A New TCAD-Based Statistical Methodology for the Optimization and Sensitivity Analysis of Semiconductor Technologies

S. Williams and K. Varahramyan, *Member, IEEE*

***Abstract*—A new TCAD-based statistical methodology for the optimization and sensitivity analysis of semiconductor technologies has been developed and demonstrated on a 0.18-μm SOI CMOS process. Two new screening techniques applicable to deterministic systems (Lenth's test and normal probability plots) were introduced and compared with the correlation analysis described in [1]. The graphical nature of the new techniques provided easier analysis of the screening results by clearly displaying which process factors surpass predefined significance limits. A multiresponse steepest ascent analysis was developed to locate regions of improved process performance before beginning response surface experimentation. To perform the analysis, a composite function representing the response criteria was constructed using desirability functions and incorporated within the steepest ascent methodology. Locating the region of improved performance allowed smaller experimental designs to be used for the response study significantly improving model accuracy. The response models were used to optimize the SOI CMOS process and perform sensitivity analyzes on both the baseline and optimized processes. Optimization resulted in a 15% increase in $I_{d\mathrm{sat}}$ without violating any other criteria. The results of the sensitivity analyzes, which showed the greatest benefit from the increased model accuracy, indicated no conspicuous device performance degradation caused by anticipated manufacturing variations.**

***Index Terms*—Design for manufacturability, semiconductor technology optimization, sensitivity analysis, statistical methodology, TCAD.**

I. Introduction

GROWING competition within the semiconductor industry has forced process engineers to develop advanced optimization strategies to reduce device development time while increasing process yield. A great deal of focus is also being placed on designing robust processes insensitive to fluctuations in the process conditions during manufacturing. This concept is known as design for manufacturability (DFM). In the last ten years, process designers have begun to incorporate statistical design of experiment methods within the TCAD environment to generate simple empirical models, referred to as response surface models (RSM's), relating process input conditions to key device parameters. Generated over a small design space, such models provide a mathematically efficient tool for performing process optimization and sensitivity analyzes.

Manuscript received March 4, 1999; revised October 25, 1999.

S. Williams is with Silvaco Data Systems, Austin, TX 78759 USA.

K. Varahramyan is with the Electrical Engineering Program, Louisiana Tech University, Ruston, LA 71272 USA (e-mail: kody@coes.latech.edu).

Publisher Item Identifier S 0894-6507(00)03547-8.

Previous papers have discussed basic response surface methodology [2]–[8]; however, specific application issues encountered when incorporating these methodologies within the TCAD environment have not been sufficiently addressed. Many of the statistical techniques used in conjunction with RSM were derived to analyze stochastic systems in which experimental error is present. TCAD process and device simulations, however, are deterministic in nature, and therefore, such statistical techniques may not be directly applied. In the past, process designers have avoided these issues by only using those RSM methods directly applicable to deterministic systems. Other designers have attempted to work around these problems by adding experimental error to their number of necessary simulations and raises questions regarding how to accurately simulate experimental error distributions. Another approach would be to assume that certain higher order effects are negligible and use the sum of their effect estimates to approximate experimental error. In the early stages of a response surface study, it may not be clear if such an assumption is valid. Additional design points would also be needed to estimate these higher order effects.

In this paper, we propose a systematic, statistically based approach to ULSI process optimization and manufacturing sensitivity analysis through the application of advanced response surface methodologies within the TCAD environment. Our approach is sequential involving four primary steps. First, statistical screening methods are used to identify those input conditions having the greatest effect on the device parameters. We introduce two new screening techniques directly applicable to deterministic systems and compares these techniques with one reported previously [1]. Second, the method of steepest ascent is employed to locate the region of improved process performance. We have developed a new multiresponse steepest ascent analysis to locate such a region for processes or systems involving more than one response. The third and fourth steps consist of second-order response modeling and process optimization and sensitivity analysis, respectively. We provide a general overview of these techniques and discuss some previously unexamined design considerations specific to deterministic studies. The effectiveness of the proposed approach is illustrated through the design and analysis of a 0.18-μm partially depleted SOI CMOS technology representative of technology currently under consideration by industry.

II. Methodology

A. Factor Screening

Time constraints often limit the feasibility of performing a response surface study in which all possible input conditions are included. Statistical factor screening methods may be employed at the onset of a response surface study to identify those conditions having the greatest effect on the output parameters. By considering only those conditions showing a significant effect, the subsequent response surface study may be streamlined, and fewer runs or tests required [9]. Because of the deterministic nature of TCAD simulation, analysis of variance techniques, commonly used to screen stochastic systems, may not be applied to simulated data. Because of the widespread assimilation of TCAD tools throughout the semiconductor industry, it is important that factor screening methods applicable to deterministic systems be identified.

This work introduces two new screening techniques directly applicable to deterministic systems, Lenth's test, and normal probability plots and compares them to the correlation analysis in [1]. Though only one method is needed to successfully screen an experiment, employing a second or third method may be useful in verifying results or aiding in decision-making. Because the methods presented here only consider the main effect of each input factor when determining significance, a two-level statistical design is used to efficiently explore the experimental design space. Two-level factorial and fractional factorial designs are ideally suited for such experiments [3]. To help minimize the number of experimental points, Plackett–Burman designs may be employed for processes involving a large number of input factors. Such designs should be used with care because of their complex alias structure. For an in-depth discussion of two-level fractional factorial and Plackett–Burman designs, see Montgomery [10].

Lenth's test is a parametric test for determining factor significance. Significance is established graphically by plotting calculated main factor effects against defined cutoff limits. The design analyst may easily assess the size and significance of each effect from this plot [11]. A main factor effect represents the change in output produced by a change in level of an individual input factor. For a two-level experimental design, the main effect of factor P on the output C is [10]

$$C_P = \frac{2}{n}\left(\sum_{P_{\mathrm{HIGH}}} Y - \sum_{P_{\mathrm{LOW}}} Y\right) \tag{1}$$

where Y is the output value and n is the number of experimental runs. Mathematically, this equation represents the average value of C for all experimental runs with P set high minus the average value of C for all experimental runs with P set low. To conduct Lenth's test, the main factor effects of each process condition are calculated for every device parameter, and cutoff limits for factor significance are determined as follows. First, let

$$s_o = 1.5 \times \mathrm{median}|C_p| \tag{2}$$

where p is the number of process conditions and C_p are the p main factor effects. The pseudostandard error of the effect estimates is then defined as [11]

$$\mathrm{PSE} = 1.5 \times \mathrm{median}|C_p| \qquad \text{for } |C_p| < 2.5 s_o. \tag{3}$$

The margin of error is defined as [11]

$$\mathrm{ME} = t_{0.025,\,df} \times \mathrm{PSE} \tag{4}$$

where $t_{0.025,\,df}$ is the 0.975th quantile of the t-distribution (a statistical sampling distribution [10]) and df represents the experimental degrees of freedom. Lenth suggests using $df = p/3$. A more stringent measure of margin of error, the simultaneous margin of error, is also defined as [11]

$$\mathrm{SME} = t_{\gamma,\,df} \times \mathrm{PSE} \tag{5}$$

where

$$\gamma = \frac{(1 + 0.95^{1/p})}{2} \tag{6}$$

and $t_{\gamma,\,df}$ is $(1-\gamma)$th quantile of the t-distribution on df. Graphically, the main factor effects are presented as a bar graph with overlaying reference lines representing $\pm$ME and $\pm$SME. Any effect not exceeding $\pm$SME is strongly significant. Any effect not exceeding $\pm$ME are insignificant. Those effects lying between the two reference lines cannot be clearly identified and may be included in further experimentation [11].

