

# Direct Sampling Methodology for Statistical Analysis of Scaled CMOS Technologies

Michael Orshansky, *Student Member, IEEE*, James C. Chen, *Member, IEEE*, and Chenming Hu, *Fellow, IEEE*

**Abstract**—The continued scaling of CMOS technologies introduces new difficulties to statistical circuit analysis and invalidates many of the methodologies developed earlier. The analysis of device parameter distributions reveals multiple sources of parameter correlations, some of which exhibit mutually opposing trends. We found that applying principal component analysis (PCA) to such heterogeneous statistical data may lead to confounding of data and result in underestimation of the total parameter variance. This imposes considerable constraints on the use of several methods of statistical circuit analysis based on PCA. Also, the highly nonlinear relationships between the device parameters become more pronounced and cannot be approximated as linear even in the differential range. As a result, the response surface models based on the linear expansion of the performance variable around the nominal point of the device model parameters may lead to significant prediction errors. To address these difficulties, we propose a conceptually simple and accurate approach of direct sampling that treats the extracted SPICE parameter sets and their physical locations as an inseparable set and thus bypasses the dangerous stage of statistical inferences. We illustrate the methodology by applying it to the statistical analysis of a production CMOS process.

## I. INTRODUCTION

THE trend of the semiconductor industry to rapidly scale CMOS device dimensions continues unabated [1]. The increasing difficulty of improving the manufacturing tolerances leads to the greater relative variance of device parameters around the nominal technology point. Therefore, accurate statistical modeling becomes more important than ever.

In this paper, we address the problem of relating the stochastic manufacturing variations describable on the device level to the resulting variations in the performance of integrated circuits. Solving this problem is important for accurate statistical characterization and prediction of such circuit performance variables as speed, power consumption, and noise margins and can also be used for predicting the yield of the circuit. The fluctuations during the processing lead to variability in device characteristics both within the die and between the die. This work addresses only the interdie variations because they have traditionally been considered the more important ones. Recently, however, some researchers have also pointed out the importance of analyzing and modeling the intradie variations [2].

Manuscript received August 31, 1998; revised August 16, 1999. This work was supported in part by the SRC.

M. Orshansky and C. Hu are with the Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA 94720 USA.

J. C. Chen is with BTA Technology Inc., San Jose, CA 95112 USA.

Publisher Item Identifier S 0894-6507(99)09245-3.

Because analysis of circuit behavior is carried out by means of circuit simulations, a way of reflecting the device-level variations in terms of device model parameters is required. The major problem of statistical circuit analysis is how to relate the device-level variation to circuit-level variation in an accurate and efficient way. Over the years, many different approaches to constructing this link have emerged. The simplest approach is the method of worst case corners. Among others are gradient analysis [3] and a family of methods based upon principal component analysis [4]. We found, however, that when applied toward the analysis of deep submicron technologies, the methods referred to may yield significantly erroneous results. Here, we address two potentially problematic issues. The first lies in the highly nonlinear behavior of various device characteristics associated with the scaled technologies. The second issue is related to the possibility of heterogeneous data patterns produced by the complex processing conditions of deep submicron technologies.

Section II discusses these findings in more detail. First, we address the issue of highly nonlinear device parameter behavior and its effect on the use of statistical methodologies involving linear model construction, such as gradient analysis. Second, we discuss the implications of the combination of systematic and stochastic sources of variations in the processing phase on the use of methods of multivariate statistics, such as principal component analysis. Last, in Section III, we propose a methodology that can serve as a conceptually simple and accurate alternative to the previous methods, and we illustrate it by applying it to analysis of a production CMOS technology.

