A STATISTICAL PERFORMANCE SIMULATION METHODOLOGY FOR VLSI CIRCUITS

Michael Orshansky, James C. Chen, and Chenming Hu
Department of Electrical Engineering and Computer Science,
University of California at Berkeley, Berkeley, CA, 94720

ABSTRACT

A statistical performance simulation (SPS) methodology for VLSI circuits is presented. Traditional methods of worst-case corner analysis lack accuracy and Monte-Carlo simulations cannot be applied to VLSI circuits because of their complexity. SPS methodology is accurate because no statistical information about the device parameter variation is lost. It achieves efficiency by analyzing the smaller circuit blocks and generating the performance distribution for the entire circuit. Circuit evaluation at any specified performance level is possible.

1 INTRODUCTION

Deep sub-micron technologies have made the problem of statistical modeling of the device and circuit behavior more critical. Shrinking dimensions make device characteristics more sensitive to stochastic process variation. Thus the relative spread of device and circuit behaviors is broadened. Combined with the continuing reduction of the design cycle, this creates an urgent need for methodologies that can accurately model and predict the statistical variations of circuit performances, such as speed and power consumption.

The industry's de facto standard in the field of statistical modeling is still limited to worst- and best-case analysis. The worst- and best-case SPICE models are produced by independently combining the outmost device parameter values. A circuit designer then uses this set of models to verify the design at the extremes of the process variations. What would be much more valuable, however, is the ability to analyze the circuit at any performance level, e.g. 25%, 75%, and 95%, and not only at the ill specified "best" and "worst" extremes. Moreover, this approach completely neglects parameter inter-correlations, and produces circuit performance predictions that are overly pessimistic (or optimistic).

A number of approaches have been taken toward solving this problem. The most straightforward of them are methods that identify the important device parameters and perform Monte-Carlo simulations for a given circuit [1]. However, the complexity of VLSI circuits makes Monte-Carlo based methods prohibitive. In addition to being very time-intensive, Monte-Carlo methods either assume that the device parameters are independent, or necessitate building elaborate models to account for the complex correlation of the parameters. Alternatively, the correlated device parameters can be expressed in terms of independent factors, or components, using statistical techniques such as principal component analysis [2]. While improving the simple Monte Carlo based methods, this approach is susceptible to the problems of statistical transformation techniques that are discussed below.

To overcome these limitations, a methodology based upon expressing the circuit characteristics in terms of the device model parameters has been proposed [3]. This approach reduces the number of device parameters under consideration using principal component analysis (PCA), and then employs response surface methodology to generate the statistical circuit performance characteristics. However, the use of PCA on device parameters from the deep sub-micron technologies is undesirable because it may fail to capture the complex correlations of the device parameters [5].

This is due to the fact that for the deep sub-micron CMOS technologies a combination of device physics, die location-dependence, optical proximity effect, microloading in etching and deposition, etc., may lead to heterogeneous and non-monotonic relationships among the device parameters that cannot be captured by the PCA-based device parameter characterization. For example, the same industrial technology can produce two opposite correlations between the threshold voltage and the channel length (Figure 1). One trend is purely stochastic: devices with shorter effective channel lengths have slightly higher Vt following the reverse short-channel effect dependence. The other trend is deterministic, i.e. die-location dependent: probably, due to variation of the lateral doping profile near wafer edge. PCA is a statistical technique that transforms correlated variables into uncorrelated variables, which graphically amounts to a rotation of the original axis. In the presence of variation arising from two or more mechanisms, a phantom averaging effect occurs leading to underestimation of the real level of parameter fluctuations. This can be clearly seen from Figure 2, where the larger number of stochastic data points 'outweighs' the systematic variation.

The proposed SPS methodology attempts to find a solution that avoids the drawbacks and problems mentioned above. It allows generation of statistical information at any specified

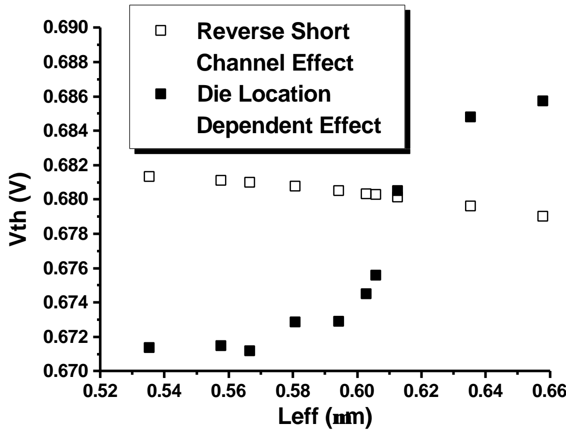


Figure 1. Different trends of Vth vs Leff dependence are found in deep sub-micron technologies.

