# Gate Oxide Leakage Current Analysis and Reduction for VLSI Circuits

Dongwoo Lee, *Student Member, IEEE*, David Blaauw, *Member, IEEE*, and Dennis Sylvester, *Member, IEEE*

*Abstract*—**In this paper we address the growing issue of gate oxide leakage current ($I_{\rm gate}$) at the circuit level. Specifically, we develop a fast approach to analyze the total leakage power of a large circuit block, considering both $I_{\rm gate}$ and subthreshold leakage ($I_{\rm sub}$). The interaction between $I_{\rm sub}$ and $I_{\rm gate}$ complicates analysis in arbitrary CMOS topologies and we propose simple and accurate heuristics based on lookup tables to quickly estimate the state-dependent total leakage current for arbitrary circuit topologies. We apply this method to a number of benchmark circuits using a projected 100-nm technology and demonstrate accuracy within 0.09% of SPICE on average with a four order of magnitude speedup. We then make several observations on the impact of $I_{\rm gate}$ in designs that are standby power limited, including the role of device ordering within a stack and the differing state dependencies for NOR versus NAND topologies. Based on these observations, we propose the use of pin reordering as a means to reduce $I_{\rm gate}$. We find that for technologies with appreciable $I_{\rm gate}$, this technique is more effective at reducing total leakage current in standby mode than state assignment, which is often used for $I_{\rm sub}$ reduction.**

## I. Introduction

Feature size reduction in MOSFETs has been the key enabler to the continuation of Moore's law. Just as significant as effective channel length ($L_{\rm eff}$) reduction has been the shrinking of the gate oxide layer thickness ($T_{\rm ox}$). Early indications of 90-nm CMOS technologies set to come online in 2003 call for $T_{\rm ox}$ values in the range of 12–16 Å (1.2–1.6 nm), or approximately 4–5 atomic layers of $\rm SiO_2$ [1]–[3]. While aggressive scaling of $T_{\rm ox}$ is required to provide substantial current drive at reduced voltage supplies and to suppress short-channel effects such as drain-induced barrier lowering (DIBL), it results in the presence of significant gate tunneling leakage current ($I_{\rm gate}$).

$I_{\rm gate}$ arises due to the finite (nonzero) probability of an electron directly tunneling through the insulating $\rm SiO_2$ layer. The probability, and hence $I_{\rm gate}$ itself, is a strong exponential function of $T_{\rm ox}$ as well as the voltage potential across the gate oxide. A difference in $T_{\rm ox}$ of just 2 Å can lead to an order of magnitude change in $I_{\rm gate}$, making it the most sensitive device performance parameter with respect to any physical dimensions. Although gate oxides are very well controlled (often ±4%) compared to other dimensions such as $L_{\rm eff}$ and metal linewidth, this heightened sensitivity makes $I_{\rm gate}$ highly variable across a wafer. Another key point is that $I_{\rm gate}$ for a pMOS device is typically one order of magnitude smaller than an nMOS device with identical $T_{\rm ox}$ and $V_{dd}$ when using $\rm SiO_2$ [4]. This is due to the much higher energy required for hole tunneling in $\rm SiO_2$ and the fact that there are very few electrons associated with a pMOS device. However, in alternate dielectric materials the energy required for electron and hole tunneling can be completely different. In the case of nitrided gate oxides, in use today in some processes, pMOS $I_{\rm gate}$ can actually exceed nMOS $I_{\rm gate}$ depending on the nitrogen concentration (higher nitrogen content increases pMOS $I_{\rm gate}$ relative to nMOS) [5], [6].

For $T_{\rm ox} > 20$ Å, $I_{\rm gate}$ is typically very small in comparison to other forms of leakage current, specifically subthreshold leakage ($I_{\rm sub}$) which arises due to the partial formation of a conducting channel even at $V_{gs} = 0$ V. In recent generations, $I_{\rm sub}$ has been seen to rise by a factor of 3 to 5 × per generation under normal scaling theory. On the other hand, $T_{\rm ox}$ is 30% thinner in each new process technology and for an initial $T_{\rm ox}$ of 20 Å, this results in a 1000 × rise in $I_{\rm gate}$ in a subsequent process with $T_{\rm ox}$ of 14 Å (it will be somewhat smaller due to a $V_{dd}$ reduction). It is clear that $I_{\rm gate}$ either will, or in some cases already has, caught up to $I_{\rm sub}$ in magnitude. An example is NEC's 100 nm process with $T_{\rm ox} = 16$ Å [2]. High- $V_{th}$ (mid-performance) devices exhibit an $I_{\rm sub}$ of 0.3 $\rm nA/\mu m$ of gate width. nMOS $I_{\rm gate}$ for this process is 0.65 $\rm nA/\mu m$ with 1 V on the gate, exceeding $I_{\rm sub}$.

This NEC process uses a nitrided gate oxide (also called oxynitride) to raise the dielectric constant of the gate insulator from 3.9 to $\sim 4.1 - 4.2$. Even this small increase in the dielectric constant can yield an order of magnitude reduction in $I_{\rm gate}$ for the same $C_{\rm ox}$ value (since $T_{\rm ox}$ can be increased by about 5%–10% along with $\varepsilon_{\rm ox}$). Oxynitrides represent the first move toward high-k materials that will supplant thermal $\rm SiO_2$ as the gate insulator of choice in nanometer CMOS. High-k materials are typically metal oxides such as hafnium oxide $\rm HfO_2$ and zirconium oxide $\rm ZrO_2$ that provide dielectric constants in the range of 25–50. There are numerous process integration problems with such high-k materials; in particular their compatibility with Si and the resulting mobility degradation which reduces drive current. As a result, the introduction of true high-k materials (beyond oxynitrides) is not expected before the 65-nm node in 2007 [3]. Even this projection may be optimistic as the introduction of new materials has traditionally proven a much slower process than very aggressive scaling of already existing solutions. An example of the former is the use of low-k dielectrics for interconnections—the adoption of such materials has been much slower than anticipated in the 1997

Manuscript received February 28, 2003; revised June 30, 2003. This work was supported in part by the National Science Foundation, the Semiconductor Research Corporation (SRC), and the Microelectronics Advanced Research Corporation/The Defense Advanced Research Project (MARCO/DARPA).

The authors are with the University of Michigan, Ann Arbor, MI 48109 USA (e-mail: dongwool@umich.edu; blaauw@umich.edu; dmcs@umich.edu).

Digital Object Identifier 10.1109/TVLSI.2003.821553

and 1999 technology roadmaps. Thus, circuit designers may be forced to use devices with an $SiO_2$ based gate insulator for five or more years which brings with it a large $I_{\rm gate}$ and new design challenges.

There has been extensive work in the analysis and minimization of $I_{\rm sub}$ based on the understanding that it poses a fundamental scaling limit to traditional CMOS design [7]–[15]. However, $I_{\rm gate}$ has been growing much faster and to this point has almost solely received attention from device engineers and not circuit designers, EDA tool developers, etc. In [16] and [17], the authors examined the impact of gate leakage on circuit functionality but did not address its contribution to leakage power. In [18], the authors contribute the first circuit design concepts to reducing the impact of gate leakage—these focus on leveraging the lower $I_{\rm gate}$ in pMOS devices by using p-type domino circuits rather than n-type as well as pMOS sleep transistors for standby modes. Other papers addressing gate oxide tunneling current provide quantum-mechanical based models for computing $I_{\rm gate}$ in an individual MOS device [19], [20]. While useful, they do not provide insight into the impact of $I_{\rm gate}$ in actual circuits and their standby current.

Circuit level analysis of $I_{\rm gate}$ is complicated by two important factors: 1) state dependency and 2) the interaction of $I_{\rm sub}$ and $I_{\rm gate}$. The state dependence of $I_{\rm sub}$ is fairly well understood, especially in the context of the stack effect and there are efficient models to compute $I_{\rm sub}$ based on the number of off transistors in a stack [12]. However, there are different considerations with gate tunneling current since on, or conductive, devices are most responsible for $I_{\rm gate}$ in contrast to $I_{\rm sub}$. Furthermore, total gate leakage current is not always the sum of $I_{\rm sub}$ and $I_{\rm gate}$. In some states the currents interact at internal nodes (for gates with two or more inputs), altering the node voltages and complicating the analysis. Finally, the role of $I_{\rm gate}$ in the total leakage of a reasonably sized circuit block ($\sim$10000 gates) has not been determined—does it render standby modes based on the well-understood state dependency of $I_{\rm sub}$ useless?

In this paper, we make two primary contributions. First is the development of a fast approach for total leakage power analysis that considers both $I_{\rm gate}$ and $I_{\rm sub}$. We consider the interaction between these two sources of current and make several observations about the nature of the standby current problem when $I_{\rm gate}$ is no longer negligible. We categorize the state dependence of a transistor stack into cases where 1) only $I_{\rm sub}$ or $I_{\rm gate}$ occurs, 2) $I_{\rm sub}$ and $I_{\rm gate}$ sum, and 3) $I_{\rm sub}$ and $I_{\rm gate}$ interact in a complex fashion. We partition these cases based on the on/off states of devices within a stack. We then build precharacterized tables for individual device $I_{\rm gate}$ and $I_{\rm sub}$ currents, apply our state dependence heuristic, and compute the total leakage current on a gate by gate basis. We apply this method to a number of benchmark circuits in a predictive 100-nm CMOS technology to demonstrate the accuracy of the proposed method. We also gain insight on the role of $I_{\rm gate}$ in standby current within large circuit blocks. For instance, we find that the spread in total leakage for a given gate over its input state space is drastically reduced for NOR structures when considering $I_{\rm gate}$ but increased for NANDs. The second contribution of this paper is the proposed use of pin reordering as a new method for reducing $I_{\rm gate}$. While pin reordering is relatively ineffective for $I_{\rm sub}$, we exploit the dependence of $I_{\rm gate}$ on the node voltages in the stack and show that $I_{\rm gate}$ can be significantly reduced by placing transistors that are off at the bottom of the stack. We then demonstrate how this method can be combined with state assignment targeted at reducing $I_{\rm sub}$ during standby mode, as well as for runtime reduction of $I_{\rm gate}$. We present several heuristic solutions to this new optimization problem and demonstrate results on a large set of benchmark circuits.

