2-1-2010

Statistical Approach for Yield Optimization for Minimum Energy Operation in Subthreshold Circuits Considering Variability Issues

Md. Waliullah Khan Nomani
University of South Carolina - Columbia, khannoma@email.sc.edu

Mohab Anis
University of Waterloo

Goutam Koley
University of South Carolina - Columbia, koley@engr.sc.edu

Publication Info

Published in *IEEE Transactions on Semiconductor Manufacturing*, Volume 23, 2010, pages 77-86.
http://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=66

Statistical Approach for Yield Optimization for Minimum Energy Operation in Subthreshold Circuits Considering Variability Issues

Md. Waliullah Khan Nomani, *Student Member, IEEE*, Mohab Anis, *Senior Member, IEEE*, and Goutam Koley, *Senior Member, IEEE*

***Abstract*—The supply voltage ($V_{\rm dd}$) and threshold voltage ($V_{\rm th}$) are two significant design variables that directly impact the performance and power consumption of circuits. The scaling of these voltages has become a popular option to satisfy performance and low power requirements. Subthreshold operation is a compelling approach for energy-constrained applications where processor speed is less important. However, subthreshold designs show dramatically increased sensitivity to process variations due to the exponential relationship of subthreshold drive current with $V_{\rm th}$ variation and drastically growing leakage power. If there is uncertainty in the value of the threshold or supply voltage, the power advantages of this very low-voltage operation diminishes. This paper presents a statistical methodology for choosing the optimum $V_{\rm dd}$ and $V_{\rm th}$ under manufacturing uncertainties and different operating conditions to minimize energy for a given frequency in subthreshold operation while ensuring yield maximality. Unlike the traditional energy optimization, to find the optimal values for the voltages, we have considered the following factors to make the optimization technique more acceptable: the application-dependent design constraints, variations in the design variables due to manufacturing uncertainty, device sizing, activity factor of the circuit, and power reduction techniques. To maximize the yield, a two-level optimization is employed. First, the design metric is carefully chosen and deterministically optimized to the optimum point in the feasible region. At the second level, a tolerance box is moved over the design space to find the best location in order to maximize the yield. The feasible region, which is application dependent, is constrained by the minimum performance and the maximum ratio of leakage to total power in the $V_{\rm dd}$-$V_{\rm th}$ plane. The center of the tolerance box provides the nominal design values for $V_{\rm dd}$ and $V_{\rm th}$ such that the design has a maximum immunity to the variations and maximizes the yield. The yield is estimated directly using the joint cumulative distribution function over the tolerance box requiring no numerical integration and saving considerable computational complexity for multidimensional problems. The optimal designs, verified by Monte Carlo and SPECTRE simulations, demonstrate significant increase in yield. By using this methodology, yield is found to be strongly dependent on the design metrics, circuit switching activity, transistor sizing, and the given constraints.**

***Index Terms*—Optimization, process variations, subthreshold circuits, tolerance design, yield modeling.**

Manuscript received May 07, 2009; revised September 25, 2009. First published December 22, 2009; current version published February 03, 2010.

Md. W. K. Nomani and G. Koley are with the Department of Electrical Engineering, University of South Carolina, Columbia, SC 29208 USA (e-mail: khannoma@email.sc.edu).

M. Anis is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSM.2009.2039184

I. Introduction

As very large-scale integration (VLSI) technology is scaled down toward the nanoscale regime, drastically growing leakage power and variations in device characteristics have motivated a significant investigation into the optimum supply and threshold voltage design for minimizing energy or power for a given performance constraint. Although energy dissipation has improved with each new technology node, as systems on chip are integrating tens of millions of devices on-chip following Moore's law, the energy expended per operation has become a critical issue in analog and digital integrated circuit design. Subthreshold operation is emerging as an energy-saving approach that involves scaling voltages below the device thresholds [1]. In this region, the energy per operation can be reduced by an order of magnitude compared to conventional operation, at the cost of circuit performance. In order to satisfy the performance requirements demanded by the applications, the threshold voltage of the device should also be lowered to have both low power operation and high performance. However, there is a cost of higher static power dissipation due to large leakage currents [2].

It is well known that subthreshold designs have dramatically increased sensitivity to process variations since drive current becomes exponentially dependent on threshold voltage [3]. Process variations and manufacturing uncertainties cause design variables to deviate from their nominal values. Moreover, the manufacturing process of nanoscale transistors and structures has introduced several new sources of variation that have made the control of process variation more difficult [4]. Thus the robustness of the design has come up with enormous challenges that must be resolved by the designers [5].

The variations in $V_{\rm dd}$ may arise from nonuniformity in the distribution of the power supply and for different switching activity of the circuit. With leakage power becoming a significant contributor to total power dissipation, leakage current flowing through the grid can result in significant supply-voltage drops. The increase in the current density with the scaling in the technology and increase in rate of switching make it more challenging to retain the traditional bound on the supply-voltage noise [6]. Device threshold voltage is dependent on a number of process parameters such as channel doping concentration and gate length. As the critical dimension is scaled down, the number of dopant atoms becomes less and hence small variations in their number and position result in a large variation in device performance. Moreover, variations in the doping level

cause variations in the base transit time in bipolar junction transistors [7], further increasing the variability in mixed signal design. Increased $V_{\rm th}$ variability and lower $V_{\rm th}$ values can result in functional failures in dynamic logic design. A number of successful subthreshold designs have been presented in the literature [1], [3]. Using subthreshold design, it is expected that energy efficiency in the range of 1 pJ/instruction can be achieved [8]. In addition, wide-range dynamic voltage scaling has been proposed where the processor can scale from high-performance superthreshold operation to ultra-low-power subthreshold operation depending on workload [9].

In previous work, a minimum energy voltage $(V_{\rm min})$ for complementary metal–oxide–semiconductor (CMOS) subthreshold region was demonstrated [9], [10]. However, the proposed analysis did not account for the impact of process variation. It has been shown that minimum-size devices are theoretically optimal for minimizing energy in subthreshold operation [11]. But minimum-size devices have increased sensitivity to $V_{\rm th}$ variations. Therefore, variability must be considered when analyzing the minimum energy operating point. Zahi *et al.* addressed intradie variation by providing statistical models for energy and delay of an inverter chain in subthreshold [12]. Calhoun *et al.* derived analytical expressions for optimum $V_{\rm dd}$ to minimize energy in subthreshold operations and its dependence on design characteristics and operating conditions for a given technology [13]. However, functional yield was not considered until [14].

Gonzalez *et al.* considered process variation for minimizing energy-delay product (EDP) [15]. Analytical expressions for the optimum $(V_{\rm dd}, V_{\rm th})$ point to minimize power at a given performance are shown for transregional models based on fitted [16] or physical [17] parameters. Measurements of a test chip with adaptive $V_{\rm dd}$ and adaptive body bias demonstrate the minimum power point for a given performance, but they also show forward-bias diode currents can make the theoretical optimum unreachable [18]. Expressions of the sensitivities of energy and delay to different parameters give a guideline for optimum energy circuits [19]. Optimizing subthreshold circuits has not been studied enough. Inclusions of uncertainty in subthreshold region of operation do provide a guideline for designing under variations. Nevertheless, each design needs to satisfy constraints that are application-dependent. Therefore, an unconstrained optimum point, suggested by this literature, does not satisfy the constraint and thus seems to be futile.