Factor significance may also be determined by analyzing the normal probability plots of the main factor effects. This method assumes all insignificant effects are normally distributed about a mean zero with a constant variance. Significant effects share the zero mean, but they have an inflated variance [15]. When prepared as a normal probability plot, insignificant effects lie along a straight line, whereas significant effects appear as outliers deviating from the line [10]. This method has found wide use for the analysis of unreplicated experimental designs because it is relatively easy to prepare and its results can be presented graphically. A normal probability plot of the main factors effects for an individual response may be generated by: 1) calculating all main effects; 2) sorting the main effects in ascending order; 3) calculating the cumulative frequency for each effects (y coordinate); 4) plotting the data pairs; and 5) assessing the significance of the plotted main effects [15]. The cumulative frequency for the main effects may be calculated as [10]

$$y_k = 100(k - 0.5)/10 \tag{7}$$

where k varies from one to p. The straight line, chosen subjectively, is drawn through the plotted points representing the normal line [10].

As presented in [1], if two variables are linearly related, as is assumed during screening experimentation, correlation coefficients may be used to quantify the relationship present between the variables. By comparing the relative coefficient values, we can determine the relative significance of the process conditions. Mathematically, correlation coefficients are expressed as [12]

$$R_{ij} = \frac{1}{n}\sum_{k=1}^{n} Q_k^i Q_k^j \tag{8}$$

where

n number of experimental runs;
Q_k^i kth standardized value of the ith process condition;
Q_k^j kth standardized value of the jth condition.

The process condition values are standardized to eliminate scale effects introduced into experimentation by the units of measure [13]. Process factor P is standardized by [12]

$$Q_i = \frac{P_i - \mu(P)}{\sigma(P)} \tag{9}$$

where

P_i ith process factor value;
$\mu(P)$ mean value of P;
$\sigma(P)$ standard deviation of P.

Correlation coefficients vary from negative one to positive one, with the coefficient magnitude indicating the strength of correlation. A value of positive one indicates perfect positive correlation; negative one indicates perfect negative correlation [14]. Factor significance is based on two criteria: 1) significant correlation ($R > 0.13$) between an input condition and a majority of the output responses, or 2) strong correlation ($R > 0.5$) between an input condition and a single output response.

B. Process Improvement by Multiresponse Steepest Ascent Analysis

When using a response surface study to both optimize a process and conduct a sensitivity analysis, a compromise must be made in the size of the design area. Using a large area will help ensure the optimal operating conditions fall within the design space, but it will reduce response surface model accuracy, which is detrimental to the sensitivity analysis. In this work, a multiresponse steepest ascent analysis was developed to locate the general region of improved process performance before the response surface experiment was conducted. Steepest ascent analysis uses a series of simple statistical experiments involving linear response models to move the design space toward the region of improved performance. Using such a procedure allows the experiment designer to construct a smaller RSM design space and maintain model accuracy. The multiresponse approach we developed incorporates a composite response function, representative of all selected performance criteria, within the steepest ascent methodology to locate the region of improved process performance.

Because of the sequential nature of the method of steepest ascent, design economy and simplicity are important. For this reason, it is initially assumed that a linear response model is an adequate approximation of process behavior, which typically exhibits linear behavior away from a maximum or minimum point [9]. A resolution III or higher fractional factorial experiment (see Montgomery [10]) is conducted about the initial design point. This point may represent the base processing conditions established during screening experimentation, or it may be arbitrarily set by the experiment designer. From the simulated experimental data, a linear response model approximating the composite response function is generated and a path of steepest ascent is computed. A series of single-run experiments are conducted along the path until an approximate maximum, is located. The experimental process is repeated about the approximate maximum and a new linear model is generated. A second path of steepest ascent based on the new model is computed, and another set of single-run experiments is conducted along its path until a second approximate maximum is located. This entire procedure is continued until any noticeable response improvement diminishes. At each stage of experimentation, model adequacy is checked to ensure that the linear assumption holds. Violation of this assumption indicates that the region of improved performance has been located and that the point of optimum performance is near. The subsequent response model experimentation is then centered about the last approximate maximum found.

Desirability functions were used to construct the composite response function for our multiresponse steepest ascent analysis. The composite response function is defined by combining individual desirability functions (d) representing the optimization criteria for each response into a joint desirability function (D) [16]. The individual desirability functions are determined depending on whether the response is to be maximized or minimized, or if a target response value is desired. The two-sided desirability function given by [9]

$$d = \begin{cases} \left(\dfrac{y-A}{B-A}\right)^{\alpha}, & A \le y \le B \\ \left(\dfrac{y-C}{B-C}\right)^{\beta}, & B \le y \le C \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

is used to locate target response values. Here, y represents the output value, A is the lower bound, B is the response target value at which $d = 1$, and C is the upper bound. The exponents, α and β, are set to weight the individual response. A graphical depiction of the two-sided desirability function is presented in Fig. 1 [9]. One-sided desirability functions are used for maximizing or minimizing a response. They are obtained by taking the appropriate side of the two-sided function and defining any response outside that region as zero or one accordingly. The joint desirability value is usually computed as the geometric mean of the individual desirability values at the ith experimental design point, or [9]

$$D_i = \left(d_i^1 \cdot d_i^2 \cdot \cdots d_i^m\right)^{1/m} \tag{11}$$

where m is the number of output responses. A joint desirability function is generated by fitting a linear response model to the computed values of D.

Though the definition of the individual desirability functions stems from nonlinear functions, at the onset of experimentation, it is not clear whether the joint desirability will behave linearly. We begin by assuming linear behavior, fit the linear response model, and then use statistical measures-of-fit to verify the assumption. Should the assumption fail, two options are available. First, a nonlinear data transformation, such as $\log(y)$, could be applied to the joint desirability data before fitting the linear model, or secondly, an alternate definition for joint desirability could be explored. For example, joint desirability could be calculated from the average of the individual desirability values rather than by finding their geometric mean. If a good linear

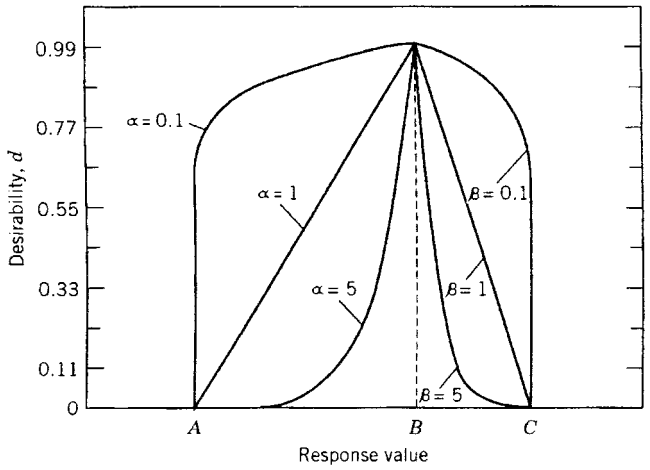


Fig. 1. Two-sided desirability functions with varying values of exponents α and β [9].

approximation cannot be found, the multiresponse steepest ascent analysis cannot be used and the response surface experiment begun at the initial design point.

The path of steepest ascent is directly dependent on the magnitude and direction of the linear response model coefficients. Movement along the path is proportional to the magnitude of the coefficients and acts in the direction of their signs. The path of steepest ascent is determined by finding the partial derivatives of a Lagrange multiplier associated with D with respect to x_j $(j = 1, 2, \cdots, k)$ of [9]

$$L = b_o + b_1x_1 + b_2x_2 + \cdots + b_kx_k - \lambda\left(\sum_{i=1}^{k} x_i^2 - r^2\right) \quad (12)$$

where

- b_i ith coefficient value of the linear response model representing D;
- x_i ith process input factor;
- λ constant (defined later);
- r^2 model error.