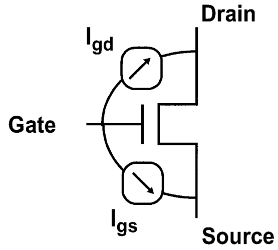


Fig. 1. Macro model for transistor gate leakage.

It is important to recognize the difference between standby mode leakage current, when the circuit is idle, and active leakage current, when the circuit is fully operating. In this work, our main focus is on standby mode leakage analysis and reduction methods. We also extend our approach to active leakage (also called runtime) reduction using input switching statistics. In addition, it should be noted that the proposed standby mode leakage reduction methods can be applied to reduce leakage in active mode when used in conjuction with clock gating [15].

The remainder of this paper is organized as follows. In Section II, we discuss the model and technology parameters used in our SPICE simulations considering $I_{\rm gate}$. In Section III, we present our proposed circuit level analysis of $I_{\rm gate}$ and $I_{\rm sub}$. In Section IV, we discuss the impact of $I_{\rm gate}$ on circuit operation and propose a method for reducing $I_{\rm gate}$ using pin reordering. Finally, in Section V we present results of the proposed $I_{\rm gate}$ analysis and reduction methods on benchmark circuits, and in Section VI we draw conclusions.

## II. Oxide Leakage Model

For simulation purposes, an oxide leakage model was incorporated in an existing 100 nm BSIM3v3 (level 49) model generated using the Berkeley Predictive Technology Model (BPTM) technique [21]. Since BSIM3 does not model oxide leakage[1], voltage dependent current sources from the gate to source $(I_{gs})$ and from the gate to drain $(I_{gd})$ were implemented in the macromodel, as shown in Fig. 1. The dependence of these currents on gate to source voltage $(V_{gs})$ and gate to drain voltage $(V_{gd})$ is given by the following two expressions:

$$i_{gs} = \frac{127.04 \times L_{\rm eff} \times e^{(5.606\,25 \times V_{gs} - 10.6 \times {T_{\rm ox}}^{-2.5})}}{2} \quad (1)$$

$$i_{gd} = \frac{127.04 \times L_{\rm eff} \times e^{(5.606\,25 \times V_{gd} - 10.6 \times {T_{\rm ox}}^{-2.5})}}{2} \quad (2)$$

[1] Although BSIM4, which incorporates an $I_{\rm gate}$ model, was recently released, reliable $I_{\rm gate}$ model parameters are currently not publicly available.

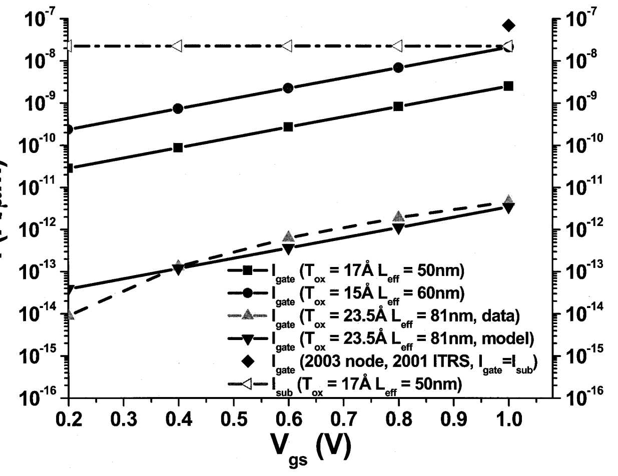


Fig. 2. Fit of macro model to industrial gate leakage measurements.

where $T_{\rm ox}$ and $L_{\rm eff}$ are given in nanometers and $i_{gs}$ and $i_{gd}$ are given in $\mu$A per $\mu$m of transistor width (assuming minimum channel length). Equations (1) and (2) are based on an empirical model of total gate leakage fit to IBM data on thin $\mathrm{SiO_2}$ dielectrics that was used in the 2001 ITRS. The model was further refined to fit data from an industrial 0.13 $\mu$m process over the full range of $V_{ds}$ and $V_{gs}$. The model was also found to maintain good stability during SPICE simulation.

Since our analysis focuses on bulk technology, we do not consider the tunneling current from gate to bulk since this current is expected to be several orders of magnitude less than the gate to channel tunneling current. However, it is important to note that in partially depleted SOI technology, the gate to body leakage current could have a significant impact on the body voltage and hence on the leakage of the device. When our analysis is applied to PD SOI devices, the tunneling current component from the gate to the body should be added in the macro model.

For leakage current estimation, we further assume that the leakage current is independent of the load of a gate. In runtime mode, the loading of a gate will influence the decay time of the output voltage of the gate, and should be considered in the analysis. However, the main focus of this paper is on standby model leakage where there is no signal switching activity. Therefore, the impact of gate loading is limited to the impact of reverse leakage current from source/drain to gate on the driving gate. This reverse tunneling current was found to have negligible impact on the leakage of the driving gate and hence was not included in the analysis.

As seen in Fig. 2, a reasonable correlation between the industrial data and the experimental data for the oxide leakage was obtained. The percentage error between the data and the empirical model of gate leakage current increases as $V_{gs}$ is decreased from 1.0 to 0.4 V from approximately 10%–40%. However, since the total gate current is significantly reduced for small $V_{gs}$, this error has a negligible effect on the total predicted leakage current for a CMOS gate. In digital circuits, the typical cases of interest are when $V_{gs} \approx V_{dd}$ with $V_{ds}$ equal to either 0 or $V_{dd} - V_{th}$, for which the empirical model shows good accuracy.

To determine the impact of $I_{\rm gate}$ on circuit behavior and to develop a fast and accurate total leakage model, two 100-nm technology files were generated—the first has a $T_{\rm ox}$ of 17 Å and $L_{\rm eff}$ of 50 nm, while the second has a $T_{\rm ox}$ of 15 Å and $L_{\rm eff} = 60$ nm. $V_{th}$ in both technologies is approximately 200 mV. The goal in using two processes is to examine the role of $I_{\rm gate}$ in total leakage for a range of $I_{\rm gate}/I_{\rm sub}$ ratios. In the 17-Å process, $I_{\rm gate}$ is roughly 1/9 of $I_{\rm sub}$ under worst case biasing conditions while in the 15 Å process $I_{\rm gate}/I_{\rm sub} = 2/3$. $I_{\rm sub}$ values are in the range of 20–40 nA$/\mu$m of gate width at room temperature which is slightly below the ITRS projected value of 70 nA$/\mu$m at 100 nm (see Fig. 2). While both oxide thicknesses are in the higher end of the range specified for 100-nm devices by the ITRS (year 2003), we also assume the use of $\mathrm{SiO_2}$ and not an oxynitride since $I_{\rm gate}$ models are more readily available for the former. To compensate for the higher expected $I_{\rm gate}$ in $\mathrm{SiO_2}$, we select conservative $T_{\rm ox}$ values to provide more realistic $I_{\rm gate}/I_{\rm sub}$ ratios. $V_{DD}$ is 1 V for both cases and all results in this work are for room temperature ($I_{\rm sub}$ is highly temperature dependent while $I_{\rm gate}$ is not).

## III. Efficient Leakage Analysis Method

Based on the proposed gate tunneling current model, SPICE simulation can be performed to obtain the total leakage current for a circuit consisting of multiple gates. However, for large circuits consisting of tens to hundreds of thousands of gates, SPICE simulation becomes infeasible. We therefore describe a new analysis method that achieves an average error of 0.04% compared with SPICE with a four order of magnitude run time improvement.

Standby current estimation is complicated by the state dependence of both the $I_{\rm gate}$ and $I_{\rm sub}$ currents. The state dependence of subthreshold leakage current has been extensively studied and exhibits the so-called *stack effect*, where multiple transistors that are off in series have a significantly reduced subthreshold leakage current. Similarly, gate tunneling current has state dependence, as well as dependence on the device type. As mentioned, pMOS devices typically exhibit gate tunneling currents that are approximately one order of magnitude lower than those of nMOS devices [4]. Hence, we ignore the pMOS gate current and focus only on nMOS transistors in our analysis. However, our analysis method can be easily extended to include pMOS-based $I_{\rm gate}$, as would be necessary when nitrided gate oxides are used.

Gate tunneling current furthermore has a strong dependence on the $V_{gs}$ and $V_{gd}$ of a device, leading to state dependence. To examine this dependence, we first consider a simple inverter circuit shown in Fig. 3. The maximum gate tunneling current occurs when the input is at $V_{dd}$ and $V_s = V_d = 0$ V for the nMOS device. In this case, $V_{gs} = V_{gd} = V_{dd}$ and the gate tunneling current is at its maximum with equal current flowing to the source and drain nodes. At the same time, the pMOS device exhibits subthreshold leakage current.