This paper describes a novel simultaneous optimization of supply and threshold voltage scaling for subthreshold circuits in the leakage dominant era. Design optimization will be done in two different steps: finding the design variable that optimizes the objective function of the product and finding a tolerance region that provides the most yields.

To accomplish this, a design-specific feasible region is defined by constraints on minimum performance, maximum achievable performance, and numerically solving the ratio of leakage to total power in the $V_{\rm dd}$-$V_{\rm th}$ plane. However, due to manufacturing imperfections and technology shifting, it may not be possible to realize the nominal design value exactly. Therefore, design variables $V_{\rm dd}$ and $V_{\rm th}$ are assumed to be random whose probability distribution may be known. A tolerance box that represents the variations in the voltages $(\mu \pm 3\sigma)$ is moved over the design space. It attempts to place the tolerance box in such a way that the portions of the box with higher yield lie in the feasible region. The final location of the box and its center provides the nominal design values for $V_{\rm dd}$ and $V_{\rm th}$ such that the design has a maximum immunity to the variations. This center maximizes parametric yield and optimizes energy for application-specific designs.

II. Application Areas

Swanson *et al.* predicted ultralow voltage CMOS logic operating at a supply voltage of $(8kT/q) = 200$ mV at room temperature and derived the fundamental limits of voltage scaling [20]. This was a key result for low-voltage logic design, which acts as a catalyst for many low-power applications. Subthreshold circuits have been used quite extensively in analog design [21] but not in digital domain. Subthreshold operations are suitable for specific applications, which do not need high performance but require extremely low power consumption such as hearing aids, pacemakers [22], wearable computing [23], and self-powered devices [24]. Emerging ultra-low-power application such as distributed sensor network is a natural fit with subthreshold circuits [13]. Subthreshold circuits can also be applied to applications where circuits remain idle for an extended period of time. This type of application appears in almost every design, including the high-performance microprocessor. So the recent explosion in applications that benefit from low-energy operation has created two classes of applications for which subthreshold circuits are well suited. The first is severely energy-constrained systems. The second class is called burst-mode applications that require high performance for part of the time but spend significant fractions of their operation doing non-performance-critical tasks.

III. Subthreshold Operation

There are three different sources of power dissipation in CMOS circuits: the dynamic power, the static power, and the short-circuit power. The dynamic power results from switching capacitive loads between different voltage levels, whereas the static power is mainly due to subthreshold leakage between power supply and ground. The short-circuit power is dissipated during a switching event when there is a direct path from supply to the ground. In sub-100 nm technologies, there is also tunneling through the gate oxide due to reduced oxide thickness [25]. The contribution of power from gate leakage is significant when circuits remain idle for an extended period of time, i.e., at standby operation. For this, it can be included in the static power. Similarly, the short-circuit power can be modeled by the dynamic power equation, as it occurs only when the circuit is switching. Therefore, the total power can be written as a summation of these two powers as

$$P_{\rm total} = P_{\rm dynamic} + P_{\rm static}. \tag{1}$$

For a CMOS gate, the dynamic power is [15]

$$P_{\rm dynamic} = aCV_{\rm dd}{}^2 f \tag{2}$$

where a is the activity factor of the output node, C is the total capacitance of the output node, $V_{\rm dd}$ is the supply voltage, and f is the operating frequency. If the circuit performs one operation per cycle, then the energy per operation is [15]

$$E_{\rm dynamic} = aCV_{\rm dd}{}^2. \tag{3}$$

For a complex chip, the total dynamic power is simply the sum of the dynamic power of all gates. To determine the delay of an inverter, we use the alpha power model [26]

$$T_g = \frac{CV_{\rm dd}}{I} \tag{4}$$

and the maximum operating frequency of the chip is given by

$$f = \frac{1}{L_d T_g} \tag{5}$$

where L_d is the logic depth and I is the drain current. The subthreshold current for $V_{\rm dd}$ in the subthreshold region is [15]

$$I_l = I_s \exp\left[\frac{V_{gs} - V_{\rm th}}{\gamma V_0}\right]\left(1 - \exp\left[-\frac{V_{ds}}{\gamma V_0}\right]\right) \tag{6}$$

where γ is the subthreshold slope factor, V_0 is (kT/q), and I_s is the zero threshold leakage current. The thin-oxide leakage of a single gate can be modeled as [27]

$$I_{\rm oxide} = J_g A \tag{7}$$

where

$$J_g = \exp\left[\frac{P - t_{\rm ox}}{Q}\right] \tag{8}$$

where J_g is the current density, A is the area of the gate, P and Q are the technology constants, and $t_{\rm ox}$ is the oxide thickness. The power consumption of a chip can be represented as

$$P_{\rm total} = aCV_{\rm dd}{}^2 f + N(1-a)I_s \exp\left[\frac{V_{gs} - V_{\rm th}}{\gamma V_0}\right] \times \left(1 - \exp\left[-\frac{V_{ds}}{\gamma V_0}\right]\right) V_{\rm dd} + NJ_g A V_{\rm dd} \tag{9}$$

where N is the total number of gate of the circuit. Gate tunneling current has a strong dependence on the voltage across the gate, so it decreases with supply voltage much more quickly than subthreshold current. As a result, gate leakage become negligible in subthreshold region except when $V_{\rm th}$ is high and $V_{\rm dd}$ is very low, which is not a good choice from subthreshold design perspective [28]. So from a realistic assumption, we have considered subthreshold leakage only in the leakage power term while plotting Fig. 1. The power-delay product (PDP) metric could be written as follows:

$$\text{PDP} = P_{\rm total} \times \text{Delay}. \tag{10}$$

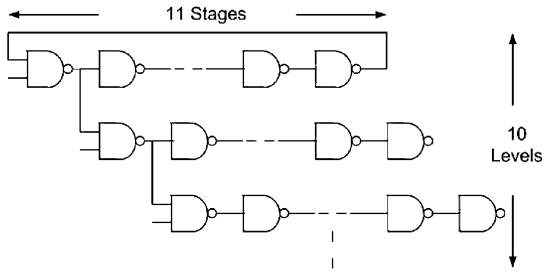

Fig. 1. Characterization circuit for energy and performance analysis in the subthreshold region [28].

Low-power consumption in high-performance circuits is highly desirable. Scaling of $V_{\rm dd}$ is the most effective way to decrease power consumption since CMOS dynamic power quadratically depends on $V_{\rm dd}$ but degrades the performance of the circuit as well. This performance degradation can be compensated by decreasing $V_{\rm th}$, but then the subthreshold leakage power increases exponentially. Therefore, there is an optimum set of $V_{\rm dd}$ and $V_{\rm th}$ that ensures the low-power and high-performance circuit design [29], [30]. Given that there is a tradeoff between power and performance, we investigate their combined effect with recent trend in CMOS scaling in the $V_{\rm dd}$-$V_{\rm th}$ plane. Different metrics have been used to study supply and threshold voltage scaling such as EDP, energy or power-delay product, and power-energy product (PEP) [25]. The EDP metric gives a higher priority to performance than power. It is more suitable when performance is the primary concern. To prioritize power, PEP is a useful metric to be used as an objective function in optimizing power consumption and quality of a design. In this metric, power has a higher geometric weighting than delay and thus produces a lower power solution than the other two metrics. The energy metric gives balanced weighting to both. In this yield optimization for subthreshold design, to have both low-power operation and satisfactory performance, we choose PDP as our design metric with $V_{\rm dd}$ and $V_{\rm th}$ to be two design variables to make the optimization technique more acceptable.