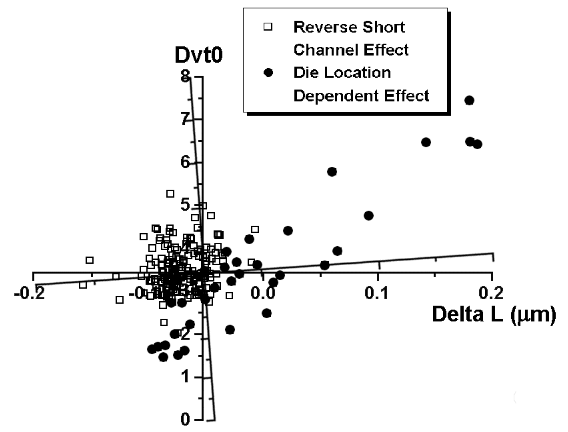


Figure 2. PCA-based methods lead to erroneous results when data coming from different sources of variation is present.

performance level. It does not employ any statistical techniques that may lead to the inadvertent loss of parameter correlations arising from the complex physical processing of the deep sub-micron technologies. This is accomplished with a direct-sampling method [5]. In order to avoid a large number of simulations of the whole circuit, and thus overcome the deficiency of Monte-Carlo based methods, it constructs a circuit model in terms of performance variables of the simpler blocks constituting the circuit. The computational effort is greatly reduced because circuits normally consist of only a limited number of distinctly behaving 'primitives', and only one representative of every such primitive needs to be analyzed. Also, because simulation time increases more than linearly with the number of nodes, this approach results in additional simulation time saving. A concise comparison of advantages and disadvantages of some of the reported methods and SPS methodology is summarized in Table 1.

2 METHODOLOGY

The proposed SPS methodology makes no assumptions about the degree of device parameter independence. This is achieved by using the direct sampling method (DSM). The main feature of DSM is the insistence that all sampled device characteristics from a test die be kept as a set, and be directly applied to all further analysis. No statistical operation is performed that can inadvertently lead to the effect of averaging out the device fluctuations arising from different mechanisms, which can happen if principal component analysis or factorial analysis are used [5]. In DSM a statistically significant number of dies is sampled and kept as a set that typically contains the device I-V parameters, and gate, overlap, junction, and interconnect capacitances. Because all the parameters from the same location are kept together, neither additional information nor analysis is required concerning their mutual correlation. The larger is N, the finer is the resolution of the performance distribution. Therefore, a large number of SPICE device parameters need to be efficiently generated. One possible approach is to extract SPICE parameters directly from Electrical-Test data, using an equation solver technique [6].

The flow chart of the sequence of steps in SPS methodology is given in Figure 3. The basic premise of SPS is the idea that a large circuit can be analyzed in terms of simpler blocks that

Method	Advantages	Disadvantages
Traditional Worst-Case Model File Generation	- Straightforward - Computationally efficient	- Not accurate - No statistical yield information
Monte-Carlo [1]	- Provides statistical performance information. - Accurate	- Very computationally expensive. Impractical for large circuits
RSM Model Construction In Terms Of Device Parameters [3]	- Provides statistical yield information - Computationally efficient	- Not suitable for complex correlation of deep-submicron device parameters
Optimization Approaches [4]	- Reasonably efficient - Evaluates circuit performance	- Provides no statistical yield information
SPS – Model Construction In Terms Of Circuit Blocks (Current Paper)	- Provides statistical yield information - Computationally more efficient than Monte-Carlo	- Must isolate the "building" blocks of the larger circuit - Approximate substitute for Monte-Carlo

Table 1. A comparison of different approaches to statistical circuit modeling.

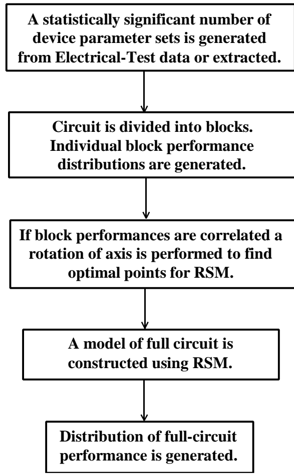


Figure 3. Steps involved in SPS methodology.

constitute the circuit (Figure 4). The possibility of such structural subdivision is important in two respects.

