

Accurate Stacking Effect Macro-modeling of Leakage Power in Sub-100nm Circuits

†Shengqi Yang, †Wayne Wolf, †N. Vijaykrishnan, ‡Yuan Xie and ¶Wenping Wang

†Department of Electrical Engineering, Princeton University, Princeton, NJ, 08544

‡Microsystems Design Lab, The Penn State University, University Park, PA, 16802

¶Microelectronics Department, Peking University, Beijing, P.R.C, 100871

{shengqiy, wolf}@princeton.edu, {vijay, yuanxie}@cse.psu.edu, {wangwp}@ime.pku.edu.cn

Abstract—An accurate and efficient stacking effect macro-model for leakage power in sub-100nm circuits is presented in this paper. Leakage power, including subthreshold leakage power and gate leakage power, is becoming more significant compared to dynamic power when technology scaling down below 100nm. Consequently, fast and accurate leakage power estimation models, which are strongly dependent on precise modeling of the stacking effect on subthreshold leakage and gate leakage, are vital for evaluating optimizations. In this work, making use of the interactions between subthreshold leakage and gate leakage, we focus our attention on analyzing the effects of transistor stacking on gate leakage between the channel and the gate and that between the drain/source and the gate. The contribution of the latter has been largely ignored in prior work, while our work shows that it is an important factor. Based on the stacking effect analysis, we have proposed a new best input vector to reduce the total leakage power; and an efficient and accurate leakage power estimation macro-model which achieves a mean error of 3.1% when compared to HSPICE.

I. INTRODUCTION

Being consistent with Moore's law, the feature size of VLSI circuit is decreasing at a rate of $0.7\times$ per generation. According to the International Technology Roadmap for Semiconductors (ITRS) 2003 [1], it will reach 65nm by 2007 and 45nm by 2010. Continuous decrease in feature size results in bigger and denser chips with more transistors per unit area and more logic modules. These changes in density and integration ability of VLSI circuit make power consumption to be a more important and critical issue [2], [3].

Before the deep-submicron era, most of the attention has focused on reducing dynamic power P_d , which is given by,

$$P_d = \frac{1}{2} \cdot \alpha \cdot C_L \cdot V_{dd}^2 \cdot f \quad (1)$$

where α is the switching activity, C_L is the capacitive load, V_{dd} is the power supply voltage, and f is the clock frequency. From Equation (1) we can see that reducing V_{dd} is a very efficient mechanism of reducing P_d due to its quadratic contribution. However, reducing the ratio of V_{dd} to threshold voltage, V_{th} , increases the delay, D , of a circuit, as given by,

$$D \propto \frac{C_L V_{dd}}{(V_{dd} - V_{th})^\beta} \quad (2)$$

where β is an experimental constant that varies from 1.4 to 2 for sub-100nm technologies. To overcome this problem, V_{th} is reduced as well. This reduction in the threshold voltage causes a $5\times$ increase in subthreshold leakage current per generation. This trend indicates a significant change in power perspective as technology scales down to the sub-100nm era, i.e., subthreshold leakage power becomes increasingly important as compared to dynamic power [4].

As technology scales, the gate oxide thickness needs to be scaled as well in order to keep the driving capacity of the gate on a considerable level. When the thickness reaches 3nm and below, not only is subthreshold leakage current important, but gate direct tunneling current becomes significant [5], [6] as it increases exponentially with decreasing oxide thickness [7]. Furthermore, the tunneling current occurs in both transistor on and off states. This introduces two significant consequences with respect to off-state power consumption: (i) an increase in the number of transistors contributing to the total

off-state power consumption of the chip and (ii) an increase in the conventional off-state power (i.e. subthreshold leakage power). Due to its exponential dependence on oxide thickness, gate direct tunneling current has the potential to become the dominant factor in sub-100nm technologies. In order to reduce gate leakage, high- κ gate dielectric [8] and metal gate electrodes [9] have been proposed. However, high- κ gate dielectric and metal gate electrodes are not compatible with conventional CMOS process and early availability of manufacturing-worthy materials is expected to be delayed until 2008 according to ITRS 2003. Among the six leakage components of a transistor [10] (subthreshold leakage, gate leakage due to direct tunneling, reverse-biased pn junction band-to-band tunneling (BTBT) leakage, gate leakage due to hot carrier injection, gate induced drain leakage, and channel punchthrough leakage), reverse-biased junction BTBT leakage is also an important contributor to the total leakage. It is caused by high substrate doping density and the application of halo/pocket structure, and will be significant under 25nm technology [11]. However, highly doped halo/pocket structure can be substituted by some more advanced techniques, for example, ultra-thin-body SOI MOSFET, to overcome its shortcomings in sub-100nm technologies [12]. As a result, the reverse-biased junction BTBT leakage is not much severe compared with gate and subthreshold leakage. In our work, we focus on gate and subthreshold leakage.