## II. DIFFICULTIES OF STATISTICAL ANALYSIS FOR DEEP SUBMICRON TECHNOLOGIES

### A. Nonlinearities in Device Functionality

It has been routinely assumed that for the purpose of response surface modeling, the device parameters can be approximated as linear functions of one another in the differential range of variation. For the devices with long channel lengths, the linear approximation indeed worked reasonably well. With the scaling of device dimensions, however, highly nonlinear effects, such as threshold voltage rolloff, came to prominence in determining the device behavior. The term “V<sub>t</sub> roll-off” describes the sharply decreasing V<sub>t</sub> of the device as  $L_{\text{eff}}$  is reduced (Fig. 1). Because the available analytical models describe the V<sub>t</sub> behavior in the sharply falling region as exponential, it is clear that the linear function approximation would work poorly [5]. The stochastic variation of  $L_{\text{eff}}$  due to

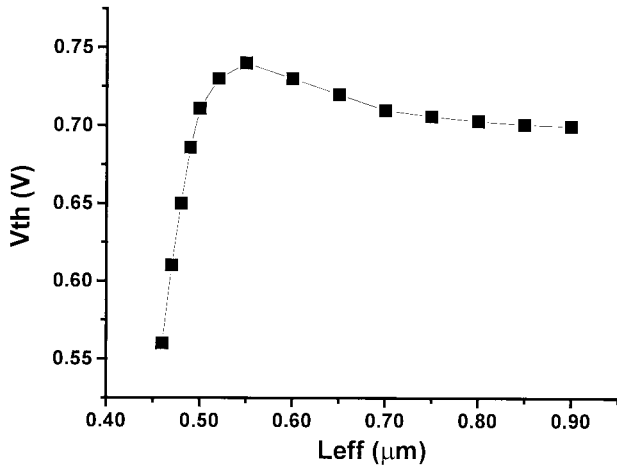


Fig. 1. The short-channel effects, such as  $V_t$  rolloff, lead to the highly nonlinear device parameter dependencies, which are propagated further and affect other characteristics.

the various steps of the process is often quite significant [2]. In addition, the device design is often centered near the peak  $V_t$  point, increasing the nonlinear  $V_t$  response as a result of small variations of  $L_{\text{eff}}$ .

The nonlinear  $L_{\text{eff}}$  dependence of  $V_t$  is further propagated to affect other important device parameter relationships. To illustrate the point, we consider the variation of NMOS saturation current ( $I_{D\text{sat}}$ ) in response to the  $\pm 3\sigma$  variation of  $L_{\text{eff}}$ . The value of  $\sigma$  was determined by characterizing 252 samples taken from an industrial  $0.5 \mu\text{m}$  CMOS process. Fig. 2 confirms that the ( $I_{D\text{sat}}$ ) response to the  $\pm 3\sigma$  variation of  $L_{\text{eff}}$  around its nominal point for a device with the nominal  $L_{\text{dr}} = 1.2 \mu\text{m}$  is linear. On the other hand, the ( $I_{D\text{sat}}$ ) response of the device with the nominal  $L_{\text{dr}} = 0.5 \mu\text{m}$  is apparently nonlinear.

Gradient analysis (GA) has been proposed as one simple technique for relating device-level to circuit-level variations [3]. As such, it could be used to perform the worst case analysis on existing technologies as well as to predict the distributions of circuit performance variables based upon the specifications of a process still under development. The circuit performance variable (such as transistor saturation current) is related to device parameters through an analytical linear model that is constructed using the response surface method. This technique uses a linear expansion of a performance variable in terms of device parameters around the nominal point. The standard deviation of the circuit performance  $P(x_1, x_2, \dots, x_n)$  can then be calculated as

$$\sigma_P = \sqrt{\sum_{i=1}^n \left( \frac{\partial P}{\partial x_i} \right)^2 \sigma_{x_i}^2 + \sum_{i=1}^n \sum_{j=1}^n \left( \frac{\partial P}{\partial x_i} \right) \left( \frac{\partial P}{\partial x_j} \right) \sigma_{x_i} \sigma_{x_j} \rho_{ij}}$$

where  $\rho_{ij}$  are the members of the  $n \times n$  correlation matrix,  $\sigma_{x_i}$  are the measured or predicted device parameter standard deviations, and  $\partial P / \partial x_i$  are the gradients or sensitivities of  $P$  to  $x_i$ .

