Design and Optimization of Dual-Threshold Circuits for Low-Voltage Low-Power Applications

Liqiong Wei, *Student Member, IEEE,* Zhanping Chen, *Student Member, IEEE,* Kaushik Roy, *Senior Member, IEEE,* Mark C. Johnson, Yibin Ye, *Member, IEEE,* and Vivek K. De, *Member, IEEE*

***Abstract*—Reduction in leakage power has become an important concern in low-voltage, low-power, and high-performance applications. In this paper,[1] we use the dual-threshold technique to reduce leakage power by assigning a high-threshold voltage to some transistors in noncritical paths, and using low-threshold transistors in critical path(s). In order to achieve the best leakage power saving under target performance constraints, an algorithm is presented for selecting and assigning an optimal high-threshold voltage. A general leakage current model which has been verified by HSPICE simulations is used to estimate leakage power. Results show that the dual-threshold technique is good for leakage power reduction during both standby and active modes. For some ISCAS benchmark circuits, the leakage power can be reduced by more than 80%. The total active power saving can be around 50% and 20% at low- and high-switching activities, respectively.**

***Index Terms*—CMOS, critical-path, delay, high performance, low-power design, low voltage, power estimation.**

I. Introduction

With the growing use of portable and wireless electronic systems, reduction in power consumption has become more and more important in today's very large scale integration (VLSI) circuit and system designs [1], [2].

In CMOS digital circuits, power dissipation consists of dynamic and static components. Since dynamic power is proportional to the square of supply voltage V_{dd} and static power is proportional to V_{dd}, lowering V_{dd} is obviously the most effective way to reduce power consumption. With the scaling of supply voltage, transistor threshold voltage ($V_{\rm th}$) should also be scaled in order to satisfy the performance requirements. Unfortunately, such scaling leads to an increase in leakage current which becomes an important concern in low-voltage high-performance circuit designs.

Multiple thresholds can be used to deal with the leakage problem in low-voltage high-performance CMOS circuits. This technique has commonly been used in DRAM chips by raising threshold voltages of the array devices with a fixed body bias [4]. For large scaled integration (LSI) circuits, multithreshold-voltage CMOS (MTCMOS) circuit technology was proposed by inserting high-threshold devices in series to normal circuitry [14], [16]. However, only the standby leakage power can be reduced and the large inserted MOSFET's will increase the area and delay. Moreover, the data retention must also be considered [18].

For a logic circuit, a higher threshold voltage can be assigned to some transistors in noncritical paths so as to reduce leakage current, while the performance is maintained due to the low-threshold transistors in the critical path(s). Therefore, no additional transistors are required, and both high performance and low power can be achieved simultaneously. This dual-threshold technique is good for leakage power reduction during both standby and active modes.

Dual-threshold voltages can be achieved by body biasing [17]. A source to well reverse bias can be applied to some transistors to achieve high thresholds. Recently, a dual-$V_{\rm th}$ MOSFET process was developed [5], which makes the implementation of dual-$V_{\rm th}$ logic circuits more feasible.

However, due to the complexity of a circuit, not all the transistors in noncritical paths can be assigned a high-threshold voltage, otherwise, the critical path may change, thereby increasing the critical delay. We presented a breadth-first search (BFS)-based algorithm for selecting and assigning an optimal high-$V_{\rm th}$ in [12]. In this paper, a levelization back-tracing algorithm is given to achieve the best leakage power saving under performance constraints. A leakage model which has been verified by HSPICE simulations is used to estimate leakage power of a circuit. Standby leakage power, active leakage power, dynamic power, and total active power of single-$V_{\rm th}$ and dual-$V_{\rm th}$ circuits are compared and analyzed.

II. Delay Model

A. Definitions

A combinational circuit can be represented as a directed acyclic graph $G(V, E)$. Each node (except for primary inputs and outputs) in the graph maps to a logic gate in the circuit while each edge maps to a path.

The propagation delay through node x, denoted as $t_p(x)$, defines how quickly the output responds to a change in the input. The propagation delay of a path π_j is the sum of the propagation delays $t_p(i)$ of each node i along this path, which can be expressed as $Pd(\pi_j) = \Sigma t_p(i)$.

The arrival time $[T_a(x)]$ is the propagation delay of each fan-in path of node x. Among all the fan-in paths, there exists a path (or paths) which has a maximum propagation delay

Manuscript received December 12, 1997; revised May 15, 1998. This work was supported in part by the Defense Advanced Research Projects Agency under Contract F33615-95-C-1625, by the National Science Foundation Career Award 9501869-MIP, and by Intel Corporation.

L. Wei, Z. Chen, K. Roy, and M. C. Johnson are with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA.

Y. Ye and V. K. De are with Microcomputer Research Laboratories, Intel Corporation, Hillsboro, OR 97124 USA.

Publisher Item Identifier S 1063-8210(99)01552-8.

[1]See the Guest Editorial of the Special Section on Low-Power Electronics and Design of the IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 6, pp. 518–519, Dec. 1998.

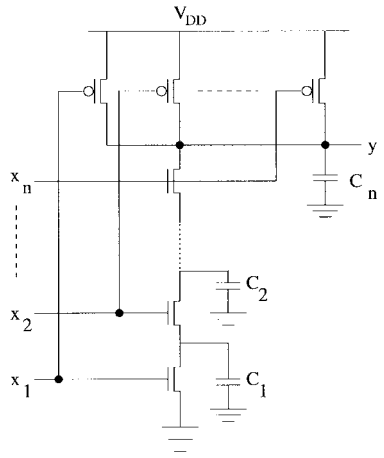


Fig. 1. n-input NAND gate.

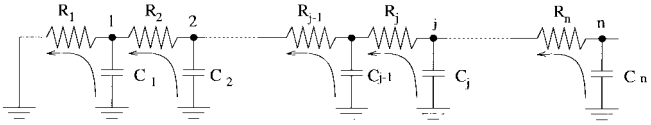


Fig. 2. Equivalent PDN of n-input NAND gate.

$T_{\max}(x)$, where

$$T_{\max}(x) = \max_{i \in \text{all fanins}} \Big\{ T_a(x)[i] \Big\}. \tag{1}$$

The departure time ($T_l(x)$) of node x is defined as

$$T_l(x) = T_{\max}(x) + t_p(x). \tag{2}$$

The path which determines the maximum speed of the circuit is called the critical path. There may be more than one critical path. Critical delay (T_{critical}) is the delay along the critical path.

B. Elmore Delay Model

Let us look at an n-input NAND gate (Fig. 1). The NAND gate can be analyzed using an equivalent RC network. Each MOS transistor has an equivalent on-resistance R_j, and each node in the n-input NAND gate has a capacitance C_j (j varies from 1 to n). The equivalent RC network of the pull-down network (PDN) is shown in Fig. 2.

