Short Papers

On a Generalized Framework for Modeling the Effects of Process Variations on Circuit Delay Performance Using Response Surface Methodology

B. P. Harish, Navakanta Bhat, and Mahesh B. Patil

***Abstract*—A generalized methodology for modeling the effects of process variations on circuit delay performance is proposed by directly relating the variations in process parameters to variations in delay metric of a digital circuit. The 2-input NAND gate is used as a library element for 65 nm gate length technology, whose delay is extensively characterized by mixed-mode simulations. This information is then used in a general-purpose circuit simulator SEQUEL, by incorporating appropriate templates for the NAND gate library. A 4-bit $\times$ 4-bit Wallace tree multiplier circuit, consisting of about 300 2-input NAND gates, is used as a representative combinational circuit to demonstrate the proposed methodology. The variation in the multiplier delay is characterized by an extensive Monte Carlo analysis. To extend this methodology for a generic technology library with a variety of library elements, modeling of NAND gate delays by response surface methodology (RSM), in terms of process parameters, is carried out using design of experiments (DOE). A simple piecewise quadratic model, based on the least squares method (LSM), is proposed for one-parameter variation to address significant cubic effects observed in the delay response function. Then, a hybrid model for gate delays is generated by superimposing the interaction terms of DOE–RSM model upon the quadratic model of one-parameter variation to address the generalized case of simultaneous variations in multiple process parameters. The proposed methodology has been demonstrated for NAND gate library with 266 gates, and the simplicity and generality of the approach make it equally applicable to a large library of cells for both statistical timing analysis and statistical circuit simulation at the gate level.**

***Index Terms*—Delay distribution, design of experiments (DOE), hybrid model, least squares method (LSM), mixed-mode simulations, Monte Carlo analysis, process sensitivity, response surface methodology (RSM).**

I. Introduction

The design of high-performance digital application-specific integrated circuits (ASICs), in the deep-submicrometer (DSM) regime calls for incorporating the performance fluctuations caused by process-induced parameter variations. Variability is fast emerging as a major challenge that threatens to impact the yield at its best and the circuit functionality, at its worst. Even as device dimensions are being aggressively scaled to achieve faster and more complex integrated circuits, the process tolerances are not becoming tight enough, resulting in increased effects of process variations on device and circuit characteristics. The problem is not the amount of variability, but the variability turned into uncertainty if it is not modeled and cancelled out by design techniques, as uncertainty can only be handled by vastly over guard-banding of designs. It is expected that performance variances caused by this mismatch in short-channel MOS circuits may ultimately introduce a limitation for device scaling in integrated circuits [1].

Manuscript received July 4, 2005; revised January 26, 2006. This work was supported by the Department of Science and Technology, Government of India. This paper was recommended by Associate Editor S. Saxena.

B. P. Harish and N. Bhat are with the Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore 560 012, India (e-mail: harish@ece.iisc.ernet.in; navakant@ece.iisc.ernet.in).

M. B. Patil is with the Department of Electrical Engineering, Indian Institute of Technology, Bombay, Mumbai 400 076, India (e-mail: mbpatil@ee.iitb.ac.in).

Digital Object Identifier 10.1109/TCAD.2006.883910

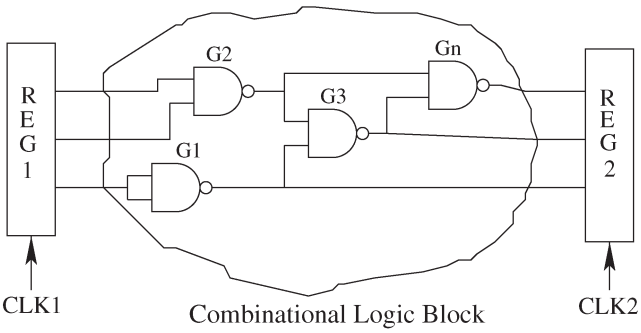

Fig. 1. Timing closure in a digital circuit.

The factors that cause mismatch can be broadly classified as systematic variations and random variations. Processing gradients across the wafer introduce systematic variations related to wafer maps during manufacturing, which are independent of device size. The random variations include extrinsic and intrinsic variations and are dependent on device size. The extrinsic variations are caused by variations in implant dose, implant energy, oxidation, and annealing temperature, all of which are equipment related. The intrinsic variations are due to random fluctuations in channel dopant number, gate oxide thickness, interface charge, oxide charge, and interlevel dielectric permittivity. The corner SPICE parameters of the transistors are derived taking into account both extrinsic and intrinsic variations, which encompass the complete fluctuations of the processes, under worst case. In other words, they correspond to variations in transistor parameters across the wafer lots in the fabrication environment. However, the variations in parameters of transistors located within a die have some correlation, as enunciated by Pelgrom's classical mismatch model [2]. This is because the variations in individual processes have short-range and long-range orders; in addition, some of the processes have nonzero correlations.

The typical timing closure issue in digital ASIC design is illustrated in Fig. 1. The data from a register REG1 of pipeline stage goes through combinational logic block and is latched in REG2. The typical design closure criterion of the slowest arrival time of inputs at REG2 due to worst case logic delay and the fastest arrival time of CLK2 due to worst case clock skew can result in a very pessimistic design. The notion of probabilistic design has been introduced recently to address this issue [3]. Furthermore, algorithms for statistical timing analysis have been suggested for efficient timing sign-off of digital integrated circuits [4].

In this context, it would be very desirable to have a direct relationship between the gate delay of various library elements and the underlying manufacturing process parameters. This will facilitate the realistic evaluation of circuit delay variability considering the actual short-range and long-range orders of individual processes, as well as the correlations. The design will then become more robust and less conservative.

Response surface modeling has been used to identify and relate significant process parameters to device parameters that are taken as response variables, by a quadratic model, to optimize device design [5]. A device design methodology using the second order response surface modeling by multivariable optimization using process sensitivity considerations has been demonstrated [6]. The design of experiments–response surface methodology (DOE–RSM) has been used for optimizing semiconductor process to help reduce design/analysis cycle time [7]. An integrated system called DOE/Opt has

been prototyped for performing DOE, response surface modeling, and optimization, using coupled process and device simulations for process control modeling and statistical process optimization [8]. Similarly, there has been prior work using RSM to relate circuit parameters to variations in SPICE parameters [9], [10]. Monte Carlo simulation based second order polynomial modeling approach has been proposed to correlate statistical variations in device parameters to random variations in process parameters [11]. Recently, the mixed-mode simulation approach was presented to correlate the inverter delay variations to implant dose variations [12]. In this paper, we present a methodology to evaluate the effect of process variations on the delay variations of a complex digital circuit. An optimal second order "hybrid model" is obtained using DOE–RSM modeling and least squares method (LSM) technique for gate delays directly in terms of multiple process parameters. We demonstrate that the worst case design approach is overly pessimistic, and the hybrid model based statistical design approach results in robust circuit design.

