Simultaneous Subthreshold and Gate-Oxide Tunneling Leakage Current Analysis in Nanometer CMOS Design

Dongwoo Lee, Wesley Kwong, David Blaauw, Dennis Sylvester
University of Michigan, Ann Arbor, MI
{dongwool,wkwong,blaauw,dmcs}@umich.edu

Abstract

In this paper we develop a fast approach to analyze the total leakage power of a large circuit block, considering both gate leakage, I_{gate}, and subthreshold leakage, I_{sub}. The interaction between I_{sub} and I_{gate} complicates analysis in arbitrary CMOS topologies. We propose simple and accurate heuristics to quickly estimate the state-dependent total leakage current considering the interaction between I_{sub} and I_{gate}. We apply this method to ISCAS benchmark circuits in a projected 100nm technology and demonstrate excellent accuracy compared to SPICE simulation with a 20,000X speedup on average.

1 Introduction

In the aggressive scaling of MOSFETs seen over the past several decades, the shrinking of the gate oxide layer thickness (T_{ox}) has been just as significant as effective channel length (L_{eff}) reduction. CMOS processes in the 90nm technology node will have T_{ox} values of 12-16 Angstroms (1.2-1.6nm) [1][2][3]. While continued scaling of T_{ox} is necessary to provide substantial current drive at reduced voltage supplies, it leads to significant gate tunneling leakage current (I_{gate}).

I_{gate} arises due to the small probability of an electron directly tunneling through the insulating SiO_2 layer. Both this probability and I_{gate} itself are strong exponential functions of T_{ox} as well as functions of the voltage potential across the gate oxide. A difference in T_{ox} of just 2 Angstroms (A) can lead to an order of magnitude change in I_{gate}, making it extremely sensitive to process fluctuations. Another key point is that I_{gate} for a PMOS device is typically one order of magnitude smaller than an NMOS device with identical T_{ox} and V_{dd} when using SiO_2 [4]. This is due to the much higher energy barrier seen by holes in a MOSFET channel in an Si-SiO_2 system. However, dielectric materials other than SiO2 present different barrier heights to electrons and holes such that PMOS I_{gate} may not always be negligible. In the case of nitrided gate oxides, in use today in some processes, PMOS I_{gate} exceeds NMOS I_{gate} for higher nitrogen concentrations [15][16].

Some modern processes use a nitrided gate oxide (also called oxynitride) to raise the dielectric constant of the gate insulator from 3.9 to ~ 4.1-4.2 and yield an order of magnitude reduction in I_{gate} for the same C_{ox} value. More aggressive high-k materials, such as hafnium oxide (HfO_2), provide dielectric constants in the range of 25-50 and will greatly diminish the significance of I_{gate}. However, there are numerous process integration problems with such high-k materials As a result, the introduction of true high-k materials (beyond oxynitride) is not expected before the 65nm node in 2007 [3].

There has been extensive work in the analysis and minimization of I_{sub} based on the understanding that it poses a fundamental scaling limit to traditional CMOS design. However, I_{gate} has been growing much faster and to this point has been addressed primarily by device engineers and not circuit designers, EDA tool developers, etc.

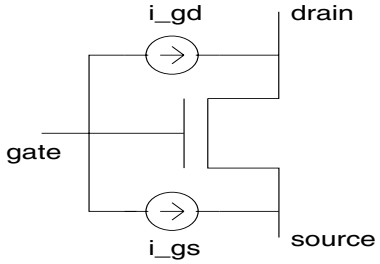

Figure 1. Macro model for transistor gate leakage.

In [5] and [6], the authors examined the impact of gate leakage on circuit functionality but did not address its contribution to leakage power. In [7], the authors contribute the first circuit design concepts to reducing the impact of gate leakage – these focus on leveraging the lower I_{gate} in PMOS devices by using p-type domino circuits rather than n-type as well as PMOS sleep transistors for standby modes.

Circuit level analysis of I_{gate} is complicated by 1) state dependency and 2) the interaction of I_{sub} and I_{gate}. The state dependence of I_{sub} is well understood, especially in the context of the stack effect. Efficient models to compute I_{sub} based on the number of off transistors in a stack have been proposed [8]. However, gate tunneling current is primarily driven by ON-devices in contrast to I_{sub}, which changes the problem substantially. In addition, total leakage current is not always the sum of I_{sub} and I_{gate}. For certain input states the currents become inter-dependent, which affects the internal node voltages and complicates the analysis.