IV. Problem Formulation

The computation of parametric yield is based on constructing a feasible region of operation where yield is defined as the percentage of chips that satisfy the constraints of the design metrics. The method uses the aspects of advanced first-order second-moment (AFOSM) reliability method in probabilistic design to find a linearized feasible region. The frequency lines determine two boundaries of the feasible region, while the power ratio curve determines the other boundary. Fig. 1 shows the variable activity factor characterization circuit used in this analysis. The circuit consists of an 11-stage two-input NAND ring oscillator with nine additional 11-stage NAND delay chains driven by the ring oscillator. This circuit is also used to model the pipeline delays for current microprocessor design.

Fig. 2 shows the minimum energy point and contribution of active and leakage power to total energy in a ring oscillator circuit. Most circuits must meet specific performance targets. Fig. 2 also shows the relationship between switching energy

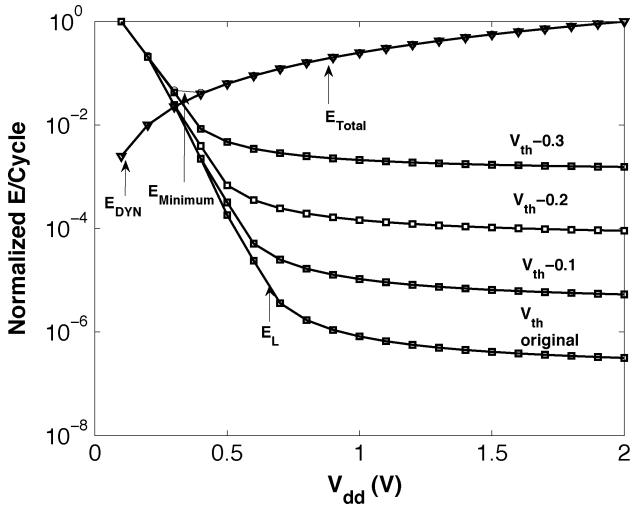

Fig. 2. Effect of lowering $V_{\rm th}$ on energy per operation in the subthreshold region [13]. This figure also indicates that the minimum energy point occurs in subthreshold region.

and leakage energy in the characterization circuit as a function of $V_{\rm dd}$ with $V_{\rm th}$ originally fixed at 500 mV and an activity of one. Mathematically, the minimum energy point occurs where the slopes of the leakage energy and the active energy are equal in magnitude and opposite in sign. The figure shows that the minimum energy point occurs in the subthreshold region at $V_{\rm dd} = 370$ mV.

Markers indicate SPECTRE simulation results in 90 nm technology, and lines indicate analytical results. To choose different $V_{\rm th}$ for different simulations, the device model was modified and new models were generated. As $V_{\rm th}$ decreases, the subthreshold current increases exponentially, which is shown by the rise in leakage energy above $V_{\rm th}$. When $V_{\rm th}$ decreases too far, then $V_{\rm dd} > V_{\rm th}$, and so subthreshold operation becomes invalid. From the figure, we see that leakage energy exceeds the dynamic energy for extremely low $V_{\rm th}$ in this ring oscillator example. Of course, the advantage of lowering $V_{\rm th}$ is increased performance in the subthreshold region for the same energy per operation, but it also increases the leakage energy dissipation, which puts a limit on the scaling of $V_{\rm th}$.

Fig. 3 shows the effect of activity factor on the optimal $V_{\rm th}$ and $V_{\rm dd}$ for minimum energy operation. Markers indicate SPECTRE simulation results in 90 nm technology, and lines indicate analytical results. If we vary the workload, the active energy decreases in proportion to the decrease in workload, but the leakage energy remains constant. So with the increase in workload, the minimum energy of operation increases and optimal supply voltage moves to the lower supply voltage, which leads to increase in variability and decrease in yield. The reduced gate performance is one factor that is keeping supply voltages from scaling down too quickly. Therefore, submicrometer technologies with low threshold voltages should be in demand for low-power applications. By simply moving to a low $V_{\rm th}$ process, a designer could reduce the supply voltage and power without requiring a major change of the design, as the performance remains constant. But the irony arises from increasing leakage power in presence of process and operating point variation.

Fig. 4 is obtained by numerically solving the ratio of leakage power to total power. The feasible region is constructed by the following constraints: the maximum achievable performance within the subthreshold design, the performance at the minimum energy point of operation, and the contour of leakage to total power ratio. Performance constraints are normalized by the performance contour that runs approximately through the optimal point. Any point in F satisfies the constraints and is expressed as

$$F = \{x \in \Re^2 | h_i(x) \geq 0, i = 1, 2, \ldots m\} \tag{11}$$

where x is the design variable vector $(V_{\rm dd}, V_{\rm th})$ and $h_i(x)$ is the performance functions and index i (here it is three) represents the $i^{\rm th}$ constraint. These constraints can assume different values from one application to another. Here it is assumed that the minimum frequency is that corresponding to the minimum energy point, and the ratio of leakage power to total power is about 0.5. The first constraint ensures subthreshold operation with maximum achievable performance

$$V_{\rm dd} \leq V_{\rm th}. \tag{12}$$

The circuit clock frequency must exceed a given minimum value $(f_{\rm min})$. This constraint ensures the minimum frequency of operation

$$f \geq f_{\rm min}. \tag{13}$$

Finally, the following constraint comes from the power budget:

$$\frac{P_{\rm leak}}{P_{\rm total}} \leq 0.5. \tag{14}$$

The methods construct a polyhedral approximation of the feasible region by taking first-order approximation of each $h_i(x)$ at the expansion point x^*

$$h_i(x) \approx h_i(x^*) + g_i(x^*)^T (x - x^*) \tag{15}$$

where $g_i(x^*)$ is the gradient vector of h. The point x^* is on the surface $h_i(x)$ and has the minimal distance from the center of the initial tolerance box x^c. The superscript c stands for the center of the initial tolerance box. The shortest distance to the constraint is found by solving the minimization problem

$$\min\ \beta = \left[(x - x^c)^T (x - x^c)\right]^{\frac{1}{2}} \quad \text{subject to } h_i(x) = 0. \tag{16}$$

This minimization problem can be solved by the Newton or the gradient method. We use an iterative formula based on

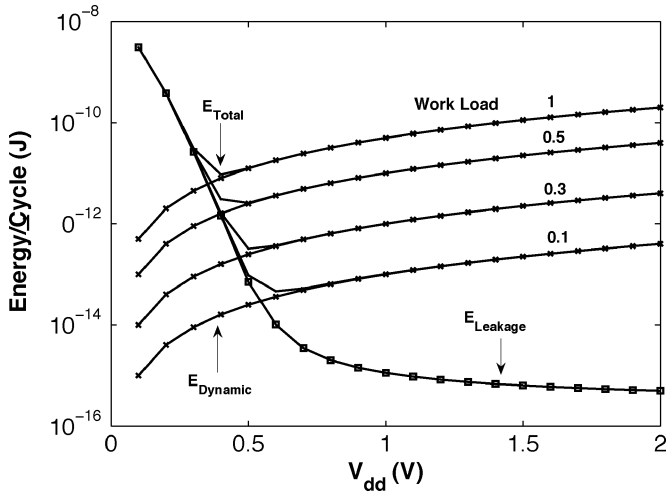

Fig. 3. Effect of varying workload on dynamic, leakage, and total energy [28].