We perform mixed-mode simulations, which bring the process-simulated devices directly into the netlist of the circuit, wherein both circuit and device equations are solved simultaneously. Process/device simulation is considered appropriate to the study of process sensitivity as it enables the precise control of process variations that are difficult to achieve experimentally. A commercial technology computer-aided design (TCAD) tool suite from Integrated Systems Engineering (ISE) has been used for process and device simulations [13]. The general-purpose circuit simulator, i.e., A Solver for circuit EQuations with User-defined ELements (SEQUEL), has been used for circuit simulations [14].

Section II discusses the design of 65 nm gate length transistors and the process sensitivity at the device/circuit level and delay characterization of NAND gate. Section III describes the principles of statistical modeling methodology. Section IV presents the SEQUEL simulations for delay distributions of 4-bit $\times$ 4-bit multiplier circuit for a set of process parameters, with individual variation and simultaneous variations. Section V concludes with a summary of results.

II. Process Sensitivity and Delay Characterization of NAND Gate

The overall flow of events that transform the process variations to relevant delay distributions using various simulation tools and models is illustrated in Fig. 2. The nominal NMOS and PMOS devices with 65 nm physical gate length are designed and optimized for an off-state leakage current constraint of 10 nA/μm at $V_{\rm dd} = 1.2$ V. The disposable spacer process sequence [15], with pocket halo and super steep retrograde channel (SSRC) implants, has been used for source/drain and channel engineering. Fig. 3 shows the 65 nm gate length NMOS/PMOS devices generated using process and device simulation based design approach. The contours of doping concentration are shown along with their values in atoms per cubic centimeter. A set of process parameters, whose variability has a significant impact on device parameters, is identified based on our simulations and published literature [1], [11]. They include the gate length (L_g), gate oxide thickness ($T_{\rm ox}$), halo dose, SSRC dose, halo tilt angle, and source/drain anneal temperature. The process parameter variations are assumed to have a Gaussian distribution with a $\pm 3\sigma$ variation of $\pm 10\%$ of the nominal value, except for the anneal temperature for which it is taken as ± 10 °C. These sigma levels are in accordance with [1] and [11]. A set of NMOS/PMOS devices with these assumed variations in each of the six process parameters, taken one at a time, is generated by process simulations using the DIOS process simulator. All the devices are simulated with drift–diffusion transport model to obtain I_d–V_g and I_d–V_d characteristics, and their respective saturation

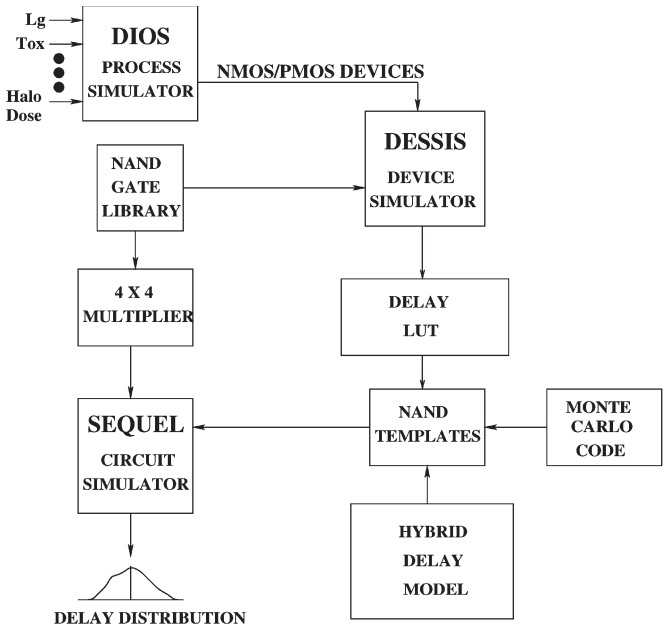

Fig. 2. Block diagram of simulation flow.

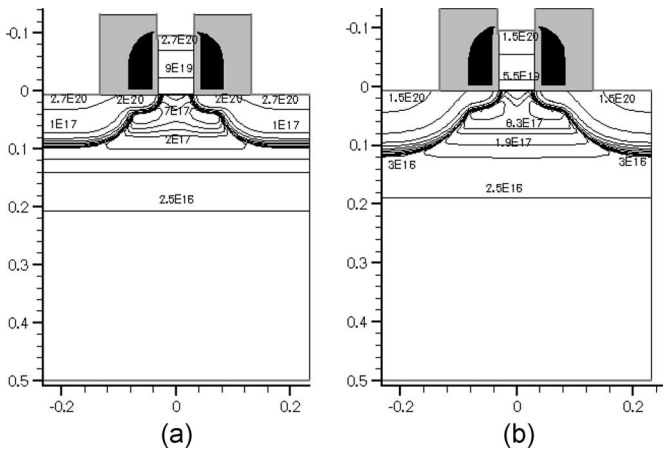

Fig. 3. Process-simulated nominal devices: (a) NMOS and (b) PMOS.

currents $I_{\rm on}$ are measured. For device simulations, QCVandort channel quantization model, band-to-band recombination model, mobility models for doping, normal-field dependence, and high-field saturation (velocity saturation) are included.

The percentage variation in the saturation current $I_{\rm on}$ at the device level, with variations in process parameters considered, is presented in Table I. The relative deviation of any parameter x about its nominal value $x_{\rm nom}$ is calculated as $\Delta x = (x - x_{\rm nom})/x_{\rm nom}$. It can be seen that the variations in L_g, $T_{\rm ox}$, and halo dose have the maximum impact on drive current variations for both NMOS and PMOS, which is in accordance with the published literature.