A. Related Work

A great deal of previous work analyzed the subthreshold leakage power [7], [13], [14] and provided techniques to reduce it. Some of them are input vector control [15], multiple threshold voltage CMOS [16], [17] and supply voltage gating [18], [19]. Gate direct tunneling current was also studied in [5], and minimization techniques were proposed in [5], [6]. All the listed work falls into two categories. They either ignored subthreshold leakage power or gate leakage power, or they treated these two leakage components as they are equally important under all technologies. However, factoring the contributions of both components and also weighing in their relative importance to overall leakage are essential for devising appropriate leakage optimizations.

Recently, comprehensive analysis of the total leakage including subthreshold leakage and gate leakage under 100nm was carried out by S. Mukhopadhyay *et.al.* [11], [20], [21] and D. Lee *et.al.* [22], [23]. The methodology of Mukhopadhyay's work neglected the interactions between subthreshold leakage and gate leakage, which can result in inaccurate leakage estimation for complex circuit structures. While D.Lee's work was based on two important assumptions. The first is that reverse tunneling current between the drain/source and the gate can be ignored compared with the tunneling current between the channel and the gate; another is that gate leakage current of PMOS need not be considered when compared against that of NMOS. Our experiments will show that the first assumption is not valid for sub-100nm technologies because the ignored reverse tunneling current between the drain/source and the gate is not only comparable to that between the channel and the gate, but is also much greater than the subthreshold leakage current. For the second assumption, the gate leakage of PMOS should not be neglected any more, because

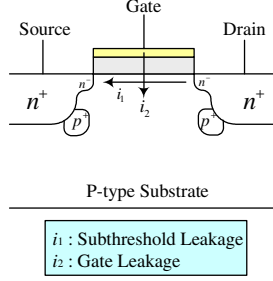


Fig. 1. Significant leakage components for sub-100nm technologies.

this leakage is much larger than the subthreshold leakage and is comparable to the gate leakage of NMOS. Although the subthreshold leakage is smaller than the gate leakage of NMOS and PMOS, we cannot neglect it because the internal voltage is affected by it. In summary, all the ignored leakage components in these two assumptions should not be overlooked any more for sub-100nm technologies. Their existences have a strong effect on the stacking effect analysis, which is a critical factor in determining the best input state and circuit structure for integrated leakage reduction technique. After discarding these assumptions, we should refresh our memory of the total leakage power properties. This paper embarks on a comprehensively quantitative approach, trying to make the following contributions.

B. Paper Contributions

Considering the interactions between subthreshold leakage and gate leakage, we are the first to comprehensively analyze the novel stacking effects on gate leakage current, including the often ignored reverse tunneling current from the drain/source to the gate, which has strong effects on the stacking analysis in sub-100nm nodes. Based on this analysis, a new best input vector is proposed to reduce the total leakage power, and a simple and accurate leakage power estimation macro-model is put forward, which can be integrated into a leakage power estimator in the future work. Further, novel stacking effect on logically equivalent circuit structures is discussed and optimal structure is selected for circuit synthesis.

The rest of this paper is organized as follows. Section II provides some preliminaries. Section III analyzes the novel stacking effects on gate leakage. Section IV concludes this work.

II. PRELIMINARIES

In this section, we will elaborate on some background material that is helpful for understanding the remainder of this paper and further facilitates the stack effect analysis. Specifically, we discuss mechanisms and compact models for significant leakage components, including subthreshold leakage and gate direct tunneling leakage as shown in Fig. 1.

A. Subthreshold Leakage Current

Subthreshold or weak inversion conduction current between source and drain in an MOS transistor occurs when gate voltage is below the threshold voltage. According to BSIM4 model, the equation governing this subthreshold leakage current can be expressed as following:

$$I_{sub} = I_0 e^{\frac{V_{gs} - V_{th}}{n k T / q}} (1 - e^{-\frac{V_{ds}}{k T / q}}) \quad (3)$$

where $I_0 = \mu_0 C_{ox} (W/L) (\frac{kT}{q})^2 e^{1.8}$, W and L are the transistor channel width and length, μ_0 is the low field mobility, C_{ox} is the gate oxide capacitance, k is the Boltzmann constant, q is the electronic charge, V_{gs} and V_{ds} are the gate to source and the drain to source voltages, n is the subthreshold swing factor. From the above equation, we can get the dependence between subthreshold leakage current and V_{ds} :

$$\frac{\partial I_{sub}}{I_{sub}} = \frac{q}{kT(e^{qV_{ds}/kT} - 1)} \cdot \partial V_{ds} \quad (4)$$