The worst case occurs when all C_j's are discharged simultaneously. Based on the Elmore delay model [6], the worst case delay (t_{PHL}) of the PDN is given by

$$t_{\text{PHL}} = 0.69 \sum_{j=1}^{n} \left(C_j \sum_{k=1}^{j} R_k \right). \tag{3}$$

The capacitance of each internal node j ($j = 1, \cdots, n-1$) in the n-input NAND gate is given as follows:

$$C_j = 2\, C_{dN} \tag{4}$$

where C_{dN} is the diffusion capacitance of an nMOSFET. The capacitance of the gate at output is given by

$$C_n = F_O\,(C_{\text{gP}} + C_{\text{gN}}) + F_I\, C_{dP} + C_{dN} + F_O\, C_{\text{int}} \tag{5}$$

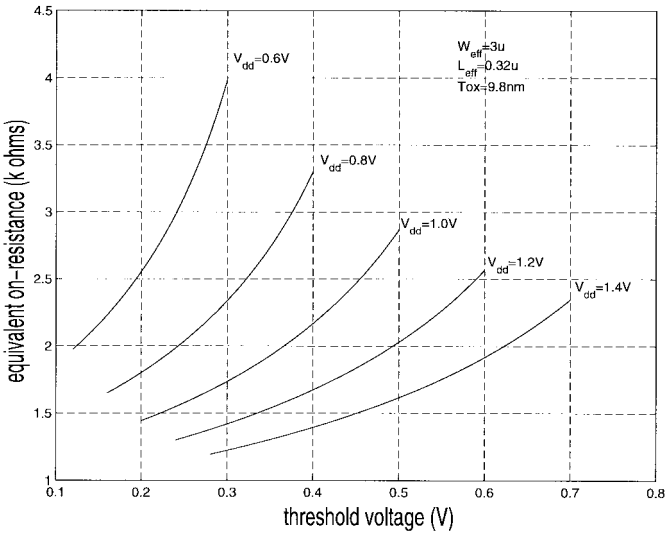


Fig. 3. Relationship between R_N and V_{th}.

where C_{dP} is the diffusion capacitance of a pMOSFET. C_{gP} and C_{gN} are the gate capacitances of pMOS and nMOS transistors, respectively. C_{int} represents the interconnect capacitance per fan-out. F_O is the number of fan-outs of the gate, while F_I represents the number of fan-ins. For an n-input NAND gate, we have $F_I = n$.

Assuming that each nMOSFET has the same on-resistance, the worst-case delay of the PDN can be simplified as follows:

$$t_{\text{PHL}} = 0.69[R_N\, C_{dN}\, F_I\, (F_I - 1) + F_I\, R_N\, C_n]. \tag{6}$$

Although the on-resistance depends on the operation point and varies during the switching transient, we still can make a reasonable approximation by using a fixed value. This value is the average of the resistances at the end points of the transitions [6]. The on-resistance of an nMOSFET is given by

$$\begin{aligned} R_N &= \frac{R_{\text{nMOS}}|_{V_{\text{out}}=V_{dd}} + R_{\text{nMOS}}|_{V_{\text{out}}=V_{dd}/2}}{2} \\ &= \frac{1}{2}\left(\frac{V_{\text{DS}}}{I_D}\bigg|_{V_{\text{out}}=V_{dd}} + \frac{V_{\text{DS}}}{I_D}\bigg|_{V_{\text{out}}=V_{dd}/2} \right) \\ &= \frac{V_{dd}}{k_N (V_{dd} - V_{TN})^{\alpha}} + \frac{V_{dd}}{k_N \left[2(V_{dd} - V_{TN}) V_{dd} - \frac{V_{dd}^2}{2} \right]} \end{aligned} \tag{7}$$

where V_{TN} is the threshold voltage of an nMOSFET and k_N is the gain factor. The constant α is 2 and 1.3 for long-channel and short-channel MOSFET's, respectively. The relationship between R_N and V_{TN} at different supply voltages is shown in Fig. 3. For a pMOSFET, the on-resistance (R_P) can be evaluated similarly. For simplicity, we assume that $|V_{TN}| = |V_{TP}| = |V_{\text{th}}|$ and $R_N = R_P$.

For the pull-up network (PUN), the worst case occurs when only one pMOS transistor is "on." The worst case delay (t_{PLH}) can be expressed by

$$t_{\text{PLH}} = 0.69 R_P\, C_n. \tag{8}$$

The worst-case propagation delay of a CMOS gate is

$$t_p = (t_{\text{PHL}} + t_{\text{PLH}})/2. \tag{9}$$

Following a similar procedure, we can get the worst-case propagation delay of the other gates.

Process variation will introduce variations in the transistor parameters such as threshold voltage, which will influence the circuit performance. For the worst-case propagation delay, threshold variation can be considered by changing $V_{\rm th}$ to $V_{\rm th}+\Delta V_{\rm th}$ in the on-resistance equation, where $\Delta V_{\rm th}$ is the threshold variation. Since the total threshold variation can be controlled to be within 20 mV and the local variations are significantly smaller than global variations [20], we assume that $\Delta V_{\rm th}$ is 20 mV. When the supply voltage is 1 V and threshold voltage is 0.2 V, the propagation delay variation can be less than 5%.

III. Power Estimation

For a CMOS circuit, total power includes dynamic and static components at active mode. It can be expressed as $P_T = P_{\rm dyn}+P_{{\rm leak}_a}$, where $P_{\rm dyn}$ and $P_{{\rm leak}_a}$ are dynamic power and active leakage power. In standby mode, the power dissipation ($P_{{\rm leak}_s}$) is mainly because of the standby leakage current. In this section, we will present the power-estimation methods used in our simulation: a Monte Carlo-based statistical method to estimate dynamic power and an accurate leakage power-estimation method, which considers circuit topology as well as signal levels.

A. Dynamic Power Estimation

Ignoring power dissipation due to direct-path short-circuit current, dynamic power of a CMOS circuit is due to the charging and discharging of load capacitances and internal-node capacitances, which can be evaluated as follows:

$$\begin{aligned} P_{\rm dyn} &= P_{{\rm dyn}_o}+P_{{\rm dyn}_j} \\ &= \frac{1}{2}f\left(V_{dd}^2\sum_i(\alpha_i C_{L_i})+V_{dd}\sum_i\sum_j(\alpha_{ij}C_{ij}V_{ij})\right) \end{aligned} \tag{10}$$

where $P_{{\rm dyn}_o}$ and $P_{{\rm dyn}_j}$ are the dynamic power due to the load capacitances and the dynamic power due to the internal-node capacitances, respectively. f is the clock frequency. i represents the gate i and j denotes the jth internal node in a gate. V_{ij} is the voltage swing of the jth internal node of gate i, which equals to $V_{dd}-V_{\rm th}$. α_i and α_{ij} are the switching activities (the probability of switching) at gate i and at the jth internal node of gate i, respectively. C_{L_i} and C_{ij} are the load capacitance and the jth internal-node capacitance of gate i, respectively.