Digital circuits are usually well structured and can be broken down into the smaller building blocks such as inverters, NAND and NOR gates, clock generators, latches, etc… Thus, a large circuit normally is a combination of a limited number of statistically distinct building blocks. Statistically distinct blocks are those that produce different performance distributions for the identical set of input SPICE files. Two statistically distinct blocks, for example, would produce a performance (e.g. delay) value lying at a different level of statistical distribution, i.e. the same SPICE model file would lead to a simulation result at the 95th percentile and the 60th percentile of the performance distribution. It is reasonable to assume, as a first-order approximation, that the statistically identical blocks remain behaving similarly even when placed into the network of connections within the larger circuit. This assumption allows a great reduction in analysis effort, because now only a representative block needs to be analyzed, leaving aside a large number of its replications. Clearly, this advantage can be utilized only when a hierarchical block approach similar to SPS methodology is used.

In addition, because the simulation time of the SPICE-based

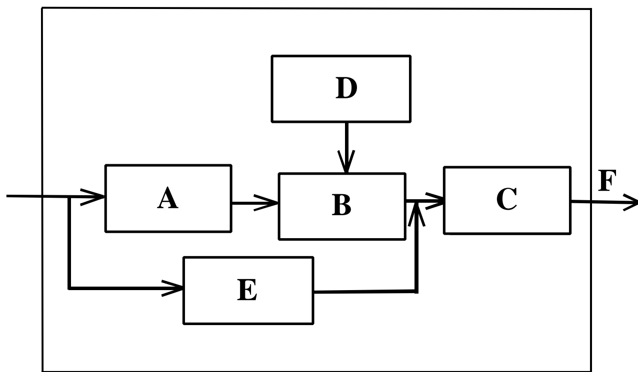


Figure 4. Large circuit is analyzed in terms of its building blocks.

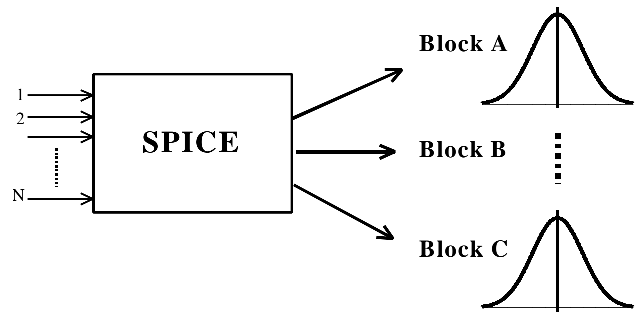


Figure 5. Generation of block performance distributions. N is the number of SPICE model files.

circuit simulators is on the order of O(M$^{1.2}$) with M being the number of nodes, it is advantageous to break up a single set of nodes into a number of smaller sub-sets [7]. If $T(M)$ is the simulation time of M nodes, then:

$$T(M) > T_1(M_1)+T_2(M_2)+\ldots+T_n(M_n),$$

where $M_i<M$, and n is the number of blocks. With the increase of M, the advantage of breaking up the large circuit into blocks becomes greater.

The model is to be constructed using response surface methodology (RSM). Because the number of simulations of the whole circuit required to generate an accurate model using RSM is on the order of 2^n, it is beneficial to minimize the number of distinct blocks n. For example, several simple digital blocks, such as inverters, NAND and NOR gates, have been shown to be highly correlated with respect to speed [8]. (Even though they can behave differently with respect to another circuit performance characteristics, such as noise margins [9].) The blocks that are a priori known to be statistically similar n be excluded from the analysis; thus, reducing the required number of full circuit simulations.

Once the simpler blocks are identified, they are evaluated with respect to a particular circuit performance variable, e.g. speed, or power dissipation. Each block is simulated N times using SPICE, where N is a statistically significant number of samples from DSM. The distribution of random variables corresponding to each block is thus generated (Figure 5).

We next wish to find the distribution of the performance variable for the whole circuit based upon the generated block performance distributions. A model relating block performances A,B,C… and the whole circuit performance F is needed: F=f(A,B,C…). The analytic form of a function f (A,B,C…) is not available and has to be estimated by using response surface method. RSM involves running a limited number of full circuit simulations, and then fitting a function relating full circuit performance to block performances. The coefficients of the model are calculated using linear regression.