B. Gate Leakage Current

The high electric field across oxide coupled with low oxide thickness results in tunneling of electrons from substrate to gate and also from gate to substrate through the gate oxide for NMOS (holes tunneling for PMOS), which is referred as the gate oxide tunneling current. To analytically describe this current, many compact models were put forward, among which is the BSIM4 gate tunneling model [7]. But the original BSIM4 gate leakage model is complex and a more simple model to capture the dependence between leakage current and gate oxide or gate voltage is desirable for fast estimations. Here, we simplify the BSIM4 model and explain the gate leakage current as below:

$$I_{gate} = (A \cdot C) \cdot (W \cdot L) e^{-B \cdot \frac{t_{ox}}{V_{gs}} \cdot \alpha} \quad (5)$$

where $A = q^3 / 8\pi h \phi_b$, $B = 8\pi \sqrt{2m_{ox}} \phi_b^{3/2} / 3hq$, $C = (V_{gs} / t_{ox})^2$, t_{ox} is the thickness of gate oxide, α is a parameter which is ranged from 1 to 0.1 (a typical value is 0.22867) and dependent on the voltage drop across the oxide. For the above terms, h is the Planck's constant, ϕ_b is the barrier height for electrons/holes in the conduction/valence band (for electron it is 3.1eV, for hole it is 4.5eV), m_{ox} is the effective mass of electron/hole. Based on these numbers, we get $A = 4.9589 \times 10^{-7} A/V^2$ (for NMOS), $B = 6.6795 \times 10^{10} (Kg/F \cdot Sec^2)^{0.5}$ (for NMOS). Because V_{gs} and t_{ox} are scaled roughly by the same factor, and the leakage current is dominated by the exponential part, here we approximate parameter C as $0.7225 \times 10^{18} V^2/m^2$. Using this simplified model, we can derive the dependence between gate leakage current and gate oxide thickness or gate voltage:

$$\begin{aligned} \frac{\partial I_{gate}}{\partial t_{ox}} &= (AC)(WL) \left(\frac{-B\alpha}{V_{gs}} \right) e^{-\frac{B\alpha}{V_{gs}} t_{ox}} \\ \frac{\partial I_{gate}}{I_{gate}} &= \left(\frac{-B\alpha}{V_{gs}} \right) \cdot dt_{ox} \end{aligned} \quad (6)$$

$$\begin{aligned} \frac{\partial I_{gate}}{\partial V_{gs}} &= (AC)(WL) \left(\frac{Bt_{ox}\alpha}{V_{gs}^2} \right) e^{-\frac{B\alpha}{V_{gs}} t_{ox}} \\ \frac{\partial I_{gate}}{I_{gate}} &= \left(\frac{Bt_{ox}\alpha}{V_{gs}^2} \right) \cdot dV_{gs} \end{aligned} \quad (7)$$

$$\frac{\partial I_{gate}}{I_{gate}} = \left(\frac{-B\alpha}{V_{gs}} \right) \cdot dt_{ox} + \left(\frac{Bt_{ox}\alpha}{V_{gs}^2} \right) \cdot dV_{gs} \quad (8)$$

For the above equation, we only need to care about the second term in the following stacking effect analysis.

III. ACCURATE STACKING EFFECT MODELING AND ANALYSIS

This section embarks on a quantitative approach, instead of a qualitative approach as much previous work did, and gives a refreshed analysis of the gate leakage power, considering the novel stack effect when the often ignored gate leakage components are included. It first analyzes the gate leakage current of single transistors under different bias conditions. Based on this fundamental analysis, novel stacking effect on gate leakage is elaborately addressed. A new best input vector is deduced to reduce the total leakage power and a simple and accurate leakage power estimation methodology is put forward, which can be integrated into a leakage power estimator in the future work. Stacking effect on logically equivalent circuit structures is discussed and optimal structure for leakage reduction is selected.

A. Experimental Settings

In the following experiments, we use *IGCMOD* and *IGBMOD*, which are control parameters in HSPICE, to turn on and turn off the gate leakage measurements. The operating temperature for NMOS and PMOS is set to be 80°C. The threshold voltage is extracted when the device is working in the saturation region. All the parameters are listed in Table I.