Using these devices, a two-stage 2-input NAND gate (Fig. 4) is simulated to evaluate its transient behavior. The mixed-mode simulation approach is used with the DESSIS device simulator. Both NMOS and PMOS are simulated at full device level. An area factor for PMOS devices is considered to be twice that of NMOS, to account for the difference in carrier mobility, as reflected in their drive currents. An input pulse $V_{\rm in}$ with a rise and fall time of 1 ps is applied, and the stage delay of the first stage at its output Y is monitored, when loaded by an identical second stage. Delay values for rising and falling edge

TABLE I
PERCENTAGE VARIATION IN SATURATION CURRENT I_{on} FOR NMOS AND PMOS. THE NOMINAL VALUES OF CURRENTS ARE $I_{on} = 723\ \mu$A (NMOS) AND $I_{on} = 362\ \mu$A (PMOS)

Process variation	NMOS						PMOS					
	L_g	T_{ox}	Halo	SSRC	Halo tilt	Anneal temp	L_g	T_{ox}	Halo	SSRC	Halo tilt	Anneal temp
-10%	+5.1	+12.2	+3.8	+1.9	+3.4	-1.8	+0.8	+1.1	+2.8	+3.4	+1.6	-6.9
-5%	+4.6	+1.7	+2.0	+1.25	+1.8	-0.8	+0.4	+0.3	+0.8	+2.5	+0.03	-4.0
+5%	-0.8	-0.5	-1.15	-0.1	-0.95	+1.8	-0.9	-5.8	-3.0	-1.8	-2.4	+1.6
+10%	-0.95	-3.0	-2.6	-0.7	-2.2	+3.4	-9.4	-8.6	-4.8	-4.2	-3.7	+3.6

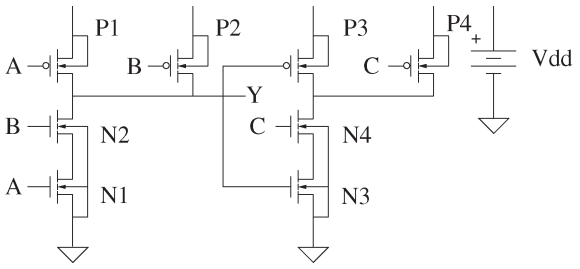

Fig. 4. Two-stage NAND gate.

transitions at the output due to all possible input combinations are obtained using transient analysis.

The percentage variation in the stage delay with respect to the nominal, with variations in process parameters considered, is presented in Tables II–IV. The gate delay A_{up} denotes the rising edge delay at output Y caused due to A input. The gate delay B_{up} denotes the rising edge delay at output Y caused due to B input. The gate delay AB_{up} indicates the rising edge delay when both A and B inputs are tied together. Similarly, A_{down}, B_{down}, and AB_{down} denote the respective falling edge gate delays.

Fig. 5 shows the delay variations for A_{down} and A_{up} transitions for different processes. The gate delays are most sensitive to variations in L_g and T_{ox} for both rising and falling edge transitions. A 5.1% increase in I_{on} of NMOS due to -10% variation in its L_g results in 10.4% decrease in A_{down}. The corresponding delay variations are 11% and 10% decrease in B_{down} and AB_{down}, respectively. On the same lines, a 9.4% decrease in I_{on} of PMOS due to $+10\%$ variation in its L_g results in 18.4% increase in A_{up}. The corresponding delay variations are 16.5% and 16.1% increase in B_{up} and AB_{up}, respectively. It is observed that an increase in device drive current due to process parameter variations does not necessarily translate into faster circuits when these process parameter variations also result in higher device capacitances, as observed in the case of T_{ox}.

A lookup table of NAND gate delay transitions A_{up}, A_{down}, B_{up}, B_{down}, AB_{up}, and AB_{down} for nominal, $\pm5\%$, and $\pm10\%$ variations in L_g is shown in Table V. Similar lookup tables are generated for variations in other process parameters such as T_{ox}, halo dose, SSRC dose, halo tilt angle, and annealing temperature. This completes the delay characterization of NAND gate.

III. MODELING METHODOLOGY

An analytical model proposed in [16] to relate NAND gate delays and device saturation currents based on CV/I metric, although efficient in tracking variations in one process parameter at a time, fails to capture the effects of interaction between process parameters when multiple parameters are simultaneously varied. However, these interactions are too significant to be ignored [2]. Hence, this paper attempts to model the dependence of gate delays upon process parameter variations by statistical methods.

To model the relationship between the gate delay with simultaneous variations in multiple process parameters, the statistical technique of DOE is used. The DOE is performed, and second order models are built for rising and falling edge gate delays using RSM [17]–[19].

A 3-level face centered central composite (FCCC) design of resolution VI [20] for six process parameters is designed with 52 experimental runs. This design is a highly fractionated 3-level DOE for fitting second order response surfaces. In FCCC design, the star points are at the center of each face of the factorial space, with $\alpha = \pm 2.38$. The parameter α indicates the distance of the axial point from the center point in the normalized parameter space. To ensure that the design is rotatable, the value of α is selected as $\alpha = [2^{6-1}]^{1/4}$, where 2^{6-1} is the number of factorial portion of runs [20]. For six factors, a fraction of 52 experiments have been chosen out of a large set of 3^6 $(= 729)$ full factorial design. The 2^{6-1} fractional factorial design, which results in 32 factorial runs, is augmented with 12 axial star points and 8 replicated center points to yield a total of 52 experiments to be conducted in the FCCC design for 6 factors. The process parameters under consideration are varied by $\pm10\%$ except for anneal temperature, which is varied by ±10 °C about their nominal values. It is assumed that $\pm10\%$ or ±10 °C corresponds to $\pm3\sigma$ variation in the process under study.

The second order models that have been obtained by the regression technique using simulation results are long polynomials of the form

$$\begin{aligned} y = {} & \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_6 x_6 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 \\ & + \cdots \beta_{23} x_2 x_3 + \beta_{24} x_2 x_4 + \cdots + \beta_{56} x_5 x_6 \\ & + \beta_{123} x_1 x_2 x_3 + \cdots + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \cdots + \beta_{66} x_6^2 \end{aligned} \quad (1)$$

where β_0 is a constant, x_i are the normalized process parameters varying between -1 and $+1$, and β_i are the corresponding regression coefficients determined by the data obtained from the response surface DOE, for $i = 1, \ldots, 6$.

An optimum second order model is obtained by removing all insignificant effects and recomputing the regression. The NAND gate delays that have been modeled are the rising and falling edge delays caused due to A input, B input, and when A and B inputs are tied together.