Note, the coefficient and process input factor values used during the steepest ascent analysis are standardized [see (8)] and are therefore unitless. Setting the partial derivatives to zero, the coordinates of the path of steepest ascent are given by [9]

$$x_j = \frac{b_j}{2\lambda} \quad (13)$$

where $1/2\lambda$ is a constant of proportionality we will call ρ. The constant ρ is arbitrarily selected and used to set the experimental step size along the path of steepest ascent.

C. Response Model Generation

Because a large amount of work (e.g., Myers [9]) is readily available on response model generation, we will only provide a brief overview. For most applications, RSM's are limited to second-order polynomials of the form [9]

$$Y \approx b_0 + \sum_{i=1}^{p} b_ix_i + \sum_{i=1}^{p} b_{ii}x_i^2 + \sum_{i=1}^{p}\sum_{j\neq i}^{p} b_{ij}x_ix_j \quad (14)$$

where

- p number of input factors;
- x_i ith input factor;
- b_i ith regression coefficient.

Response models are limited to second-order polynomials for two reasons: Higher order polynomials exhibit oscillatory behavior between data points, and lower order models permit the use of more efficient experimental designs. Such designs require fewer experimental points, greatly reducing the total number of simulations [4]. Model coefficient values are determined using a least-squares fit [9].

The CCD is perhaps the most popular design for fitting second-order response models and has been the predominate choice of experimenters in the semiconductor field. Combining a two-level full factorial or fractional factorial design of n_f runs with $2p$ axial runs and n_c center runs, they provide five input levels for each process input factor and offer great flexibility to the experiment designer [10]. We will not go into great detail about CCD construction (see Myers *et al.* [9] or Montgomery [10]), but one topic should be addressed. For CCD's, the input factors levels for the factorial portion of the experiment are set to ± 1 (standardized, see Section II-A), and the input levels for the axial points are set according to $\alpha = (n_f)^{1/4}$ to maintain model "rotatability" [10]. For deterministic experiments, however, in which only one center point exists, setting the axial levels to ± 1, a face center cube (FCC), provides better design stability by distributing the prediction variance more evenly. This is illustrated in Fig. 2 [9], which compares the scaled prediction variance distributions of a rotatable CCD design and a FCC design with only one center point each.

D. Process Optimization and Manufacturing Sensitivity Analysis

Process optimization is achieved through numerical analysis of the generated response surface models. Optimal operating conditions are determined by minimizing an objective function of the form [12]

$$G(\xi) = \frac{\sum_{i=1}^{m}\left\{\dfrac{w_i[y_{oi} - y_i(\xi)]}{\Delta_i}\right\}^2}{\sum_{i=1}^{m} w_i^2} \quad (15)$$

with respect to the weighted difference between the performance targets and predicted device parameters. Here, m represents the number of output responses, w_i is the assigned weight for the ith response, y_{oi} is the ith designated response target, $y_i(\xi)$ is the ith model value, ξ is the vector of input factors, and Δ_i is the response magnitude calculated as $3.92\sigma_i$ (σ_i represents the standard deviation of the ith response). The process conditions are constrained within the experimental design space boundaries [5], [12], [16].

Manufacturing sensitivity is determined by statistical analysis of the generated RSM's. We use a similar approach to that proposed by Hasnat *et al.* [6] and further developed by Angelo *et al.* [8], in which Monte Carlo techniques are employed to ascertain the statistical distribution for each device parameter re-

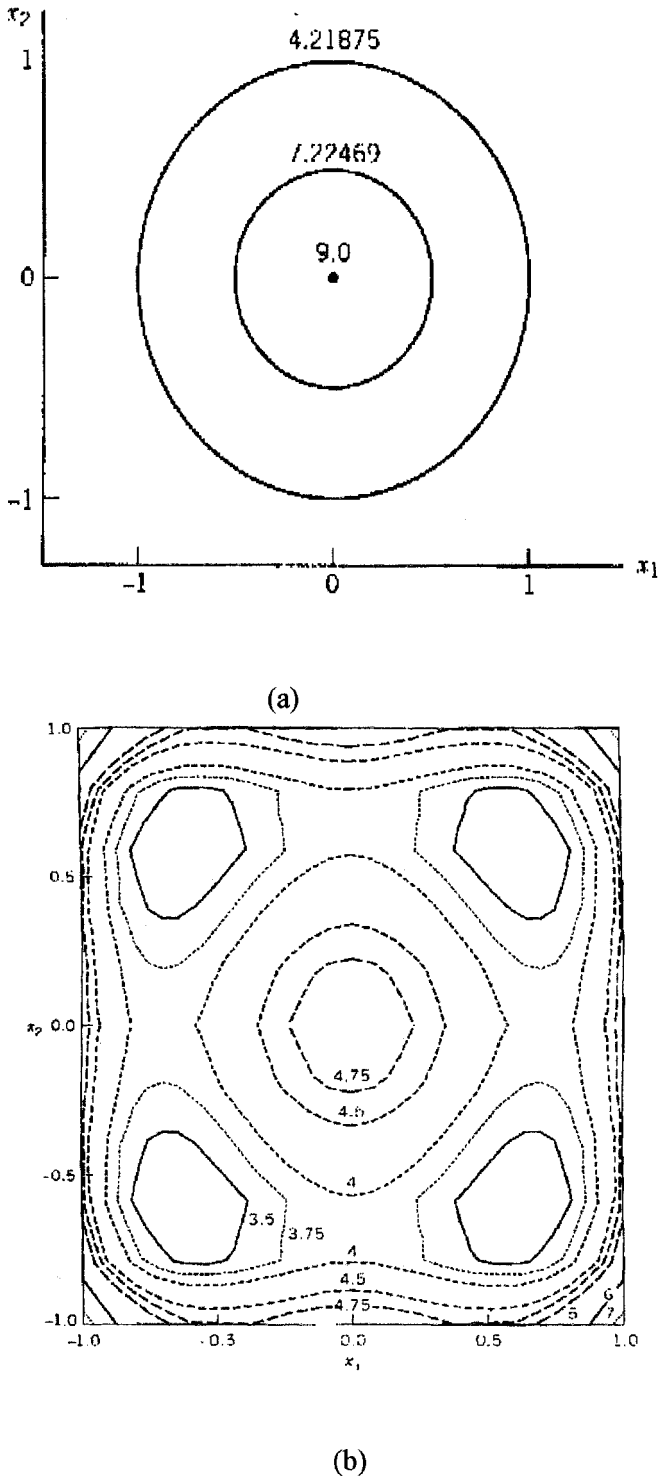


Fig. 2. Scaled prediction variances for (a) a two-factor central composite design (CCD) with one center point and (b) a two-factor face center cube (FCC) with one center point [9].

sulting from random variations in the process input conditions. The input conditions are assumed to vary normally about their specified process condition value, or mean. Their standard deviation is set as a percentage of this value. The resulting statistical distributions provide an estimate of expected performance variation under normal manufacturing conditions [17].

III. Application to 0.18-μm SOI CMOS Technology

Silicon-on-insulator-based technologies have emerged as a promising alternative to bulk CMOS technologies because of their reduced short-channel effects, higher transconductances, and lower power consumption. SOI wafers consist of three distinct layers. The top layer is a thin film of single-crystalline silicon on which devices are fabricated. Sandwiched between this layer and the silicon substrate rests an insulating layer usually composed of silicon dioxide. This layer commonly referred to as the buried oxide, or BOX, layer allows for dielectric isolation between neighboring devices.