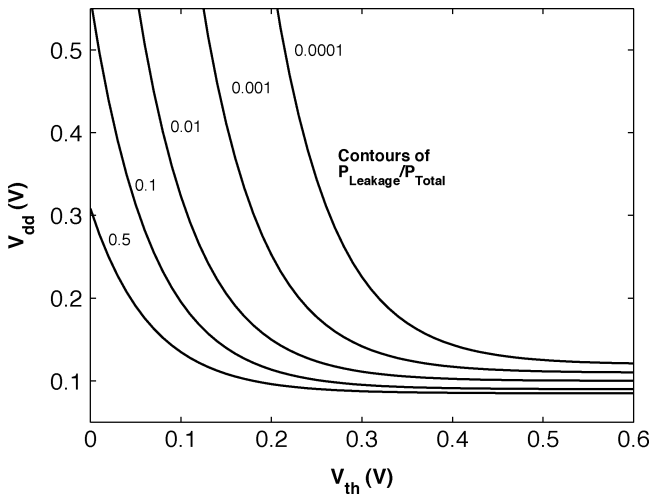

Fig. 4. Contours of ratio of leakage to total power in subthreshold region. This is an extension of plot presented in [15] for above threshold operation.

Madsen *et al.* that uses the Lagrangian of the above with a fixed-point method and is given below by [31]

$$x^{k+1} = x^c - \frac{(g_i^k)\left[(g_i^k)^T(x^c - x^k) + g_i^k\right]}{(g_i^k)^T(g_i^k)}. \tag{17}$$

The superscripts k and $k+1$ refer to the index of iteration and to the gradient at the k^{th} iteration. This formula attempts to solve $h_i(x) = 0$ indirectly. Convergence of (14) depends on continuity and convexity of $h_i(x)$, and the solution may not be unique. The method finds a linear approximation for each constraint in the form of (13). Such approximations, together with some extra bounds on the design variables, form a polyhedral feasible region. The polytope P is defined by

$$P = \left\{x | Ax \geq C, x_j^{\min} \leq x_j \leq x_j^{\max}\right\}. \tag{18}$$

The ith row of A is g_i^T, where all the partial derivatives are evaluated at x^*, found by (15) and $C_i = (g_i^*)^T x_i^*$. The lower and upper bounds on the design variable are denoted by $x_j^{\min}$ and $x_j^{\max}$, where index j represents the jth design variable. In this case, the design variables are V_{dd} and V_{th} of the device. So far, the design space has been constructed. For the purpose of yield optimization, distribution of design variables is modeled.

Simulations results indicate that V_{dd} and V_{th} variations can be modeled as normal distributions, which is a standard statistical model [32], [33]. But the normal distribution does not have a closed-form cumulative distribution function (cdf), which is necessary for the yield estimation. In our method, yield is estimated directly using the joint cdf over the tolerance box requiring no numerical integration and thus saving considerable complexity for multidimensional problems. For this reason, instead of the standard model, we have considered Kumaraswamy's distribution [34], double-bounded probability density function (DB-pdf), with the following form:

$$f(z) = abz^{a-1}(1 - z^a)^{(b-1)} \tag{19}$$

$$z = \frac{x - x^{\min}}{x^{\max} - x^{\min}}, x^{\min} \leq x \leq x^{\max}. \tag{20}$$

Depending on the choice of parameters a and b, DB-pdf can take various shapes. It can be used to approximate uniform, triangular, Gaussian, highly kurtic, and many other single modal distributions. Another advantage of DB-pdf is that its integral, needed for yield evaluation, is available in closed form

$$F(z) = 1 - (1 - z^a)^b. \tag{21}$$

V. Two-Level Optimization

To start the two-level optimization problem, a design metric (PDP, EDP, PT^3, etc.) is initially selected, based on the priority that needs to be given to the power as opposed to the performance in a given application. In the first level, the respective metric is optimized, subject to the constraints, and the solution of the constrained optimization is adopted as the initial solution for the yield optimization in the second level. At the second level, a tolerance box, with its center initially located on the deterministic optimum point obtained in the first level, is moved over the design space to find the best location in order to maximize the yield. The location of the box must not only satisfy the maximum yield objective but also be as close as possible to the deterministic optimum point obtained in the first level. The final location and the center of the box indicate a design that has the highest immunity against the variations in V_{dd} and V_{th}. In addition to the design robustness, this center must provide the best possible tradeoff between power and performance. The size and location of the tolerance box depend on the allowed percentage of tolerance, the yield, the shape of the feasible region, and the distribution of the design variables. By minimizing the objective function β in (16), the tolerance box is moved over the feasible region to maximize the yield and minimize the distance between the deterministic optimum point and the center of the tolerance box in its final location. Fig. 5 depicts the reversed normalized energy contours at an average of 42 °C for 90 nm CMOS technology. The optimal value shown here is the nominal PDP value

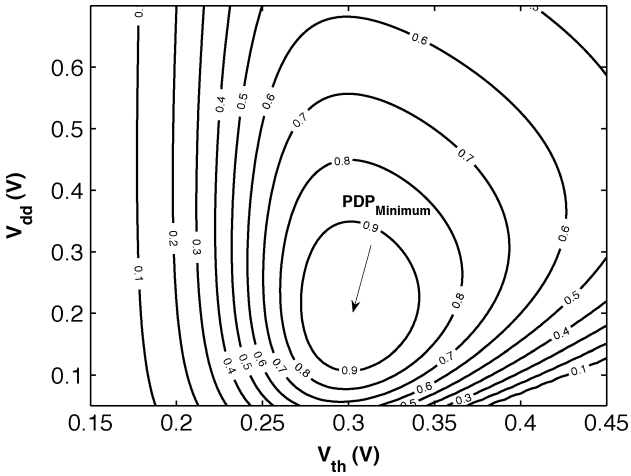

Fig. 5. Normalized contours of energy obtained through deterministic optimization showing minimum energy (PDP) point in the $V_{\rm dd}$-$V_{\rm th}$ plane.

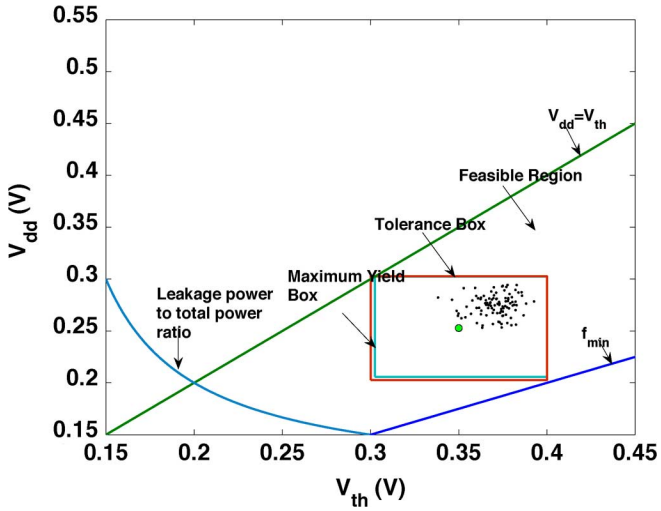

Fig. 6. The final location of the tolerance box over which the yield is maximized.

of unconstrained minimization of PDP equation. Note that the contours are normalized by dividing the minimum energy by the calculated energy for any pair of $V_{\rm dd}$ and $V_{\rm th}$.