At this stage, the objective is to minimize the number of full circuit simulations without sacrificing the model's accuracy. The training points should be chosen to cover as much area as possible in the space of block performances, and a simple way to do it is to "sample" the space at the extremes of the distribution. If block distributions are correlated the choice of the optimal training point is not a trivial task. Therefore, the algorithm used to select such points for RSM requires that the block performances be uncorrelated. To address this issue,

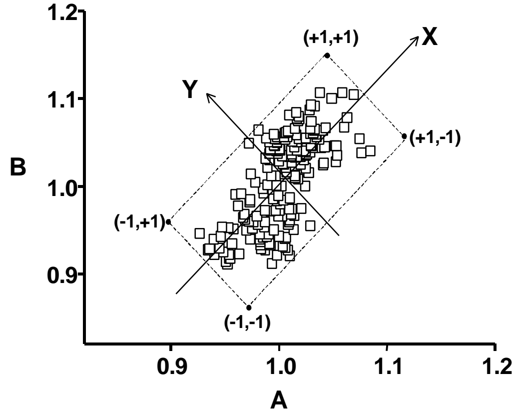


Figure 6. Rotation of axis may be necessary to determine the optimal training points for RSM.

after the distributions of individual blocks are generated, an analysis is performed to see whether the blocks are correlated. If some of them are significantly correlated a rotation of axis is performed in such a way that the resultant variables are mutually independent (Figure 6). The training points are easily identified as corresponding to the extremes of the distribution relative to the new rotated axis.

Design of experiment theory helps to determine the optimal set of training points for the simulations in such a way that the best model is generated [10,11]. The most straightforward way is to perform the simulations with all possible combinations of the extremes of block performances. This approach is called full factorial analysis and requires a full set of 2^n simulations. It is possible to reduce this number. Fractional factorial designs require 2^{n-k} simulations (n- number of blocks, k- degree of fraction). We consider only two level factors, and denote the maximum and minimum values of each factor (block performance variable) as +1 or −1. Then, Table 2 lists the set of combinations of variables required by full factorial design and indicates which combinations should be used in half-factorial design (n=3, k=1).

It is important to see the limitations of fractional factorial design. Reducing the number of full-circuit simulations results in the loss of some accuracy. The main effects (coefficients of A, B, and C) and two-factor interaction effects (coefficients of $A\times B$, $B\times C$, and $A\times C$) can still be accurately represented. Only the three-factor interaction term is lost but since in most cases such a high-order interaction is negligible this loss is permissible [11].

The selection of optimal training points is achieved by considering the uncorrelated variables after the axis are rotated. To find the actual SPICE model file corresponding to a selected point no backward transformation is needed, however. Even though the coordinates were rotated, the correspondence between the points and the actual model file was kept. The rotation of axis was needed only to identify which of N models corresponds to the optimal training points for RSM. Figure 6 helps to illustrate this idea.

The SPICE model files corresponding to the chosen training points are used to simulate the whole circuit, resulting in 2^{n-k} values of full-circuit delay. Using this set of data points together with the already simulated 2^{n-k} values for each of the building blocks, the coefficients of the model can be generated using linear regression. Once the model is generated, the distribution of the full-circuit performance can be easily calculated by the simple substitutions of N sets of block performance values of A,B,C… into the model. From the cumulative probability plot of full circuit distribution, the performance analysis can be straightforwardly performed. Because to each point on this plot there is a corresponding SPICE model, one can very easily use it to evaluate other performance characteristics of the whole circuit (such as power dissipation, noise margin, etc.) at a given level of speed distribution. In the cases of linear or quadratic models, the expected value and variance of the full-circuit distribution can also be directly calculated from the knowledge of the mean values and variance-covariance matrix of the block performances.

Simulation	A	B	C	A*B	B*C	A*C	A*B*C	Half-fraction
1	-1	-1	-1	+1	+1	+1	-1	
2	-1	-1	+1	+1	-1	-1	+1	*
3	-1	+1	-1	-1	+1	-1	+1	*
4	-1	+1	+1	-1	-1	+1	-1	
5	+1	-1	-1	-1	-1	+1	+1	*
6	+1	-1	+1	-1	+1	-1	-1	
7	+1	+1	-1	+1	-1	-1	-1	
8	+1	+1	+1	+1	+1	+1	+1	*

Table 2. Choice of training points for the half-factorial design.