The percentage variation of gate delays as a function of -10% to $+10\%$ variation of process parameters is shown in Fig. 5. The delay variation is calculated at each X value as a percentage of its value for the nominal design. To detect and fit the cubic effect seen in the falling edge delay response due to variation in gate oxide thickness

TABLE II
NAND GATE DELAY VARIATIONS FOR RISING EDGE (IN PERCENT). THE NOMINAL VALUES OF DELAY ARE $A_{\rm up} = 7.7$ ps AND $B_{\rm up} = 6.7$ ps

Process variation	A_{up}						B_{up}					
	L_g	T_{ox}	Halo	SSRC	Halo tilt	Anneal temp	L_g	T_{ox}	Halo	SSRC	Halo tilt	Anneal temp
-10%	-1.7	+4.5	-2.1	-1.9	-0.4	+4.4	-2.3	+4.7	-1.4	-1.7	-0.5	+4.7
-5%	-1.3	+1.5	+0.04	-1.3	+0.9	+2.8	-1.8	+0.7	+0.01	-1.2	+0.9	+3.0
+5%	+5.9	+6.1	+3.1	+2.1	+2.6	+0.5	+4.5	+6.1	+3.0	+2.0	+3.1	+0.5
+10%	+18.4	+7.7	+4.85	+5.1	+3.6	+0.35	+16.5	+7.5	+5.3	+5.4	+3.6	+0.3

TABLE III
NAND GATE DELAY VARIATIONS FOR FALLING EDGE (IN PERCENT). THE NOMINAL VALUES OF DELAY ARE $A_{\rm down} = 8.5$ ps AND $B_{\rm down} = 7.6$ ps

Process variation	A_{down}						B_{down}					
	L_g	T_{ox}	Halo	SSRC	Halo tilt	Anneal temp	L_g	T_{ox}	Halo	SSRC	Halo tilt	Anneal temp
-10%	-10.4	-9.3	-1.5	+0.6	-1.7	-0.5	-11.0	-9.8	-2.0	+0.6	-2.1	-0.6
-5%	-6.6	-2.9	-0.7	+0.8	-0.8	-0.15	-7.5	-3.6	-0.8	+0.9	-0.75	-0.2
+5%	+2.4	-3.0	+1.2	+0.45	+0.9	+0.45	+1.5	+3.5	+0.9	+0.5	+1.0	+0.2
+10%	+7.2	+0.03	+2.0	+1.0	+1.8	+0.4	+5.8	-0.1	+1.8	+0.8	+1.9	+0.2

TABLE IV
DELAY VARIATIONS OF NAND GATE AS INVERTER FOR RISING AND FALLING EDGES (IN PERCENT). THE NOMINAL VALUES OF DELAY ARE $AB_{\rm up} = 4.4$ ps AND $AB_{\rm down} = 9.4$ ps

Process variation	AB_{up}						AB_{down}					
	L_g	T_{ox}	Halo	SSRC	Halo tilt	Anneal temp	L_g	T_{ox}	Halo	SSRC	Halo tilt	Anneal temp
-10%	-2.5	+4.0	-1.1	-1.0	-0.3	+4.3	-10.0	-8.8	-1.7	+0.5	-1.9	-0.25
-5%	-1.8	+1.3	+0.05	-0.9	+0.9	+2.9	-6.8	-3.0	-0.9	+0.4	-0.9	-0.01
+5%	+5.3	+5.3	+3.0	+2.4	+2.65	+0.9	+2.3	-2.2	+1.1	+0.5	+1.1	+0.07
+10%	+16.1	+6.9	+4.5	+4.7	+3.8	+0.7	+7.3	+0.78	+2.1	+1.0	+2.1	-0.07

$T_{\rm ox}$ in Fig. 5(a), the minimum number of levels needed for factor settings is 4. Thus, a 3-level FCCC design will not fit these response functions. However, a 4-level DOE–RSM modeling is prohibitive in terms of number of experimental runs and, hence, in terms of cost and computational effort. Under these circumstances, to achieve this response fit with only three levels for factor settings and a resultant quadratic model, the -10% to $+10\%$ range is split into two regions: -10% to 0% and 0% to $+10\%$. This piecewise modeling is justified because the nominal device in sub-100 nm technologies is typically designed very aggressively. As a result, from the device perspective, we expect that the roll-off in any given device response below the nominal design would be significantly different from the one above the nominal design. A simple quadratic model is obtained for the gate delay response as a function of every process parameter in each of these regions. The piecewise quadratic model considered is of the form

$$y = \beta_0' + \beta_1' x_i + \beta_2' x_i^2 \tag{2}$$

where β_0' is a constant, x_i are the normalized process parameters, and β_i' are the corresponding regression coefficients determined by performing nonlinear regression analysis using LSM. The LSM model does not contain any interaction terms between process variables as it is carried out taking one process variable at a time and for all process variables independently.

The two-way and three-way interaction terms obtained from the DOE–RSM modeling are then superimposed on the quadratic models obtained by the LSM method to generate a hybrid model equation of the form set out in (1), whose constant, linear, and quadratic terms come from the LSM model, and the interaction terms come from the DOE–RSM model. This hybrid modeling approach ensures that one-parameter variations are tracked closely and, at the same time, interaction effects between process parameters are effectively captured. The significant computational benefits of this approach come from the selection of minimum number of levels needed for factor settings as 3 instead of 4. Also, the limitation of quadratic model

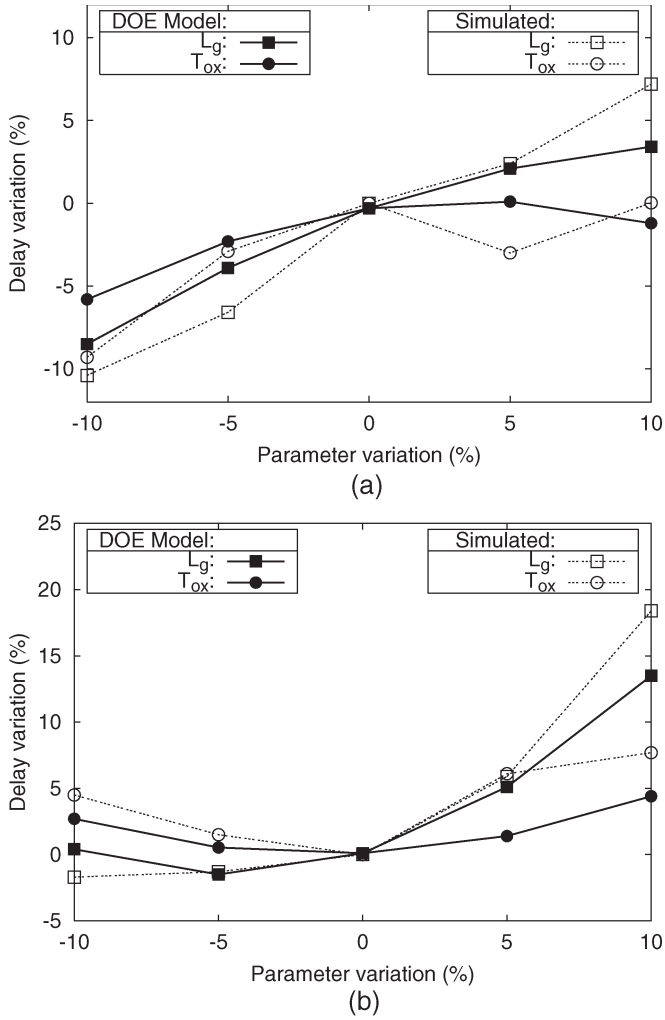

Fig. 5. Comparison of simulated and DOE-modeled delay variation with respect to process parameter variation. (a) Falling edge. (b) Rising edge.