Fig. 6 depicts the final location of the optimum tolerance box for 90 nm technology and all three constraints of the design. The outer box represents the tolerance range, and the inner box is the tolerance box with the maximum yield in the feasible region. If the variations in the design variable can be controlled so that the size of the outer box is reduced to that of the inner box, the yield becomes 100%. The yield is expressed as a function of the lower and upper limit of the design variables. The problem reduces to the search for a box over which the yield is maximized and is contained in the polytope P defined by (13)

$$R(x^l, x^u) = \{x \in \Re^2 | x^l \le x \le x^u\} \tag{22}$$

where $R(x^l, x^u)$ is the inner optimum box contained in the feasible region and x^l and x^u are lower and upper bounds of the box, respectively. The containment requirement $R \subseteq P$ is equivalent to

$$A^+ x^u - A^- x^l \le C. \tag{23}$$

Recall that A_i is the transpose of the gradient vector g_i, obtained by linearization of the performance constraint h_i at a given x. A^+ and A^- are the upper and lower bounds of the same performance constraint, and C refers to the constant terms in the linearization. We choose the reference point x^r to be greater than or equal to $x^{\min}$ to define the location of the larger tolerance box. The left bottom corner is x^r, the top right corner is x^{r+t}, and the range t is given for the jth width by $t_j = x_j^{\max} - x_j^{\min}$. The variables x^l and x^u together place the location and the size of the optimal box that corresponds to the maximum yield. The variables to be optimized are x^r, x^l, and x^u. The yield function is given by

$$\begin{aligned}\text{Yield}\,(x^r, x^l, x^u) &= \prod_{i=1}^{2} \Pr\left\{x_i^l \le x_i \le x_i^u\right\} \\ &= \left[F\left(\frac{V_{\rm th}^u - V_{\rm th}^r}{t_{Vth}}\right) - F\left(\frac{V_{\rm th}^l - V_{\rm th}^r}{t_{Vth}}\right)\right] \\ &\quad \times \left[F\left(\frac{V_{\rm dd}^u - V_{\rm dd}^r}{t_{Vdd}}\right) - F\left(\frac{V_{\rm dd}^l - V_{\rm dd}^r}{t_{Vdd}}\right)\right].\end{aligned} \tag{24}$$

The yield model (22) estimates the yield for given values of x^r, x^l and x^u. The objective of optimization, formulated below, is to move the tolerance box in such a way that the yield is maximized. The optimization model is given by

$$\begin{aligned}&\max\ \text{Yield}\,(x^r, x^l, x^u) \\ &\text{subject to :} \\ &A^+ x^u - A^- x^l \le C \\ &x^r \ge x^{\min} \\ &x^l \ge x^r \\ &x^u - x^l \ge t \\ &x^r + t \le x^{\max}.\end{aligned} \tag{25}$$

VI. Result and Discussion

We used the sequential quadratic programming in MATLAB to solve the optimization problem. The MOSFET model for 90 nm technology is adapted from the BSIM 4 model. Monte Carlo simulations have been done to investigate an optimal operating region within which a circuit could function optimally and to verify its yield maximality. The result of yield optimization for different tolerances and different DB-PDFs is presented

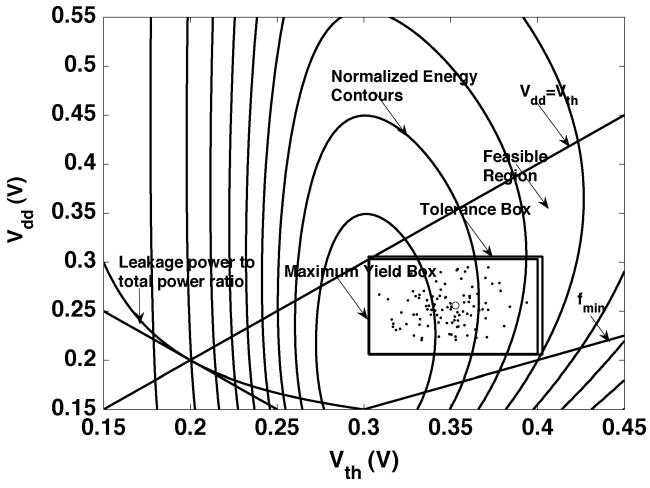

Fig. 7. The final location of the tolerance box for Gaussian distribution of design variables over which the yield is maximized considering 10% tolerance for variations.

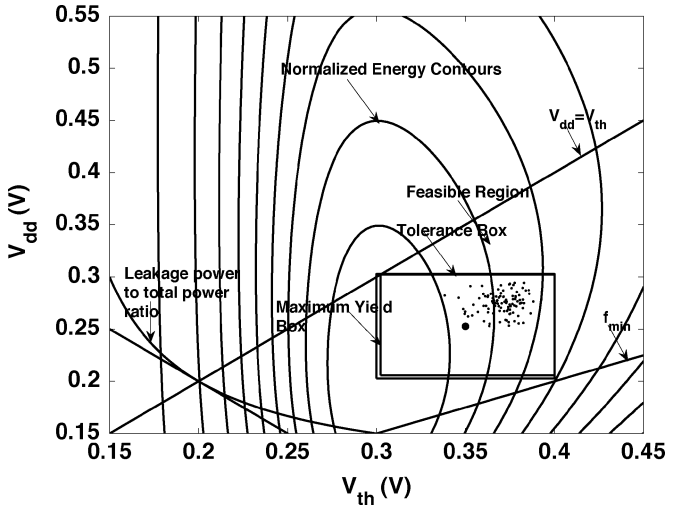

Fig. 9. The final location of the tolerance box for highly kurtic distribution of design variables over which the yield is maximized considering 10% tolerance for variations.

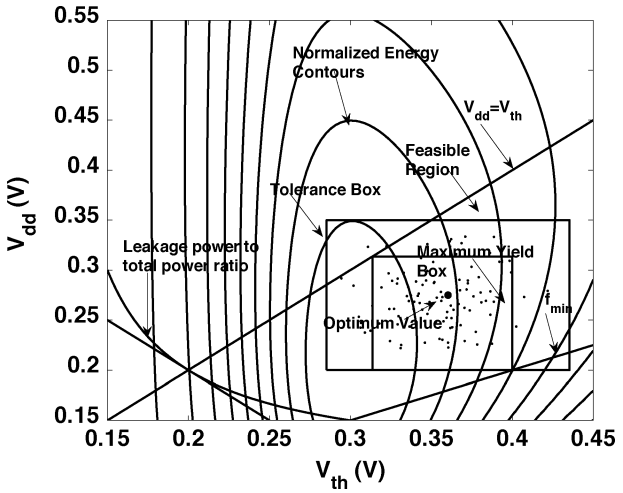

Fig. 8. The final location of the tolerance box for Gaussian distribution of design variables over which the yield is maximized considering 15% tolerance for variations.