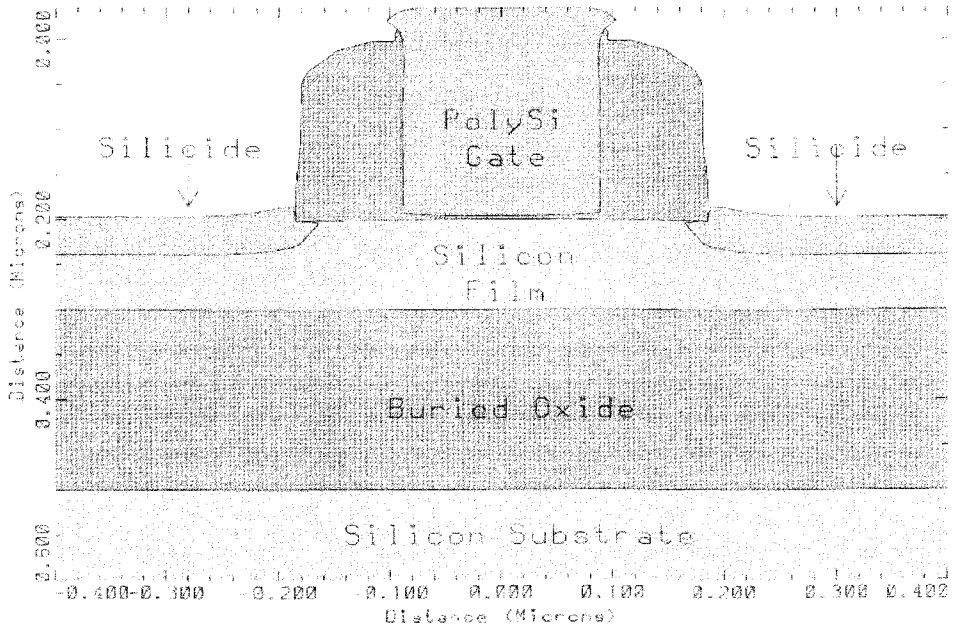
The 0.18-μm SOI CMOS process was implemented with the aid of two-dimensional (2-D) process (TSUPREM-4) and device (MEDICI) simulations. The structural features of the devices include a silicon film thickness of 0.10 μm, a gate oxide thickness of 3.5 nm, a MDD source/drain structure, and separate threshold voltage adjust and punch-through implants. The source/drain structures for the separate NMOS and PMOS devices were formed using As and BF_2, respectively. The two respective channel implants were formed using a combination of BF_2 and B for the NMOS device and P and As for the PMOS. The source, drain, and gate electrodes were silicided using a self-aligned titanium disilicide ($TiSi_2$) process. Fig. 3(a) and (b) show the simulated cross sections for the 0.18-μm SOI NMOS and PMOS transistors, respectively.

The TMA WorkBench (TWB) [12] was used to integrate the process and device simulation tools, construct the experimental designs, and automate simulation of the experimental points. TWB also provides utilities for generating the RSM's and performing process optimization. TSUPREM-4 and MEDICI are incorporated within TWB as "device drivers" to perform the necessary 2-D process and device simulations. A separate program was written to perform Lenth's test and generate simplified response surface models. The Monte Carlo analysis was conducted using Mathsoft's Mathcad, a general-purpose mathematics program.

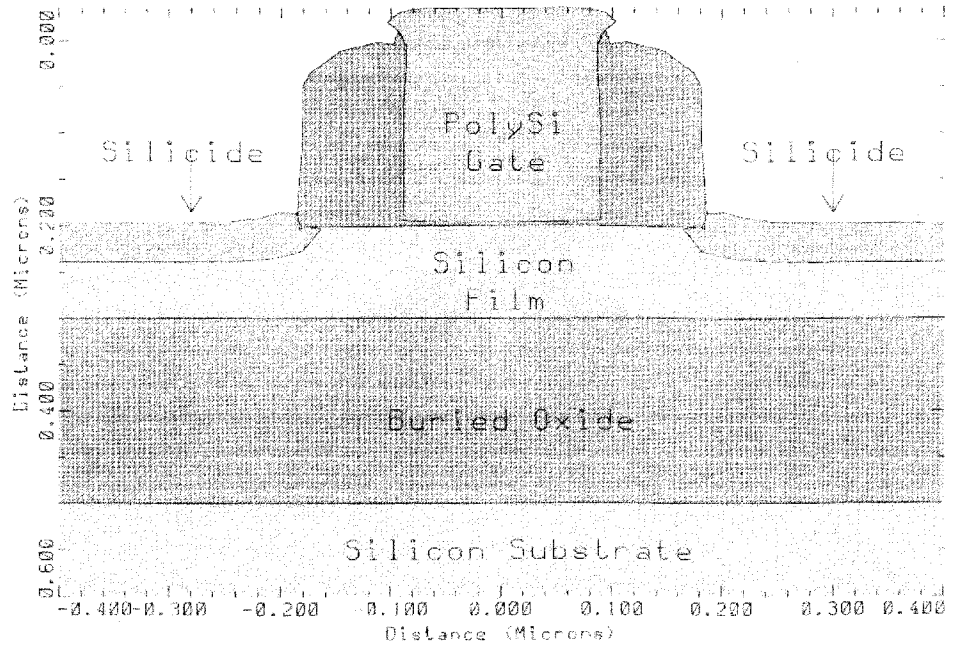
A comprehensive effort was undertaken in realizing process and device models for both NMOS and PMOS devices, with an emphasis on defining and using the necessary models for realistic simulation of the process and device conditions. As part of this effort, carrier and lattice heating effects were accounted for using a self-consistent solution of the Poisson, carrier continuity, energy-balance, and heat flow equations. An important part of this work consisted of the incorporation of quantum mechanical effects within the simulation modules to more accurately determine the device characteristics [1].

IV. Results and Discussion

In evaluating the 0.18-μm PD SOI CMOS technology, five device parameters were considered: threshold voltage ($V_{\rm th}$), subthreshold slope (S), off-state current ($I_{\rm off}$), drain saturation current ($I_{d\rm sat}$), and drain-induced barrier lowering ($\Delta V_g|_{\rm DIBL}$). Threshold voltage was determined using the linear extrapolation method. Off-state current was defined as drain current (I_d) observed for gate voltage (V_g) of 0 V and drain voltage (V_d) of ± 1 V ($+$ for NMOS and $-$ for PMOS). Drain saturation current was defined as I_d for V_g and $V_d = \pm 1.0$ V.



(a)



(b)

Fig. 3. Simulated cross sections of (a) 0.18-μm partially depleted SOI NMOS and (b) 0.18-μm partially depleted SOI PMOS.

Drain-induced barrier lowering was defined as the difference in V_g determined at $I_d = \pm 1$ nA/μm for $V_g = \pm 0.05$ and ± 1.0 V.

A. Factor Screening

Thirty process conditions were initially identified from the 0.18-μm SOI CMOS process as candidates for factor screening. For statistical experimentation, the overall process was separated into its NMOS and PMOS components. The necessary statistical data was generated using two 24-run Plackett–Burman experimental designs. Following experimentation, all three factor screening methods were applied to the data. The maximum number of factors for the response surface study was limited to eight per process component.

Overall, the three screening methods showed agreement in identifying the significant input factors. Gate oxidation time and temperature, as well as gate length, were identified to be strongly significant for both process flows by all three methods. Punch-through implant energy and dose were identified for the NMOS process by all three methods, whereas punch-through energy and the threshold voltage adjust dose were identified for the PMOS. Further analysis of the individual screening results suggested including two additional input factors for the NMOS process, threshold voltage dose and spacer width, and three more for the PMOS, punch-through dose, spacer width, and the drain extension implant dose. Table I summarizes the identified significant process conditions, and indicates which screening method identified each. Fig. 4(a) and (b) present Lenth's test results and the normal probability plots for $V_{\rm th}$ of the NMOS process flow, respectively.