This approach is conceptually and computationally simple but depends critically on the linearity of  $P$  at least in the design "window" of  $\pm 3\sigma$ . While this assumption held well

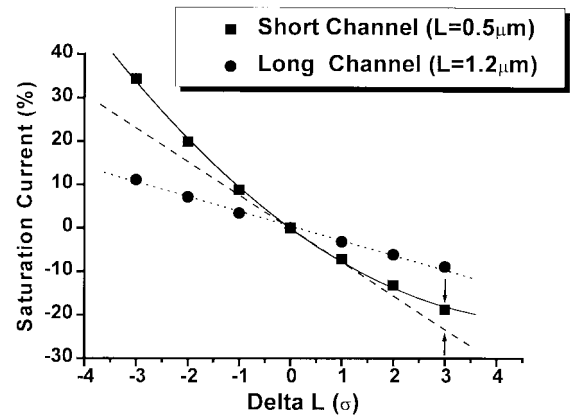


Fig. 2. Saturation current ( $I_{D\text{sat}}$ ) versus delta  $L$ . Relationship is not linear for  $L_{\text{drawn}} = 0.5 \mu\text{m}$ , leading to an error when using a linear approximation.

for earlier technologies, it may be inaccurate for technologies with shorter channel lengths, as we saw earlier. Under such conditions, using GA with its linear expansion of a circuit variable in terms of model parameters leads to a significant error of prediction (Table I).

### B. Heterogeneity of Device Parameter Patterns

Another source of difficulties for statistical circuit analysis is related to the internal structure that device parameters of the deep submicron technologies may attain, resulting in the difficulty of applying certain techniques of multivariate data analysis and transformation. Statistical technology characterization typically involves collection of a large number of device parameters from a considerable number of sites. Moreover, many of the collected device parameters are not independent from each other. The high dimensionality of the device parameter space and the intercorrelation of the parameters greatly complicates the problem of relating the device-level variations to the circuit-level variations. As a way to alleviate this problem, it has been proposed to utilize the methods of multivariate statistics such as principal component analysis (PCA) and factor analysis (FA) [4], [6], [7]. These techniques perform dimensionality reduction by transforming a large number of correlated device parameters to a small set of independent factors. These factors can then be efficiently used in the further analysis either for worst case corner generation or Monte Carlo simulation.

Statistical parameter data characterizing the devices of scaled technologies may contain some specific data patterns that complicate the application of multivariate transformation techniques, such as PCA or FA. Analyzing the data that contain device parameters for a production CMOS technology, we found that it is possible to identify two distinct and opposite trends of the threshold voltage versus channel-length relationship (Fig. 3). The larger group of data points appears to be purely stochastic, originating in the statistical fluctuations of the device channel length: the devices with shorter channel lengths have higher  $V_t$ 's due to the reverse short-channel effect (Fig. 1). The smaller group of data points has an opposite dependence: actually, devices with shorter effective channel lengths have a lower  $V_t$ . It is clear that, in contrast to the

TABLE I  
 RESULTS OF DEVICE CURRENT DRIVE PREDICTION USING DIRECT SAMPLING METHODOLOGY (DSM), GA, AND PRINCIPLE COMPONENT ANALYSIS (PCA): FOR LONGER  $L_{DRAWN}$ , GA AND PCA ARE QUITE ACCURATE. FOR  $L_{DRAWN} = 0.5 \mu\text{m}$  BOTH METHODS ARE INACCURATE, WITH ERRORS OF UP TO 28% FOR GA AND 17% FOR PCA

Channel Length - $L_{drawn}$	Measurement Sigma $I_D (\mu A)$	Exact Simulation (DSM) Sigma $I_D (\mu A)$	Gradient Analysis Sigma $I_D (\mu A)$	PCA-based Methodology Sigma $I_D (\mu A)$
1.2 $\mu\text{m}$	NA	143	145	140
0.5 $\mu\text{m}$	663	657	838	546