The switching activity can be determined by a Monte Carlo based statistical method. The basic idea is to simulate a circuit with random patterns applied to the primary inputs. Such patterns conform to the given signal probabilities (the probability of a signal being logic ONE) and activities [9]. A stopping criterion is used to determine when node activities have converged to the correct values [7], [8].

B. Static Power Estimation

The leakage power of a CMOS circuit is determined by the leakage current through each transistor, which has two main sources: reversed-biased diode-junction leakage current and subthreshold leakage current. Diode-junction leakage is very small and can be ignored [6]. Subthreshold leakage exponentially increases with the reduction of threshold voltage [3], making it critical for low-voltage circuit design. Therefore, in our simulation, we focus on subthreshold leakage power estimation.

In order to estimate leakage power accurately, a general transistor model [10], [11], which considers sub-zero gate-to-source voltage ($V_{\rm GS}$) for nMOS and super-zero $V_{\rm GS}$ for pMOS (occurs when multiple series connected transistors are turned off), body effect and drain-induced barrier lowering (DIBL), is used. The following analysis is done for nMOSFET's, but is equally applicable to pMOSFET's.

From a Berkeley short-channel IGFET (BSIM) MOS transistor model [13], the subthreshold current of a MOSFET can be modeled as

$$\begin{aligned} I_{\rm sub} = A\,e^{q/n'kT}&(V_G-V_S-V_{{\rm TH}_0}-\gamma' V_S+\eta V_{\rm DS}) \\ &\cdot(1-e^{-qV_{\rm DS}/kT}) \end{aligned} \tag{11}$$

where $A=\mu_0 C_{\rm ox}W_{\rm eff}/L_{\rm eff}(kT/q)^2e^{1.8}$. $C_{\rm ox}$ is the gate oxide capacitance per unit area. μ_0 is the zero bias mobility. n' is the subthreshold swing coefficient of the transistor. $V_{{\rm TH}_0}$ is the zero bias threshold voltage. The body effect for small values of V_S is very nearly linear. It is represented by the term $\gamma' V_S$, where γ' is the linearized body effect coefficient. η is the DIBL coefficient.

If transistors are connected in parallel and are both turned off (such as in the pull-down network of an NOR gate), then the values of $V_{\rm DS}$ and V_S are the same for each transistor. The leakage contribution of each transistor can be calculated separately and added together. However, things become more complicated if they are in series. Consider the pull-down network of an n-input NAND gate. Without loss of generality, we consider the case where all nMOS transistors are turned off. The quiescent subthreshold leakage through each transistor must be identical, given that other leakage components are negligible. Thus, we equate the current of the first (top) and second transistor. Equation (12) can be obtained by solving for $V_{{\rm DS}_2}$ in terms of V_{dd} (we assume that $V_{S_1}\ll V_{dd}$) as follows:

$$V_{{\rm DS}_2} = \frac{n'kT}{q(1+2\eta+\gamma')}\ln\left(\frac{A_1}{A_2}e^{q\eta V_{dd}/n'kT}+1\right) \tag{12}$$

$$V_{{\rm DS}_i} = \frac{n'kT}{q(1+\gamma')}\ln\left(1+\frac{A_{i-1}}{A_i}\left(1-e^{(-q/kT)V_{{\rm DS}_{i-1}}}\right)\right). \tag{13}$$

One can similarly equate the current through the $(i-1)$th and ith transistors, solving for $V_{{\rm DS}_i}$ in terms of $V_{{\rm DS}_{i-1}}$. This results in (13). (A more detailed derivation of (12) and (13) can be found in [10].) Equation (13) can be used iteratively to find $V_{{\rm DS}_i}$ ($3\le i\le N$). Finally, the voltage offset at the source of each transistor is given by $V_{S_i}=\sum_{j=i+1}^{N}V_{{\rm DS}_j}$, and $V_{{\rm DS}_1}$ can be determined by $V_{dd}-V_{S_1}$. Now (11) can be used to

calculate the quiescent leakage for any transistor in the stack, which should be the same for each transistor. Finally, the total leakage power can be determined by

$$P_{\text{leak}} = \sum_i I_{\text{DS}_i} V_{\text{DS}_i}. \tag{14}$$

The sensitivity of P_{leak} with respect to V_{th} is given by

$$\frac{\partial P_{\text{leak}}}{\partial V_{\text{th}}} = \sum_i \frac{\partial I_{\text{DS}_i}}{\partial V_{\text{th}}} V_{\text{DS}_i} = -\frac{q}{n'kT} P_{\text{leak}} \tag{15}$$

where the summation is taken for all transistors.

The general method of computing leakage power for a large circuit involves the following steps. Given a particular set of circuit input values, determine which pull-up and PDN's are turned off. Within each network, the transistors which are turned on can be treated as short circuits. Transistors that are parallel to a transistor that is turned on can be eliminated from the leakage calculation. Given the resulting simplified network, estimate V_{DS} for the remaining transistors using (12) and (13). Finally, the magnitude of leakage current and resulting leakage power can be computed.

The above method is very suitable for leakage power estimation during standby mode. In active mode, the time required for the leakage current in transistor stacks to converge to its final value is determined by the internal-node capacitance, input conditions, and subthreshold leakage current [15]. Subthreshold leakage current strongly depends on V_{th} and temperature. If the internal-node capacitance is small and temperature is high, the given method can also be used to estimate active leakage power of low-V_{th} circuits, especially at low-switching activities. Considering the fact that standby leakage current depends on input signal levels, the average leakage power can be evaluated with random patterns applied to primary inputs.

IV. Algorithm

Due to the exponential relationship between threshold voltage and drain current in the weak inversion region, a higher threshold voltage will significantly reduce leakage current, thereby reducing leakage power. However, Fig. 3 indicates that a higher threshold voltage will increase the equivalent on-resistance of each transistor, which results in a higher propagation delay. Normally, threshold voltage is empirically defined to be around 20% of supply voltage to maintain the performance of a circuit [19]. For low-supply voltage circuits, the threshold voltage could be very small, leading to a large leakage current.

This problem can be circumvented by using dual-threshold voltages. A low V_{th} is assigned to the transistors in critical path(s) in order to achieve high performance, while a high V_{th} may be assigned to some transistors in noncritical paths to reduce leakage power. The lower bound of low V_{th} is determined by noise margin. The possible high V_{th} value should be in the range from low V_{th} to $0.5V_{dd}$. However, not all the transistors in noncritical paths can be assigned the high V_{th}. Otherwise, some noncritical paths may become critical. Whether a node can be assigned a higher V_{th} depends on the value of the high threshold. If it is too small, there is little difference of propagation delay between low-V_{th} and high-V_{th} transistors. Hence, more nodes can be assigned high-V_{th} without influencing the critical delay, but the leakage current improvement for each high-V_{th} transistor would be small. On the other hand, if the high-threshold voltage is too large, the leakage current can be reduced by a large amount for each such transistor. However, fewer nodes can be modified. Hence, among the allowable values for high-threshold voltage, there exist an optimal one. In this section, a levelization back-tracing algorithm is given to select and assign the optimal high-V_{th}.