B. Process Improvement by Multiresponse Steepest Ascent Analysis

Before beginning process improvement, five performance criteria were established for process optimization: $|I_{d\rm sat}|$ maximized, $|I_{\rm off}| \leq 1$ nA/μm, $|V_g|_{\rm DIBL}| < 100$ mV, $|V_{\rm th}| \leq 350$ mV, and $S < 85$ mV/dec. Based on these criteria and the response values observed during factor screening experimentation, the following joint desirabilities were defined as

$$D_{\rm NMOS} = (d_{I_{d\rm sat}} \cdot d_{I_{\rm off}})^{1/2} \tag{16}$$

where

$$d_{I_{d\rm sat}} = \left(\frac{I_{d\rm sat} - 200}{100}\right)^2, \qquad d_{I_{\rm off}} = \left(\frac{\log(I_{\rm off}) - 1}{-1}\right)^2.$$

Initial experimentation with the PMOS device indicated that the linear model was not adequate for approximating the joint desirability data generated using the geometric mean. Experimenting with alternate definitions of joint desirability, we discovered that a good linear approximation could be fit to the averaged individual desirabilities, or

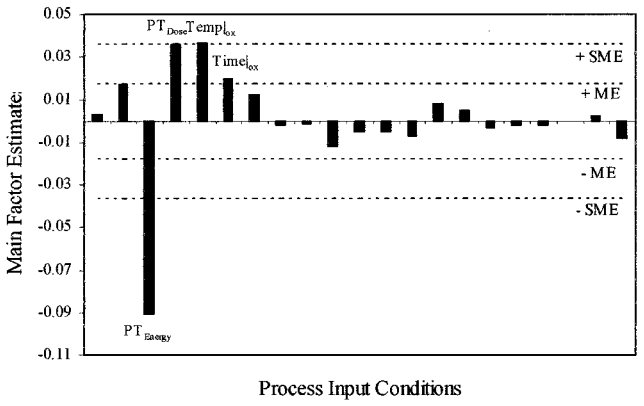
$$D_{\rm PMOS} = \frac{(d_{I_{d\rm sat}} + d_{I_{\rm off}})}{2} \tag{17}$$

where

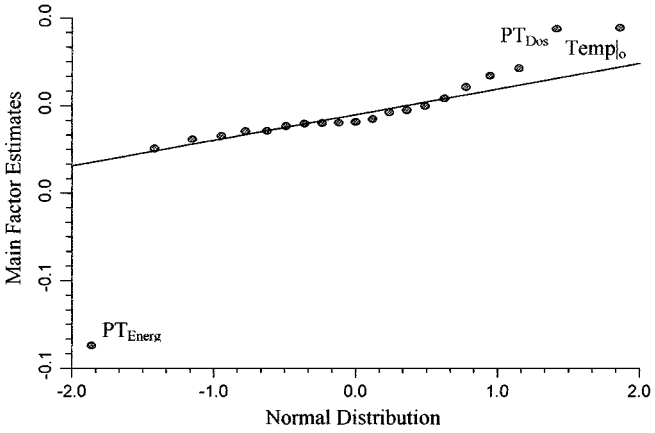
$$d_{I_{d\rm sat}} = \left(\frac{|I_{d\rm sat}| - 85}{45}\right), \quad d_{I_{\rm off}} = \left(\frac{\log|I_{\rm off}| - 1}{-2}\right).$$

Individual desirability functions for $V_{\rm th}$, S, and $\Delta V_g|_{\rm DIBL}$ were omitted to simplify the composite functions, but their values were continually monitored to ensure they did not violate the established optimization criteria.

As highlighted in Table II(a) and (b), two steepest ascent steps were found necessary for the NMOS device and three for the PMOS to locate the region of improved performance. For the given conditions, an eight-run fractional factorial design, including all seven process factors, was conducted at each step for the NMOS process, whereas an eight-run fractional design of only five factors ($\rm Temp|_{ox}$, $\rm Time|_{ox}$, and L_g were omitted) was used for the PMOS process. A complete study involving both $\rm Temp|_{ox}$ and $\rm Time|_{ox}$ would have been preferred so that shifts in the gate oxidation conditions could be explored; however, including these addition factors would require twice the number of simulations. As a compromise, $\rm Temp|_{ox}$ and $\rm Time|_{ox}$ were omitted from this portion of the PMOS study with the understanding that the NMOS results would guide the movement of these conditions. Variations in gate oxidation time and temperature were considered for both devices in the subsequent process



(a)



(b)

Fig. 4. Results from (a) Lenth's test for factor significance and (b) normal probability plot of main factor effects for $V_{\rm th}$ of the 0.18-μm SOI NMOS device.

TABLE I
IDENTIFIED PROCESS CONDITIONS DISPLAYING A SIGNIFICANT EFFECT ON 0.18-μm SOI CMOS DEVICE

NMOS		PMOS	
Process Condition	**Identifying Method**	**Process Condition**	**Identifying Method**
$V_{th}\|_{Dose}$	CC, LT, and NPP	$V_{th}\|_{Dose}$	CC, LT, and NPP
PT_{Energy}	CC, LT, and NPP	PT_{Energy}	CC, LT, and NPP
PT_{Dose}	CC, LT, and NPP	PT_{Dose}	CC and LT
$Temp\|_{ox}$	CC, LT, and NPP	$Temp\|_{ox}$	CC, LT, and NPP
$Time\|_{ox}$	CC, LT, and NPP	$Time\|_{ox}$	CC, LT, and NPP
L_g	CC, LT, and NPP	L_g	CC, LT, and NPP
W_{sp}	CC, LT, and NPP	W_{sp}	CC and LT
		MDD_{Dose}	LT

TABLE II
(a) STEP-BY-STEP PROCESS IMPROVEMENT BY MULTIFACTOR STEEPEST ASCENT ANALYSIS FOR 0.18-μm SOI NMOS. (b) STEP-BY-STEP PROCESS IMPROVEMENT BY MULTIFACTOR STEEPEST ASCENT ANALYSIS FOR 0.18-μm SOI PMOS

	Process Conditions				Device Parameters					
Step	$V_{th\|Dose}$ (cm^{-2})	PT_{Energy} (keV)	PT_{Dose} (cm^{-2})	W_{sp} (nm)	I_{dsat} (nA/μm)	I_{off} (nA/μm)	V_{th} (mV)	S (mV/dec.)	$\Delta V_{g\|DIBL}$ (mV)	D
0	1×10^{12}	30	1×10^{13}	90	240	0.724	291	78.20	66.4	0.400
1	9.6×10^{11}	33.8	1.2×10^{13}	63	277	2.116	269	76.66	68.7	0.519
2	9.8×10^{11}	31.3	1.28×10^{13}	62	264	0.729	294	77.44	71.0	0.640

(a)

	Process Conditions					Device Parameters					
Step	$V_{th\|Dose}$ (cm^{-2})	PT_{Energy} (keV)	PT_{Dose} (cm^{-2})	W_{sp} (nm)	MDD_{Dose} (cm^{-2})	I_{dsat} (nA/μm)	I_{off} (nA/μm)	V_{th} (mV)	S (mV/dec.)	$\Delta V_{g\|DIBL}$ (mV)	D
0	2×10^{12}	160	1×10^{13}	90	1.25×10^{13}	-90	-0.813	-265	77.51	-83.5	0.398
1	2.04×10^{12}	157	1×10^{13}	83	1.26×10^{13}	-105	-1.046	-263	77.46	-81.6	0.558
2	2.18×10^{12}	147	9.7×10^{12}	71	1.32×10^{13}	-109	-1.126	-276	78.11	-92.7	0.504
3	2.33×10^{13}	142	9.65×10^{12}	66	1.32×10^{13}	-114	-0.868	-287	78.80	-94.5	0.822

(b)

optimization analysis. Table II(a) and (b) summarize the process improvement results for the NMOS and PMOS devices, respectively. The final process conditions represent the design center for the subsequent response surface study.