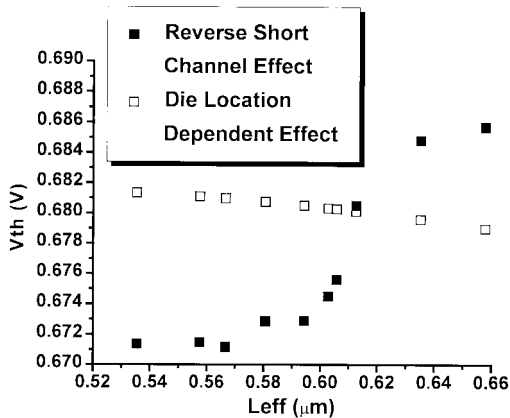


Fig. 3. One trend of  $V_t$  versus  $L_{eff}$  is due to the stochastic variations of  $L_{eff}$  and the reverse short channel effect. Another is associated with the die location of the transistor. PCA cannot capture both trends. (All points are for devices with  $L_{drawn} = 0.6 \mu\text{m}$ .)

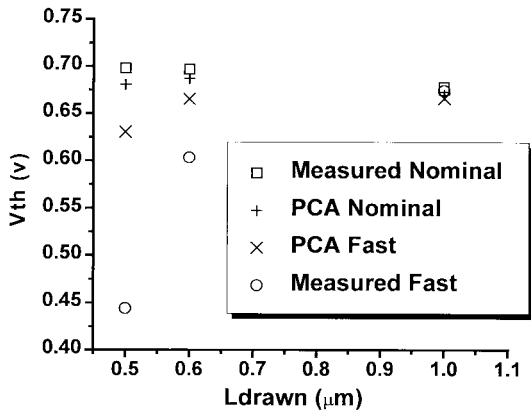


Fig. 4. Application of the principal component analysis results in averaging out the effects of two trends. The actual  $V_t$  variance is thus underestimated.

stochastic data points, the anomalous data points are due to a completely different mechanism under which the devices in certain locations failed to follow the expected reverse short-channel effect dependence. Many causes with origins in device physics, die location dependence, optical proximity effect, microloading in etching and deposition, etc., may have contributed to produce this distinct  $V_t$ - $L$  correlation.

Such heterogeneous and nonmonotonic relationships among the device parameters may not be adequately captured by

the PCA-based device parameter characterization. In a simple interpretation, PCA and FA can be graphically represented as rotations of the original axis. In the presence of nonhomogeneous multimodal data coming from two or more sources, a phantom averaging effect occurs. The larger number of stochastic points outweighs the data points of the systematic group in the process of the PCA's weighted axis rotation. However, the devices from the systematic group have a greater spread of values than the stochastic data points. Thus, the rotation performed by PCA leads to confounding of the data and a resultant underestimation of the actual level of parameter fluctuations (Fig. 4). Also, if the results of PCA are now used to predict the level of variation of a device or circuit characteristics, the error of parameter estimation is further propagated (Table I).

### III. DIRECT SAMPLING METHODOLOGY

#### A. Methodology

The difficulties described earlier lead us to believe that it is necessary to avoid the statistical methods that are based on making any purely statistical inferences and to instead rely on data itself as much as possible. Toward this end, we need to keep the measured parameters as a set, preserving the existing complex nature of parameter correlations and directly linking the data sets to their physical die locations. Such an approach we call a direct sampling methodology (DSM). When using DSM, we skip the additional level of data transformation involved in PCA and avoid the danger of data confounding because we do not need to make any assumption on the homogeneous nature of the data set. We also do not require the assumption of linearity of device relationships that is explicit in the methods relying on the linear model construction, such as GA.