TABLE V
LOOKUP TABLE OF NAND GATE DELAYS FOR VARIATIONS IN L_g (IN ps)

Process variation	A_{up}	A_{down}	B_{up}	B_{down}	AB_{up}	AB_{down}
-10%	7.54	7.58	6.56	6.79	4.26	8.49
-5%	7.57	7.90	6.60	7.06	4.29	8.80
Nominal	7.67	8.46	6.72	7.63	4.37	9.44
+5%	8.13	8.66	7.02	7.75	4.60	9.65
+10%	9.09	9.07	7.83	8.07	5.07	10.12

in fitting a cubic or higher order response observed is overcome by fitting the response with piecewise quadratic models to the two split regions. Although the model is somewhat heuristic, this novel idea proves to be computationally efficient and yet very effective, as will be demonstrated in Section IV.

The hybrid models for delay response variables have been tested for their validity to predict the response values by various residual plots, which are found to be satisfactory [20]. The plot of residuals and predicted response does not exhibit any pattern to the residuals.

TABLE VI
LIST OF HYBRID-MODELED RESPONSE VARIABLES AND THEIR CORRELATION COEFFICIENTS

Response variable	Correlation coefficient
a_{down}	0.898
b_{down}	0.886
ab_{down}	0.903
a_{up}	0.823
b_{up}	0.815
ab_{up}	0.806

Residuals are randomly scattered on either side of residual $= 0$ line, with approximately constant variance. The model R^2 statistic and R^2-adjusted statistic are found to be acceptable. Also, the original experimental response and the model-predicted response correlate reasonably well with the correlation coefficient $r > 0.80$. The list of response variables and their correlation coefficients are summarized in Table VI. This model accuracy is taken to be adequate considering that we have fitted cubic or higher order response effects with piecewise quadratic models using a 3-level FCCC design with some improvisation. The correlation plots for the gate delays along with their correlation coefficients r are shown in Fig. 6.

IV. DELAY DISTRIBUTIONS OF A DIGITAL CIRCUIT

A 4-bit $\times$ 4-bit Wallace tree multiplier circuit is designed using 2-input NAND gates as a library element. The input combination that results in worst case circuit delay is identified by full coverage of the input vector space, i.e., by applying all possible $(= 256)$ input combinations to the multiplier circuit, and then used in all subsequent simulations. The transient analysis of the circuit is carried out to obtain the circuit delay using the SEQUEL circuit simulator. The event-driven simulation capability of SEQUEL for gate-level simulation is used. The systematic variations are modeled by assuming that the process parameters vary as per Gaussian distribution with varying mean of 0%, +5%, and −5% of the nominal value. The random variations are modeled by assuming a $\pm 3\sigma$ variation of $\pm 5\%$ of the nominal value around respective mean values. The $\pm 5\%$ variation is only for the short-range process variation, and we treat it to be the subset of the worst case of $\pm 10\%$ variation. In other words, the delay variation of the NAND gates of this multiplier on any given chip is a subset of the overall NAND gate delay range, which is determined by the global variations in process parameters.

A probability distribution for the circuit delay in generating the multiplier output is obtained using the rigorous Monte Carlo simulations by randomly varying different process parameters individually and then simultaneously. A custom Monte Carlo code is written that treats each of the process parameters as an uncorrelated random input variable and is integrated with SEQUEL simulator. As a result, every NAND gate in the circuit obtains various process parameter values, as per the assumed Gaussian distribution of respective process parameters. It is presumed that the two NMOS and two PMOS devices constituting the NAND gate are closely spaced as to suffer identical process variations. The delay values for different transitions of different gates that take place in the circuit are assigned from the lookup table by applying linear interpolation. To guarantee accurate

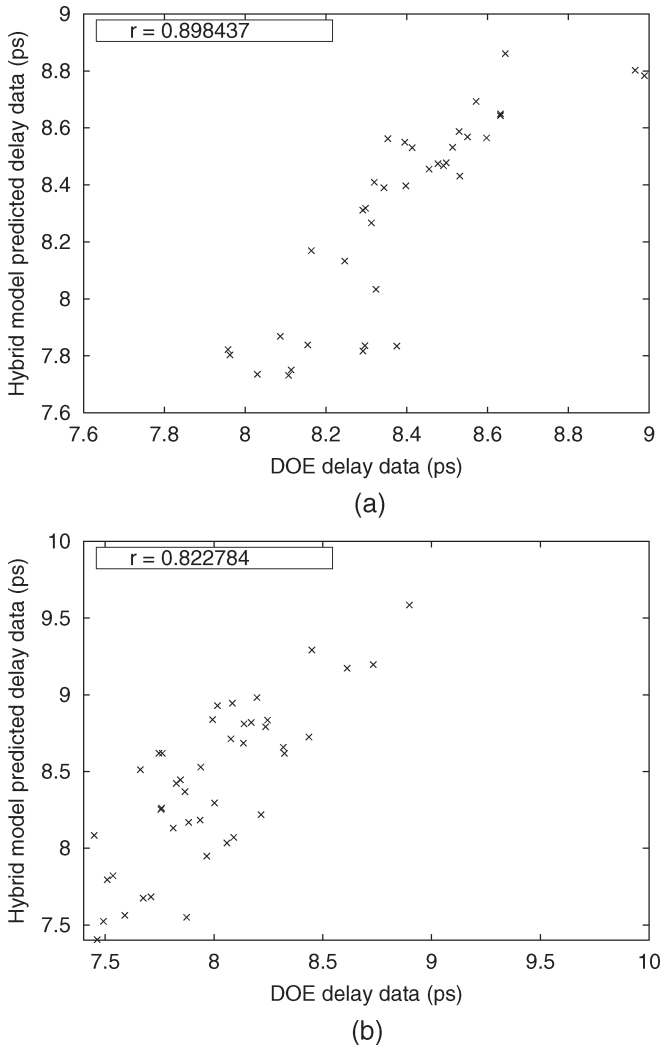

Fig. 6. Correlation plot of hybrid-model-predicted delay data versus simulated DOE data. (a) Falling edge. (b) Rising edge.