TABLE I

PARAMETERS FOR THREE TECHNOLOGY NODES, USED TO SIMULATE THE SUBTHRESHOLD LEAKAGE CURRENT AND GATE LEAKAGE CURRENT

Parameters	90nm	65nm	45nm
V _{dd} (V)	1.2	0.9	0.6
Tox(nm)	1.5	1.0	0.7
Temperature(K)	355	355	355
V _{th} (sat mV)	290/305	170/180	90/100
HSPICE LEVEL	54	54	54
IGCMOD(sub/gate)	0/1	0/1	0/1
IGBMOD(sub/gate)	0/1	0/1	0/1

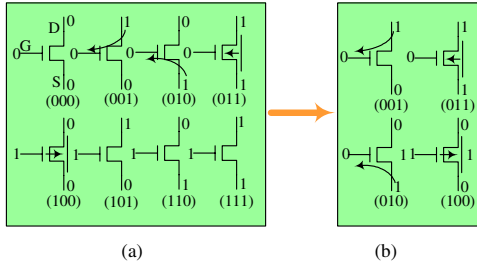


Fig. 2. (a) All eight possible bias conditions for a NMOS transistor in a CMOS logic circuit. (b) significant contribution states. S, D, and G represent source, drain and gate respectively.

B. Gate Leakage Current Analysis

Before the analysis of stacking effects on gate leakage current, we first need to get deep understanding and quantitative feel of the gate leakage current of a single transistor. Until now, state dependence of gate leakage has not been comprehensively studied in a quantitative approach for sub-100nm technologies. This results in either the neglect of some important leakage components, such as the reverse tunneling from the drain/source to the gate, or treating them in a same way. And this means the optimal results obtained previously are not actually optimal. Now let's use an NMOS transistor as an example. We assume that the supply voltage V_{dd} , the threshold voltage V_{th} , and oxide thickness t_{ox} are fixed and depend on the technology used. Furthermore, we will consider the transistor at steady state, i.e., in-between transitions only and not during transient switching.

As we can see from Equation(7), the magnitude of the gate current is strongly dependent on the applied bias V_{gs} . In general, a transistor in a static CMOS logic gate operates in one of the 3 regions of operation, each with a significantly different amount of gate leakage. Case 1 is strong inversion with $V_{gs} = V_{dd}$; case 2 is off state with $V_{gs} = 0$; and case 3 is critical voltage $V_{gs} = V_{th}$. For case 3, a NMOS transistor operating at the threshold will leak significantly less than that in strong inversion, typically 3 to 6 orders of magnitude less [7]. To simplify the analysis, we neglect this case. For case 1 and 2, we consider all eight possible bias conditions seen by a NMOS device at steady state, as shown in Fig. 2(a). For all the eight possible bias conditions (here, we call them states), we group them into four categories. The first category includes states [000] and [111]. For these two states, there is no gate direct tunneling current, i.e., no gate leakage. States [100] and [011] belong to the second category. For state [100], electrons tunnel from the channel, source and drain directly into gate polysilicon. For state [011], the analysis is same as state [100] except that the current direction is reverse. The third category consists of states [001] and [010]. Since the CMOS transistor is a symmetric device, its source and drain can be interchanged without changing its operation, these two states are same. Under such state, there exists gate leakage current caused by electron direct tunneling from gate to source/drain. The fourth category contains states [101] and [110]. These two states belong to transient states. They cannot occur in steady state as the drain/source nodes will eventually reach logic '1' level resulting in

the [111] bias state. To sum up, there are four states, i.e., state [001], [010], [011] and [100], which are significant contributors to gate leakage current compared with other four possible bias states. Based on this conclusion, we only consider these four states, as illustrated in Fig. 2(b), when we calculate the gate leakage current.

Table II lists the gate tunneling current for NMOS and PMOS devices at four bias conditions with technology scaling. It shows that the gate leakage current for NMOS in state [001] is same as that in state [010]. And the gate leakage current for PMOS in state [110] is same as that in state [101]. This is because of the symmetrical property of MOSFET, i.e., the interchangeability of source and drain. Furthermore, the gate leakage current in state [011] for NMOS is roughly equal to the summation of the gate current in state [001] and [010]; for PMOS, the gate current in state [100] is roughly equal to the summation of current in state [110] and [101]. For NMOS, the current under state [100] is higher than that under state [011] because the channel is in strong inversion under state [100] and there are more electrons in the channel which can tunnel into the gate. While for PMOS the current under state [011] is less than that under state [100], this is caused by the fact that the holes can more easily tunnel from gate to channel, source and drain than from channel, source and drain to gate.