In this paper, we describe a fast new approach to total leakage power analysis considering both I_{gate} and I_{sub}. We account for the interaction between these two sources of current and highlight changes in the traditional standby current problem when I_{gate} is appreciable. Using table lookup and knowledge of the state dependence of I_{sub} and I_{gate}, we use a number of benchmark circuits in two similar technologies (with different T_{ox} values) to demonstrate the accuracy of the proposed method. We begin with a discussion of the gate leakage models used in this work.

2 Model for I_{gate}

An empirical gate leakage model was incorporated in a 100nm BSIM3v3 (level 49) model generated using the Berkeley Predictive Technology Model (BPTM) [13]. The gate leakage was modeled using voltage dependent current sources from gate to source (I_{gs}) and gate to drain (I_{gd}), depending on, respectively, V_{gs} and V_{gd}, as shown in Figure 1. The expressions for I_{gs} and I_{gd} are shown below:

$$i_{gd} = \frac{127.04 \times L_{eff} \times e^{(5.60625 \times V_{gd} - 10.6 \times T_{ox} - 2.5)}}{2}, \quad \text{(EQ 1)}$$

$$i_{gs} = \frac{127.04 \times L_{eff} \times e^{(5.60625 \times V_{gs} - 10.6 \times T_{ox} - 2.5)}}{2}, \quad \text{(EQ 2)}$$

where T_{ox} and L_{eff} are given in nanometers. EQ1 and EQ2 are based

Proceedings of the Fourth International Symposium on Quality Electronic Design (ISQED'03)
0-7695-1881-8/03 $17.00 © 2003 IEEE

on an empirical model of total gate leakage fit to IBM data on thin SiO_2 dielectrics that was used in the 2001 ITRS. This model was then adjusted by fitting the equation to data from an industrial 0.13 µm process. Fitting was accomplished by examining the oxide leakage current over the full range of V_{ds} and V_{gs} and then performing a non-linear curve fitting of the equation to the industrial data. This process resulted in the addition of correction factors onto various terms in the equation to obtain a reasonable average fit to the data, as shown in Figure 2. The model was also found to maintain good stability during SPICE simulation.

As seen in Figure 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 increased as V_{gs} is decreased from 1.2 to 0.4 V from approximately 10% to 40%. For $V_{gs} < 0.2$ V, the error is much larger, however since the total gate current is extremely small in these instances 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_{gate} on circuit behavior and to develop a fast and accurate total leakage model, two 100nm technology files were generated - the first with a T_{ox} of 17A and L_{eff} of 50 nm, while the second has a T_{ox} of 15 A and L_{eff} = 60 nm. V_{th} in both technologies is approximately 200mV. The goal in using two processes is to examine the role of I_{gate} in total leakage for a range of I_{gate}/I_{sub} ratios. Specifically, in the 17A process I_{gate} is roughly 1/9 of I_{sub} under worst-case biasing conditions while in the 15A process I_{gate}/I_{sub} = 2/3. I_{sub} values are in the range of 20-40nA/µm of gate width at room temperature which is slightly below the ITRS projected value of 70nA/µm at 100nm (see Figure 2). While both oxide thicknesses are in the higher end of the range specified for 100nm devices by the ITRS (year 2003), we also assume the use of SiO_2 and not an oxynitride since I_{gate} models are available for the former. To compensate for the higher expected I_{gate} in SiO_2, we select conservative T_{ox} values to provide more realistic I_{gate}/I_{sub} ratios. V_{DD} is 1V for both cases and all results in this work are for room temperature (I_{sub} is highly temperature dependent while I_{gate} is not).

3 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 10's to 100's 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.

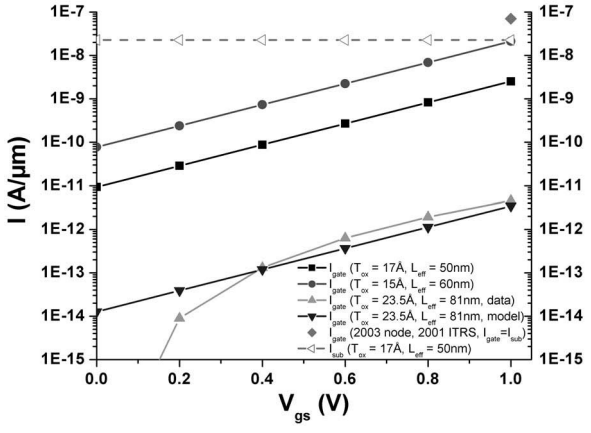

Figure 2. Fit of industrial gate leakage measurements and macro model.