As the input voltage is decreased, $I_{gs}$ decreases rapidly and is reduced by more than one order of magnitude when $V_{gs} = V_{th,\rm nmos}$, and becomes zero when $V_{gs} = 0$. As the input voltage decreases and the output voltage increases, $V_{gd}$ will become negative, resulting in a reverse gate tunneling current from the drain to the gate node. However, this reverse gate tunneling

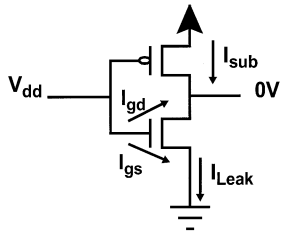


Fig. 3. Inverter circuit with nMOS oxide leakage current.

occurs when the nMOS transistor is off and tunneling is restricted to the gate-to-drain overlap region, due to the absence of a channel. Since the gate-to-drain overlap region is substantially smaller than the channel region, reverse tunneling current is much smaller than the forward tunneling current when the device is on, and hence can be ignored [22]. In addition, the corner oxide thickness can be increased by oxidizing the polysilicon after gate formation which would further suppress tunneling in the overlap regions [23].

For a simple inverter, the nMOS gate tunneling current and the nMOS subthreshold leakage current occur in mutually exclusive states, simplifying the analysis. For a high input state, the pMOS subthreshold leakage current combines with the nMOS gate tunneling current and each can be computed independently and then simply added to obtain the total leakage current $I_{\text{leak}}$ of the gate, as shown in Fig. 3. For a low input state, the nMOS transistor is off and the total leakage current of the gate is equal to the subthreshold leakage current through the nMOS device (since we are ignoring pMOS $I_{\text{gate}}$ in this discussion).

We next consider a multi-input gate with an nMOS transistor stack. If all inputs have a high state, the analysis is again similar to that of the inverter. The total standby current is equal to the sum of $I_{\text{sub}}$ through the pMOS transistors added to $I_{\text{gate}}$ through the nMOS transistors. However, for input states where at least one input is low and the gate output is $V_{dd}$, $I_{\text{sub}}$ through turned-off nMOS transistors and $I_{\text{gate}}$ through turned-on nMOS transistors occur in the same transistor stack. Both currents combine at internal stack nodes and impact the stack node voltages. $I_{\text{sub}}$ and $I_{\text{gate}}$ are therefore interdependent in these cases, and must be analyzed simultaneously.

We consider gate tunneling current in three distinct scenarios for a transistor within a transistor stack, as shown in Fig. 4. We consider the gate tunneling current through the transistor labeled $t_n$, with a high gate input state. The complementary pMOS transistors are omitted for clarity. We now discuss each scenario in more detail.

1) In the first scenario, shown in Fig. 4(a), transistor $t_n$ is positioned above zero or more conducting transistors and below one or more nonconducting transistors. In this case, the internal nodes $n_a$ and $n_b$ have a conducting path to the ground node and are at nominal 0V. The $I_{\text{gate}}$ of transistor $t_n$ therefore does not affect the voltage at nodes $n_a$ and $n_b$ and can be added to $I_{\text{sub}}$ of the stack to obtain the total leakage current of the gate.

2) In the second scenario, shown in Fig. 4(b), transistor $t_n$ is positioned above one or more nonconducting transistors and below zero or more conducting transistors. In this case, nodes $n_a$ and $n_b$ are connected to the output of the logic gate through conducting nMOS transistors and will be held at $V_{dd} - V_{th,\text{nMOS}}$. For transistor $t_n$, $V_{gs,n}$ and $V_{gd,n}$ are therefore small; approximately one threshold voltage. Based on SPICE simulations, the $I_{\text{gate}}$ in this case is more than one order of magnitude smaller than in scenario 1 and can be safely ignored. Note that if $t_n$ is the top most transistor of the stack, a $V_t$ drop will occur only for the source node $V_{s,n}$ and $V_{gd,n} = 0$ V, thereby further reducing $I_{\text{gate}}$ in this scenario.

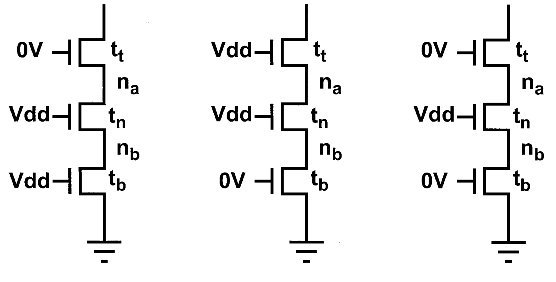


Fig. 4. Three-input nMOS stack with three scenarios of combined $I_{\text{sub}}$ and $I_{\text{gate}}$.

3) In the third scenario, shown in Fig. 4(c), there is at least one nonconducting transistor *both* above and below transistor $t_n$ in the stack. In this case, the subthreshold leakage current exhibits the stack effect and the internal nodes $n_a$ and $n_b$ have a voltage in the range of 100–200 mV. The top transistor $t_t$ is therefore strongly turned off due to its negative $V_{gs,t}$. However, since $V_{gs,n}$ and $V_{gd,n}$ for transistor $t_n$ are only slightly diminished from $V_{dd}, t_n$ will exhibit significant $I_{\text{gate}}$ current. This current combines with the $I_{\text{sub}}$ through $t_t$ and causes the node voltages at $n_a$ and $n_b$ to increase from their values when only subthreshold current is considered.

A rise in the voltage at $n_a$ and $n_b$ reduces $I_{\text{sub}}$ through $t_t$, as $V_{gs,t}$ becomes further negative, and also reduces $I_{\text{gate}}$ through $t_n$. However, the dependence of subthreshold leakage current on $V_{gs,t}$ is exponential and is much stronger than the dependence of gate tunneling current on $V_{gs,n}$ and $V_{gd,n}$ [2]. Therefore, as the voltage of $n_a$ is raised by $I_{\text{gate}}$ through $t_n$, the $I_{\text{sub}}$ through $t_t$ is diminished by a nearly equal amount. The gate tunneling current therefore effectively displaces the subthreshold current, leaving the total leakage current relatively unchanged. When $I_{\text{gate}}$ becomes sufficiently large and exceeds the original subthreshold current, the subthreshold current is effectively pinched off and becomes negligible. In this case, the total leakage current is equal to the oxide tunneling current.

This effect is illustrated in Table I, where we show the node voltages of $n_a$ and $n_b$ as well as the leakage currents for the circuit shown in Fig. 4(c) for three SPICE simulations: when only subthreshold current is present, when only gate tunneling current is present, and when both are present. For the 17 Å process, the voltages at $n_a$ and $n_b$

[2] For example, [18] states that a 0.3-V change in $V_{gs}$, $V_{gd}$ leads to a decade change in $I_{\text{gate}}$. However, a reduction in $V_{gs}$ of only $\sim$0.1 V yields a 10 × drop in $I_{\text{sub}}$.

TABLE I
SIMULATION RESULTS FOR INDIVIDUAL AND COMBINED $I_{\rm gate}/I_{\rm sub}$

| | 17Å | | | 15Å | | |
|---|---|---|---|---|---|---|
| | $I_{sub}$ only | $I_{gate}$ only | combined | $I_{sub}$ only | $I_{gate}$ only | combined |
| $V_{na}$ / $V_{nb}$ | 68mV | 95mV | 111mV | 51mV | 285mV | 285mV |
| $I_{sub}$ | 399pA | - | 65pA | 693pA | - | 32fA |
| $I_{gate}$ | - | 446pA | 407pA | - | 1.27nA | 1.27nA |
| $I_{leak}$ | 399pA | 446pA | 472pA | 693pA | 1.27nA | 1.27nA |

increase by 42 mV over the case with $I_{\rm sub}$ only when considering both $I_{\rm sub}$ and $I_{\rm gate}$, resulting in a decrease of $I_{\rm sub}$ by a factor of 6. However, the voltages at $n_a$ and $n_b$ rise by only 16 mV when the analysis is expanded from only $I_{\rm gate}$ to $I_{\rm gate}$ and $I_{\rm sub}$, resulting in a decrease of $I_{\rm gate}$ through $t_n$ by just 9%. Table I also shows SPICE results for the 15 Å process. In this case, $I_{\rm sub}$ is reduced by four orders of magnitude due to the presence of $I_{\rm gate}$, and becomes negligible.

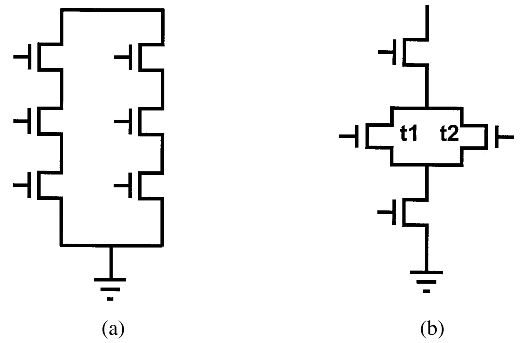
As a result, the total leakage with both $I_{\rm sub}$ and $I_{\rm gate}$ present is nearly equal to the maximum of $I_{\rm gate}$ and $I_{\rm sub}$, when they are computed independently. In our approach, we therefore find the total leakage current by computing $I_{\rm gate}$ and $I_{\rm sub}$ separately and set the total leakage current to their maximum.