Most importantly, we observed that the gate leakage current tunneling from the drain/source to the gate is comparable to and about half of that between the channel and the gate, as illustrated in Table II. This fact is more obvious for 65nm and 45nm technologies and strongly suggests stacking effect should be refreshed. This detailed analysis will help us analyze the stacking effects on gate leakage as shown in the following section. Here, we use I_{ng1} , I_{ng2} , I_{pg1} and I_{pg2} to represent the gate leakage current for NMOS at state [001] and [100], and for PMOS at state [110] and [011], respectively.

C. Stacking Effect

The effect of transistor stacking on circuit topology [14], [24], [25] was first proposed and analyzed for subthreshold leakage current. In serially connected transistors, V_{gs} is more negative when a transistor is closer to the top of the stack. In addition, the threshold voltages for top transistors are increased because of the reverse-biased body to source voltage. As a result, a stack of "OFF" transistors leaks less than a single device in the stack. However, this stacking effect on total leakage current is not exactly true when the device dimension scales down to sub-100nm and gate leakage current is becoming dominant. In sub-100nm era, detailed analysis of the novel stacking effect on gate current is necessary for assignments of optimal input vector and circuit structure to reduce the total leakage power.

C.1 Analysis of Stacking Effect

Here, we use a 3-input NAND gate and a 3-input NOR gate as typical examples to analyze this novel stacking effect, and develop a simple methodology to determine the total leakage power based on the knowledge of a single transistor, considering the internal voltage affected by the subthreshold leakage current. For a 3-input NAND gate as shown in Fig. 3, it has 8 possible input vectors or patterns. For input vector [000], the subthreshold leakage current is determined by the 3 serial NMOS transistors, and it is equal to $1/15I_{nsub}$ according to [13] (here, we use I_{nsub} to represent the subthreshold leakage of an NMOS biased at off state); the gate leakage current is contributed by three PMOS transistors which are working under [011] state, and one NMOS transistor under [010] state. For input vector [001], the subthreshold leakage current is determined by 2 serial NMOS transistors m1 and m2, and it is equal to $1/8.4I_{nsub}$ according to [13]; the gate leakage current has four sources, i.e., m4 and m5 under [011] state, m1 under [010] state and m3 under [100] state. The subthreshold leakage current will increase the voltage at the internal nodes 1 and 2. The voltages at node 1 and 2 will trigger gate leakage at m2 transistor. In our experiment,

TABLE II

GATE LEAKAGE CURRENT FOR NMOS AND PMOS AT FOUR BIAS CONDITIONS WITH TECHNOLOGY SCALING. G, D, AND S REPRESENT GATE, DRAIN AND SOURCE. THE UNIT FOR GATE CURRENT IS NA

Device	Bias(GDS) and Control Conditions	90nm(nA)	65nm(nA)	45nm(nA)
NMOS	001	74.498	578.0	1585.7
	010	74.498	578.0	1585.7
	011	148.99	1155.4	3171.3
	100	171.17	1659.5	4032.0
	IGCMOD=IGBMOD=1			
PMOS	110	7.739	212.2	1240.5
	101	7.739	212.2	1240.5
	100	15.478	424.5	2480.7
	011	12.594	347.2	1991.5
	IGCMOD=IGBMOD=1			

the gate leakage happening at m2 can be compensated by the gate leakage of m3 under full voltage [100]. For input vector **[010]**, the subthreshold leakage current is determined by m1 and m3, and it is same as that under input vector **[001]**. For gate leakage, m2 is biased under on-state, the voltages at node 1 and node 2 are similar and about half of V_{dd} . As a result of the exponential dependence between gate leakage and gate to source/drain voltage, we neglect the gate leakage current happening in m2 and m3 and only count that in m1 under state **[010]** plus gate leakage current from m4 and m6. For input vector **[011]**, m1, m2 and m3 are under states of **[010]**, **[100]** and **[100]** respectively after neglecting the trivial voltage change of internal nodes due to subthreshold leakage current, the calculations of subthreshold leakage and gate leakage are simple and direct. For input vector **[100]**, the subthreshold leakage is same as that under input vector **[001]**. For this case, the voltage at node 2 is about $2/3V_{dd}$, the gate leakage of m2 is calculated to be $1/3I_{ng1}$ according to Equation (7). For input vector **[101]**, the subthreshold leakage is determined by m2 transistor and it is I_{nsub} ; because the bias condition of m2 is similar as that under input vector **[100]**, the gate leakage current is calculated as $(1/3I_{ng1} + I_{ng2})$ for NMOS transistors. For input vector **[110]**, the subthreshold leakage and gate leakage can be directly achieved. For input vector **[111]**, all the NMOS transistors are under on-state, the subthreshold leakage current is determined by the 3 parallel connected PMOS transistors; while the gate leakage current is contributed by all the three serial connected NMOS transistors under **[100]** state. For 3-input NOR gate as shown in Fig. 3(b), the analysis process is same as 3-input NAND gate. We list all the results for NAND and NOR gates in the following descriptions. In order to calculate the leakage power, we multiply these current expressions by the supply voltage which is specified in Table I for different technology nodes.