Standby current estimation is complicated by the state dependence of both the I_{gate} and I_{sub} currents. The state dependence of subthreshold leakage current has been extensively studied and exhibits the so-called *stack effect*. Similarly, gate tunneling current has state dependence, as well as dependence on the device type. As mentioned, PMOS devices 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_{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 Figure 3. The maximum gate tunneling current occurs when the input is at V_{dd} and $V_g = V_d = 0$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 1 order of magnitude when $V_{gs} = V_{th,nmos}$, 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 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 [12]. In addition, the corner oxide thickness can be increased by subsequently oxidizing the polysilicon after gate formation which would further suppress tunneling in the overlap regions [14].

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_{leak} of the gate, as shown in Figure 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.

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_{sub} through the PMOS transistors added to I_{gate} through the NMOS

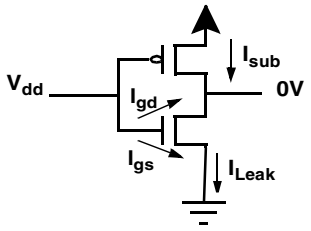

Figure 3. Inverter circuit with NMOS oxide leakage current.

transistors. However, for input states where at least one input is low and the gate output is V_{dd}, I_{sub} through turned-off NMOS transistors and I_{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_{sub} and I_{gate} are therefore interdependent in these cases, and must be analyzed simultaneously.

We consider gate tunneling current in three distinct scenarios for a transistor in a transistor stack, as shown in Figure 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 Figure 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_{gate} of transistor t_n therefore does not affect the voltage at nodes n_a and n_b and can be added to the I_{sub} of the stack to obtain the total leakage current of the gate.
2. In the second scenario, shown in Figure 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} - $V_{th,nmos}$. For transistor t_n, $V_{gs,n}$ and $V_{gd,n}$ are therefore small; approximately one threshold voltage. Based on SPICE simulations, the I_{gate} in this case is more than one order of magnitude smaller than in scenario 1 and can be safely ignored.
3. In the third scenario, shown in Figure 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 100-200mV. 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_{gate} current. This current combines with the I_{sub} through t_t and causes the node voltages at n_a, n_b to increase from their value with only subthreshold current.

   A rise in the voltage at n_a and n_b reduces I_{sub} through t_t, as $V_{gs,t}$ becomes further negative, and also reduces I_{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 $V_{gd,n}$.[1] Therefore, as the voltage of n_a is raised by I_{gate} through t_n, the I_{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_{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.

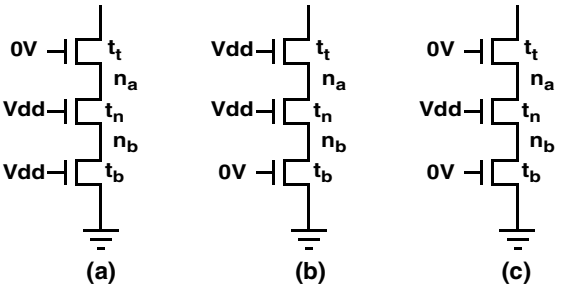

Figure 4. Three input NMOS-stack with three scenarios of combined I_{sub} and I_{gate}.

This effect is illustrated in Table 1, where we show the node voltage of n_a, n_b and the leakage currents for the circuit shown in Figure 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 17A process, the voltages at n_a and n_b increase by 42mV over the case with I_{sub} only when considering both I_{sub} and I_{gate}, resulting in a decrease of I_{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_{gate} to I_{gate} and I_{sub}, resulting in a decrease of I_{gate} through t_n by just 9%. Table 1 also shows SPICE results for the 15A process. In this case, I_{sub} is reduced by 4 orders of magnitude, and becomes negligible.

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

Table 1. Simulation results for individual and combined I_{gate}/I_{sub}.

	17A			15A		
	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

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 [8] and may be used. In this paper, we use an empirical model in which the total subthreshold leakage current is expressed as follows:

$$I_{sub,k} = I_{sub,1} * S_k * s_t, \quad \text{(EQ 3)}$$

where $I_{sub,1}$ 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 $I_{sub,1}$ 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 $V_{th,nmos}$ 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 in our approach by constructing an additional set of tables where a conducting transistor is placed above the off-transistor stack.