The idea of direct sampling is conceptually simple and computationally efficient. First, a statistically significant number of test sites need to be fully characterized, resulting in a device model parameter set per site. The test sites should be in the scribe line so that they cover the entire wafer rather than only a few fixed die locations. This way we can guarantee that all the variation patterns will be accurately modeled. Each physically distinct site possesses a corresponding SPICE-compatible model parameter set. For the methodology to be

most accurate, each test die would be characterized in terms of its I-V parameters, device, and parasitic capacitances. If one is sure that a particular variation component is well controlled or that its variation does not contribute much to the overall circuit performance variability, the characterization set-up may be correspondingly simplified. The device model parameters pertaining to the I-V may be efficiently extracted from electrical-test (E-T) data that is routinely collected in the fab. A large number of device parameter sets may thus be extracted in an efficient way using an equation-solver routine [8]. The model parameters are guaranteed to produce the accurate simulation results for the particular die through a direct substitution of several physically meaningful E-T parameters ( $T_{ox}$ ,  $\Delta L$ ,  $\Delta W$ ) and optimization of the fitting model parameters to conform to the measured electrical targets ( $I_{d,sat}$ ,  $V_t$  for short-channel devices). The extraction procedure is efficient (30 s/die) and accurate (average error of fit is 4.5%).

Once a significant number ( $N$ ) of device parameter sets is generated, the statistical analysis proceeds by the direct  $N$  SPICE simulations of one or a few small sample circuits, e.g., an inverter, generating their full continuous performance distribution. Then, the statistical SPICE models can be easily extracted from a cumulative probability plot for an arbitrary performance level (i.e. 2%, 5%, 95%, 98%), thus improving on the current approach of considering only worst, best, and typical points. These model files can then be used to simulate a complex circuit being designed in order to assess its overall statistical performance and yield. In doing this, we assume that the statistical SPICE models chosen on the basis of some simple circuit behave statistically similarly when applied to various larger circuits. This assumption implies, for example, that a model parameter set corresponding to a 95% point of an inverter speed distribution also corresponds to the 95% point of a multiplier speed distribution. Simulation studies show that this assumption is usually justified for a variety of common digital circuits [9]. If necessary, the statistical SPICE models for a more complex circuit can be accurately generated by the direct simulation of such circuit. This approach does not require the assumption of statistical correlation, but may be computationally prohibitive if the circuit is sufficiently large.

Direct sampling also allows one to perform other interesting analysis of the statistical properties. Because each simulated data point can be easily traced to the corresponding E-T parameter set, data can be used to analyze the circuit performance sensitivity to different device and process parameters. This information may then be used for circuit optimization, reducing the circuit's sensitivity to a particular source of variation. It can also be used for process optimization. We can also easily construct two-dimensional wafer maps with regard to variation of device and circuit characteristics of interest. Such information is often valuable for process optimization and yield analysis.

### B. Experiment

To illustrate the methodology, we applied it to the statistical characterization of a production 0.5  $\mu\text{m}$  CMOS process. We extracted the full SPICE model files for 252 dies of five wafers

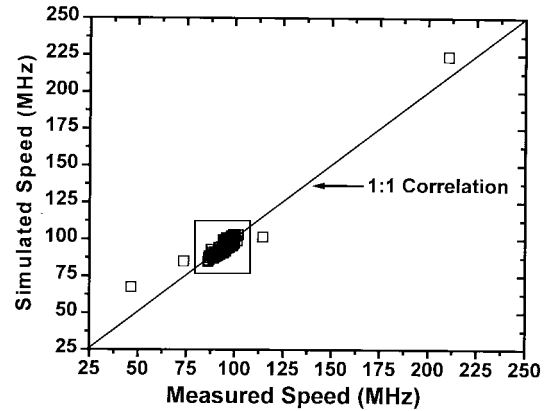


Fig. 5. An excellent match between the measured and simulated ring oscillator speeds over an entire speed range (4 $\times$ ) is achieved using DSM.