NAND: total leakage current;

$$\begin{aligned}
 [000] & 1/15I_{nsub} + I_{ng1} + 3I_{pg2}; \\
 [001] & 1/8.4I_{nsub} + I_{ng1} + I_{ng2} + 2I_{pg2}; \\
 [010] & 1/8.4I_{nsub} + I_{ng1} + 2I_{pg2}; \\
 [011] & I_{nsub} + I_{ng1} + 2I_{ng2} + I_{pg2}; \\
 [100] & 1/8.4I_{nsub} + 1/3I_{ng1} + 2I_{pg2}; \\
 [101] & I_{nsub} + 1/3I_{ng1} + I_{ng2} + I_{pg2}; \\
 [110] & I_{nsub} + 1/3I_{ng1} + I_{pg2}; \\
 [111] & I_{psub} + 3I_{ng2}.
 \end{aligned}$$

NOR: total leakage current;

$$\begin{aligned}
 [000] & 3I_{nsub} + 3I_{ng1} + 3I_{pg2}; \\
 [001] & I_{psub} + I_{ng2} + I_{pg1}; \\
 [010] & I_{psub} + I_{ng2} + I_{pg1} + I_{pg2}; \\
 [011] & 1/8.4I_{psub} + 2I_{ng2} + 3I_{pg2}; \\
 [100] & I_{psub} + I_{ng2} + I_{pg1} + 2I_{pg2}; \\
 [101] & 1/8.4I_{psub} + 2I_{ng2} + 3I_{pg1}; \\
 [110] & 1/8.4I_{psub} + 2I_{ng2} + 3I_{pg1} + I_{pg2}; \\
 [111] & 1/15I_{psub} + 3I_{ng2} + 3I_{pg1}.
 \end{aligned}$$

The above simple methodology is not only very efficient to calculate the total leakage power of CMOS gates, but also very accurate compared with the HSPICE simulation results. Table III and

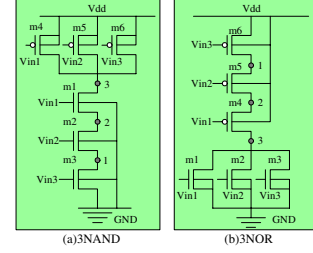


Fig. 3. (a)3-input NAND gate; (b)3-input NOR gate. V_{dd} is the supply voltage; m_i represents different transistor; Vin1, Vin2 and Vin3 are gate inputs, 1, 2 and 3 are voltage nodes.

TABLE V

VALIDATION OF OUR SIMPLE METHODOLOGY VS. HSPICE SIMULATION RESULTS FOR THREE TECHNOLOGY NODES. IN THE TABLE, M.ERROR REPRESENTS MEAN ERROR, ST.DEV. MEANS STANDARD DEVIATION AND W.STATE MEANS WORST PREDICTION STATE.

		90nm(%)	65nm(%)	45nm(%)
3NAND	M.Error	3.33	-4.18	-2.02
	St.Dev.	11.3	6.43	11.7
	W.State	[110]	[100]	[011]
3NOR	M.Error	3.94	7.16	10.3
	St.Dev.	3.21	6.45	13.1
	W.State	[011]	[011]	[011]

Table IV list the gate leakage power, subthreshold leakage power and the total leakage power for 3-input NAND and 3-input NOR gates under 90nm, 65nm and 45nm technology nodes, and these results were obtained from our model. Table V shows mean error, standard deviation and worst prediction states of our simple methodology compared with the HSPICE simulation results. As shown in Table V, the model underestimates the total leakage power for 3NAND gate with 65nm and 45nm gate length; while it overestimates the leakage power for 3NOR gate. This is because the model does not provide very good estimation when the determination of voltage at the internal nodes is complicated by the subthreshold current, for example, the **[110]** state for 3NAND and the **[011]** state for 3NOR. In those cases, we can modulate the model by multiplying a coefficient to adjust the calculated leakage power to almost same as the simulated results. Under other input cases, the prediction based on our model is very accurate. The mean error for three future technologies is about 3.1%.