Note that in a transistor stack each conducting transistor will fall into one of the three discussed scenarios. Based on the three scenarios, we propose the following simple table-based leakage estimation method for arbitrary gate structures. First, we determine the subthreshold leakage current of the circuit, without consideration of gate tunneling current. A number of approximate analytical solutions have been proposed for this purpose [12] and may be used. In this paper, we use an empirical model in which the total subthreshold leakage current is expressed as follows:

$$I_{\rm sub,k} = I_{\rm sub,1} * S_k * s_t \tag{3}$$

where $I_{\rm sub,1}$ is the leakage current for a single off-transistor of unit size, $S_k$ is the stack factor for a stack with $k$ off-transistors in series, and $s_t$ is the size of the transistor. Both $I_{\rm sub,1}$ and $S_k$ are precharacterized using SPICE for stacks with different size transistors and stored in a table. In the presence of one or more conducting nMOS transistors above a stack of $k$ off-transistors, the voltage across the off-transistors is diminished by the $V_{th,{\rm nMOS}}$ voltage drop across the conducting transistors (including body effect). This reduces the subthreshold leakage current by approximately 35% in our technology and is accounted for by constructing an additional set of tables where a conducting transistor is placed above the off-transistor stack.

Next, we measure $I_{\rm gate}$ for a single transistor of unit size in each of the three discussed scenarios when $I_{\rm sub}$ is eliminated. In scenario 3, the $I_{\rm gate}$ current is dependent on the number of off-transistors below transistor $t_n$. We therefore specify the gate tunneling current as $I_{\rm gate,l}$, where $l$ indicates the number of off-transistors below $t_n$, and characterize $I_{\rm gate,l}$ for different value of $l$ in a table. Note that the current $I_{\rm gate,0}$ corresponds to the gate tunneling current in scenario 1.



(a) (b)

Fig. 5. Leakage current computation for series/parallel structures.

The total leakage current, as well as its $I_{\rm gate}$ and $I_{\rm sub}$ components, are then computed as follows. First, the total number of off-transistors in the stack is determined and the $I_{\rm sub}$, in the absence of $I_{\rm gate}$, is found using (3). Next, the tunneling currents $I_{\rm gate,l}$ of the on-transistors in scenarios 1 and 3 are determined based on precharacterized table values and are multiplied by their transistor size. The total leakage current $I_{\rm total}$, and its tunneling and subthreshold components $I_{\rm gate}$ and $I_{\rm sub}$, are then determined as follows:

$$I_{total} = \sum_{l=0} I_{\rm gate,l} + Max\left(\sum_{l>0} I_{\rm gate,l}, I_{\rm sub,k}\right) \tag{4}$$

$$I_{\rm gate} = \sum_{l} I_{\rm gate,l} \tag{5}$$

$$I_{\rm sub} = \begin{cases} I_{\rm sub,k} - \sum_{l>0} I_{\rm gate,l} & \text{if } \left(I_{\rm sub,k} > \sum_{l>0} I_{\rm gate,l}\right) \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

The first term in (4) corresponds to the $I_{\rm gate}$ current of transistors in scenario 1, which is independent of the other currents in the stack. The second term of (4) corresponds to the $I_{\rm gate}$ of transistors in scenario 3 which displaces the $I_{\rm sub}$ of the stack. Hence, the current for this term is the maximum of these two currents. Equations (5) and (6) express the total $I_{\rm sub}$ and $I_{\rm gate}$ in the transistor stack.

For the analysis of series/parallel nMOS structures, such as and-or-invert (AOI) and or-and-invert (OAI) gates, we use the following rules to compute the total leakage current. Given multiple parallel transistor stacks, such as those shown for the AOI stacks in Fig. 5(a), we compute the leakage current of each stack separately and then add them to obtain the total leakage of the gate. For parallel transistors within an nMOS stack, such as transistors $t_1$ and $t_2$ for the OAI gate in Fig. 5(b), we first collapse the two parallel transistors using the following rules.

1) If the two parallel transistors $t_1$ and $t_2$ have the same gate input state, they are replaced with a single transistor with transistor size equal to the sum of their sizes.
2) If the two parallel transistors $t_1$ and $t_2$ have different input states, the off-transistor impacts neither $I_{\rm gate}$ nor $I_{\rm sub}$ and is neglected during leakage current computation.

TABLE II
LEAKAGE ESTIMATION FOR THREE-INPUT NAND GATE WITH 15 Å OXIDES

| State | Estimated current [nA] | | | SPICE [nA] | % error |
|---|---|---|---|---|---|
| | $I_{sub}$ | $I_{gate}$ | $I_{total}$ | | |
| 000 | 0.382 | 0.000 | 0.382 | 0.382 | 0.11% |
| 001 | 0.709 | 6.339 | 7.048 | 7.047 | 0.02% |
| 010 | 0.709 | 1.275 | 1.275 | 1.292 | -1.25% |
| 011 | 5.626 | 12.677 | 18.303 | 18.295 | 0.04% |
| 100 | 0.676 | 0.000 | 0.676 | 0.675 | 0.18% |
| 101 | 3.804 | 6.339 | 10.143 | 10.140 | 0.03% |
| 110 | 3.804 | 0.000 | 3.804 | 3.641 | 4.48% |
| 111 | 28.273 | 19.015 | 47.288 | 47.278 | 0.02% |

TABLE III
LEAKAGE ESTIMATION FOR THREE-INPUT NAND GATE WITH 17 Å OXIDES

| State | Estimated current [nA] | | | SPICE [nA] | % error |
|---|---|---|---|---|---|
| | $I_{sub}$ | $I_{gate}$ | $I_{total}$ | | |
| 000 | 0.196 | 0.000 | 0.196 | 0.197 | -0.29% |
| 001 | 0.402 | 0.761 | 1.163 | 1.163 | -0.07% |
| 010 | 0.446 | 0.399 | 0.446 | 0.477 | -5.51% |
| 011 | 6.774 | 1.522 | 8.295 | 8.291 | 0.05% |
| 100 | 0.382 | 0.000 | 0.382 | 0.383 | -0.42% |
| 101 | 3.720 | 0.761 | 4.481 | 4.482 | -0.02% |
| 110 | 3.720 | 0.000 | 3.720 | 3.471 | 7.17% |
| 111 | 31.971 | 2.282 | 34.253 | 34.248 | 0.02% |

After collapsing parallel devices in a transistor stack using the above two rules, we compute the gate tunneling and subthreshold leakage current using (4).

To demonstrate the accuracy of the proposed leakage estimation method, we show the analysis results for a three-input NAND gate under all possible input states in Tables II and III for both 15- and 17-Å gate oxide thicknesses. The leakage current obtained from SPICE simulation and using the proposed analysis method is shown and has an average error of 1.2% over all input states. The maximum error occurs for state 110 with 17-Å gate oxide thickness. However, the total leakage current in this case is small and hence the error, in terms of absolute current, is acceptable. Conversely, states with the largest total leakage such as 010, 101, and 111 tend to show extremely small errors—this will translate to very good overall estimation of leakage in large circuit blocks.

As mentioned earlier in Section III, we do not consider pMOS gate leakage current and reverse gate tunneling current from source/drain to gate in our SPICE macro model. Hence, if these leakage current components were considered in the SPICE simulation, a greater difference between the estimated total leakage current using the proposed method and SPICE simulation would be observed. If this error is significant, it may be necessary to extend the proposed approach to include such current components.

## IV. GATE LEAKAGE REDUCTION METHODS

TABLE IV
IMPACT OF $I_{\text{gate}}$ ON STATE DEPENDENCE WITH $I_{\text{leak}}$

| Gate type | Average $I_{leak}$ [nA] | | max $I_{leak}$ / min $I_{leak}$ across all states | |
|---|---|---|---|---|
| | w/o $I_{gate}$ (15Å / 17Å) | w/ $I_{gate}$ (15Å / 17Å) | w/o $I_{gate}$ (15Å / 17Å) | w/ $I_{gate}$ (15Å / 17Å) |
| NAND2 | 7.25 / 8.05 | 12.0 / 8.62 | 26.6 / 53.00 | 44.40 / 56.85 |
| NAND3 | 5.5 / 5.97 | 11.1 / 6.61 | 74.0 / 162.8 | 123.8 / 174.4 |
| NAND4 | 3.8 / 3.99 | 9.9 / 4.73 | 138 / 327.7 | 231.4 / 351.0 |
| NOR2 | 7.3 / 7.84 | 13.6 / 8.60 | 7.57 / 19.50 | 1.40 / 6.10 |
| NOR3 | 5.7 / 5.79 | 15.2 / 6.93 | 21.26 / 59.00 | 1.48 / 9.28 |
| NOR4 | 4.1 / 3.93 | 16.8 / 5.45 | 21.26 / 120.5 | 1.94 / 12.37 |

In this section, we propose a method for reducing $I_{\text{gate}}$ through simultaneous pin reordering and state assignment. Traditionally, state assignment has been used to reduce standby mode $I_{\text{sub}}$ by setting the output of each flip-flop to a known state during standby mode such that $I_{\text{sub}}$ is minimized. The standby mode state is chosen so that the stack effect occurs in as many gates as possible [24]. Although the logic correlation between gates prevents all gates from being in a low $I_{\text{sub}}$ state, reasonable reductions in subthreshold leakage currents have been obtained using this method for circuit blocks [12]. Furthermore, the area and delay penalty incurred by the additional transistors required for forcing the output of a flip-flop to a given sleep state is minor [25]. However, the presence of significant $I_{\text{gate}}$ affects the state dependence of the total leakage and must be considered. In this section, we first discuss the impact of $I_{\text{gate}}$ on standby mode state assignment in general and then propose a new method that combines state assignment with pin reordering for more effective total leakage reduction.