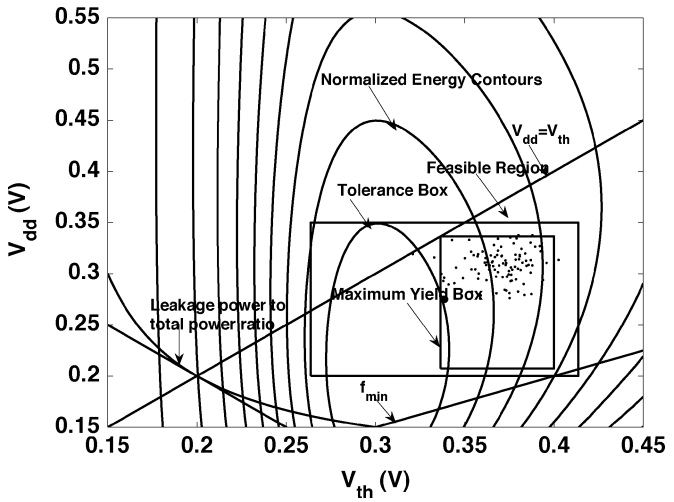

Fig. 10. The final location of the tolerance box for highly kurtic distribution of design variables over which the yield is maximized considering 15% tolerance for variations.

in Table I. By using the methodology, optimized yield (Final M-C) is found to be significantly higher than the yield of initial nominal values (Initial M-C). For simple cases, the designer can estimate the initial design values close to the optimal ones, but for a subthreshold circuit when the slightest change in design variables can cause considerable change in performance functions, the proposed yield optimization method can give better yield. If the distribution of design variables (DB-PDF) changes, the location of the tolerance box also changes. Figs. 7–10 show the result of yield optimization for Gaussian and highly kurtic distribution of design variables and for different values of tolerances. Any pair of $V_{\rm dd}$ and $V_{\rm th}$ in the feasible region satisfies the constraints. As seen from the figures, the distribution of variables in a specific part of feasible region causes the tolerance box to move in such a way that it crosses the performance functions, but the design variables are still inside the feasible region. The advantage of first-order approximation for feasible region is to find a tolerance box that corresponds to the yield equal to one and then modify the process in a way that provides the expected tolerances for design variables to produce 100% acceptable products. As in our case, all performance functions are convex, so the estimated feasible region is always inside the exact feasible region. Yield increases when tolerance decreases, so there is a good tradeoff between increase in yield and the cost of design and manufacturing. So the financial data due to tolerances (manufacturing) plus cost due to yield (wastage cost) need to be evaluated.

Table II lists the yield, Mean value, and standard deviation of the design variable for tight constraints considering Gaussian distribution. From Table II, we see that due to the variations in the design variables, design metric PDP has statistical measures. The mean (μ) and standard deviation (σ) of PDP, at the maximum yield point, are also calculated for two different standard deviations in design variables, which is mentioned in Table II. To set tight constraints, the maximum allowed frequency can be lowered or the acceptable ratio of leakage to total power

TABLE I
YIELD ESTIMATION FOR DIFFERENT JOINT DISTRIBUTION OF DESIGN VARIABLES

Type of Distribution ($Tolerance\%$)	Yield		
	Initial M-C	Estimated	Final M-C
Gaussian(15)	0.62	0.78	0.90
Gaussian(10)	0.62	0.86	0.98
Non Symmetric(15)	0.60	0.79	0.94
Non Symmetric(10)	0.60	0.88	0.98
Highly Kurtic(15)	0.86	0.96	1.00
Highly Kurtic(10)	0.86	0.96	1.00
Uniform(15)	0.64	0.72	0.80
Uniform(10)	0.64	0.78	0.84

TABLE II
YIELD OPTIMIZATION FOR VARIATIONS IN $V_{\rm dd}$ AND $V_{\rm th}$ WITH TIGHT CONSTRAINTS

Parameters		Value
Variations of 6σ	V_{dd}	20% 10%
	V_{th}	20% 10%
V_{th}	Mean(μ)	0.3415 0.3222
	STD(σ)	0.032 0.016
V_{dd}	Mean(μ)	0.2344 0.2286
	STD(σ)	0.030 0.015
Normalized PDP	Mean (μ)	0.768 0.842
	STD(σ)	0.048 0.03
$Yield(\%)$	Estimated	42 68
	Monte-Carlo	62 84

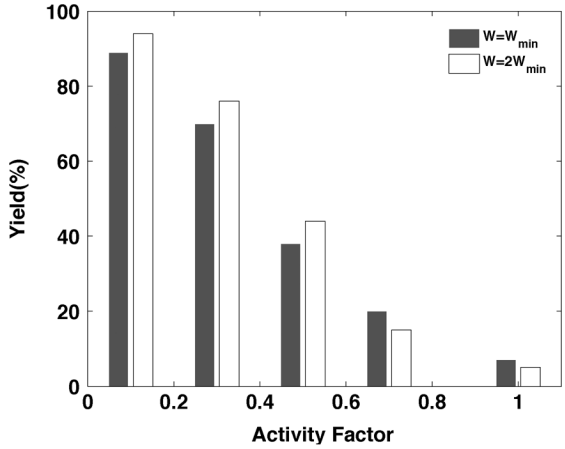

Fig. 11. Effect of the activity factor of the circuit and the device sizing on the yield.

TABLE III
OBTAINED OPTIMUM VALUES FOR DESIGN VARIABLES

Type of Distribution	$V_{dd}(V)$		$V_{th}(V)$	
(Tolerance%)	15%	10%	15%	10%
Gaussian	0.272	0.235	0.361	0.352
Highly Kurtic	0.275	0.238	0.352	0.337
Non Symmetric	0.277	0.243	0.325	0.292
Uniform	0.282	0.258	0.364	0.348

can be reduced. In this paper, to see the effect of tight constraints, we take THE maximum allowed frequency to be equal to 80% of maximum frequency achievable with equal $V_{\rm dd}$ and $V_{\rm th}$. For minimum frequency of operation, we assumed $f_{\rm min} = 1.1 \times f_{\rm PDP_{min}}$. Again, the constraint on leakage to total power ratio was set to 0.4. This causes the area of the feasible region to shrink, which results in increase in the yield loss. It is very natural that due to tight constraints, many designs fail to satisfy either constraint.

From Tables I and II, it is clear that, with tight constraints, if we consider Gaussian distribution of design variables, yield decreases considerably from 98% to 84% for 10% tolerance (i.e., $6\sigma = 10\%$) of variations of the design variables. To increase the yield, the process and environmental variations must be controlled so that the tolerance box (outer box) is resized to match the 100% yield box (inner box). To achieve this, we need to increase the precision of the equipment in the fabrication process at a higher cost. We can reduce the voltage noise by controlling the voltage drop within the chip. The variations of the design variables can be systematic or random in nature, and they fall into the following categories: fab-to-fab, lot-to-lot, interwafer, die-to-die, and even within-die. Systematic variations depend on the position of the devices on a die and the layout environment surrounding the devices [35]. It is possible to compensate the systematic variations from lithographic, etching, and layout information [36]. But the random uncertainties such as dopant number and location, however, cannot be predicted and cause all of the devices in close proximity to exhibit different characteristics. Therefore, to increase the yield, the constraints must be relaxed. This occurs when the feasible region is increased with a new set of constraints and the yield loss approaches zero.