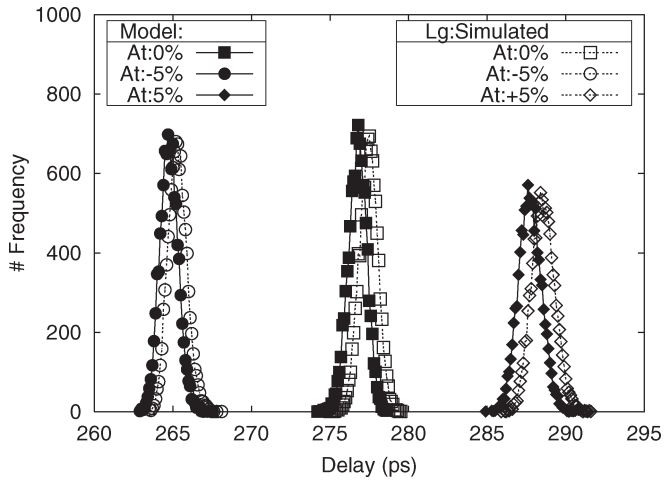

Fig. 7. Simulated and hybrid-modeled delay distributions for variations in gate length.

results at minimum computational cost, 10 000 Monte Carlo trials are performed. Figs. 7–9 show the delay distribution for variations in gate length L_g, oxide thickness T_{ox}, and halo dose, respectively, with each process parameter acting individually. The distribution obtained

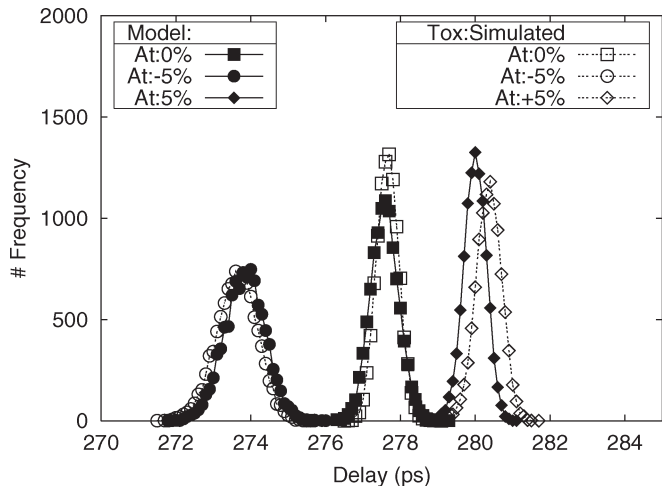

Fig. 8. Simulated and hybrid-modeled delay distribution for variations in gate oxide thickness.

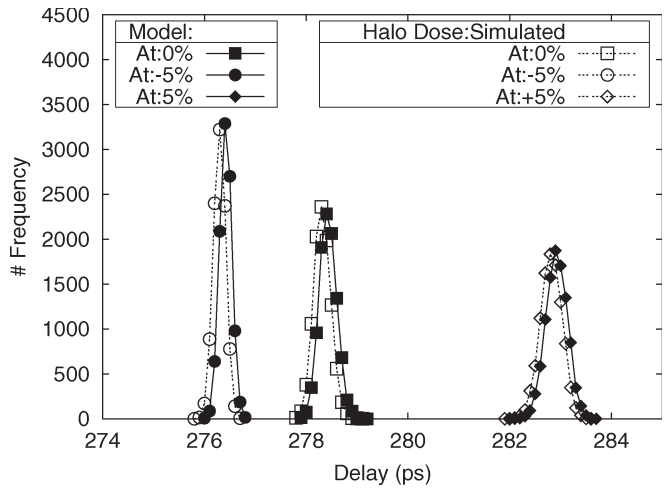

Fig. 9. Simulated and hybrid-modeled delay distributions for variations in halo dose.

using rigorous mixed-mode simulation delay values is overlaid with the distribution obtained using hybrid model approach. We observe a fairly good match for these two distributions. The statistics for variations in L_g, obtained by analyzing the resulting distributions, are presented in Table VII. Nominal delay is the circuit delay obtained when all the devices in the circuit have the nominal process parameter values. Similarly, best and worst delays are obtained when all the devices in the circuit have the best or worst process parameter values, respectively. The model statistics track the actual statistics extremely well, thus validating the hybrid model approach.

The methodology has been generalized to consider simultaneous variations in multiple process parameters. To begin with, for simplicity, simultaneous variations in two dominant process parameters are considered, and a large lookup table is generated that contains 25 delay values, corresponding to 25 device/circuit splits with nominal, $\pm 5\%$, and $\pm 10\%$ variations for two parameters. This is realized by generating all 25 pairs of NMOS/PMOS devices and performing mixed-mode simulations of NAND gates using these devices. Then, Monte Carlo simulations are performed by generating two uncorrelated random numbers for every NAND gate in the circuit, one for each parameter, as per the assumed statistics of process parameters. The delay values for different transitions of different gates that take place in the circuit are assigned from the lookup table by applying two-dimensional

TABLE VII
STATISTICS OF DELAY DISTRIBUTION OF L_g (IN PICOSECONDS)

Statistics	Simulated			Hybrid Model		
	At -5%	At 0%	At +5%	At -5%	At 0%	At +5%
Nominal delay	264.5	277.9	287.3	264.5	277.9	287.2
Distribution mean	265.3	277.4	288.6	264.8	276.8	287.7
Median	265.3	277.4	288.6	264.8	276.8	287.7
Std. deviation	0.5935	0.5877	0.7455	0.5866	0.5885	0.7347
Best delay	258.2	264.5	277.9	258.2	264.5	277.9
Worst delay	277.9	287.3	308.1	277.9	287.3	308.1

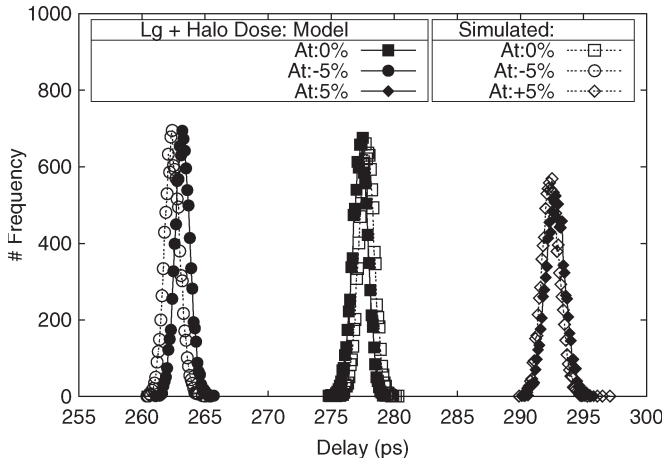

Fig. 10. Simulated and hybrid-modeled delay distribution for simultaneous variations in halo dose and gate length.