1. For example, [7] states that a 0.3V change in V_{gs}, V_{gd} leads to a decade change in I_{gate}. However, a reduction in V_{gs} of only ~0.1V yields a 10X drop in I_{sub}.

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

The total leakage current, as well as its I_{gate} and I_{sub} components, are then computed as follows. First, the total number of off-transistors in the stack is determined and the I_{sub}, in the absence of I_{gate}, is found using EQ3. Next, the tunneling currents $I_{gate,l}$ 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_{total}, and its tunneling and subthreshold components I_{gate} and I_{sub}, are then determined as follows:

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

$$I_{gate} = \sum_{l} I_{gate,l} \quad \text{(EQ 5)}$$

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

The first term in EQ4 corresponds to the I_{gate} current of transistors in scenario 1, which is independent of the other currents in the stack. The second term of EQ4 corresponds to the I_{gate} of transistors in scenario 3 which displaces the I_{sub} of the stack. Hence, the current for this term is the maximum of these two currents. EQ5 and EQ6 express the total I_{sub} and I_{gate} in the transistor stack.

For the analysis of series/parallel NMOS structures, such as AOI and 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 Figure 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 Figure 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_{gate} nor I_{sub} and is neglected during leakage current computation.

After collapsing parallel devices in a transistor stack, we compute the gate tunneling and subthreshold leakage current using EQ4.

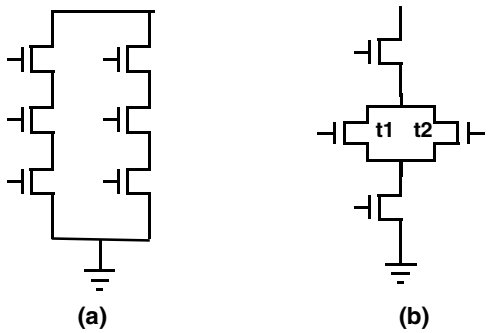

Figure 5. Leakage current computation for series/parallel structures.

To demonstrate the accuracy of the proposed leakage estimation method, we show the analysis results for a 3-input NAND gate under all possible input states in Tables 2 and 3 for both 15A and 17A gate oxide thicknesses. The leakage current obtained from SPICE simulation using the proposed analysis method is also shown and has an average error of 1.2% over all input states. The maximum error occurs for state 110 with 17A gate oxide thickness. However, the total leakage current in this case is small and hence the error, in terms of absolute current, is acceptable.

Table 2. Leakage estimation for 3 input NAND gate with 15A oxides.

State	estimated current			SPICE	%diff
	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 3. Leakage estimation for 3 input NAND gate with 17A oxides.

State	estimated current			SPICE	%diff
	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%

4 Impact of I_{gate} on Circuit Leakage Behavior

In this section we discuss the role of gate tunneling current in overall circuit leakage behavior, with emphasis on how I_{gate} alters conventional notions of standby power analysis and minimization. We begin by examining the differing state dependences of I_{gate} and I_{sub}. The strong state dependence of I_{sub} forms the basis of standby modes that exercise a specific input vector to minimize the total I_{sub} in a circuit block [8][17]. These approaches rely on the stack effect by turning off multiple series connected devices in as many gates as possible. However, the consideration of I_{gate} complicates the state dependence analysis. Table 4 shows the change in the average leakage current over all possible input states when considering I_{gate}. Both technologies are characterized for 2, 3, and 4-input NOR and NAND gates. Even with a relatively low I_{gate} value for the T_{ox} = 17A technology, the average leakage over all states in the gates studied increases by 10-35% when considering both I_{gate} and I_{sub} together. In the more aggressive 15A technology, the rise in average leakage is 65-160% for NANDs and up to 310% for 4-in NOR gates.

Table 4. Impact of I_{gate} on state dependence with I_{leak}

gate type	Average I_{leak}		max I_{leak} / min I_{leak} across all states	
	w/o I_{gate} (15A / 17A)	w/ I_{gate} (15A / 17A)	w/o I_{gate} (15A / 17A)	w/ I_{gate} (15A / 17A)
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

Note that the increase in total leakage due to I_{gate} can be larger than the relative magnitude of I_{gate} to I_{sub} due to the differing state dependencies - that is, states exhibiting low I_{sub} values may exhibit large I_{gate} values.