3 RESULTS AND DISCUSSION

To verify the proposed SPS methodology, a case study of a low-swing bus driver for low-power applications was performed [12]. This specific circuit was chosen because it combines blocks of digital as well as analog circuitry (Figure 7). We consider a part of the circuit dividing it into 3 blocks: the sense-amplifier (Sense), the latch (Latch), and the clock generator (Clock). The clock is generated through the use of ring oscillators.

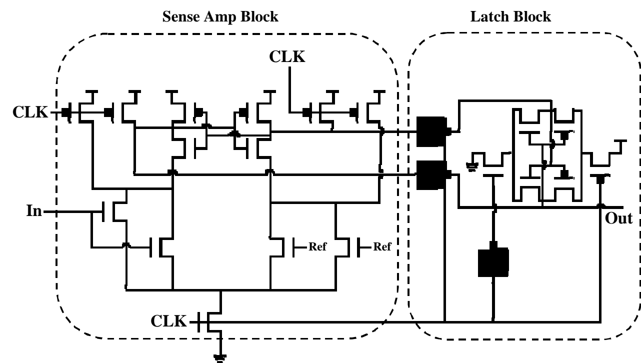
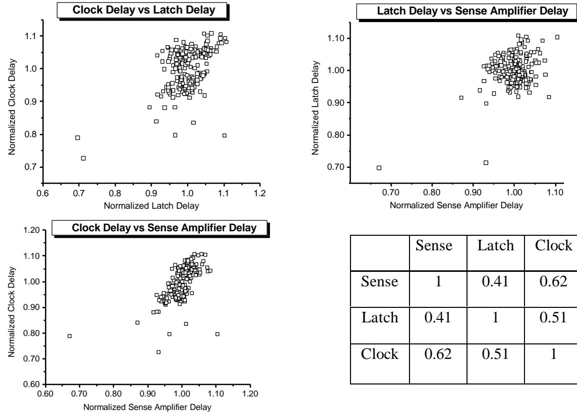


Figure 7. A bus driver circuit. Clock generator is not shown.

The individual block performance distributions were generated by considering the delay of each block, independently from other blocks. The total delay (Total) was then measured, being defined as time from arrival of the clock to the output signal. The simulations were performed using HSPICE [13].

The methodology was tested using a $0.5\,\mu m$ CMOS production technology. BSIM3v3 IV device model parameters from two lots were efficiently generated from Electrical Test data using an equation solver [6]. In addition, device gate, overlap, and junction capacitances were extracted for each die, and kept as a set together with IV parameters. Altogether, N=214 sets of device parameters were prepared for use with direct sampling method.

Each block was then simulated N=214 times to generate the block performance distributions. Out of 214 simulated data points, 8 were filtered because of simulation failures. The scatter-plots of the 3 block performance distributions and their correlation matrix are shown in Figure 8. The characteristics of the individual distributions are presented in Table 3. The scatter-plots show that the sense amplifier and latch are correlated very weakly, while there is a stronger correlation between the clock generator and the sense-amp. In both cases the blocks are neither uncorrelated nor absolutely correlated. This finding justifies treating them as statistically distinct, and explicitly accounting for their contribution.



	Sense	Latch	Clock
Sense	1	0.41	0.62
Latch	0.41	1	0.51
Clock	0.62	0.51	1

Figure 8. Scatter plots of individual block distributions and their correlation matrix.

	Mean	Standard Deviation
Sense	$9.16 \cdot 10^{-10}$	$3.63 \cdot 10^{-11}$
Latch	$5.63 \cdot 10^{-10}$	$2.8 \cdot 10^{-11}$
Clock	$1.35 \cdot 10^{-8}$	$8.6 \cdot 10^{-10}$
Total	$1.07 \cdot 10^{-9}$	$4.4 \cdot 10^{-11}$

Table 3. Characteristics of block and full circuit distributions.

Because the block distributions are not independent, the rotation of axis is necessary in order to simplify identifying the optimal training points for RSM. Employing a half-factorial design, four training points, picked out of possible 8, were selected. These points correspond to the ones identified in Table 2. The whole circuit was simulated using the SPICE models corresponding to these 4 points.