TABLE VIII
STATISTICS OF DELAY DISTRIBUTION OF TWO PARAMETERS: L_g AND HALO DOSE (IN PICOSECONDS)

Statistics	Simulated			Hybrid Model		
	At -5%	At 0%	At +5%	At -5%	At 0%	At +5%
Nominal delay	261.4	277.9	291.3	263.0	277.9	292.3
Distribution mean	262.4	277.8	292.3	263.2	277.3	292.8
Median	262.4	277.8	292.4	263.2	277.3	292.8
Std. deviation	0.6086	0.6099	0.7758	0.6049	0.6117	0.7793
Best delay	253.1	261.4	277.9	253.5	263.0	277.9
Worst delay	277.9	291.3	315.6	277.9	292.3	316.9

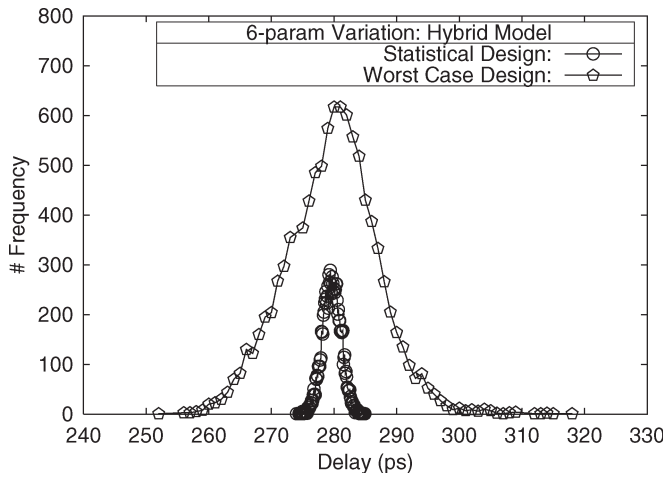

Fig. 11. Hybrid-modeled delay distribution with statistical design and worst case design for simultaneous variations in six parameters for $\pm 3\sigma = \pm 10\%$ at nominal.

interpolation. Then, delay distribution plots are obtained by the previous method and by using the hybrid model equations to generate gate delay values for simultaneous variations in L_g and halo dose. This parameter combination is selected as they are the significant process parameters from the perspective of variability. Delay distribution plots for simultaneous variations in L_g and halo dose are shown in Fig. 10, and their statistics are given in Table VIII. The results of hybrid model differ from that of mixed-mode simulation by less than 0.3% in terms of mean and by less than 0.6% in terms of standard deviation at their worst. This demonstrates that the hybrid model yields reasonably accurate results with less computational requirements, apart from being scalable to variations in multiple process parameters. The model statistics track the actual statistics extremely well, thus validating the hybrid model approach for simultaneous variations in two process parameters.

A rigorous verification for simultaneous variations in more than two process parameters requires 5^n device/circuit splits, where n is the number of process parameters that are varying simultaneously. Hence, the number of device/circuit splits required increases in a power series fashion, as n increases. With simultaneous variations in two $(n = 2)$ process parameters, the predictive ability of the hybrid model has been demonstrated. In other words, the hybrid model has adequately captured the interaction effects between process parameters. Since all interaction terms have come from the same single step of DOE–RSM modeling, it stands to reason to extend this methodology to multiple process variables.

The hybrid model can be used to gain some useful insight in timing analysis for design closure. Suppose that the worst case design methodology is used to obtain the delay spread of the multiplier circuit. Then, for simultaneous variations in six process parameters, the delay spread will be from 243.2 to 337.7 ps. Furthermore, the delay distribution obtained using Monte Carlo analysis is as shown in Fig. 11. The process parameters are assumed to vary by $\pm 10\%$. In the worst case methodology, all the NAND gates in the multiplier take identical set of process parameters for any given trial in the Monte Carlo loop. On the other hand, if we take the statistical design approach, each of the NAND gate can take a random set of process parameters in every trial. The distribution obtained using this methodology is overlaid in Fig. 11. We see that the worst case approach gives a standard deviation of 7.3 ps, whereas the statistical approach gives 1.5 ps. For simultaneous six-parameter variations, normalized delay variation is $(337.7 - 243.2)/6 = 15.75$ ps with traditional worst case design using the hybrid model. However, the corresponding value with statistical design using hybrid model would be $(284.9 - 274.0)/6 = 1.82$ ps. Clearly, the delay distribution is tighter by almost an order of magnitude, demonstrating the significance of statistical design with the hybrid model. The statistics are summarized in Table IX.

TABLE IX
STATISTICS OF DELAY DISTRIBUTION OF SIX PARAMETERS FOR $\pm 3\sigma = \pm 10\%$ AT NOMINAL (IN PICOSECONDS)

Statistics	Worst Case Design	Statistical Design
Nominal delay	277.9	277.9
Distribution mean	279.64	279.64
Median	279.97	279.63
Std. deviation	7.3080	1.4626
Best delay	243.2	274.0
Worst delay	337.7	284.9

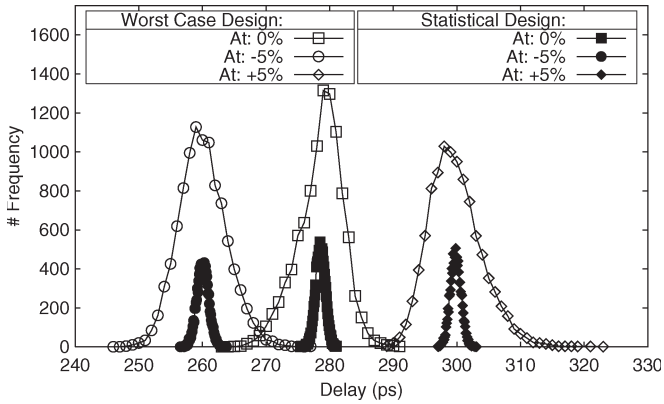

Fig. 12. Hybrid-modeled delay distribution with statistical design and worst case design for simultaneous variations in six parameters for $\pm 3\sigma = \pm 5\%$.

It should be noted that $\pm 10\%$ variation in process parameters includes systematic and random variations. Thus, using such a wide variation in itself is very pessimistic for the timing closure problem illustrated in Fig. 1. It is more realistic to consider that the process could be centered around either -5%, 0%, or $+5\%$ points, and there would be a random variation of $\pm 5\%$ around these points. We have again performed Monte Carlo analysis for this scenario to obtain delay distribution. Fig. 12 shows the distribution using worst case analysis and the statistical methodology. The worst case methodology results in very pessimistic design with a significantly higher standard deviation compared to statistical approach. Table X summarizes these results.