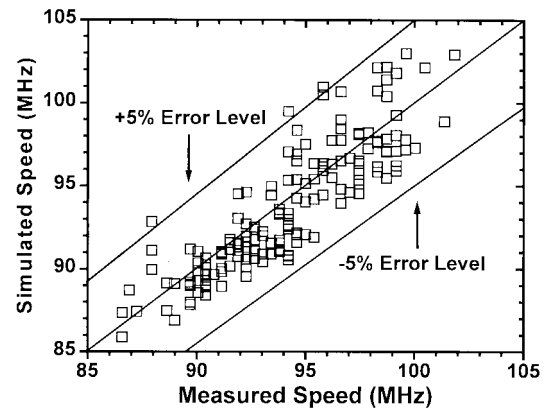


Fig. 6. Zoom in of the squared region in Fig. 5. Most data points are within the 5% error boundary.

TABLE II  
MEASURED AND SIMULATED RING OSCILLATOR (RO) SPEEDS UNDER DIFFERENT SIMULATION CONDITIONS. EVEN THOUGH RO HAS FANOUT = 1, VARIATION OF PARASITIC JUNCTION CAPACITANCES (CJ) DOES NOT HAVE MUCH IMPACT

Simulation Condition	Mean (MHz)	Sigma (MHz)
Measured Ring Oscillators	94.44	10.57
IV+Cox Varying	94.20	11.04
IV+Tox+Cj+Cov Varying	94.51	11.12

coming from two different lots. The I-V device model parameters were extracted from E-T database using the equation solver. The capacitance model parameters, which included the junction bottom and sidewall capacitances and the overlap capacitances, were extracted from all 252 dies. Ring oscillator speeds were also collected from the same dies.

To assess the accuracy of the methodology, the extracted 252 SPICE model files were then used to perform the simulation of the ring oscillator circuits. The overall match between the ring oscillator speeds simulated using DSM, and the measured data is good over the entire speed range (4 $\times$ ), as shown in Fig. 5. The amount of variation explained by the predictor is 92% as inferred from the value of  $R^2$  statistics. The error is typically within 5% (Fig. 6). Mean and standard deviation are also accurately predicted (Table II). We consid-

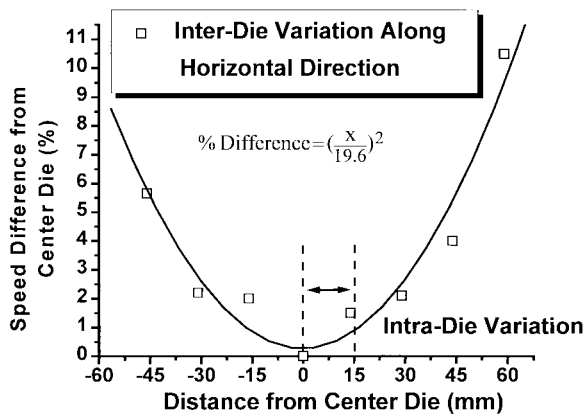


Fig. 7. Systematic variations across the wafer’s diameter result in significant speed differences.

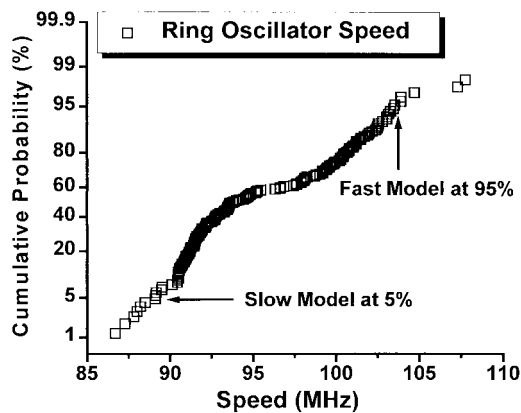


Fig. 8. Cumulative probability plots can be used to pick the SPICE model files corresponding to a desired performance level.