This leads to another key observation in Table 4; the worst-case leakage states of common CMOS gates, behave differently when both I_{sub} and I_{gate} are considered. When only I_{sub} is relevant, the worst-case leakage state for NAND structures is typically 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 leads to all NMOS devices leaking concurrently. For these two cases, we now include I_{gate}. For NAND gates with inputs all tied high, the NMOS devices in the pull-down stack all exhibit *worst-case* I_{gate} which adds to the large I_{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_{gate}, the overall impact will be small. Meanwhile, the parallel pull-down devices exhibit only reverse edge direct tunneling which we assume to be negligible. As a result of the these trends, we find that the *range* of total leakage current across states is broadened for NAND gates and compressed for NORs. In fact, for NOR gates we find that with a reasonable magnitude of I_{gate} (compared to I_{sub}), the range of leakage current over all input states is extremely small since I_{gate} and I_{sub} are complementary over the input state space (in this sense, complementary means that states with large I_{gate} have small I_{sub} and vice versa). This is shown in Table 4 where the ratio of maximum to minimum leakage current over all possible states is reduced from 21.3X in a 3-input NOR to 1.48X when considering I_{gate}. Interestingly, the minimum I_{sub} leakage state of all high inputs for NORs becomes the *highest* total leakage state for 3 and 4-input NORs when T_{ox} = 15A.

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_{sub} [18]. In [7], the authors addressed the impact of I_{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 non-activity. Thus, in the normal configuration (NMOS sleep device) when the sleep transistor is OFF, V_{gs} = 0 and V_{gd} floats towards -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_{gate} at V_{gs}=V_{dd} and V_{ds}=0 [12]. While this reduction is not as substantial as the several orders of magnitude drop in I_{sub} realized with MTCMOS, it is still beneficial. Since in the sleep mode I_{gate} will likely be dominant, two approaches may be considered: 1) Reduce the V_{th} of the sleep device somewhat (e.g. 100mV) to minimize the delay penalty associated with an extra series device. This allows the use of smaller sleep devices to simultaneously reduce I_{gate}, dynamic power, and layout area. 2) Incorporate a multi-T_{ox} process to allow the sleep devices to reduce I_{gate} in addition to I_{sub}. A limited (and practical) form of a multi-T_{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 [11].

5 Results

The proposed method for gate tunneling and subthreshold leakage current estimation was implemented and tested for ten benchmark circuits (nine of which are ISCAS85 circuits [9]). All circuits were synthesized with a 0.18 µm Artisan library using Synopsys Design Compiler and scaled to a 100nm technology (results in this section use the 15A process). For SPICE simulation, Berkeley predictive SPICE models for 100nm technology were used along with the gate tunneling current model discussed in Section 2. 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 5. 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 0.04% over all circuits and simulated circuit states, with a maximum error of 0.35%. The final column in Table 5 shows the run time for the proposed leakage estimation method (note units differ). The run time speedup compared to SPICE ranged from 6,000 to 52,000X, making it feasible to perform combined gate tunneling and subthreshold leakage estimation for large designs.

Table 5. Leakage estimation results for benchmark circuits.

circuit	# gates	estimated leakage current, µA (avg)		SPICE leakage current (avg)	% error (avg/max)	run time	
		w/o I_{gate}	w/ I_{gate}			proposed method (ms)	SPICE (s)
C432	121	1.71	2.82	2.81	0.12/0.32	0.18	9.36
C499	517	6.44	10	10	0.01/0.02	2.4	38.38
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	41.39
C1908	425	5.55	8.61	8.61	0.01/0.04	2.7	35.84
C2670	750	9.45	14.46	14.46	0.02/0.06	3.9	60.55
C3540	890	11.77	18.99	18.98	0.04/0.08	6.3	100.2
C5315	1524	20.49	32.28	32.28	0.01/0.02	11.1	180.79
C6288	2388	32.82	54.54	54.53	0.02/0.04	34.4	971.3
alu64	1791	25.83	40.58	40.63	0.14/0.35	42.6	244.95

Figure 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 5, 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. Finally, Figure 7 shows the resulting histogram of leakage current both with and without I_{gate} for 10000 random input states for the C5315 circuit. The range of the distribu-

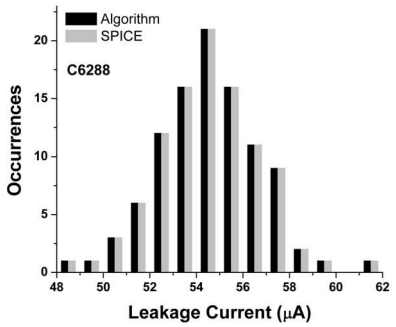

Figure 6. I_{leak} histograms for c6288 over 100 input states using SPICE and our approach show perfect match (1μA bin size).

tion (maximum leakage - minimum leakage) grows in relation to the average leakage when considering I_{gate}.