### A. Impact of $I_{\text{gate}}$ on Circuit Leakage Behavior

In general, the worst case and best case leakage states of common CMOS gates behave differently when both $I_{\text{sub}}$ and $I_{\text{gate}}$ are considered compared to $I_{\text{sub}}$ alone. Table II showed that when only $I_{\text{sub}}$ is considered, the worst case leakage state for NAND structures occurs when all inputs are high as the pMOS devices leak in parallel and sum. For NOR structures, the reverse is true: all inputs set to low causes all nMOS devices to leak concurrently in parallel. For these two cases, we now include $I_{\text{gate}}$. In NAND gates with all inputs tied high, the nMOS devices in the pull-down stack all exhibit *worst-case* $I_{\text{gate}}$ which adds to the large $I_{\text{sub}}$ of the pMOS devices to create a large total leakage current. In the NOR gate with all inputs set to low, the pMOS devices have $V_{gd} = V_{gs} = V_{dd}$ but since pMOS devices show very small $I_{\text{gate}}$, the overall impact will be small. Meanwhile, the parallel pull-down devices exhibit only reverse edge direct tunneling which is negligible. As a result of these trends, we find that the *range* of total leakage current across states is broadened for NAND gates and compressed for NORs.

This is illustrated in Table IV where the average leakage and the ratio of max/min leakage over all possible input states is shown for NAND and NOR gates. Results for 15 Å and 17 Å technologies are shown both with and without considering $I_{\text{gate}}$. Columns 2 and 3 show that even with a relatively low $I_{\text{gate}}$ value for the $T_{\text{ox}} = 17$ Å technology, the average leakage over all states in the gates studied increases by 10–35% when considering both $I_{\text{gate}}$ and $I_{\text{sub}}$ together. In the more aggressive 15-Å technology, the rise in average leakage is 65–160% for NANDs and up to 310% for four-input NOR gates. The last two columns show that the presence of $I_{\mathrm{gate}}$ significantly reduces the range of leakage current for NOR gates, while at the same time, it increases this range for NAND gates. For the 15-Å technology, the ratio of maximum to minimum leakage current over all possible states is reduced from 21.3 × in a three-input NOR to 1.48 ×. On the other hand, the max/min leakage ratio for NAND gates increases by approximately 2 × in the 15-Å technology since the same states that exhibit maximum $I_{\mathrm{sub}}$ also exhibit maximum $I_{\mathrm{gate}}$.

In general, standby-mode leakage in the presence of significant $I_{\mathrm{gate}}$ can be addressed with similar methods as used for $I_{\mathrm{sub}}$ leakage current. However, state assignment can be significantly more effective for circuits constructed predominantly from NAND gates, as opposed to NOR gates. Since in most of our benchmark circuits NAND gates outnumbered NORs 2-to-1, we found that the overall spread of total leakage current is typically increased slightly when $I_{\mathrm{gate}}$ is considered.

A common approach to reduce subthreshold leakage current is the use of multiple-threshold CMOS (MTCMOS) which gates a high- $V_{th}$ transistor with a sleep mode signal to virtually eliminate $I_{\mathrm{sub}}$ [13]. In [18], the authors addressed the impact of $I_{\mathrm{gate}}$ on MTCMOS and advocated a pMOS-based sleep device as opposed to nMOS which has a lower parasitic resistance. However, during normal operation (sleep device is ON) leakage power is not a major concern since the design is intended to use the sleep mode during long periods of nonactivity. Thus, in the normal configuration (nMOS sleep device) when the sleep transistor is OFF, $V_{gs} = 0$ and $V_{gd}$ floats toward-$V_{dd}$. Again, this biases the device to conduct gate current from the gate-to-drain overlap region to the gate, which is approximately an order of magnitude smaller than the worst case gate-to-channel $I_{\mathrm{gate}}$ at $V_{gs} = V_{dd}$ and $V_{ds} = 0$ [22]. While this reduction is not as substantial as the several orders of magnitude drop in $I_{\mathrm{sub}}$ realized with MTCMOS, it is still beneficial. Since in the sleep mode $I_{\mathrm{gate}}$ will likely be dominant, two approaches may be considered: 1) reduce the $V_{th}$ of the sleep device somewhat (e.g., 100 mV) to minimize the delay penalty associated with an extra series device; this allows the use of smaller sleep devices to simultaneously reduce $I_{\mathrm{gate}}$, dynamic power, and layout area while not penalizing standby mode leakage since $I_{\mathrm{sub}} \ll I_{\mathrm{gate}}$ or 2) incorporate a multi-$T_{\mathrm{ox}}$ process to allow the sleep devices to reduce $I_{\mathrm{gate}}$ in addition to $I_{\mathrm{sub}}$. A limited (and practical) form of a multi- $T_{\mathrm{ox}}$ process was proposed in the form of a boosted-gate MOS version of MTCMOS in which the sleep device is a thick-oxide, higher voltage device that is commonly used for $I/O$ circuitry [26].

### *B. Reduction of $I_{\mathrm{gate}}$ Through Pin Reordering*

A key difference between the state dependence of $I_{\mathrm{sub}}$ and $I_{\mathrm{gate}}$ is that the magnitude of $I_{\mathrm{sub}}$ primarily depends of the *number* of on vs. off transistors in a stack, while $I_{\mathrm{gate}}$ also depends strongly on the position of the on/off transistors. We consider a three-input NAND gate with input combinations 110 and 101 (where the first input value corresponds to the topmost nMOS), as shown in Table II for the 15-Å process. When $I_{\mathrm{gate}}$ is neglected, the leakage current in these two states is the same, equal to 3.8 nA. When including $I_{\mathrm{gate}}$ in the analysis, the total leakage in the 101 state increases to 10.14 nA whereas the leakage current in state 110 is unchanged. Furthermore, in state 011, $I_{\mathrm{sub}}$ is increased by approximately 30% to 5.6 nA, while $I_{\mathrm{gate}}$ is doubled, yielding a total leakage of 18.3 nA. This dependence is a consequence of the different leakage of on-transistors in scenario 1, where $I_{\mathrm{gate}}$ is negligible, and scenario 2, where $I_{\mathrm{gate}}$ sums with $I_{\mathrm{sub}}$ as discussed in Section III.

The dependence of $I_{\mathrm{gate}}$ on the position of the on-transistors in the stack suggests a combined approach where state assignment is used for reducing $I_{\mathrm{sub}}$ while pin reordering is targeted at $I_{\mathrm{gate}}$ reduction. Since pin reordering and state assignment are inter-dependent, this requires solving a combined optimization problem where a state-assignment and pin ordering is determined for the entire circuit that minimizes the total standby leakage current. A number of heuristic methods for state-assignment alone have been proposed in the literature [12], [27] using branch-and-bound methods. We therefore extend such a branch-and-bound method to incorporate simultaneous pin reordering. An input state search tree is first formulated using the approach presented in [27] and is traversed using the branch-and-bound traversal algorithm. This algorithm is augmented such that each time a leaf node is reached, and the input state of the circuit is completely defined, we apply pin reordering by placing all off transistors at the bottom of the stack for each gate. This substantially decreases $I_{\mathrm{gate}}$ while also slightly decreasing $I_{\mathrm{sub}}$. We then update the total leakage for that leaf solution with the new $I_{\mathrm{gate}}$ and $I_{\mathrm{sub}}$ leakage and continue the traversal of the state tree. Despite the pruning that is performed during the traversal, the search space is very large and an exhaustive traversal of the tree is not possible. We therefore place a limit on the run time of the algorithm and report the best solution found by the search within this allotted time.

In addition to the branch-and-bound approach, we also implemented a simple random search approach. For each randomly generated input state, the state of each transistor in a stack is determined and optimal pin reordering is performed. The input state/pin reordering combination with minimum total leakage is then recorded. In Section V, we show a comparison between the two approaches. Since pin reordering can affect the circuit performance, it must be restricted to stack inputs that are not timing critical. However, the delay impact of pin reordering is relatively small and was ignored in our implementation.

Finally, we apply pin reordering for the purpose of runtime leakage reduction. Since $I_{\mathrm{sub}}$ depends on the number of off transistors in series, it is difficult to reduce $I_{\mathrm{sub}}$ during runtime since the state of the circuit cannot be changed. However, the probability of being in a high state (referred to as the *state probability*) is significantly lower for certain nodes in the circuit than others. We use this information to place nodes with a low state probability at the bottom of the transistor stack. Based on given state probabilities for the primary inputs (PIs), we compute the state probability of each node in the circuit using the method described in [28]. We then order the transistors in a stack from top to bottom in decreasing order of their state probabilities. In this manner, the likelihood of scenarios 2 and