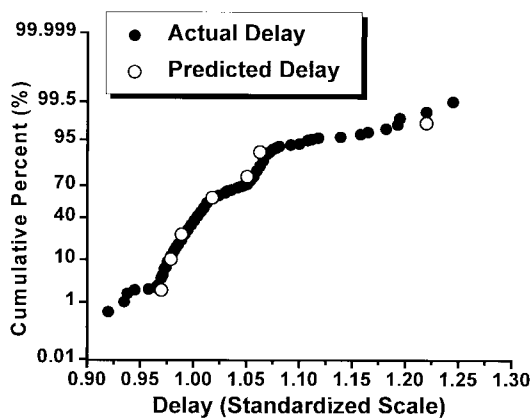


Fig. 9. SPICE models selected on a basis of a sample circuit can predict performances of larger circuits.

ered two levels of device characterization complexity: with and without characterizing the junction capacitance variation. Because the extraction procedure for the junction capacitance was not automated, characterization of this model parameter was quite time consuming. The experimental results, however, imply that the variation of the junction capacitance is not statistically important (Table II), and for this technology, it may be skipped. Even for a ring oscillator with FO = 1, where

the effect of junction capacitance is the strongest, accounting for  $C_j$  and  $C_{jsw}$  did not have much influence.

The circuit performance variation across the wafer was analyzed by measuring the ring oscillator speeds at several locations across the diameter of the various wafers. A systematic variation of speed as a function of die location is shown in Fig. 7. A simple empirical quadratic model is proposed

$$\% \text{ Difference} \cong \left( \frac{x}{19.6 \text{ mm}} \right)^2.$$

This model can be used to predict the circuit performance variation as a function of distance away from the center of the wafer ( $x$ ). Thus, if the model continues to be appropriate, we can expect the difference in circuit speed to be as much as 25% for a 200 mm wafer.

A cumulative probability plot of 252 simulated ring oscillator speed values was used to extract seven SPICE model files for specific delay percentiles (Fig. 8). The seven files correspond to 2%, 10%, 25%, 50%, 75%, 90%, and 98% performance percentiles. These model files were then used to simulate a four-bit adder circuit. The predicted specific delay values can then be compared with the full distribution of 252 simulated delays of the four-bit adder. The prediction is quite accurate (Fig. 9). Because of a close link between the simulated data points and the corresponding E-T parameter set, we can easily analyze the circuit performance sensitivity to any device and process parameters.

#### IV. CONCLUSION

In this paper, we discuss some difficulties related to statistical circuit analysis for characterizing the deep submicron technologies. One issue is the effect of the highly nonlinear device behavior on statistical methods. Another is the danger of data confounding when linear transformation techniques are applied to multiple source and heterogeneous device parameter data characteristic of the deep submicron technologies. We propose a conceptually straightforward alternative to the existing methods, the direct sampling methodology. We illustrate its use by applying it to a production industrial process, and show its accuracy and flexibility.

#### ACKNOWLEDGMENT

The authors gratefully thank Dr. D. Wan and Dr. P. Bendix of LSI Logic Inc. for their help in carrying out this project.

#### REFERENCES

- [1] B. Davari, “CMOS technology scaling, 0.1 um and beyond,” in *Proc. IEDM*, 1996, p. 555.
- [2] Z. Krivokapic, A. Minvielle, and W. Heavlin, “Intrafield effects and device manufacturability: A statistical simulation approach,” in *Proc. Third Int. Workshop Statistical Metrology*, 1998, p. 36.
- [3] M. Bolt, M. Rocchi, and J. Angel, “Realistic statistical worst-case simulations of VLSI circuits,” *Trans. Semiconduct. Manufact.*, vol. 4, no. 3, pp. 193–198, 1991.
- [4] J. Power *et al.*, “An approach for relating model parameter variabilities to process fluctuations,” in *Proc. ICTMS*, 1993, p. 63.
- [5] Z.-H. Liu, C. Hu, J.-H. Huang, T.-Y. Chan, M.-C. Jeng, P. K. Ko, and Y. C. Cheng, “Threshold voltage model for deep-submicrometer MOSFET’s,” *IEEE Trans. Electron Devices*, vol. 40, pp. 86–95, Jan. 1993.