The first step in our algorithm is to initialize a circuit with a single low V_{th}. During the initialization procedure, each node is assigned a level number. The level of each primary input is defined to be zero. The level of a node x, denoted as $l(x)$, can be calculated as $l(x) = 1 + \max\{l(j)\}$, where j varies for all fan-in nodes of node x. For each primary input x, $t_p(x) = 0$, $T_a(x) = 0$, $T_l(x) = 0$, and $T_{\max}(x) = 0$. For each node x in level 1, $T_a(x) = 0$, $T_{\max}(x) = 0$, and $T_l = t_p(x)$. Therefore, level by level, the parameters ($t_p(x)$, $T_a(x)$, $T_l(x)$, and $T_{\max}(x)$) associated with each node x can be computed by (1) and (2) during the initialization procedure. By checking all the primary outputs and then back-tracing, the critical delay and critical path(s) can be identified using a first-in-first-out (FIFO) queue Q.

The pseudo-code for the initialization procedure is shown below. Note that, PO does not map to a gate in a circuit, and each PO has only one fan-in gate [$\text{fanin}(PO)$].

```
Initialization () {
  Assign a level number to each node
  Calculate the propagation delay t_p(x) of each node x
  Calculate T_max(x) and T_l(x) of each node x level by level
  Identify T_critical by checking the maximum T_l(fanin(PO))
  For each primary output PO {
    If (T_l(fanin(PO)) = T_critical)
      Mark fanin(PO) as a node in critical path
      Add node fanin(PO) into a FIFO queue Q
  }
  While (Q not empty) {
    Remove node x from Q
    For each fan-in y of node x {
      If [(T_l(y) = T_max(x)) && (y is not a primary input)]
        Mark x as a node in critical path
        Add node y into queue Q
    }
  }
```

The next step is to assign a high threshold to some transistors on noncritical paths under performance constraints. This is performed by back-tracing the slack of each node level by level. Slack of a node [$T_\delta(x)$] denotes the amount by which the gate can be slowed down without affecting the circuit performance. For the nodes in critical path(s), slack is zero. For a PO,

$$T_\delta(PO) = T_{\text{critical}} - T_l(f_{\text{anin}}(PO)). \tag{16}$$

For any other node x, $T_\delta(x)$ can be expressed by

$$T_\delta(x) = \min_{\forall y = \text{fanout}(x)} \{(T_\delta(y) + T_{\max}(y) - T_l(x))\} \tag{17}$$

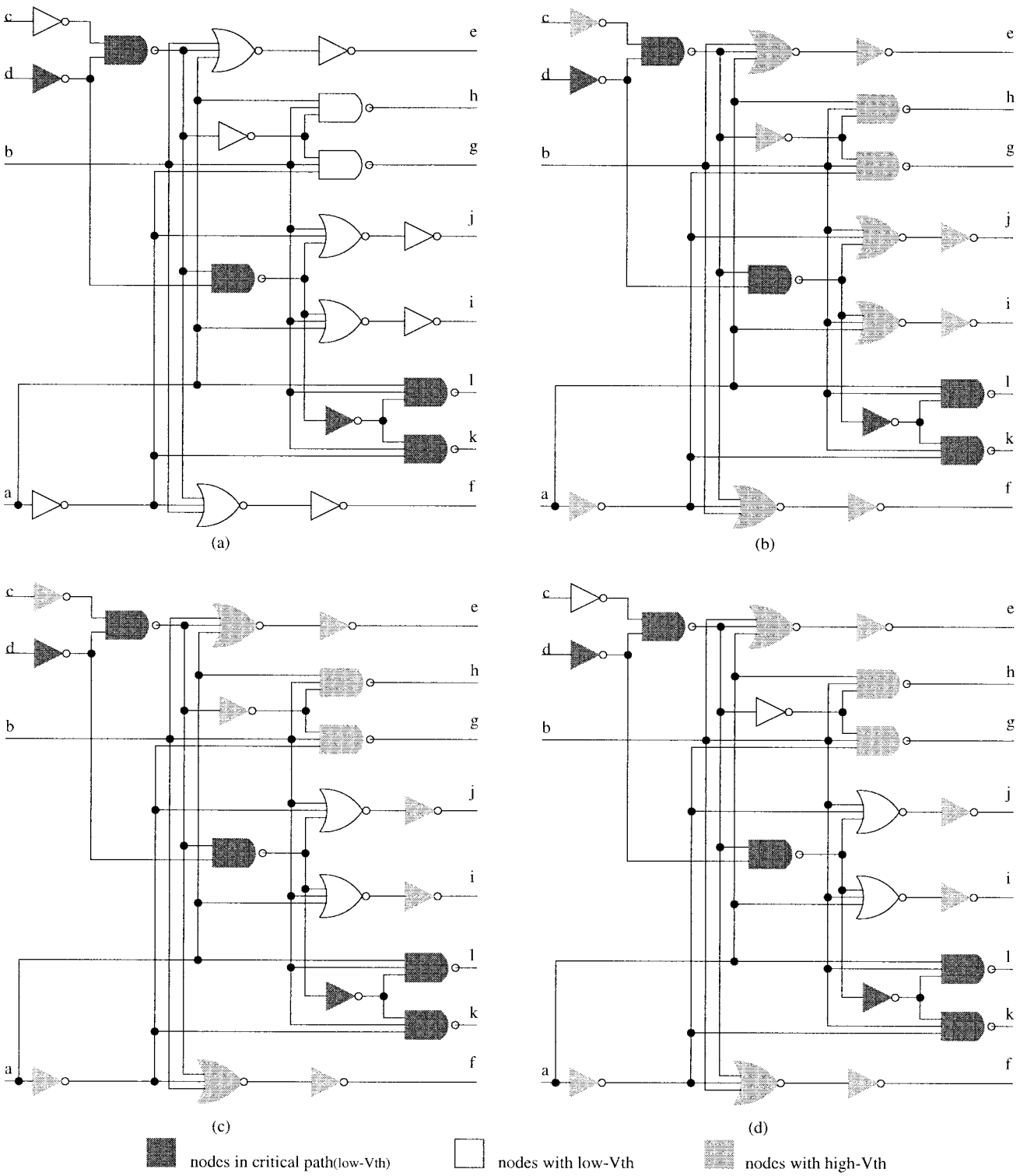


Fig. 4. An example circuit. (a) Original circuit $V_{dd} = 1$ V, $V_{\text{th}_1} = 0.2$ V. (b) $V_{\text{th}_2} = 0.25$ V. (c) $V_{\text{th}_2} = 0.395$ V. (d) $V_{\text{th}_2} = 0.46$ V.

where $f_{\text{anin}}(x)$ and $f_{\text{anout}}(x)$ are the fan-in nodes and fan-out nodes of node x, respectively. Since the nodes are traversed backward level by level, when we deal with x, the slacks of its fan-out nodes are already known. Equation (17) ensures that the propagation delay of the path(s) through x is no greater than the critical delay.