TABLE V
LEAKAGE ESTIMATION RESULTS FOR BENCHMARK CIRCUITS

| Circuit | Number of gates | Estimated leakage current, [μA] (avg) | | SPICE leakage current [μA] (avg) | % error (avg/max) | Run time | |
|---|---|---|---|---|---|---|---|
| | | w/o $I_{gate}$ | w/ $I_{gate}$ | | | Proposed method (ms) | SPICE (s) |
| C432 | 121 | 1.71 | 2.82 | 2.82 | 0.13/0.32 | 0.18 | 9.36 |
| C499 | 517 | 6.44 | 9.99 | 9.99 | 0.01/0.02 | 2.4 | 38.4 |
| C880 | 325 | 4.49 | 7.08 | 7.08 | 0.06/0.14 | 1.5 | 27.8 |
| C1355 | 478 | 6.36 | 10.22 | 10.22 | 0.02/0.06 | 2.5 | 41.4 |
| C1908 | 425 | 5.55 | 8.61 | 8.61 | 0.01/0.04 | 2.7 | 35.8 |
| C2670 | 750 | 9.48 | 14.46 | 14.46 | 0.02/0.06 | 3.9 | 60.6 |
| C3540 | 890 | 11.76 | 18.98 | 18.98 | 0.04/0.08 | 6.3 | 100.2 |
| C5315 | 1524 | 20.49 | 32.28 | 32.28 | 0.01/0.02 | 11.1 | 180.8 |
| C6288 | 2388 | 32.82 | 54.54 | 54.53 | 0.02/0.04 | 34.4 | 971.3 |
| C7552 | 1916 | 25.86 | 39.67 | 39.69 | 0.04/0.07 | 14.5 | 207.8 |
| alu64 | 1791 | 25.29 | 40.58 | 40.63 | 0.14/0.35 | 42.6 | 245.0 |
| i1 | 39 | 0.45 | 0.69 | 0.69 | 0.11/0.42 | 0.4 | 2.0 |
| i2 | 95 | 0.91 | 1.89 | 1.88 | 0.36/0.67 | 0.8 | 9.7 |
| i3 | 92 | 1.25 | 1.89 | 1.88 | 0.17/0.48 | 0.5 | 6.1 |
| i4 | 160 | 2.33 | 3.81 | 3.81 | 0.03/0.08 | 1.2 | 11.0 |
| i5 | 198 | 2.61 | 4.04 | 4.04 | 0.01/0.02 | 1.3 | 10.1 |
| i6 | 359 | 5.02 | 8.11 | 8.13 | 0.22/0.44 | 1.5 | 26.6 |
| i7 | 450 | 6.15 | 10.02 | 10.04 | 0.24/0.45 | 2.2 | 36.2 |
| i8 | 725 | 10.40 | 16.73 | 16.74 | 0.07/0.22 | 3.7 | 67.0 |
| i9 | 459 | 6.48 | 10.54 | 10.56 | 0.19/0.39 | 2.1 | 38.3 |
| i10 | 1794 | 24.33 | 38.42 | 38.43 | 0.04/0.05 | 15.0 | 747.9 |
| Avg. | | | | | 0.09/0.21 | | |

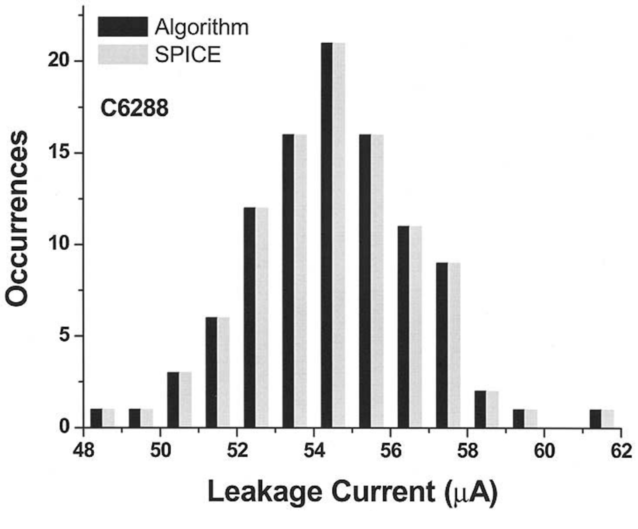


Fig. 6. $I_{\text{leak}}$ histograms for c6288 over 100 input states using SPICE and our approach (1-$\mu$A bin size).

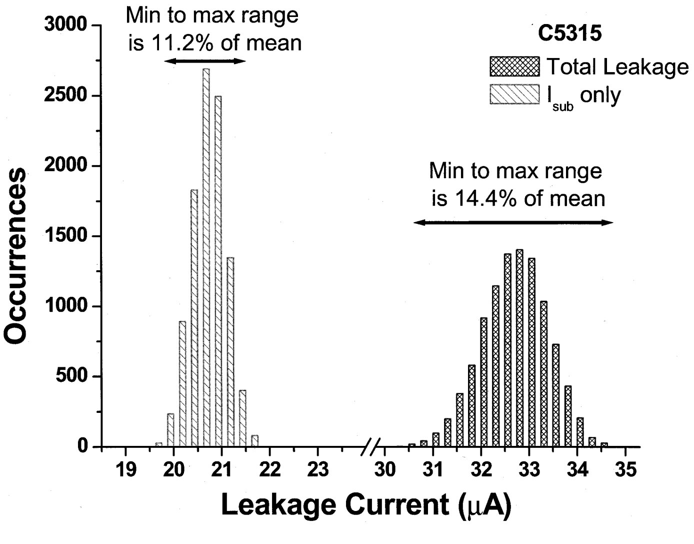


Fig. 7. The consideration of $I_{\text{gate}}$ yields a somewhat broader leakage distribution over 10 000 random input states.

3 (from Section III) occurring during normal circuit operation is increased while the occurrence of scenario 1 is reduced and, hence, the total $I_{\text{gate}}$ for the circuit is diminished. This method is not as effective at reducing $I_{\text{gate}}$ as combined state assignment and pin reordering. However, runtime approaches to leakage reduction (i.e., approaches that do not rely on the use of standby modes) will become increasingly important in the future due to shrinking $I_{on}/I_{\text{leak}}$ ratios in nanometer MOSFETs.

## V. RESULTS

The proposed method for gate tunneling and subthreshold leakage current estimation was implemented and tested for 21 benchmark circuits. These circuits include ten ISCAS85 circuits [29], ten MCNC benchmark circuits [30], and one 64-bit ALU benchmark circuit. All circuits were synthesized with a 0.18-$\mu$m Artisan library using Synopsys Design Compiler and were scaled to a 100 nm technology for the purpose of leakage

TABLE VI
PIN REORDERING RESULTS FOR SLEEP-MODE LEAKAGE REDUCTION

| Circuit | Max. reduction (%) due to state assignment | | | | Max. reduction (%) w/ state assignment & pin reordering | | | |
|---|---|---|---|---|---|---|---|---|
| | $I_{leak}$ | | $I_{gate}$ | | $I_{leak}$ | | $I_{gate}$ | |
| | Random search | Branch and bound | Random search | Branch and bound | Random search | Branch and bound | Random search | Branch and bound |
| C432 | 13.47 | 15.90 | 19.87 | 25.60 | 25.88 | 28.32 | 46.88 | 59.52 |
| C499 | 5.64 | 6.96 | 8.04 | 9.16 | 11.81 | 11.98 | 25.79 | 28.25 |
| C880 | 15.54 | 13.24 | 20.89 | 23.78 | 25.13 | 25.05 | 42.52 | 45.43 |
| C1355 | 5.95 | 8.16 | 8.05 | 10.18 | 17.79 | 19.94 | 32.24 | 33.50 |
| C1908 | 4.19 | 5.84 | 6.14 | 9.23 | 12.66 | 13.67 | 25.93 | 27.76 |
| C2670 | 6.55 | 13.21 | 12.59 | 22.43 | 15.07 | 17.90 | 29.90 | 33.80 |
| C3540 | 5.70 | 7.33 | 6.66 | 6.58 | 19.93 | 19.00 | 36.02 | 36.40 |
| C5315 | 6.48 | 9.13 | 9.91 | 10.28 | 16.65 | 17.39 | 31.68 | 34.34 |
| C6288 | 11.31 | 14.80 | 8.01 | 12.02 | 28.64 | 29.35 | 46.63 | 45.86 |
| C7552 | 3.47 | 6.94 | 6.21 | 8.86 | 12.88 | 16.74 | 25.85 | 28.10 |
| alu64 | 13.32 | 16.34 | 20.05 | 32.86 | 23.15 | 28.64 | 38.80 | 49.23 |
| i1 | 19.09 | 21.49 | 33.23 | 39.03 | 19.09 | 21.50 | 33.23 | 40.19 |
| i2 | 11.13 | 17.60 | 22.81 | 55.42 | 13.92 | 22.20 | 27.62 | 69.68 |
| i3 | 12.91 | 25.05 | 17.20 | 22.19 | 12.91 | 25.05 | 17.74 | 22.19 |
| i4 | 10.72 | 28.84 | 17.64 | 34.63 | 22.96 | 25.31 | 38.39 | 32.09 |
| i5 | 4.02 | 10.65 | 5.16 | 12.22 | 20.47 | 38.40 | 35.58 | 55.93 |
| i6 | 43.16 | 56.32 | 51.12 | 65.12 | 52.55 | 62.92 | 73.37 | 82.12 |
| i7 | 42.52 | 58.46 | 60.90 | 70.52 | 44.93 | 61.37 | 66.77 | 75.81 |
| i8 | 24.79 | 16.11 | 27.14 | 25.32 | 37.16 | 26.82 | 54.55 | 48.84 |
| i9 | 36.31 | 25.63 | 39.46 | 44.35 | 48.97 | 33.91 | 64.67 | 61.87 |
| i10 | 4.61 | 9.00 | 6.93 | 12.36 | 14.05 | 16.27 | 26.30 | 28.94 |
| Avg. | 14.33 | 18.43 | 19.43 | 26.28 | 23.65 | 26.75 | 39.07 | 44.76 |

estimation (results in this section use the 15-Å process from Section II). Each benchmark circuit was synthesized using inverters, NAND gates, and NOR gates with a maximum of four inputs for any gate. For SPICE simulation, Berkeley predictive SPICE models for 100 nm technology were used in conjunction with the gate tunneling current model discussed in Section II. The total leakage current for each circuit was determined for 100 random input states using the proposed leakage estimation method and also using SPICE simulation. The results are shown in Table V. For each circuit, the average leakage current with and without gate tunneling current is shown. The estimated total leakage current is also compared with SPICE. The proposed method had an average error of 0.09% over all circuits and simulated circuit states, with a maximum error of 0.67% across any circuit/input state combination. The final column in Table V shows the run time for the proposed leakage estimation method (note units differ). The run time speedup compared to SPICE ranged from 5,000 to 52 000 ×, making it feasible to perform combined gate tunneling and subthreshold leakage estimation for large designs.