The effect of activity factor on yield is also analyzed in this paper. The circuit in Fig. 1 consists of ten levels, with 11 stages in each level. The first level is a NAND ring oscillator, and other levels are NAND chains. An activity factor of "0.1" is controlled by assigning "1" to select zero input and "0" to the subsequent select inputs. If more select inputs are set to "1," activity increases. To verify the methodology, Cadence SPECTRE simulations are performed for this circuit using 90 nm CMOS technology. As the activity of the circuit increases, the switching of the internal nodes increases and, consequently, the power consumption increases. Simulation results demonstrate that with the increase in workload, the minimum energy of operation increases to higher energy, as seen from Fig. 3, and optimal supply voltage moves to the lower supply voltage. For this, variability increases and yield decreases. Fig. 11 shows the effect of activity factor of the circuit on parametric yield. We also investigate the effect of the transistor sizing on the yield, which is shown in Fig. 11. The variations in the threshold voltage are slightly reduced by the increase in the device size, which in turn improves the yield if low activity of the circuit is considered. However, in an application for which activity of the circuit is high, the increase in the size of the transistors reduces the yield because the transistors' parasitic capacitance and, therefore, their power consumption are increased by increasing their sizes. Consequently, similar to those of the higher activity, optimal supply voltage moves to the lower supply voltage, where it results in higher yield loss. Table III lists the values of the center of the boxes at its final location obtained from the methodology. Table IV shows the process and circuit parameters from [15] used in this paper.

TABLE IV

Variable	Value
C_{eff}	$1.0fF$
K	$155E-6$
I_s	$1\mu A$
α	1.3
L_d	30
V_0	$\frac{1.3KT}{q}$

VII. Conclusion

An application-dependent design approach is presented for selecting the most appropriate pair of values for the supply voltage and threshold voltage, for which the yield is maximized and a near optimal tradeoff between power and performance is achieved. Parameter variations pose a major challenge in the design optimization of subthreshold VLSI circuits, especially for sub-100 nm technology. As the technology scales, design variables are more sensitive to process and environmental variations and different operating conditions of the circuit. The robustness and reliability of the design of integrated circuit is now emerging as a critical challenge in the variability-aware design especially for subthreshold circuits. Unlike the traditional deterministic PDP optimization, to find the optimal values for $V_{\rm dd}$ and $V_{\rm th}$ under uncertainty, we have considered the following factors to make the optimization technique more reliable, efficient, and complete: the application-dependent power and performance constraints, variations in the design variables due to manufacturing uncertainty, device sizing, activity factor of the circuit, and power reduction techniques. For an efficient design, a metric that is application-dependent must be chosen carefully. The chosen metric is first deterministically optimized to find the optimum point in the feasible region. Then, a design center, which is the most immune to the variations, must be identified to maximize the yield. In addition to the design robustness, this center provides the best possible tradeoff between power and performance. The design-specific feasible region is constrained by minimum performance and maximum leakage to total power ratio in the $V_{\rm dd}$-$V_{\rm th}$ plane. In order to maximize the yield for double-bounded probability distribution functions, the AFOSM method was employed. We obtained the tolerance box corresponding to 100% yield and, for the worst case design, the best location for an existing tolerance box to maximize the yield. The sensitivity of the parametric yield to the activity factor was investigated, and we see that yield decreases as activity factor increases. We showed that with the increase in device size, yield is increased with lower activity of the circuit, but for the same device size, yield is decreased with higher activity of the circuit. The yield is found to be different for different values of tolerances and for different types of distribution of design variables.

Acknowledgment

One of the authors, K. Nomani, gratefully acknowledges J. Lowder of Georgia Tech and all the reviewers for their useful comments and suggestions. The authors would also like to thank Dr. K. Ponnambalam of the University of Waterloo for providing the preliminary MATLAB codes of the optimization problem.

References

[1] A. Wang and A. Chandrakasan, "A 180 mV FFT processor using subthreshold circuit techniques," *Proc. IEEE*, pp. 301–327, 2004.
[2] A. Chandrakasan and R. Brodersen, *Low Power Digital CMOS Design*. Norwell, MA: Kluwer Academic, 1995.
[3] C. H.-I. Kim *et al.*, "Ultra-low-power DLMS adaptive filter for hearing aid applications," *IEEE Trans. VLSI Syst.*, vol. 11, pp. 1058–1067, Dec. 2003.
[4] S. R. Nassif, "Analysis and mitigation of variability in sub-threshold design," in *Proc. ISLPED*, Aug. 2005, pp. 20–25.
[5] International Technology Roadmap for Semiconductors 2005 [Online]. Available: http://public.itrs.net/
[6] S. G. Narendra, "Challenges in design choices in nanoscale CMOS," *ACM J. Emerg. Technol. Comput. Syst.*, vol. 1, no. 1, Apr. 2005.
[7] M. M. S. Hassan and Md. W. K. Nomani, "Base transit time model considering field dependent mobility for BJTs operating at high-level injection," *IEEE Trans. Electron Devices*, vol. 53, pp. 2532–2539, Oct. 2006.
[8] L. Nazhandali *et al.*, "Energy optimization of sub-threshold voltage sensor network processors," in *Proc. ACM ISCA*, 2005.
[9] B. Zahi, D. Blaauw, D. Sylvester, and K. Flaunter, "Theoretical and practical limits of dynamic voltage scaling," in *Proc. DAC*, 2004, pp. 868–873.
[10] B. H. Calhoun and A. Chandrakasan, "Characterizing and modeling minimum energy operation for sub-threshold circuits," in *Proc. ISLPED*, 2004, pp. 90–95.
[11] B. H. Calhoun *et al.*, "Device sizing for minimum energy operation in sub-threshold circuits," in *Proc. CICC*, Oct. 2004, pp. 95–98.
[12] B. Zhai, S. Hanson, D. Blaauw, and D. Sylvester, "Design for variability in DSM technologies," in *Proc. Int. Symp. Quality Electron. Design*, 2000, pp. 451–454.
[13] B. H. Calhoun, A. Wang, and A. Chandrakasan, "Modeling and sizing for minimum energy operation in sub-threshold circuits," *IEEE J. Solid-State Circuits*, vol. 40, pp. 1778–1786, Sep. 2005.
[14] J. Chen *et al.*, "Robust design of high fan-in/out sub-threshold circuits," in *Proc. IEEE Int. Conf. Comput. Design (ICCD)*, Oct. 2005, pp. 405–410.
[15] R. Gonzalez, B. Gordon, and M. Horowitz, "Supply and threshold voltage scaling for low power CMOS," *IEEE J. Solid-State Circuits*, vol. 32, pp. 1210–1216, Aug. 1997.
[16] K. Nose and T. Sakurai, "Optimization of VDD and VTH for low power and high-speed applications," in *Proc. Design Autom. Conf.*, 2000, pp. 469–474.
[17] A. Bhavnagarwala, B. Austin, K. Bowman, and J. D. Meindl, "A minimum total power methodology for projecting limits on CMOS GSI," *IEEE Trans. VLSI Syst.*, vol. 8, pp. 235–251, Jun. 2000.
[18] J. Kao, M. Miyazaki, and A. Chandrakasan, "A 175-mV multiply-accumulate unit using an adaptive supply voltage and body bias architecture," *IEEE J. Solid-State Circuits*, vol. 37, pp. 1545–1554, Nov. 2002.
[19] R. Brodersen *et al.*, "Methods for true power minimization," in *Proc. IEEE/ACM Int. Conf. Computer-Aided Design*, 2002, pp. 35–42.
[20] R. M. Swanson and J. D. Meindl, "Ion-implanted complementary MOS transistor in low-voltage circuits," *IEEE J. Solid-State Circuits*, vol. SC-7, pp. 146–153, Apr. 1972.
[21] E. Vittoz and J. Fellrath, "CMOS analog integrated citcuit based on weak inversion operation," *IEEE J. Solid-State Circuits*, vol. SC-12, pp. 224–231, Jun. 1977.
[22] L. A. Geddes, "Historical highlights in cardiac pacing," *IEEE Eng. Med. Biol. Mag.*, pp. 12–18, 1990.
[23] T. Starner, "Human-powered wearable computing," *IBM Syst. J.*, vol. 35, pp. 618–629, 1977.
[24] R. Amirtharajah and A. P. Chandrakasan, "Self-powered signal processing using vibration-based power generation," *IEEE J. Solid-State Circuits*, vol. 33, pp. 687–695, May 1998.
[25] D. Sengupta and R. Saleh, "Generalized power-delay metrics in deep submicron CMOS designs," *IEEE Trans. Computer-Aided Design*, vol. 26, pp. 183–189, Jan. 2007.
[26] T. Sakurai and A. R. Newton, "Alpha-power law MOSFET model and its application to CMOS inverter delay and other formulas," *IEEE J. Solid-State Circuits*, vol. 25, pp. 584–594, Apr. 1990.
[27] W.-C. Lee and C. Hu, "Modeling gate and substrate current due to conduction and valence band electron and hole tunneling," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2000, pp. 198–199.
[28] A. Wang, B. H. Calhoun, and A. Chandrakasan, *Sub-Threshold Design for Ultra Low-Power Syst.*. Berlin, Germany: Springer, 2006.