As our model and experimental results show, the stacking effect on total leakage power, specifically on gate leakage power is quite different from that on subthreshold leakage power. Two new phenomenon are observed based on the above detailed analysis, as shown in the following sections.

C.2 New Input Pattern Dependence

For CMOS gates, many researches have made evident the influence of input pattern on circuit subthreshold leakage behavior, which is a consequence of the stacking effect [14], [24], [25]. The goal can be expressed as finding the input pattern that maximizes the number

TABLE III

GATE LEAKAGE POWER, SUBTHRESHOLD LEAKAGE POWER AND THE TOTAL LEAKAGE POWER FOR 3-INPUT NAND GATE UNDER 90NM, 65NM AND 45NM TECHNOLOGY NODES. THE UNIT IS nWATT.

leakage current states	90nm(nWatt)			65nm(nWatt)			45nm(nWatt)		
	Gate	Sub	Total	Gate	Sub	Total	Gate	Sub	Total
000	134.88	0.97	135.84	1458.36	12.31	1470.69	4544.17	36.17	4580.34
001	324.86	1.78	326.64	2617.20	19.70	2636.91	5607.95	54.73	5662.68
010	127.18	1.78	128.95	1260.09	19.78	1279.89	3654.22	55.94	3710.16
011	511.57	14.69	526.26	3555.90	51.92	3607.83	5665.80	117.54	5783.34
100	56.83	1.72	58.55	885.17	19.60	904.77	2935.23	56.43	2991.66
101	244.87	10.36	255.23	2032.56	49.96	2082.51	3955.68	122.82	4078.50
110	36.31	9.86	46.18	534.83	49.90	584.73	1633.32	130.80	1764.12
111	642.02	29.51	671.53	4867.47	115.08	4982.58	8370.60	348.60	8719.20
Best State	110	000	110	110	000	110	110	000	110

TABLE IV

GATE LEAKAGE POWER, SUBTHRESHOLD LEAKAGE POWER AND THE TOTAL LEAKAGE POWER FOR 3-INPUT NOR GATE UNDER 90NM, 65NM AND 45NM TECHNOLOGY NODES. THE UNIT IS nWATT.

leakage current states	90nm(nWatt)			65nm(nWatt)			45nm(nWatt)		
	Gate	Sub	Total	Gate	Sub	Total	Gate	Sub	Total
000	313.54	44.26	357.78	2488.36	161.06	2649.42	6099.54	414.06	6513.60
001	208.36	5.00	213.36	1597.51	35.18	1632.69	2755.98	104.10	2860.10
010	223.92	5.47	229.39	1919.75	35.60	1955.34	3941.64	99.96	4041.60
011	414.83	0.40	415.24	3108.20	11.48	3119.67	5262.05	36.85	5298.90
100	244.86	9.80	254.66	2291.19	37.47	2328.66	5093.76	99.72	5193.50
101	421.66	0.42	422.08	3224.93	11.56	3236.49	5832.23	36.97	5869.20
110	435.22	0.42	435.64	3489.28	11.54	3500.82	6725.07	36.33	6761.40
111	625.52	0.20	625.72	4672.43	6.85	4679.28	8011.92	22.08	8034.00
Best State	001	111	001	001	111	001	001	111	001

of disabled transistors in all stacks across the unit. Once this vector is found, the input vector can be switched to this minimum leakage input when the units are idle for a period of time. Some previous efforts were done just for the purpose of reducing the subthreshold leakage power when the gate leakage is not significant. With the feature size scaling down to sub-100nm dimension, the gate leakage is exponentially increasing along with the oxide thickness decreasing and becomes the dominant factor of the total leakage power, as illustrated in the above section. Recently, a few papers [22], [23] calculated the dependence between input pattern and gate leakage, but neglecting some important leakage components, which results in these input vectors not being the least leakage state for sub-100nm technologies. New best input vectors should be found to reduce the gate leakage power including the ignored reverse tunneling current from drain/source to gate.