6 Conclusions

In this paper we presented an approach for determining the total leakage current, with both gate and subthreshold leakage components, in a circuit block. We point out a number of state-dependent scenarios in which I_{gate} and I_{sub} interact in different ways -- these scenarios can be identified and the total leakage predetermined on a state-by-state basis. Then using table lookup, we showed that a fast leakage analysis approach can provide near-exact accuracy compared to SPICE with an average run time improvement of 20000X. We then discussed the impact of I_{gate} on the standby power of large CMOS circuits. Due to the differing state dependencies of I_{gate} and I_{sub}, NAND gates exhibit a broader range of leakage over all input combinations when considering I_{gate} whereas total leakage of NOR gates becomes almost insensitive to input state.

REFERENCES

[1] S. Thompson *et al*., "A 90nm logic technology featuring 50nm strained silicon channel transistors, 7 layers of Cu interconnects, Low k ILD, and 1μm^2 6-T SRAM cell," *Proc. IEDM*, in press, 2002.

[2] A. Ono, *et al*., "A 100nm node CMOS technology for practical SOC application requirement," *Proc. IEDM*, pp. 511-514, 2001.

[3] 2001 International Technology Roadmap for Semiconductors.

[4] B. Yu, *et al*., "Limits of gate oxide scaling in nano-transistors," *Proc. Symp. VLSI Tech.*, pp. 90-91, 2000.

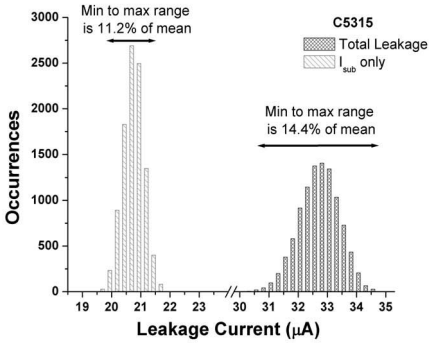

Figure 7. The consideration of I_{gate} yields a somewhat broader leakage distribution over 10000 random input states.

[5] 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*, pp. 2823-2829, Dec. 2001.

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

[7] F. Hamzaoglu and M.R. Stan, "Circuit-level techniques to control gate leakage for sub-100nm CMOS," *Proc. ISLPED*, pp. 60-63, 2002.

[8] M.C. Johnson, *et. al*., "Models and algorithms for bounds on leakage in CMOS circuits," *IEEE Trans. CAD*, pp. 714-725, June 1999.

[9] F. Brglez and H.Fujiwara, "A Neutral Netlist of 10 Combinatorial Benchmark Circuits", *Proc. ISCAS*, 1985, pp.695-698.

[10] S. Sirichotiyakul, *et al.*, "Standby power minimization through simultaneous threshold voltage and circuit sizing," *Proc. DAC*, pp. 436-441, 1999.

[11] T. Inukai, *et al.*, "Boosted Gate MOS (BGMOS): Device/circuit cooperation scheme to achieve leakage-free giga-scale integration," *Proc. CICC*, 2000.

[12] 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-2nm gate oxides," *IEEE Trans. Electron Devices*, pp. 1636-1644, Aug. 2000.

[13] http://www-device.eecs.berkeley.edu/~ptm

[14] Y. Taur, "CMOS design near the limit of scaling," *IBM J. R&D*, pp. 213-222, March/May 2002.

[15] Y.-C. Yeo, *et al.*, "Direct tunneling gate leakage current in transistors with ultra thin silicon nitride gate dielectric," *IEEE Electron Device Letters*, pp. 540-542, Nov. 2000.

[16] Q. Xiang, *et al.*, "Very high performance 40nm CMOS with ultra-thin nitride/oxynitride stack gate dielectric and pre-doped dual poly-Si gate electrodes," *Proc. IEDM*, pp. 860-862, 2000.

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

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