[29] D. Liu *et al.*, "Trading speed for low power by choice of supply and threshold voltages," *IEEE J. Solid-State Circuits*, vol. 28, pp. 10–17, Jan. 1993.
[30] P. Pant, V. K. De, and A. Chatterjee, "Simultaneous power supply, threshold voltage, and transistor size optimization for low-power operation of CMOS circuits," *IEEE Trans. VLSI Syst.*, vol. 6, pp. 538–545, Dec. 1998.
[31] H. O. Madsen, S. Krenk, and N. C. Lind, *Methods of Structural Safety*. Englewood Cliffs, NJ 07632: Prentice-Hall, 1986.
[32] S. Zhang, V. Wason, and K. Banerjee, "A probabilistic framework to estimate full-chip subthreshold leakage power distribution considering within-die and die-to-die P-T-V variations," in *Proc. ISLPED*, Jun. 2004, pp. 156–161.
[33] A. Devgan, S. Narendra, D. Blaauw, F. Najim, and K. Banerjee, "Leakage issues in IC design: Trends, estimation and avoidance," in *Proc. ICCAD*, 2003.
[34] P. Kumaraswamy, "A generalized probability density function for double-bounded random processes," *J. Hydrol.*, vol. 46, pp. 79–88, 1980.
[35] K. Agarwal and S. Nassif, "Characterizing process variation in nanometer CMOS," in *Proc. DAC*, 2007, pp. 396–397.
[36] J. Watts, N. Lu, C. Brittner, S. Grundon, and J. Oppold, "Modeling FET variation within a chip as a function of circuit design and layout choices," in *Proc. Nanotech Workshop Compact Model.*, 2005, pp. 87–92.

Md. Waliullah Khan Nomani (S'09) received the B.Sc. and M.Sc. degrees in electrical and electronic engineering from Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh, in 2004 and 2006, respectively. He is currently working towards the Ph.D. degree in electrical engineering at the University of South Carolina, Columbia.

From 2004 to 2007, he was an Assistant Professor with the Department of Electrical and Electronic Engineering, BUET. During 2007–2008, he was a Research Assistant with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. His research interests include variability aware subthreshold design, microfabrication of MEMS-based sensors for biomedical applications, modeling, simulation, and characterization of novel nanoscale devices. He is currently working with highly sensitive graphene-based sensors for a wide range of sensing applications.

Mr. Nomani is the recipient of the prestigious Graduate School Fellowship for the years 2008 to 2010.

Mohab Anis (S'98–M'03–SM'09) received the B.Sc. degree (Hon) in electronics and communication engineering from Cairo University, Cairo, Egypt, in 1997, and the M.A.Sc. and Ph.D. degrees in electrical engineering from the University of Waterloo, Waterloo, ON, Canada, in 1999 and 2003, respectively. He received the Masters of Business Administration with concentration in Innovation and Entrepreneurship from Wilfrid Laurier University, Waterloo, and the Masters of Management Sciences from University of Waterloo.

He joined the Department of Electrical and Computer Engineering, University of Waterloo in 2003, and is currently a tenured Associate Professor. During 2009–2010, he was with the Department of Electronics Engineering at the American University, Cairo. He has authored/coauthored over 100 papers in international journals and conferences and is the author of *Multi-Threshold CMOS Digital Circuits - Managing Leakage Power* (Kluwer, 2003) and *Low-Power Design of Nanometer FPGAs - Architecture and EDA* (Morgan Kaufmann, 2009). He is the Co-founder of Spry Design Automation Inc. and has published a number of papers on technology transfer. His research interests include integrated circuit design and design automation for VLSI systems in the deep submicrometer regime.

Dr. Anis was awarded the 2009 Early Research Award from Ontario's Ministry of Research and Innovation, the 2004 Douglas R. Colton Medal for Research Excellence in recognition of excellence in research leading to new understanding and novel developments in microsystems in Canada, and the 2002 International Low-Power Design Contest Award. He is an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS-I, the *Microelectronics Journal*, *Journal of Circuits, Systems and Computers*, and the ASP *Journal of Low Power Electronics*. He also served as an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II (2008-2009) and *VLSI Design* (2008-2009). He is a member of the program committee for several IEEE conferences.

Goutam Koley (S'99–M'03–SM'09) received the B.Tech. degree from the Indian Institute of Technology, Kharagpur, in 1998, the M.S. degree from the University of Massachusetts, Lowell, in 1999, and the Ph.D. degree from Cornell University, Ithaca, NY, in 2003.

Immediately after his graduation, he joined the Department of Electrical Engineering, University of South Carolina, Columbia, as an Assistant Professor, and was promoted to the rank of Associate Professor with tenure in 2009. His current research interests include MEMS- and NEMS-based sensors, bioimplantable sensors, nanoelectronic devices, and scanning probe microscopy. He did pioneering work on Kelvin probe-based characterization of wide-bandgap semiconductor materials and devices as his doctoral work. He has authored or coauthored 33 journal articles, more than 50 conference presentations/publications, and one book chapter. He has several patents and invention disclosures and has started two companies.

Dr. Koley won the prestigious National Science Foundation CAREER Award in 2009, which will support his work on NEMS-based sensors.