- [6] E. Felt, S. Zanella, C. Guardiani, and A. Sangiovanni-Vincetelli, "Hierarchical statistical characterization of mixed-signal circuits using behavioral modeling," in *Proc. IC-CAD*, 1996, p. 374.
- [7] S. J. Press, *Applied Multivariate Analysis*. New York: Holt, Rinehard and Winston, 1972.
- [8] J. C. Chen, C. Hu, D. Wan, P. Bendix, and A. Kapoor, "E-T base statistical modeling and compact statistical circuit simulation methodologies," in *Proc. IEDM*, 1996, p. 635.
- [9] J. C. Chen, C. Hu, Z. Liu, and P. K. Ko, "Realistic worst-case SPICE file extraction using BSIM3," in *Proc. Custom Integrated Circuits Conf.*, 1995, p. 375.

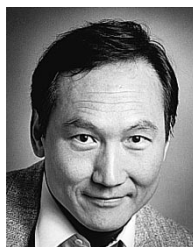
**Michael Orshansky** (S'96) received the B.S. (with honors) and M.S. degrees in electrical engineering and computer science from the University of California at Berkeley, in 1996 and 1998, respectively. Since 1996, he has been pursuing the Ph.D. degree at the University of California, Berkeley and plans to graduate in 2000.

He has held internships with Integrated Device Technology in San Jose, CA, and Advanced Micro Devices in Sunnyvale, CA. His research interests are in the area of statistical modeling for deep sub-micron CMOS technologies.

Mr. Orshansky is a 1998 Fellow of the Semiconductor Research Corporation and Advanced Micro Devices.

**James C. Chen** (S'96-M'98) received the B.S., M.S., and Ph.D. degrees in electrical engineering and computer science at the University of California, Berkeley in 1993, 1995, and 1998 respectively.

He is presently Product Marketing Manager at BTA Technology Inc., San Jose, CA, where he is working on providing new EDA solutions for the Analog/Mixed-Signal market in terms of mismatch and mixed-mode simulation algorithms. He holds several patents and has published more than 20 papers in statistical and interconnect modeling.



**Chenming Hu** (S'71-M'76-SM'83-F'90) received the B.S. degree from National Taiwan University, Taiwan, and the M.S. and Ph.D. degrees in electrical engineering from the University of California, Berkeley.

He is Chancellor's Professor of Electrical Engineering and Computer Sciences at the University of California, Berkeley. He was an Assistant Professor at the Massachusetts Institute of Technology, Cambridge, for three years. He was the Board Chairman of the East San Francisco Bay Chinese School and is a frequent advisor to industry and educational institutions. His present research areas include microelectronic devices, thin dielectrics, circuit reliability simulation, and nonvolatile memories. He has authored or coauthored four books and over 600 research papers and supervised 60 doctoral students. He leads the development of the MOSFET model BSIM3v3 that has been chosen as the first industry standard model for IC simulation by the Electronics Industry Association Compact Model Council and was given a Research and Development 100 Award as one of the 100 most technologically significant new products of the year, in 1996.

Dr. Hu is a member of the U.S. National Academy of Engineering, an Adjunct Professor of Peking University, and an Honorary Professor of the Chinese Academy of Science. In 1991, he received the Excellence in Design Award from Design News and the inaugural Semiconductor Research Corporation Technical Excellence Award for leading the research of the IC reliability simulator, BERT. He received the SRC Outstanding Inventor Award in 1993 and 1994. IEEE awarded him the 1997 Jack A. Morton Award for his contributions to the physics and modeling of MOSFET reliability. Also in 1997, he received the University of California, Berkeley's highest honor for teaching, the Distinguished Teaching Award. In 1998, he received the Monie A. Ferst Award of Sigma Xi for encouragement of research through education.