V. CONCLUSION

In the DSM regime, the deterministic circuit design approach may not be adequate to produce robust designs in the presence of severe process variations, and it becomes imperative that the circuit design adopts statistical approach. This paper presents one such statistical circuit design approach that takes into account variability in any number of process parameters, if their statistics are known. Two-input NAND gate has been used as a library element, and its delay is extensively characterized through mixed-mode simulations. An optimal second order hybrid model is obtained for gate delays directly in terms of process parameters through response surface modeling using DOE and LSM. The delay of a large digital circuit is characterized in statistical terms by taking a 4-bit $\times$ 4-bit multiplier as a representative circuit. We demonstrate that the worst case design approach is very pessimistic, whereas the hybrid model based statistical design approach can result in robust design. The proposed methodology has been demonstrated for NAND gate library with 266 gates, and the simplicity and generality of the approach make it equally applicable to a large library of cells for both statistical timing analysis and statistical circuit simulation at the gate level. This paper attempts to efficiently bridge the gap between the TCAD and design CAD through process simulations, mixed-mode device simulations, RSM, and a general-purpose circuit simulator.

TABLE X
STATISTICS OF DELAY DISTRIBUTION OF SIX PARAMETERS FOR $\pm 3\sigma = \pm 5\%$ (IN PICOSECONDS)

	Worst Case Design			Statistical Design		
Statistics	At -5%	At 0%	At +5%	At -5%	At 0%	At +5%
Nominal delay	260.1	277.9	299.3	260.1	277.9	299.3
Distribution mean	260.0	278.6	299.8	260.1	278.5	299.8
Median	259.9	278.9	299.4	260.1	278.5	299.8
Std. deviation	3.7337	3.5332	4.1289	0.9259	0.7615	0.8401
Best delay	243.2	260.1	277.9	256.5	275.4	297.0
Worst delay	282.3	305.0	325.9	263.8	281.1	303.1

ACKNOWLEDGMENT

The authors would like to thank D. Vinay Kumar, Department of Electrical Engineering, Indian Institute of Technology, Bombay, for the useful discussions regarding the SEQUEL simulator.

REFERENCES

[1] *The International Technology Roadmap for Semiconductors (ITRS) 2006 Updates-Process Integration, Devices and Structures (PIDS)*. [Online]. Available: http://public.itrs.net
[2] M. J. M. Pelgrom, A. C. J. Duinmaijer, and A. P. G. Welbers, "Matching properties of MOS transistors," *IEEE J. Solid-State Circuits*, vol. 24, no. 5, pp. 1433–1440, Oct. 1989.
[3] T. Karnik, S. Borkar, and V. De, "Sub-90 nm technologies—Challenges and opportunities for CAD," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Des.*, Nov. 2002, pp. 203–206.
[4] J. A. G. Jess, K. Kalafala, S. R. Naidu, R. H. J. M. Otten, and C. Visweswariah, "Statistical timing for parametric yield prediction of digital integrated circuits," in *Proc. Des. Autom. Conf.*, Jun. 2003, pp. 932–937.
[5] A. R. Alvarez, B. L. Abdi, D. L. Young, H. D. Weed, J. Teplik, and E. R. Herald, "Application of statistical design and response surface methods to computer-aided VLSI device design," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 7, no. 2, pp. 272–288, Feb. 1988.
[6] M. Rodder, A. Chatterjee, D. S. Boning, and I. C. Chen, "Transistor design with TCAD tuning and device optimization for process/device synthesis," in *Proc. Tech. Papers, Int. Symp. VLSI Technol., Syst. and Appl.*, 1993, pp. 29–33.
[7] G. J. Gaston and A. J. Watson, "The integration of simulation and response surface methodology for the optimization of IC processes," *IEEE Trans. Semicond. Manuf.*, vol. 7, no. 1, pp. 22–33, Feb. 1994.
[8] D. S. Boning and P. K. Mozumdar, "DOE/Opt: A system for design of experiments, response surface modeling and optimization using process and device simulation," *IEEE Trans. Semicond. Manuf.*, vol. 7, no. 2, pp. 233–244, May 1994.
[9] C. Michael and M. Ismail, "Statistical modeling of device mismatch for analog MOS integrated circuits," *IEEE J. Solid-State Circuits*, vol. 27, no. 2, pp. 154–166, Feb. 1992.
[10] M. Ismail and T. Fiez, *Analog VLSI: Signal and Information Processing*. New York: McGraw-Hill International Edition, 1994.
[11] P. M. Zeitzoff, A. F. Tasch, W. E. Moore, S. A. Khan, and D. Angelo, "Modeling of manufacturing sensitivity and of statistically based process control requirements for a 0.18 μm NMOS device," in *Proc. Int. Conf. Charact. and Metrol. ULSI Technol.*, Nov. 1998, pp. 73–81.

[12] H. C. Srinivasaiah and N. Bhat, "Mixed-mode simulation approach to characterize the circuit delay sensitivity to implant dose variations," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 22, no. 6, pp. 742–747, Jun. 2003.

[13] ISE TCAD Release 8.0, *DIOS and DESSIS Manuals, Integrated Systems Engineering*, Zurich, Switzerland. [Online]. Available: http://ise.ch

[14] M. B. Patil. (2001, May). *SEQUEL User's Manual–Digital Basic and Compound Elements*. [Online]. Available: http://www.ee.iitb.ac.in/~microel/faculty/mbp/sequel1.html

[15] J. R. Pfiester, L. C. Parrillo, M. Woo, H. Kawasaki, B. Boeck, E. Travis, and C. Gunderson, "An integrated 0.5 μm CMOS disposable TiN LDD/salicide spacer technology," in *Proc. IEDM Tech. Dig.*, Dec. 1989, pp. 781–784.

[16] B. P. Harish, N. Bhat, and M. B. Patil, "Analytical modeling of CMOS circuit delay distribution due to concurrent variations in multiptle processes," *Solid State Electron*, vol. 50, no. 7–8, pp. 1252–1260, Jul./Aug. 2006.

[17] G. E. P. Box and N. R. Draper, *Empirical Model-Building and Response Surfaces*. New York: Wiley International Edition, 1987.

[18] G. E. P. Box, W. G. Hunter, and J. S. Hunter, *Statistics for Experimenters: An Introduction to Design, Data Analysis and Model Building*. New York: Wiley International Edition, 1978.

[19] R. H. Myers and D. C. Montgomery, *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. New York: Wiley International Edition, 2002.

[20] NIST/SEMATECH, *e-Handbook of Statistical Methods,* ch. 5, Process Improvement. [Online]. Available: http://www.itl.nist.gov/div898/handbook/index.htm