The procedure for choosing a high threshold works as follows. Since the circuit has been levelized during the initialization procedure, from the nodes on maximum level, the program will explore every node level by level to determine its slack. By definition, for each node in a single threshold circuit, its slack (T_δ) is no less than zero. Increasing the threshold voltage of a node can result in a higher propagation delay and departure time of this node. Therefore, slack will decrease. Whether a node should be assigned to a high threshold depends on whether its slack is still positive if its threshold is changed to high threshold. If the slack is still positive, this node will be assigned to the high threshold. Since the slack of each node on

critical path is zero, the threshold voltage of these transistors will not be changed and, hence, the performance is maintained. The pseudo-code of this subroutine is shown below:

High-$V_{\rm th}$-Assignment ($V_{\rm th_2}$) {
  present_level = maximum level
  while (present_level > 0) {
    For each node x in present_level {
      Calculate $t_p(x)$, $T_l(x)$, and $T_\delta(x)$ for high threshold $V_{\rm th_2}$
      if $T_\delta(x) \geq 0$
        Assign $V_{\rm th_2}$ to x
        Assign $t_p(x)$, $T_l(x)$, and $T_\delta(x)$ for $V_{\rm th_2}$ to x
      else
        Keep $t_p(x)$, $T_l(x)$, and $T_\delta(x)$ for initial low $V_{\rm th}$ for x
    }
    present_level = present_level - 1
  }

Using the method described in Section III, the dynamic power and leakage power of the circuit corresponding to different high-threshold voltages can be evaluated. By comparing the values of leakage power, an optimal high threshold ($\rm opt\_V_{th_2}$) can be found. After updating the network for $\rm opt\_V_{th_2}$, the circuit can be transferred into a SPICE net list and simulated using HSPICE to verify some of the results. The procedure is outlined below:

Optimal-High-$V_{\rm th}$ () {
  For each high $V_{\rm th}$ v of a set in ($0.2V_{dd}$, $0.5V_{dd}$) {
    Initialization
    High-$V_{\rm th}$-Assignment (v)
    Estimate $P_{\rm leak}$ and $P_{\rm dyn}$
    If $P_{\rm leak}$ is the least power so far
      $P_{\rm leak_{min}} = P_{\rm leak}$
      $\rm opt\_V_{th_2} = v$
    }
  Update network with $\rm opt\_V_{th_2}$
  Transfer the network into SPICE netlist

V. Implementation and Results

The method to reduce leakage power using dual-threshold-voltage transistors has been implemented in C under the Berkeley SIS environment. In order to simplify the analysis, technology mapping was used to map the circuits to a library which contains NAND gates, NOR gates, and inverters. All the simulation results were obtained using HSPICE with the BSIM model for a 0.5-μm MOSIS process. The available MOSIS models do not include measured subthreshold characteristics, so we have estimated the subthreshold swing and related parameters from threshold voltage parameters using the technique derived by Kang *et. al.* [21]. A subthreshold swing coefficient of approximately 1.44 was estimated and incorporated into the BSIM model. In order to approximate the behavior of low-threshold devices, we modify the flat-band voltage parameter (VFB0). The effective channel length was 0.32 μm and the gate–oxide thickness was 9.8 nm. The effective channel widths for pMOSFET's and nMOSFET's were assumed to be 10.5 and 3 μm, respectively. For ISCAS benchmark circuits, we assume that the diffusion capacitance is 20% of the gate capacitance.

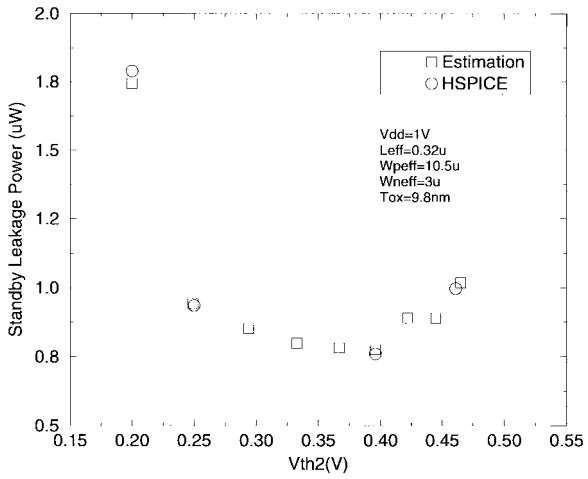


Fig. 5. Standby leakage power with different $V_{\rm th_2}$.

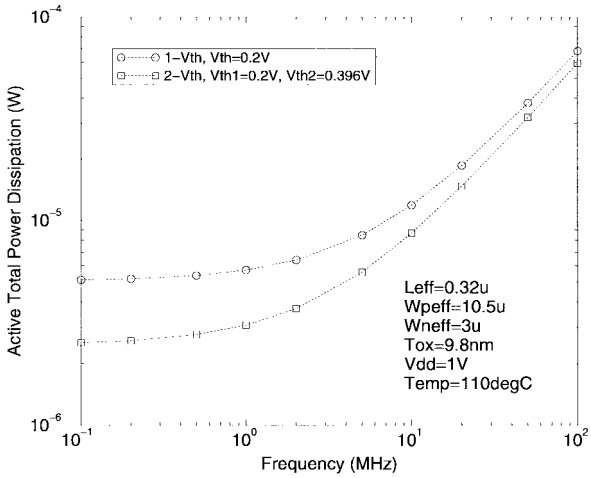


Fig. 6. Active total power dissipation at different frequencies.

Fig. 4 gives an example circuit to show how our algorithm works. Fig. 4(a) is the original single-$V_{\rm th}$ circuit, where the supply voltage is 1 V and the threshold voltage is 0.2 V. Fig. 4(b)–(d) shows the dual-$V_{\rm th}$ circuits with the low $V_{\rm th}$ of 0.2 V and the high-$V_{\rm th}$ of 0.25, 0.396, and 0.46 V, respectively. Note that the critical paths and critical delays are maintained after the assignment.

