mean squared error for the \( j \)’th impulse as given below:

\[
d(m_i, w_j) = (m_i - w_j)^2
\]  

Consequently, total distortion caused by the re-binning operator can be expressed as:

\[
\sum_{i,j : i \neq i, j \neq j} d(m_i, w_j)p_j
\]

This sum assigns the distortion parameter a weight through the \( p_j \) terms, implying that a long impulse approximated by an inaccurate impulse will cause increased distortion. The distortion equation can also be used to assume an adaptive sampling scheme, where samples are chosen so as to reduce the total distortion.

VII. Experimental Results

In order to verify the proposed techniques, the connectivity graph in Figure 6 is used. This connectivity graph belongs to 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. 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. The values used for the parameters are taken from a 0.13-μm TSMC process. The formulas are repeated here for convenience:

\[
C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}}
\]  
\[
k = \mu_0 t_{ox} \frac{W}{L}
\]  
\[
\phi_F = \frac{kT}{q} \ln \left( \frac{N_{sub}}{n_i} \right)
\]  
\[
Q_{dep} = \sqrt{4q\varepsilon_{si}}|\phi_F| N_{sub}
\]  
\[
V_{th} = \phi_{m,s} + 2\phi_F + \frac{Q_{dep}}{C_{ox}}
\]  
\[
I_D = \frac{1}{2}k(V_{GS} - V_{th})^2
\]  
\[
g_m = \sqrt{2kI_D}
\]  
\[
R_o = \frac{1}{\lambda I_D}
\]

To verify the forward probability propagation method, Gaussian distributions with 10% standard deviation have been assigned to the lowest level physical parameters.\(^3\) Figures 7 and 8 show the attained sampled probability distributions for the threshold voltage and transconductance, respectively. Transconductance has been included to show the possibility of using this approach for analog device characterization.

The validity of these sampled probability distributions has been confirmed through a Monte Carlo sampling approach in Figures 9-10 for threshold voltage and current, respectively. 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 node, the Monte Carlo based method assigns the same sample values to both 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. A probabilistic approach on a connectivity graph as in FDPP, on the other hand, takes advantage of the Bayes rule \([17]\). This means that a number of nodes can be treated as being independent given that a node having incoming edges from these nodes is separated by them from other nodes in a directed acyclic graph.

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 11-12. In both figures, FDPP has been applied with 100 samples each for \(T\) and \(N_{sub}\) to calculate the \(spdf\) of \(\phi_F\). On the other hand, 1000 samples of each have been used for the former figure and 100000 samples of each have been used for the latter figure for 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 can be of use in two ways. First, a manual calculation can be used for estimation with FDPP. 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 is also easily observed that device parameters such as current or transconductance deviate from a Gaussian distribution. For example, both methods indicate a visual deviation from Gaussian distribution in Figure 9. As a matter of fact, Gaussian approximation based parametric techniques would fail in accuracy as compared to both our method and Monte Carlo based methods.

**VIII. Summary and Conclusion**

A technique called *forward discrete probability propagation* has been presented. Distributions of device parameters can be accurately estimated through this method. Mathematical foundations for probability discretization, efficient CAD implementation algorithms and an error analysis have been presented along with experimental comparisons with Monte Carlo methods. FDPP can be considered to be an alternative to the Monte Carlo approach. The methodology we propose is preferable to Monte Carlo methods when an algebraic intuition is needed, low probability samples are important, one-to-many relationships in the connectivity graphs are present or arbitrary distributions are needed for low level parameters. This method may become prominent in deep sub-micron and sub-wavelength technologies where effects of process variations need to be integrated into the design cycle in a computationally efficient and accurate manner.

\(^3\)Notice that Gaussian approximation based methods signify methods predicting high level device parameters as Gaussian; the lowest level parameters are assumed Gaussian in all methods in this paper for a fair comparison. FDPP, however, can process arbitrary distributions assigned to low level parameters.
Fig. 7. pdf of $V_{th}$ determined through FDPP

Fig. 8. pdf of $g_m$ determined through FDPP

Fig. 9. Comparison of pdf of $V_{th}$: FDPP (solid line); Monte Carlo method (dotted line)

Fig. 10. Comparison of pdf of $I_D$: FDPP (solid line); Monte Carlo method (dotted line)

Fig. 11. Comparison of FDPP (solid line) and Monte Carlo (noisy line)

Fig. 12. Comparison of FDPP (solid line) and Monte Carlo (noisy line)

REFERENCES