Fig. 6 shows a histogram of the total leakage current for the largest benchmark circuit c6288, over 100 input states obtained from both SPICE simulation and the proposed analysis approach. As implied by the results from Table V, there is a nearly perfect match between the two leakage current distributions—in particular the state yielding the minimum leakage current for both distributions is the same, indicating that the fast analysis approach should be useful for driving sleep state assignment. Also, Fig. 7 shows the resulting histogram of leakage current both with and without $I_{\text{gate}}$ for 10 000 random input states for the C5315 circuit. The range of the distribution (maximum leakage—minimum leakage) grows in relation to the average leakage when considering $I_{\text{gate}}$.

Table VI shows the results of leakage minimization through state assignment and pin reordering for circuits in sleep mode, using the two optimization approaches discussed in Section IV: random search with 10 000 input vectors and the branch-and-bound algorithm. In columns 2–5 the leakage reduction results are shown when only state assignment is used while columns 6–9 show the results when combined state assignment and pin reordering are applied. As seen from Table VI, state assignment is less effective for large circuits (implying many levels of logic) due to functional correlations among the gates. Most of the literature focuses on comparing the minimum leakage state with the maximum possible leakage but comparing to the average state is more relevant [3] and we use that convention here. Since gate leakage is strongly dependent on the stack ordering, we also compare our results with the leakage current considering an average pin ordering. Based on the state probability of the nodes, we find the leakage under best and worst pin ordering for a

[3] Consider a circuit that does not enter a predefined standby state when sleep mode is engaged, but simply stops toggling. The leakage during sleep mode in that case depends on the prior circuit state which is random. The leakage over the course of many sleep modes will, thus, tend toward the average leakage over all possible circuit states.

TABLE VII
PIN REORDERING RESULTS FOR RUNTIME LEAKAGE REDUCTION

| Circuit | Avg. reduction (%) for 10000 input states | | | |
|---|---|---|---|---|
| | p(1)=0.5 | | p(1)=0.25 | |
| | $I_{leak}$ | $I_{gate}$ | $I_{leak}$ | $I_{gate}$ |
| C432 | 4.02 | 8.79 | 9.08 | 20.19 |
| C499 | 2.09 | 4.58 | 4.51 | 10.18 |
| C880 | 3.30 | 7.51 | 4.23 | 9.20 |
| C1355 | 3.43 | 7.19 | 6.46 | 13.62 |
| C1908 | 3.38 | 7.91 | 5.72 | 13.33 |
| C2670 | 2.17 | 5.14 | 4.23 | 10.16 |
| C3540 | 5.85 | 13.17 | 7.41 | 16.53 |
| C5315 | 1.98 | 4.39 | 5.31 | 11.77 |
| C6288 | 4.79 | 10.21 | 11.51 | 24.78 |
| C7552 | 1.71 | 4.04 | 4.54 | 10.55 |
| alu64 | 2.97 | 6.49 | 5.66 | 12.05 |
| i1 | 0.05 | 0.11 | 0.10 | 0.24 |
| i2 | 1.25 | 2.09 | 1.53 | 3.86 |
| i3 | 0.01 | 0.03 | 0.00 | 0.00 |
| i4 | 1.55 | 3.15 | 1.74 | 3.60 |
| i5 | 3.28 | 7.45 | 2.36 | 5.50 |
| i6 | 4.22 | 8.78 | 3.21 | 7.15 |
| i7 | 3.47 | 7.09 | 3.86 | 8.52 |
| i8 | 6.25 | 13.56 | 3.90 | 8.30 |
| i9 | 3.86 | 8.42 | 3.62 | 7.74 |
| i10 | 6.03 | 13.58 | 6.17 | 13.94 |
| Avg. | 3.13 | 6.84 | 4.53 | 10.06 |

circuit, and then take the average of these two leakage values. As shown in Table VI, the branch-and-bound approach performs better than random search method.[4] In the branch-and-bound approach, the average leakage reduction using only state assignment over all circuits is 18%, while the reduction in the gate leakage component of the total $(I_{\text{gate}})$ is 26%. The efficacy of state assignment is, therefore, slightly higher for $I_{\text{gate}}$ than $I_{\text{sub}}$. When performing simultaneous pin reordering and state assignment (columns 6–9), the reduction in total leakage is 27% on average over all circuits with an average reduction in the $I_{\text{gate}}$ component of 45%. The impact of pin reordering on $I_{\text{gate}}$ is pronounced, reducing $I_{\text{gate}}$ by up to 82%.

The runtime leakage reduction using pin reordering is shown in Table VII. These experiments were conducted as described in Section IV-A single pin reordering is performed based on state probabilities at all circuit nodes and 10 000 input vectors with each input having a state probability of 0.25 and 0.5 are applied to both the best and worst reordered topologies. In Table VII, we show the reduction rate between the leakage of best reordered topology and that of an average ordered circuit. The total leakage savings over all 10 000 states is 3.13% on average over all circuits for an input state probability of 0.5. Note that $I_{\text{gate}}$ is reduced by a larger factor than total leakage $(I_{\text{leak}})$, as expected; by 6.84% on average and $>10\%$ in several cases. Also, the leakage reduction is dependent on the PI state probabilities. For instance, when all PI's have state probabilities of 0.25 rather than 0.5, the average runtime $I_{\text{leak}}$ reduction becomes 4.53% over all circuits with C6288 showing an 11.51% reduction and $I_{\text{gate}}$ improvements range up to 25%. It is important to note that the improvement achieved by the proposed method for runtime leakage reduction depends on accurate information of the statistics of the PI's, which may not be available at design time. While the runtime improvements using pin reordering are not large, they do benefit power consumption at all times rather than during standby mode only. Note that i1 and i3 benchmark circuits have almost no improvement from pin reordering. While all other circuits consist of at least 50% NAND gates, only $\sim 5\%$ of the gates in these two small circuits are NAND gates. Since pin reordering is only effective for NAND gates for our implementation, the leakage improvement is negligible for circuits i1 and i3.

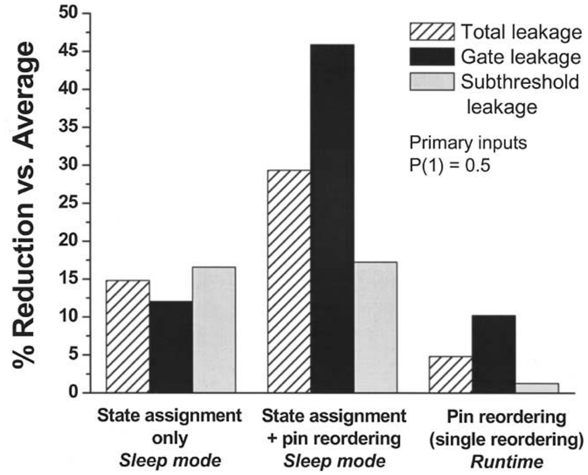


Fig. 8. Leakage reduction techniques compared to the average leakage over 10 000 random input states for C6288.

Finally, Fig. 8 summarizes the impact of state assignment and pin reordering on circuit c6288 assuming a state probability of primary inputs of 0.5 for runtime leakage reduction. The figure shows the achievable reductions in $I_{\text{gate}}$, $I_{\text{sub}}$, and $I_{\text{leak}}$ for the three different scenarios of Tables VI and VII. State assignment works equally well for $I_{\text{sub}}$ and $I_{\text{gate}}$ whereas the addition of pin reordering can be seen to provide substantial benefits for both $I_{\text{gate}}$ and $I_{\text{leak}}$ with little improvement for $I_{\text{sub}}$. Technologies with higher components of $I_{\text{leak}}$ due to $I_{\text{gate}}$ will exhibit greater improvements in both sleep mode and runtime leakage when applying pin reordering.

## VI. CONCLUSION

We developed a fast approach to computing total leakage current in large circuit blocks considering both subthreshold and gate tunneling currents. The proposed approach accurately accounts for the complex interaction between $I_{\text{gate}}$ and $I_{\text{sub}}$ in stacked MOS configurations and is based on precharacterized tables of individual leakage currents for three distinct scenarios. We applied the proposed method to benchmark circuits and demonstrated an average error of only 0.09% compared to SPICE with a four orders of magnitude runtime speedup. Based on the proposed analysis method, we found that the spread in total leakage for a given gate over its input state space is drastically reduced for NOR structures when considering

[4] The largest runtime of random search is 434 s for alu64. The branch-and-bound approach has a run time limit of 500 s for larger circuits.

$I_{\rm gate}$ but is increased for NAND gates. We also propose the use of pin reordering to effectively limit gate leakage as $I_{\rm gate}$ depends strongly on the location of off devices within a nonconducting stack. Results show 22%–82% reductions in $I_{\rm gate}$ during standby modes using pin reordering and corresponding 12–73% reductions in total leakage beyond traditional state assignment. When applied to runtime leakage, pin reordering reduces $I_{\rm gate}$ by up to 25% depending on circuit topology and input data statistics.

## References

[1] S. Thompson *et al.*, "A 90 nm logic technology featuring 50 nm strained silicon channel transistors, 7 layers of Cu interconnects, low k ILD, and 1 $\mu m^2$ 6-T SRAM cell," in *Proc. Int. Electron Devices Meeting*, 2002, pp. 61–64.

[2] A. Ono, K. Fukasaku, T. Hirai, S. Koyama, M. Makabe, T. Matsuda, M. TAkimoto, Y. Kunimune, N. Ikezawa, Y. Yamada, F. Koba, K. Imai, and N. Nakamura, "A 100 nm node CMOS technology for practical SOC application requirement," in *Proc. Int. Electron Devices Meeting*, 2001, pp. 511–514.