Fig. 5 shows the standby leakage power of the example circuit with different thresholds. The supply voltage is 1 V. At 25 °C, the original low-threshold voltage is 0.2 V and the high-threshold voltage ($V_{\rm th_2}$) varies from 0.2 to 0.5 V ($V_{\rm th_2} = 0.2$ V represents the single low-threshold circuit). The squares represent the leakage power obtained by our estimation technique while the circles denote the leakage power obtained by HSPICE. Clearly, the estimation results fit well with HSPICE simulation results. The convex point of the curve indicate that there exits an optimal high-threshold voltage (0.396 V) which leads to a 57.5% saving in standby leakage power.

Fig. 6 shows the HSPICE simulation results of the total active power dissipation of single-$V_{\rm th}$ and dual-$V_{\rm th}$ circuits

TABLE I
ACTIVE AND STANDBY LEAKAGE POWER SAVINGS FOR DUAL-$V_{\rm th}$ CIRCUITS

Circuit Chosen	PI/PO #	Max level	Gate #	$opt\_V_{th_2}$ (mV)	P_{leak_a} (μW) 1-V_{th}	2-V_{th}	%	P_{leak_s} (μW) 1-Vth	2-Vth	%
C432	36/7	23	206	367	108.62	45.02	58.55	4.41	1.73	60.77
C499	41/32	28	532	367	261.08	123.89	52.55	10.58	4.95	53.21
C880	60/26	22	353	396	179.35	25.01	86.06	7.3	0.95	86.99
C1355	41/32	28	517	367	252.09	126.45	49.84	10.18	5.04	50.49
C1908	33/25	35	615	333	301.29	67.79	77.50	12.21	2.45	79.93
C2670	233/140	26	807	396	414.30	71.78	82.67	16.87	2.77	83.58
C3540	50/22	41	1131	333	587.78	82.78	85.92	23.75	2.68	88.72
C5315	178/123	47	1778	367	887.00	126.24	85.77	36.12	4.57	87.35
C6288	32/32	123	2400	333	1364.56	796.00	41.67	56.08	31.75	43.38
C7552	207/108	47	2803	333	1466.82	224.87	84.67	59.72	7.42	87.58

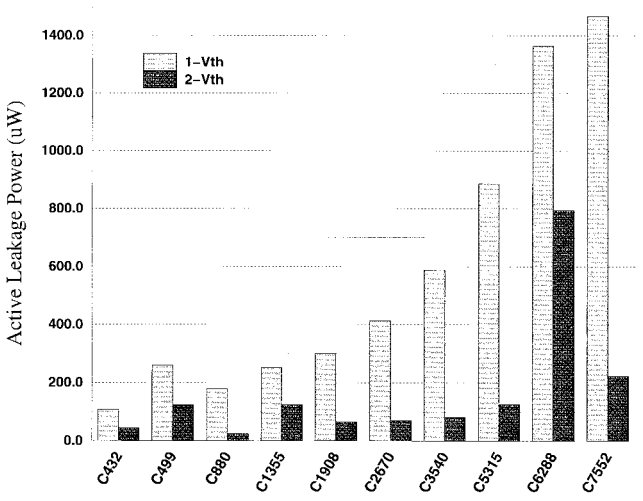


Fig. 7. Active leakage power savings for ISCAS benchmarks.

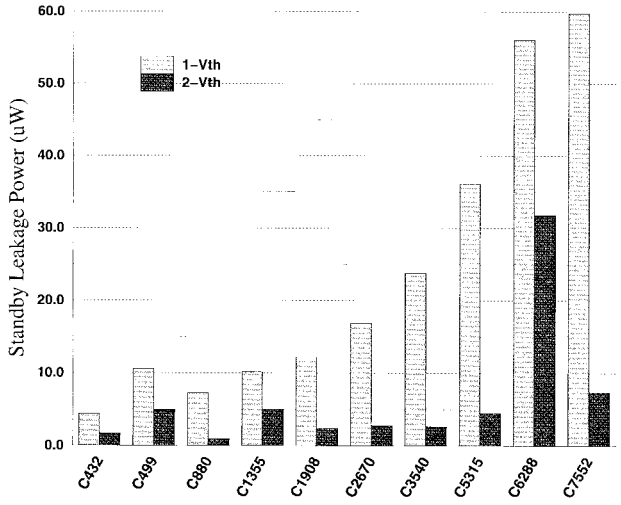


Fig. 8. Standby leakage power savings for ISCAS benchmarks.

at different frequencies. The circuits were simulated at 1-V supply voltage while the primary input switching activity is 0.5. The threshold voltage of single-$V_{\rm th}$ circuit was 0.2 V at 110 °C. The low- and high-threshold voltages of dual-$V_{\rm th}$ circuit were 0.2 and 0.396 V, respectively. In addition to leakage power saving, the dynamic power is reduced due to the reduction of internal-node voltage swing for high-threshold gates.

Table I and Figs. 7 and 8 show the optimal high $V_{\rm th}$, active, and standby leakage power savings for ISCAS benchmark circuits. In this experiment, V_{dd} was 1 V. In active mode (the circuit temperature was assumed to be 110 °C), the low $V_{\rm th}$ was 0.2 V and high $V_{\rm th}$ was the optimal value obtained from the levelization back-tracing algorithm given in Section IV. In standby mode, the circuit temperature was assumed to be 25 °C. Since $V_{\rm th}$ increases about 0.8 mV for every 1 °C decrease in temperature, $V_{\rm th}$ at standby mode is about 68 mV higher than the corresponding $V_{\rm th}$ in the active mode. Results show that both active and standby leakage power can be reduced by more than 80% for some of the circuits. Since the sensitivity of leakage power to threshold voltage is proportional to the leakage power itself, the dual $V_{\rm th}$ technique, which reduces leakage power, can reduce the sensitivity of leakage power to $V_{\rm th}$. The percentages of high-$V_{\rm th}$ transistors and gates for different dual-$V_{\rm th}$ benchmark circuits are illustrated in Fig. 9. Results indicate that the percentage of high-$V_{\rm th}$ transistors can be more than 80%. Compared to a BFS-based algorithm, which can provide 50% leakage power savings and 60% high-$V_{\rm th}$ transistors for some benchmark circuits, levelization back-tracing algorithm can achieve more leakage power savings.

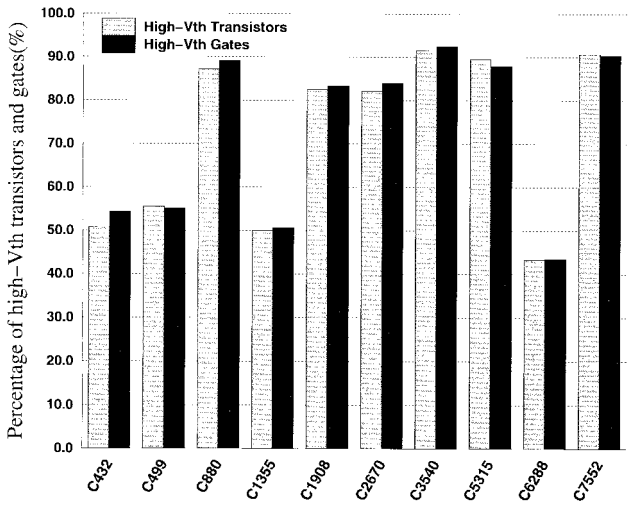


Fig. 9. Percentage of high $V_{\rm th}$ gates and transistors for ISCAS Benchmarks.