C. Response Surface Model Generation

A seven-factor, 79-run FCC design combining a resolution VII 64-run fractional factorial design, 14 axial points, and one central point was used to fit the five NMOS response models. An eight-factor, 81-run FCC design combining a 64-run resolution V fractional factorial design, 16 axial points, and one central point was used to fit the five PMOS response models (see Williams *et al.* [1]). Both experimental designs were centered about the final operating conditions determined by the steepest ascent analysis. Process condition boundaries for both designs were set to $\pm$15% of the baseline condition values with the exception of gate oxidation time and temperature. Gate oxidation time and temperature were set to $\pm$0.5 min. and $\pm10^{\circ}$C, respectively, to reflect a variation in gate oxide thickness of approximately $\pm$15%.

Following response surface experimentation, second-order response models were generated for the five output characteristics for both the NMOS and PMOS devices. To improve model performance, a logarithmic data transformation of the form

$$I'_{\text{off}} = \log_{10}(I_{\text{off}}) \tag{18}$$

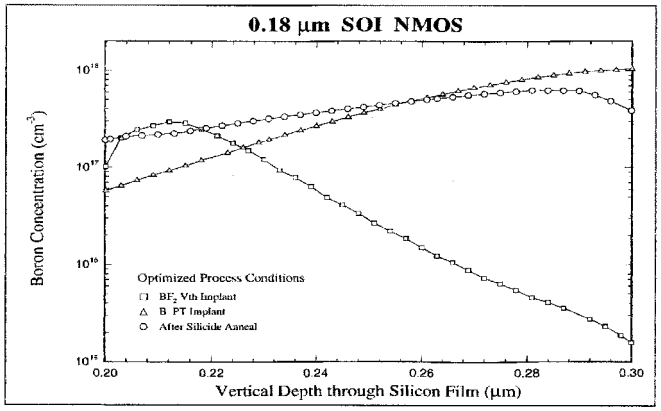
was performed on the I_{off} data for both devices before model fitting. Tests for significance of regression with respect to model lack-of-fit indicated that all models were statistically significant. Measures of the adjusted coefficient of multiple determination (R^2_{adj}) indicate that on average the response models are capable of explaining 97.4% of the variability observed in the output responses, with some models explaining as much as 99.6%. The adjusted coefficient of multiple determination is a measure of the amount of reduction in the variability of the output response that is obtained by the RSM. Comparing these results with those presented in [1], significant improvements in model accuracy are evident. These improvements can be directly attributed to the smaller design space afforded by the new multiresponse steepest ascent analysis. On average, the models generated by the new approach are nearly 12% more accurate than those fitted without the steepest ascent analysis.

D. Process Optimization and Sensitivity Analysis

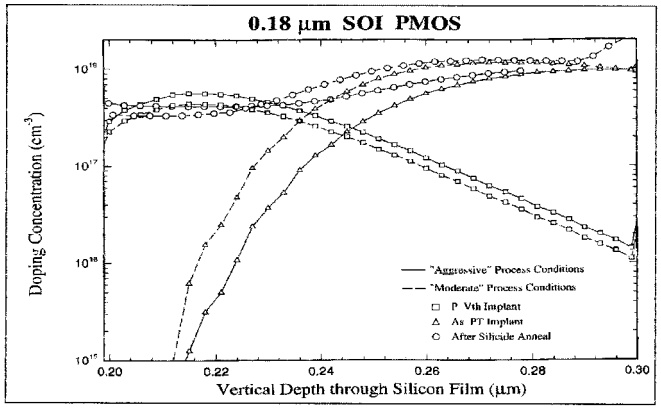
Optimization of the 0.18-μm PD SOI CMOS process was achieved by minimizing the objective function (15) with respect to the five performance criteria presented in Section IV-B. Gate oxidation time and temperature were determined such that the gate oxide thickness remained at 3.5 nm. A minimum spacer width of 72 nm was set as specified by the 1997 SIA National Technology Roadmap for Semiconductors (NTRS) for 0.18-μm technologies [18]. Because the steepest ascent analysis indicated significant performance gains could be realized for thinner spacer widths, however, a second set of optimized conditions were determined based on a more aggressive spacer width of 52 nm, as specified by the NTRS for 0.15-μm technologies. The NTRS indicates current technology is capable of producing spacer widths at that scale. Table III presents the resulting optimized processing conditions for the 0.18-μm PD SOI CMOS technology for both spacer widths, identified as "moderate" and "aggressive" values. To ensure I_{off} did not exceed 1 nA/μm while maximizing I_{dsat}, weights of 3 and 1 were assigned to I_{off} and I_{dsat}, respectively, for both devices. $\Delta V_g|_{\text{DIBL}}$ was also weighted at 1 nA/μm for the PMOS device. The remaining responses fell well inside their specified limits and therefore were not weighted.

TABLE III
OPTIMAL VALUES FOR PROCESS CONDITIONS IDENTIFIED IN TABLE II

NMOS			PMOS		
Process Condition	Moderate Value	Aggressive Value	Process Condition	Moderate Value	Aggressive Value
$V_{th\|Dose}$ (cm^{-2})	9.11×10^{11}	9.36×10^{11}	$V_{th\|Dose}$ (cm^{-2})	2.02×10^{12}	2.59×10^{12}
PT_{Energy} (keV)	30.7	30.5	PT_{Energy} (keV)	133.9	163
PT_{Dose} (cm^{-2})	1.1×10^{13}	1.1×10^{13}	PT_{Dose} (cm^{-2})	1.1×10^{13}	1.1×10^{13}
$Temp\|_{ox}$ (°C)	851	854	$Temp\|_{ox}$ (°C)	851	854
$Time\|_{ox}$ (min.)	6.9	6.5	$Time\|_{ox}$ (min.)	6.9	6.5
L_g (μm)	0.18	0.18	L_g (μm)	0.18	0.18
W_{sp} (nm)	72	57	W_{sp} (nm)	72	57
			MDD_{Dose} (cm^{-2})	1.47×10^{13}	1.35×10^{13}



(a)



(b)

Fig. 5. Channel doping profiles for (a) optimized 0.18-μm SOI NMOS device and (b) optimized 0.18-μm SOI PMOS device.

In comparing the optimization results for the two devices, little difference is seen between the moderate and aggressive processing conditions for the NMOS device, with the exception of spacer width. As seen in Fig. 5(a), which shows the vertical channel doping profiles for the optimized NMOS device (at midchannel), the channel doping profiles for both optimized conditions are essentially the same. For the PMOS device, however, significant differences exist in the channel implant conditions. As seen in Fig. 5(b), the final doping profile resulting from the aggressive conditions is more uniformly distributed than that of the moderate conditions, which is more of a retrograde profile. Both devices, however, show significant performance gains when optimized aggressively. Table IV lists the simulated device parameter values for the optimized CMOS processes.