A linear (rather than, say, quadratic) model was found to be sufficient. A simple linear regression function of the statistical package S-Plus was used to generate the coefficients of the linear model [14]. The fitted linear model is:

$$Total = 1.53 \cdot 10^{-10} + 0.034\, Clock + 0.157\, Sense + 0.58\, Latch$$

To verify the accuracy of the model, a full set of 208 simulations of the whole circuit was performed. Then the model was compared with the exact simulation result. The average error was found to be 1.33% with $\sigma = 1.22\%$. The scatter-plot of the model versus the exact simulation result is presented in Figure 9.

The model was also used to calculate the characteristics of the full-circuit performance distribution directly using the characteristics of the block performance distributions. For the particular case of a linear model, the expressions for mean and variance are of the form:

$$E[F] = C_0 + \sum_{i=1}^{3} C_i E[X_i]$$

$$D[F] = \sum_{i=1}^{3} C_i^2 D[X_i] + 2 \sum_{i<j} C_i C_j \Sigma_{ij},$$

where C_i are coefficients of the model, and $\sum_{ij}$ are elements of the variance-covariance matrix, and X_i are random variables corresponding to blocks A,B, and C (i.e. Clock, Sense, and Latch). The actual and calculated characteristics are compared in Table 4. Good accuracy can be observed.

	Mean (Exact)	Std. Dev. (Exact)	Mean (Model)	Std. Dev. (Model)
Total	$1.08 \cdot 10^{-9}$	$4.51 \cdot 10^{-11}$	$1.07 \cdot 10^{-9}$	$4.4 \cdot 10^{-11}$

Table 4. Exact characteristics of full circuit distribution and characteristics calculated directly from the model.

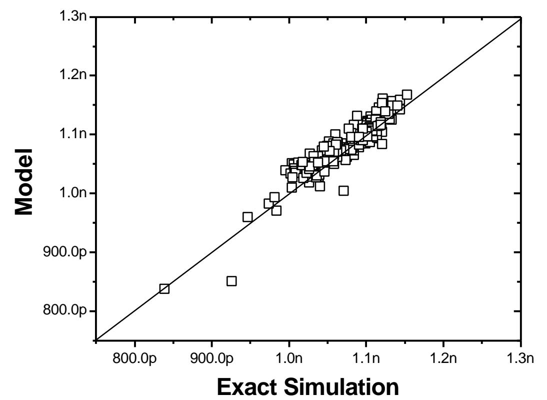


Figure 9. Good agreement between model and exact result.

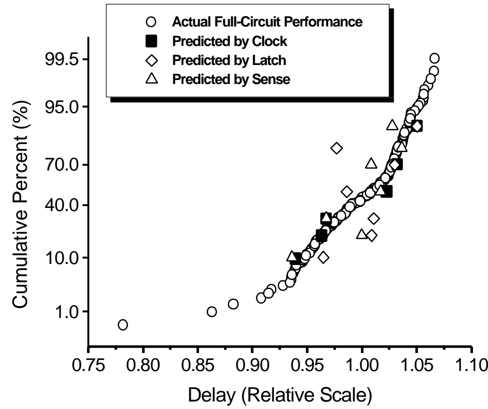


Figure 10. Blocks predict full circuit performance differently.

The model was used to generate the plot of cumulative probability function for the full-circuit delay, Figure 10. From this plot one can readily evaluate the performance of the full circuit at, for example, 2%, 10%, 50%, 90%, and 98% of speed distribution thus going further than the current approach of considering only the "worst", "best", and typical points. Because a particular SPICE model file corresponds to each data point on the plot, any other circuit performance characteristics (such as power dissipation, noise margin) can be evaluated for any specified speed performance level.

It is theoretically possible that the blocks of the circuit were well correlated, i.e. statistically identical. If it were so, statistical characterization of one of the blocks would be equivalent to the statistical characterization of the whole circuit. A SPICE file selected on the basis of the analysis of a single block at a given performance level (e.g. a latch) would correspond to the same performance level of the whole circuit. In the particular circuit considered, the blocks were only weakly correlated. Figure 10 illustrates the fact that the choice of the SPICE file at a given performance level is different for individual blocks and the whole circuit. Clearly, using a simplistic approach in this case would lead to notable prediction errors.