TABLE II
TOTAL ACTIVE POWER SAVINGS FOR DUAL-$V_{\rm th}$ CIRCUITS

Circuit Chosen	Critical delay(ns)	Input activity	P_{dyn_o} (μW)	P_{dyn_j} (μW)		P_{dyn} (μW)		P_T (μW)		
				1-V_{th}	2-V_{th}	1-V_{th}	2-V_{th}	1-V_{th}	2-V_{th}	%
C432	3.36	0.03	36.6	1.4	1.2	38.0	37.8	146.6	82.8	43.5
		0.3	248.1	10.6	9.3	258.7	257.4	367.3	302.4	17.7
C499	1.45	0.03	317.0	9.6	8.5	326.6	325.5	587.7	449.4	23.5
		0.3	1623.4	47.6	43	1671	1666.4	1932.1	1790.3	7.3
C880	1.5	0.03	126.1	6.6	5.3	132.7	131.4	312.1	156.4	49.9
		0.3	854.8	43.1	34.2	897.9	889	1077.3	914	15.2
C1355	1.63	0.03	281.3	9.1	8.8	290.4	290.1	542.5	416.55	23.2
		0.3	1304.8	44.8	40.7	1349.6	1345.5	1601.7	1472.0	8.1
C1908	2.25	0.03	238.5	8.9	7.5	247.4	246	548.7	313.8	42.8
		0.3	1209.0	45.3	38.7	1254.3	1247.7	1555.6	1315.5	15.4
C2670	2.81	0.03	188.85	5.76	4.5	194.6	193.4	608.9	265.2	56.4
		0.3	1238.8	38.8	30.9	1277.6	1269.7	1691.9	1341.5	20.7
C3540	2.95	0.03	274.1	10.7	8.7	284.8	282.8	872.6	365.6	58.1
		0.3	1645.5	65.8	55.9	1711.3	1701.4	2299.1	1784.2	22.4
C5315	2.32	0.03	572.9	17.6	14.4	590.5	587.3	1477.5	713.5	51.7
		0.03	3783.0	121.3	98.5	3904.3	3881.5	4791.3	4007.7	16.4
C6288	7.11	0.03	418.6	13.3	12.5	431.9	431.1	1796.5	1227.1	31.7
		0.3	1859.4	60.3	56.3	1919.7	1915.7	3284.3	2711.7	17.4
C7552	3.27	0.03	664.8	19.7	16.6	684.5	681.4	2151.3	906.3	57.9
		0.3	4323.3	127.8	108.2	4451.1	4431.5	5917.9	4656.4	21.3

Total active power is an important concern for a high-performance system. Table II gives the critical delays, dynamic power dissipations due to output node, and internal-node transitions ($P_{\rm dyn_o}$ and $P_{\rm dyn_j}$), total dynamic power dissipations ($P_{\rm dyn}$) and total active power dissipations (P_T) for single-$V_{\rm th}$ and dual-$V_{\rm th}$ circuits with maximum clock frequency ($1/T_{\rm critical}$) and different input activities. Since dual $V_{\rm th}$ technique can reduce the active leakage power and the dynamic power due to the internal-node capacitance, for some benchmark circuits, the total active power can be reduced by around 50% and 20% at low- and high-switching activities, respectively. For mobile systems, since the system may be idle for a long time, the standby leakage power can not be ignored. Dual $V_{\rm th}$ is a promising technique for reduction of both active and standby leakage power.

VI. CONCLUSION

In this paper, we present a method to design and optimize low-voltage dual-$V_{\rm th}$ CMOS circuits. In order to reduce leakage power under performance constraints, starting with a single low $V_{\rm th}$ circuit, an algorithm for selecting and assigning an optimal high-threshold voltage is proposed. For accurate leakage power estimation, a leakage current model, which has been verified by HSPICE simulations, is used. Results for ISCAS benchmark circuits show that both active and standby leakage power can be reduced by 80% for some circuits under performance constraints. The total active power can be reduced by around 50% and 20% at low and high-switching activities, respectively. Reduction of both active power and standby leakage power without area and delay penalty makes a dual $V_{\rm th}$ technique a good candidate of high-performance low-power applications.

REFERENCES

[1] J. D. Meindl, "Low power microelectronics: Retrospect and prospect," *Proc. IEEE*, vol. 83, p. 619, Apr. 1995.

[2] A. P. Chandrakasan, S. Sheng, and R. W. Brodersen, "Low-power CMOS digital design," *IEEE J. Solid-State Circuits*, vol. 27, p. 473, Apr. 1992.

[3] C. Hu, "Device and technology impact on low power electronics," in *Low Power Design Methodologies*, J. M. Rabaey and M. Pedram, Eds. Norwell, MA: Kluwer, 1996, pp. 21–36.

[4] B. Davari, R. Dennard, and G. Shahidi, "CMOS scaling for high performance and low power—the next ten years," *Proc. IEEE*, vol. 83, p. 595, Apr. 1995.

[5] Z. Chen *et. al.*, "0.18 μm dual V_t MOSFET process and energy-delay measurement," in *IEDM Dig.*, 1996, p. 851.

[6] J. M. Rabaey, *Digital Integrated Circuits.* Englewood Cliffs, NJ: Prentice-Hall, 1996.

[7] R. Burch, F. Najm, P. Yang, and T. Trick, "A Monte Carlo approach for power estimation," *IEEE Trans. VLSI Syst.*, vol. 1, pp. 63–71, Mar. 1993.

[8] M. G. Xakellis and F. N. Najm, "Statistical estimation of the switching activity in digital circuits," in *ACM/IEEE Design Automation Conf.*, 1994, pp. 728–733.

[9] T. L Chou and K. Roy, "Estimation of sequential circuit activity considering spatial and temporal correlations," in *IEEE Int. Conf. Computer Design*, 1995, pp. 577–583.

[10] M. C. Johnson, K. Roy, and D. Somasekhar, "A model for leakage control by transistor stacking," School Elect. Computer Eng., Purdue Univ., West Lafayette, IN, Tech. Rep. TR-ECE 97-12, 1997.

[11] Z. Chen, M. Johnson, L. Wei, and K. Roy, "Estimation of standby leakage power in CMOS circuits considering accurate modeling of transistor stacks," in *Int. Symp. Low Power Electron. Design*, 1998, pp. T5.3.1–T5.3.6.