TABLE IV
DEVICE PARAMETERS OBTAINED FOR OPTIMIZED 0.18-μm SOI CMOS PROCESS

	NMOS		PMOS	
Device Parameter	Moderate Conditions	Aggressive Conditions	Moderate Conditions	Aggressive Conditions
I_{dsat} (nA/μm)	258	276	-104	-122
I_{off} (nA/μm)	0.994	1.174	-0.835	-1.269
V_{th} (mV)	285	287	-292	-279
S (mV/dec.)	77.41	77.21	79.50	78.37
$\Delta V_{g\|DIBL}$ (mV)	66.9	70.2	-98.1	-92.5

TABLE V
SIMPLIFIED RESPONSE SURFACE MODELS FOR 0.18-μm SOI NMOS

$$I_{dsat}\ (nA/\mu m) = -1836 - 1.085\times10^{-11}\ V_{th|Dose} + 6.372\ PT_{Energy} - 9.505\times10^{-13}\ PT_{Energy}\times V_{th|Dose} - 1.675\times10^{-11}\ PT_{Dose} + 2.058\times10^{-13}\ PT_{Dose}\times PT_{Energy} + 7.206\ Temp|_{ox} - 5.13\times10^{-3}\ Temp|_{ox}^{2} + 146.000\ Time|_{ox} - 0.187\ Time|_{ox}\times Temp|_{ox} - 3051.000\ L_g - 10.110\ L_g\times PT_{Energy} + 1.264\times10^{-11}\ L_g\times PT_{Dose} + 2.430\ L_g\times Temp|_{ox} - 11.330\ W_{sp} - 1.798\times10^{-2}\ W_{sp}\times PT_{Energy} + 3.788\times10^{-14}\ W_{sp}\times PT_{Dose} + 1.094\times10^{-2}\ W_{sp}\times Temp|_{ox} + 6.253\ W_{sp}\times L_g$$

$$V_{th}\ (mV) = 2570.102 + 7.554\times10^{-11}\ V_{th|Dose} - 65.743\ PT_{Energy} + 3.229\times10^{-11}\ PT_{Dose} - 4.342\times10^{-13}\ PT_{Dose}\times PT_{Energy} - 2.765\ Temp|_{ox} + 6.573\times10^{-2}\ Temp|_{ox}\times PT_{Energy} - 228.358\ Time|_{ox} + 0.272\ Time|_{ox}\times Temp|_{ox} - 4790.469\ L_g + 27.034\ L_g\times PT_{Energy} - 3.933\times10^{-11}\ L_g\times PT_{Dose} + 5.193\ L_g\times Temp|_{ox} + 75.111\ L_g\times Time|_{ox} + 0.639\ W_{sp} - 4.207\ W_{sp}\times L_g$$

$$S\ (mV/dec.) = 26.291 + 6.586\times10^{-13}\ V_{th|Dose} + 2.145\ PT_{Energy} - 2.232\times10^{-12}\ PT_{Dose} + 7.478\times10^{-2}\ Temp|_{ox} + 2.902\times10^{-3}\ Temp|_{ox}\times PT_{Energy} + 3.068\times10^{-15}\ Temp|_{ox}\times PT_{Dose} - 18.404\ Time|_{ox} + 2.281\times10^{-2}\ Time|_{ox}\times Temp|_{ox} + 79.753\ L_g - 0.508\ L_g\times PT_{Energy} - 0.228\ L_g\times Temp|_{ox} + 299.058\ L_g^{2} - 3.958\times10^{-3}\ W_{sp} + 1.378\times10^{-3}\ W_{sp}\times PT_{Energy} + 1.378\times10^{-3}\ W_{sp}\times PT_{Dose} - 0.218\ W_{sp}\times L_g$$

$$\Delta V_{g|DIBL}\ (mV) = -972.530 + 4.062\ PT_{Energy} + 7.079\times10^{-12}\ PT_{Dose} - 2.019\times10^{-13}\ PT_{Dose}\times PT_{Energy} + 1.378\ Temp|_{ox} - 3.627\ Time|_{ox} + 1850.421\ L_g - 15.962\ L_g\times PT_{Energy} + 2.115\times10^{-11}\ L_g\times PT_{Dose} - 5.029\ L_g\times Temp|_{ox} + 6311.93\ L_g^{2} + 0.430\ W_{sp} - 5.199\times10^{-14}\ W_{sp}\times PT_{Dose}$$

$$Log(I_{off})\ (log(nA/\mu m)) = 30.525 - 1.846\times10^{-12}\ V_{th|Dose} - 0.564\ PT_{Energy} - 4.904\times10^{-13}\ PT_{Dose} + 6.392\times10^{-26}\ PT_{Dose}\times V_{th|Dose} + 2.988\times10^{-15}\ PT_{Dose}\times PT_{Energy} - 2.908\times10^{-2}\ Temp|_{ox} + 9.806\times10^{-4}\ Temp|_{ox}\times PT_{Energy} + 6.324\times10^{-2}\ Time|_{ox} + 46.007\ L_g - 0.864\ L_g\times PT_{Energy} + 9.444\times10^{-13}\ L_g\times PT_{Dose} - 0.116\ L_g\times Temp|_{ox} - 1.156\ L_g\times Time|_{ox} + 150.001\ L_g^{2} - 1.749\times10^{-2}\ W_{sp} + 7.643\times10^{-2}\ W_{sp}\times L_g$$

Before beginning the sensitivity analysis, a statistical test on the individual regression coefficients (the t-test) was conducted and the significant process conditions for each model were identified. Note, this is a statistical test of the fitted coefficients and does not require a measure of experimental error (see Myers [9]). Simplified response models were generated based on the identified factors for use in the manufacturing sensitivity analysis. Table V presents the simplified RSM's for the NMOS device. The average R^2_{adj} value for the reduced models (97.5%) is almost identical to that of the full models. Using the Monte Carlo approach described in Section II-D and the simplified RSM's, statistical distributions for each device parameter were determined for the baseline conditions and both sets of optimized conditions. Standard deviation for each input condition was set to $\pm10\%$ of its mean value with the exception of gate oxidation time and temperature. Gate oxidation time and temperature were set to $\pm7\%$ and $\pm1\%$, respectively, to maintain an approximate variation of $\pm10\%$ in gate oxide

TABLE VI
(a) SENSITIVITY ANALYSIS RESULTS FOR 0.18-μm SOI NMOS WITH STANDARD DEVIATION FOR SIGNIFICANT PROCESS CONDITIONS SET AT $\pm$10% OF SPECIFIED VALUE, EXCEPT FOR $Temp_{ox}$ AND $Time_{ox}$ AT $\pm$1% AND $\pm$7%, RESPECTIVELY. (b) SENSITIVITY ANALYSIS RESULTS FOR 0.18-μm SOI PMOS WITH STANDARD DEVIATION FOR SIGNIFICANT PROCESS CONDITIONS SET AT $\pm$10% OF SPECIFIED VALUE, EXCEPT FOR $Temp_{ox}$ AND $Time_{ox}$ AT $\pm$1% AND $\pm$1%, RESPECTIVELY

	Base Conditions		Moderate Conditions		Aggressive Conditions	
Device Parameter	Mean	σ	Mean	σ	Mean	σ
I_{dsat} (nA/μm)	258	6.130	260	4.120	274	4.380
Log(I_{off}) (nA/μm)	-0.158	0.091	0.083	0.091	0.074	0.093
V_{th} (mV)	302	6.130	282	5.660	288	5.800
S (mV/dec.)	77.75	0.802	77.28	0.797	77.37	0.800
$\Delta V_{g\|DIBL}$ (mV)	71.1	1.518	69.3	1.517	70.2	1.517

(a)

	Base Conditions		Moderate Conditions		Aggressive Conditions	
Device Parameter	Mean	σ	Mean	σ	Mean	σ
I_{dsat} (nA/μm)	-114	2.220	-112	2.030	-121	2.500
Log(I_{off}) (nA/μm)	-0.051	0.087	-0.002	0.076	0.068	0.108
V_{th} (mV)	-288	5.290	-286	5.590	-281	5.530
S (mV/dec.)	78.83	0.811	79.34	0.830	79.06	0.803
$\Delta V_{g\|DIBL}$ (mV)	-100.0	1.393	-98.5	1.268	-97.0	1.682

(b)

thickness. Table VI(a) and (b) present the resulting distribution statistics for the NMOS and PMOS devices, respectively.