[3] "2001 International Technology Roadmap for Semiconductors,", http://public.itrs.net.

[4] B. Yu, H. Wang, C. Riccobene, Q. Xiang, and M.-R. Lin, "Limits of gate oxide scaling in nano-transistors," in *Proc. Symp. VLSI Technology*, 2000, pp. 90–91.

[5] Y.-C. Yeo, Q. Lu, W.-C. Lee, T.-J. King, C. Hu, X. Wang, X. Guo, and T. P. Ma, "Direct tunneling gate leakage current in transistors with ultra thin silicon nitride gate dielectric," *IEEE Electron Device Lett.*, vol. 21, pp. 540–542, Nov. 2000.

[6] Q. Xiang, J. Jeon, P. Sachdey, B. Yu, K. C. Saraswat, and M.-R. Lin, "Very high performance 40 nm CMOS with ultra-thin nitride/oxynitride stack gate dielectric and pre-doped dual poly-Si gate electrodes," in *Proc. Int. Electron Devices Meeting*, 2000, pp. 860–862.

[7] K. Nose and T. Sakurai, "Optimization of $V_{dd}$ and $V_{th}$ for low-power and high-speed applications," in *Proc. Asia-South Pacific Design Automation Conf.*, 2000, pp. 469–474.

[8] M. Powell, S.-H. Yang, B. Falsaki, K. Roy, and T. N. Vijaykumar, "Gated-$V_{dd}$: A circuit technique to reduce leakage in deep-submicron cache memories," in *Proc. Int. Symp. Low Power Electronics Design*, 2000, pp. 90–95.

[9] D. Sylvester and H. Kaul, "Performance challenges in nanometer design," in *Proc. Design Automation Conf.*, 2001, pp. 3–8.

[10] J. Kao, A. Chandrakasan, and D. Antoniadis, "Transistor sizing issues and tool for multi-threshold CMOS technology," in *Proc. Design Automation Conf.*, 1997, pp. 409–414.

[11] R. K. Krishnamurthy, A. Alvandpour, V. De, and S. Borkar, "High-performance and low-power challenges for sub-70 nm microprocessor circuits," in *Proc. Custom Integrated Circuits Conf.*, 2002, pp. 125–128.

[12] M. C. Johnson, D. Somasekhar, and K. Roy, "Models and algorithms for bounds on leakage in CMOS circuits," *IEEE Trans. Computer-Aided Design*, vol. 18, pp. 714–725, June 1999.

[13] S. Mutoh, T. Douseki, Y. Matsuya, T. Aoki, S. Shigematsu, and J. Yamada, "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.

[14] S. Sirichotiyakul, T. Edwards, C. Oh, J. Zuo, A. Dharchoudhury, R. Panda, and D. Blaauw, "Standby power minimization through simultaneous threshold voltage and circuit sizing," in *Proc. Design Automation Conf.*, 1999, pp. 436–441.

[15] Y. Ye, S. Borkar, and V. De, "A new technique for standby leakage reduction in high-performance circuits," in *Proc. Symp. VLSI Circuits*, 1998, pp. 40–41.

[16] C.-H. Choi, K.-Y. Nam, Z. Yu, and R. W. Dutton, "Impact of gate direct tunneling on circuit performance: A simulation study," *IEEE Trans. Electron Devices*, vol. 48, pp. 2823–2829, Dec. 2001.

[17] S. Schwantes and W. Krautschneider, "Relevance of gate current for the functionality of deep submicron CMOS circuits," in *Proc. European Solid-State Device Research Conf.*, 2001, pp. 471–474.

[18] F. Hamzaoglu and M. R. Stan, "Circuit-level techniques to control gate leakage for sub-100 nm CMOS," in *Proc. Int. Symp. Low Power Electronics and Design*, 2002, pp. 60–63.

[19] C.-H. Choi, K.-H. Oh, J.-S. Goo, Z. Yu, and R. W. Dutton, "Direct tunneling current model for circuit simulation," in *Proc. Int. Electron Devices Meeting*, 1999, pp. 735–738.

[20] W.-C. Lee and C. Hu, "Modeling CMOS tunneling currents through ultrathin gate oxide due to conduction- and valence-band electron and hole tunneling," *IEEE Trans. Electron Devices*, vol. 48, pp. 1366–1373, July 2001.

[21] . [Online]. Available: http://www-device.eecs.berkeley.edu/~ptm

[22] N. Yang, W. K. Henson, and J. J. Wortman, "A comparative study of gate direct tunneling and drain leakage currents in N-MOSFETS with sub-2 nm gate oxides," *IEEE Trans. Electron Devices*, vol. 47, pp. 1636–1644, Aug. 2000.

[23] Y. Taur, "CMOS design near the limit of scaling," *IBM J. Res. Develop.*, pp. 213–222, Mar./May 2002.

[24] J. Halter and F. Najm, "A gate-level leakage power reduction method for ultra-low-power CMOS circuits," in *Proc. Custom Integrated Circuit Conf.*, 1997, pp. 475–478.

[25] A. Chandrakasan, W. Bowhill, and F. Fox, *Design of High-Performance Microprocessor Circuits*. Piscataway, NJ: IEEE Press, 2001.

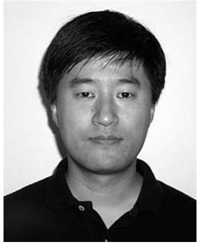
[26] T. Inukai, M. Takamiya, K. Nose, H. Kawaguchi, T. Hiramoto, and T. Sakurai, "Boosted gate MOS (BGMOS): Device/circuit cooperation scheme to achieve leakage-free giga-scale integration," in *Proc. Custom Integrated Circuit Conf.*, 2000, pp. 409–412.

[27] D. Lee and D. Blaauw, "Static leakage reduction through simultaneous threshold voltage and state assignment," in *Proc. Design Automation Conf.*, 2003, pp. 191–194.

[28] S. Ercolani, M. Favalli, M. Damiani, P. Olivo, and B. Ricco, "Estimate of signal probability in combinational logic networks," in *Proc. European Test Conf.*, 1989, pp. 132–138.

[29] F. Brglez and H. Fujiwara, "A neutral netlist of 10 combinatorial benchmark circuits," in *Proc. Int. Symp. Circuit and Systems*, 1985, pp. 695–698.

[30] Collaborative Benchmark Laboratory, http://www.cbl.ncsu.edu.



**Dongwoo Lee** (S'03) received the B.S. and M.S. degrees in electronics engineering from Korea University, Seoul, Korea, in 1994 and 1996, respectively. He is currently working toward the Ph.D. degree in electrical engineering at the University of Michigan, Ann Arbor.

From May 1996 through June 2001, he was with the Non Volatile Memory Design Team, Samsung Electronics Company, Ltd., Kyungki-Do, Korea. His current research interests include circuit analysis and optimization problems for low-power VLSI systems.

**David Blaauw** (M'93) received the B.S. degree in physics and computer science from Duke University, Durham, NC, in 1986, and the M.S. and Ph.D. degrees in computer science from the University of Illinois, Urbana, in 1988 and 1991, respectively.

He was a Development Staff Member at the Engineering Accelerator Technology Division, IBM Corporation, Endicott, NY, until August 1993. From 1993 to August 2001, he was with Motorola, Inc. Austin, TX, where he was the Manager of the High Performance Design Technology Group. Since August 2001, he has been an Associate Professor at the University of Michigan, Ann Arbor. His work has focused on VLSI design and CAD with particular emphasis on circuit analysis and optimization problems for high-performance and low-power designs.

Dr. Blaauw was the Technical Program Chair and General Chair for the International Symposium on Low Power Electronics and Design in 1999 and 2000, respectively, and was the Technical Program Co-Chair and Member of the Executive Committee for the ACM/IEEE Design Automation Conference in 2000 and 2001.

**Dennis Sylvester** (S'95–M'00) received the B.S. degree (*summa cum laude*) from the University of Michigan, Ann Arbor, in 1995, and the M.S. and Ph.D. degrees from the University of California, Berkeley, in 1997 and 1999, respectively, all in electrical engineering.

He was with Hewlett-Packard Laboratories, Palo Alto, CA, from 1996 to 1998. After working as a Senior R&D Engineer in the Advanced Technology Group of Synopsys, Mountain View, CA, he is currently an Assistant Professor of Electrical Engineering at the University of Michigan, Ann Arbor. He has published numerous papers in his field of research, which includes the modeling, characterization, and analysis of on-chip interconnect, low-power circuit design techniques, and variability-aware circuit approaches.

Dr. Sylvester received an NSF CAREER award, the 2000 Beatrice Winner-Award at ISSCC, two outstanding research presentation awards from the Semiconductor Research Corporation, and a best student paper award at the 1997 International Semiconductor Device Research Symposium. He is also the recipient of the 2003 Ruth and Joel Spira Outstanding Teaching Award in the University of Michigan College of Engineering. His dissertation research was recognized with the 2000 David J. Sakrison Memorial Prize as the most outstanding research in the Electrical Engineering and Computer Science Department of the University of California, Berkeley. He is on the technical program committee of several design automation and circuit design conferences and was the general chair for the 2003 ACM/IEEE System-Level Interconnect Prediction (SLIP) Workshop. In addition, he is part of the International Technology Roadmap for Semiconductors (ITRS) U.S. Design Technology Working Group and made significant modeling contributions to the Design and System Drivers chapters of the 2001 ITRS. He is a Member of the Association for Computing Machinery, American Society of Engineering Education, and Eta Kappa Nu.