[12] L. Wei, Z. Chen, M. C. Johnson, K. Roy, and V. De, "Design and optimization of low voltage high performance dual threshold CMOS circuits," in *ACM/IEEE Design Automation Conf.*, 1998, pp. 489–494.

[13] B. J. Sheu, D. L. Scharfetter, P. K. Ko, and M. C. Teng, "BSIM: Berkeley short-channel IGFET model for MOS transistors," *IEEE J. Solid-State Circuits*, vol. SC-22, pp. 558–566, Apr. 1987.

[14] S. Mutoh *et. al.*, "1-V power supply high-speed digital circuit technology with multithreshold-voltage CMOS," *IEEE J. Solid-State Circuits*, vol. 30, pp. 847–854, Aug. 1995.

[15] Y. Ye, S. Borkar, and V. De, "Standby leakage reduction in high-performance circuits using transistor stack effects," in *Symp. VLSI Circuits*, 1998, pp. 40–41.

[16] J. Kao, A. Chandrakasan, and D. Antoniadis, "Transistor sizing issues and tool for multi-threshold CMOS technology," in *ACM/IEEE Design Automation Conf.*, 1997.

[17] S. Thompson, I. Young, J. Greason, and M. Bohr, "Dual threshold voltage and substrate bias: Key to high performance, low power, 0.1 μm logic design," in *'97 Symp. VLSI Technol. Dig. Tech. Papers*, 1997, pp. 69–70.

[18] S. Shigematsu *et. al.*, "A 1-V high speed MTCMOS circuit scheme for power-down applications," *IEEE J. Solid-State Circuits*, vol. 32, pp. 861–869, June 1997.

[19] H. Oyamatsu *et. al.*, "Design methodology of deep submicron CMOS devices for 1 V operation," *IEICE Trans. Electron.*, vol. E79-C, no. 12, pp. 1720–1724, 1996.

[20] J. Burr, Z. Chen, and B. Baas, "Standford's ultra-low-power CMOS technology and applications," in *Low-Power HF Microelectronics: A Unified Approach*, (IEE Circuits Syst. Series 8), G. A. S. Machado, Ed. London, U.K.: Inst. Elect. Eng. Press, 1996, pp. 139–184.

[21] S. W. Kang, K. S. Min, and K. Lee, "Parametric expression of subthreshold slope using threshold voltage parameters for MOSFET statistical modeling," *IEEE Trans. Electron Devices*, vol. 43, pp. 1382–1386, Sept. 1996.

Liqiong Wei (S'97) received the B.S. and M.S. degrees in computer science and technology from Peking University, Peking, China, in 1991 and 1994, respectively, and is currently working toward the Ph.D. degree in electrical and computer engineering at Purdue University, West Lafayette, IN.

Her research interests include low-power and high-performance device/circuit design, algorithms in combinatorial optimization, and SOI device modeling.

Zhanping Chen (S'95) received the B.S. degree in computer science and technology from Peking University, Peking, China, in 1991, and is currently working toward the Ph.D. degree in electrical and computer engineering at Purdue University, West Lafayette, IN.

From 1991 to 1993, he was an ASIC Design Engineer in the Microelectronics Research Center, Beijing, China. His research interests include new techniques for power estimation, circuit design, and optimization for low-power and dual-gate SOI devices and circuits for low-power applications.

Kaushik Roy (S'83–M'83–SM'95) received the B.Tech. degree in electronics and electrical communications engineering from the Indian Institute of Technology, Kharagpur, India, and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign, in 1990.

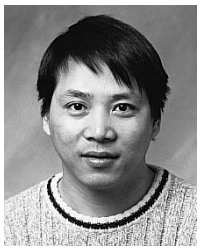
From 1988 to 1993, he was with the Semiconductor Process and Design Center, Texas Instruments, Dallas, TX, where he worked on FPGA architecture development and low-power VLSI. He joined the Electrical and Computer Engineering Faculty, Purdue University, West Lafayette, IN, in 1993, where he is currently an Associate Professor. He has authored or co–authored over 100 publications in refereed journals and conferences. His research interests include VLSI design and computer-aided design (CAD) with particular emphasis in low-power electronics, deep submicrometer design and interconnect, reconfigurable computing, and VLSI testing.

Dr. Roy is an associate editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS and *IEEE Design and Test of Computers*. He was the guest editor for a special issue on low-power VLSI in the *IEEE Design and Test of Computers*. He received the National Science Foundation Career Development Award in 1995.

Mark C. Johnson received the B.S. degree in electrical engineering from Purdue University, West Lafayette, IN, in 1983, the M.S.E.E. degree from Wichita State University, Wichita, KS, in 1991, and the Ph.D. degree in electrical engineering from Purdue University, in 1998.

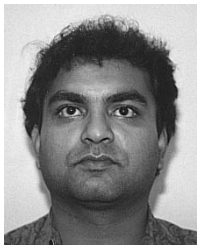
From 1983 to 1990, he was with the Design Automation Group, Boeing Airplane Company, Wichita, KS, where he provided support, integration, and development of CAD software. He served in a similar capacity at Thomson Consumer Electronics, Indianapolis, IN, from 1991 to 1994. In 1998, he joined the electrical and computer engineering faculty, Rose–Hulman Institute of Technology, Terre Haute, IN. His primary research interest is in CAD tools for electronic design, now focused on algorithms and optimizations for low-power VLSI.

Dr. Johnson is a member of Tau Beta Pi.



Yibin Ye (S'96–M'97) received the M.S. and Ph.D. degrees in electrical engineering from Purdue University, West Lafayette, IN, in 1994 and 1997, respectively.

He is currently with Microcomputer Research Laboratories, Intel Corporation, Hillsboro, OR. His current research interests include high-performance and low-power circuit techniques, logic synthesis and optimization, and algorithms in combinatorial optimization.



Vivek K. De (S'89–M'89) received the Ph.D. degree in electrical engineering from Rensselaer Polytechnic Institute, Troy, NY, in 1992.

Since April 1998, he has been a Principal Engineer and Manager of low-power circuit technology research at the Circuits Research Laboratory (CRL), Microprocessor Research Laboratories (MRL), Intel Corporation, Hillsboro, OR. From October 1997 to March 1998, he was a Senior Staff Engineer and Manager of low-power circuit technology at CRL. From November 1996 to September 1997, he was a Staff Engineer and Project Leader at CRL, where he was responsible for long-term research in low-power/high-performance circuit techniques and device structures for future generations of microprocessors. From 1992 to September 1996, he held research faculty positions at Georgia Institute of Technology, Atlanta, and at Rensselaer Polytechnic Institute. He has authored or co-authored 50 technical publications in refereed international conferences and journals. He holds nine patents in low-power circuits and devices. His primary research interests are low-power/high-performance device structures, circuit techniques, interconnect architectures, and CAD tools.

Dr. De received the Best Paper Award in the 1997 International ASIC Conference, Portland, OR.