The simulation-time saving resulting from using SPS has been evaluated. The savings come from two distinct mechanisms. First, mere breaking down the whole circuit into a combination of smaller sub-circuits reduces simulation time of a SPICE-type simulator. Second, only one representative circuit has to be simulated for each statistically distinct block, while in the full circuit the number of its replications is likely to be significant. The simulation overhead required by the methodology has also to be taken into account.

To evaluate the time saving, we consider a 16 bit version of the bus driver architecture that was analyzed. Total computational time for the direct simulation of the whole circuit necessary to generate a statistical distribution of the same resolution:

$$T_{tot}^{direct} = N \cdot T(16 \times Full)$$

where N is the number of simulations, N=208, and $T(16 \times Full)$ is time to simulate 16 replications of the full driver circuit, corresponding to 16 bits.

If only time saving due to breaking down the circuit into smaller blocks is considered:

$$T_{tot}^{I} = N\left[T(16 \times Sense) + T(16 \times Latch) + T(16 \times Clock)\right] + 4T(16 \times Full)$$

Accounting, in addition, for replication-number reduction allowed by SPS, total time is:

$$T_{tot}^{II} = N\left[T(1 \times Sense) + T(1 \times Latch) + T(1 \times Clock)\right] + 4 \cdot T(16 \times Full)$$

The simulation results show that $T_{tot}^{I} \cong 0.89 \cdot T_{tot}^{direct}$, and $T_{tot}^{II} \cong 0.04 \cdot T_{tot}^{direct}$. In other words, 11% and 96% reductions in computational time are achieved. Clearly, the dramatic simulation time reduction comes from the replication-free approach of SPS.

4 CONCLUSION

A methodology for statistical modeling of VLSI circuits has been presented. It permits performance analysis at any specified yield level, and is suitable for use with deep sub-micron CMOS technology generations. SPS considerably reduces computer simulation time required to generate the statistical distribution of a large circuit by using a novel approach of hierarchical modeling.

5 REFERENCES

[1] R.Y.Rubinstein, *Simulation and the Monte Carlo Method*, John Wiley & Sons, 1981.
[2] J. Power, B. Donnellan, A. Mathewson, W. Lane, "Relating Statistical MOSFET Model Parameter Variabilities to IC Manufacturing Process Fluctuations Enabling Realistic Worst Case Design", *IEEE Trans. Semicon. Manuf.*, August 1994, pp. 306-318.
[3] E. Felt, S. Zanella, G. Guardiani, A. Sangiovanni-Vincentelli, "Hierarchical Statistical Characterization of Mixed-Signal Circuits Using Behavior Modeling," *1996 Proc. Of IC-CAD*, pp. 374-380.
[4] A. N. Lokanathan, J. B. Brockman, "Efficient Worst Case Analysis of Integrated Circuits," *1995 Proc. of CICC*, pp. 237-240.
[5] J. Chen, M. Orshansky, C. Hu, C.-P. Wan, "Statistical Circuit Characterization for Deep-Submicron CMOS Designs", *1998 Proc. ISSCC*, pp.90-91.
[6] J. Chen, C. Hu, C.-P. Wan, P. Bendix, A. Kapoor, "E-T Based Statistical Modeling and Compact Statistical Circuit Simulation Methodologies," *1996 Proc. of IEDM*, pp. 635-638.
[7] J. Vlach, K. Singhal, *Computer Methdos for Circuit Analysis and Design*, Van Nostrand Reinhold, 1994.
[8] J. Chen, C. Hu, Z. Liu, P. Ko, "Realistic Worst-Case SPICE File Extraction Using BSIM3," *1995 Proc. of CICC*, pp. 375-378.
[9] S. R. Nassif, "Statistical Worst-Case Analysis for Integrated Circuits*", Statistical Approach to VLSI*, pp. 233-253, 1994.
[10] G. E. P. Box, W. G. Hunter, J. S. Hunter, *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*, John Wiley & Sons, 1978.
[11] S. M. Kang, Y. Leblebici, *CMOS Digital Integrated Circuits: Analysis and Design*, The McGraw-Hill Companies, Inc., 1996.
[12] M. Hiraki, H. Kojima, H. Misawa, T. Akazawa, Y. Hatano, "Data-Dependent Logic Swing Internal Bus Architecture for Ultralow-Power LSI's", IEEE Journal of Solid-State Circuits, Vol. 30, No.4, April 1995.
[13] Meta-Software, HSPICE User's Manual, 1996.
[14] MathSoft Inc, S-Plus, 1994.