The use of a smaller design area afforded by the multiresponse steepest ascent analysis had a significant impact on the results of the sensitivity analysis, which is a direct result of the improvements in model accuracy. The previous study [1] predicted much larger variations in the device characteristics caused by anticipated manufacturing fluctuations. Interestingly, the previous study also specified smaller fluctuations (5% versus 10%) for their sensitivity analysis. Such results indicate that a small experimental design space should be used when conducting a sensitivity analysis to ensure model inaccuracies do not inflate the fluctuation effects. Incorporation of a method like the multiresponse steepest ascent analysis described here then becomes necessary to help ensure the optimum operating point lies within the experimental design area.

V. SUMMARY AND CONCLUSION

A new TCAD-based statistical methodology for the optimization and sensitivity analysis of semiconductor technologies has been successfully demonstrated using a 0.18-μm PD SOI CMOS technology. Two new screening techniques were introduced and compared with a technique reported previously [1]. The three-factor screening techniques were used to test 30 process input conditions of a 0.18-μm SOI CMOS process, and their results identified 15 processing conditions significantly affecting the baseline CMOS process. The effectiveness of the newly developed multiresponse steepest ascent analysis for process improvement was illustrated on the CMOS process. Comparing the results of this work with that performed without the multiresponse steepest ascent methodology [1], we have realized on average a 12% increase in model accuracy (as measured by R^2_{adj}). Such improvements in model accuracy provide a better basis for conducting process optimization and greatly improve results of the sensitivity analysis. Following process improvement, second-order RSM's were fitted to ten separate device output characteristics using two separate central-composite experimental designs. Two sets of optimized conditions for the 0.18-μm SOI CMOS process, referred to as "moderate" and "aggressive," were determined based on deep submicron SOI technology currently under consideration by industry. Process optimization resulted in as much as a 15% increase in $I_{d\mathrm{sat}}$ over the baseline process without violating any of the other performance criteria. The Monte Carlo-based sensitivity analysis technique determined for both optimized processes that anticipated manufacturing process variations do not appear to cause conspicuous device performance degradation or deviation from the desired device performance measures.

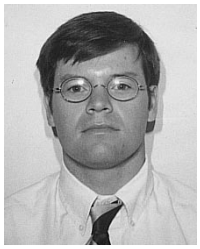
ACKNOWLEDGMENT

The authors would like to thank Avant! TCAD Business Unit (previously TMA) for providing the TCAD software used in this work.

REFERENCES

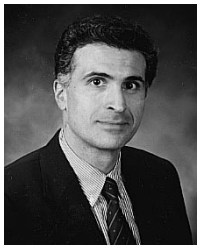
[1] S. Williams, K. Varahramyan, and W. Maszara, "Statistical optimization and manufacturing sensitivity analysis of 0.18 μm SOI MOSFET's," *Microelectron. Eng.*, to be published.
[2] Y. Aoki, H. Masuda, S. Shimada, and S. Sato, "A new design-centering methodology for VLSI device development," *IEEE Trans. Computer-Aided Design*, vol. CAD-6, pp. 452–460, May 1987.
[3] A. Alvarez, B. Abdi, D. Young, H. Weed, J. Teplik, and E. Herald, "Application of statistical design and response surface methods to computer-aided VLSI device design," *IEEE Trans. Computer-Aided Design*, vol. 7, pp. 272–288, Feb. 1988.
[4] K. Low and S. Director, "An efficient methodology for building macromodels of IC fabrication processes," *IEEE Trans. Computer-Aided Design*, vol. 8, pp. 1299–1313, Dec. 1989.
[5] G. Gaston and A. Walton, "The integration of simulation and response surface methodology for the optimization of IC processes," *IEEE Trans. Semiconduct. Manufact.*, vol. 7, pp. 22–33, Feb. 1994.
[6] K. Hasnat, S. Murtaza, and A. Tasch, "A manufacturing sensitivity analysis of 0.35 μm LDD MOSFET's," *IEEE Trans. Semiconduct. Manufact.*, vol. 7, pp. 53–59, Feb. 1994.
[7] D. Boning and P. Mozumder, "DOE/Opt: A system for design of experiments, response surface modeling, and optimization using process and device simulation," *IEEE Trans. Semiconduct. Manufact.*, vol. 7, pp. 233–244, May 1994.
[8] D. Angelo, S. Hareland, S. Khan, K. Hasnat, A. Tasch, and P. Zeitzoff, "Manufacturing sensitivity analysis of a 0.18 micron NMOSFET," *SPIE*, vol. 2875, pp. 136–145, Aug. 1996.
[9] R. Myers and D. Montgomery, *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. New York: Wiley, 1995.
[10] D. Montgomery, *Design and Analysis of Experiments*. New York: Wiley, 1997.
[11] R. Lenth, "Quick and easy analysis of unreplicated factorials," *Technometrics*, vol. 31, pp. 469–473, Nov. 1989.
[12] "TMA WorkBench, Version 2.2.5," Technology Modeling Associates, Inc. (currently Avant!), 1997.
[13] W. Dillon and M. Goldstein, *Multivariate Analysis Methods and Applications*. New York: Wiley, 1984.
[14] A. Edwards, *An Introduction to Linear Regression and Correlation*. San Francisco, CA: Freeman, 1976.
[15] W. Kasperski and H. Schneider, "Using normal probability plots to find significant factors," in *Proc. 27th Annu. Mtg. Decision Sci. Inst.*, vol. 2, Nov. 1996, pp. 1172–1174.

[16] D. Young, J. Teplik, H. Weed, N. Tracht, and A. Alvarez, "Application of statistical design and response surface methods to computer-aided VLSI device design—II: Desirability functions and Taguchi methods," *IEEE Trans. Computer-Aided Design*, vol. 10, pp. 103–115, Jan. 1991.
[17] Sobol, *A Primer for the Monte Carlo Method*. Boca Raton, FL: CRC Press, 1994.
[18] Semiconductor Industry Association, "The national technology roadmap for semiconductors,", 1997.



S. Williams received the B.S. and M.S. degrees in electrical engineering from Louisiana Tech University in 1996 and 1999, respectively.

From 1996 to 1999, he worked as a Research Assistant with the TCAD Training and Research Laboratory at Louisiana Tech University. Since 1999, he has been with Silvaco International as an Application Engineer at Silvaco's Austin Technology Center. His current research interest is in the development and incorporation of design for manufacturing tools within the TCAD environment.



K. Varahramyan (S'77–M'88) received the Ph.D. degree in electrical engineering from Rensselaer Polytechnic Institute in 1983. From 1982 to 1992, he was with IBM Microelectronics, conducting research and development in the realization of advanced semiconductor technologies. While at IBM, he also contributed to the development of FEDSS (a general-purpose microfabrication process simulator), and initiated and led the project for the realization of BEST (an integrated computer-aided system applied in semiconductor technology development). Since 1992, he has been with Louisiana Tech University, where he is the Energy Professor of electrical engineering in recognition of his teaching and research contributions in the microelectronics area. This includes the establishment of the TCAD Training and Research Laboratory, which has been serving the microelectronics teaching and research efforts at the university, and the TCAD training and research needs of the semiconductor industry.