From Table III and Table IV we can see, novel stacking effect on sub-100nm dimension devices reveals a new input dependence for leakage power, i.e., there is a big change in the best input vector control when the gate leakage power is the dominant component in the total leakage. For 3NAND gate under all three technology nodes, the best input vector to minimize the gate leakage power is state [110], and [000] for minimization of the subthreshold leakage power. The final optimal input vector for minimizing the total leakage power is [110] state. However for prior-100nm technology nodes where subthreshold leakage power is the dominant factor, the final optimal state is [000]. And we observed the similar effect of gate leakage power on the optimal input vector for 3NOR gate. Using the above simple methodology, we list all the optimal input vectors and the minimum leakage power for some typical CMOS gates in Table VI. All the results are validated by HSPICE simulations. Most importantly, we compared our results with that obtained by D.Lee *et.al.* in Table VI. In the work of D.Lee *et.al.*, they did consider the gate leakage power, but ignoring the very important leakage component for sub-100nm nodes, i.e., the reverse tunneling current from drain/source to gate. Our model and experiments show that including it will result different best input vector.

C.3 Optimal Gate Structure

The structure of CMOS gates definitely has strong effects on the total leakage power consumption [6]. But when the gate leakage power dominates the total leakage power for a CMOS circuit, novel stacking

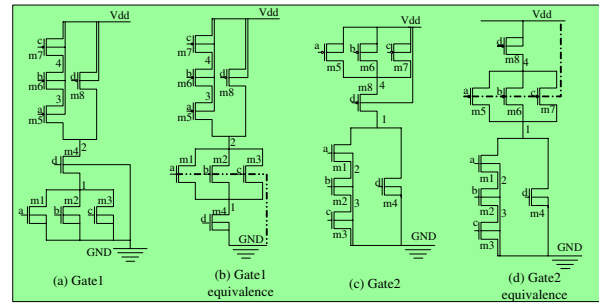


Fig. 4. Two logically equivalent circuits with different pull-down structure(a)Gate1 (b)Gate1 equivalence; And two logically equivalent circuits with different pull-up structures (c)Gate2 (d)Gate2 equivalence

effect makes this structure dependence more complex. Guindi and Najm [6] studied this structure dependence in a qualitative approach, which neglects the interactions between subthreshold leakage and gate leakage and cannot tell the detailed information for best input vector. Fig. 4 shows two logically equivalent circuits with different pull-down structures and two logically equivalent circuits with different pull-up structures. The difference between gate1 Fig. 4(a) and gate1 equivalence Fig. 4(b) is the location of m4 transistor; while the location of m8 transistor distinguishes the gate2 Fig. 4(c) from gate2 equivalence Fig. 4(d). Based on our macromodeling methodology, Fig. 5 shows the minimal, maximal and average total leakage power for gate1, gate1 equivalence, gate2 and gate2 equivalence. All the results calculated from our model and validated by HSPICE under three technology nodes reveal that the average leakage power of equivalent gates is less than that of the original gates. For gate1 equivalence, its minimal leakage power is less than that of gate1, while its maximal leakage power is same as gate1, and finally from the view of average leakage power, it is superior than gate1. For gate2 equivalence, its minimal leakage power is same as gate2, its maximal leakage power is more than that of gate2, however its average leakage power is less than that of gate2.

For gate1, the subthreshold leakage current is determined by three parallel connected transistors m1, m2 and m3; while it is determined by m4 for gate1 equivalence. This fact illustrates that there are more

TABLE VI
RESULTS OF THE BEST STATE FOR EACH CMOS TYPICAL GATE UNDER THREE TECHNOLOGY NODES.

Typical gates	Gate leakage (our work) vs (D.Lee)	Best state Sub leakage	Total leakage (our work) vs (D.Lee)	Min. Leakage(nWatt)	
				90nm	45nm
3NAND	[110]/[000]	[000]	[110]/[000]	46.18/135.84	1764.12/4580.34
3NOR	[001]/[111]	[111]	[001]/[111]	213.36/625.72	2860.10/8034.01
3AND	[110]/[000]	[000]	[110]/[000]	270.70/684.31	5043.90/9612.92
3OR	[001]/[111]	[111]	[001]/[111]	332.62/751.68	5144.42/9813.63
Inverter	[0]/[1]	[1]	[0]/[1]	119.26/224.52	2284.32/3279.78

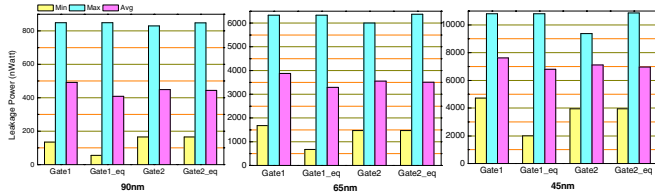


Fig. 5. Minimal, maximal and average total leakage power for gate1, gate2, gate1 equivalence and gate2 equivalence under three technology nodes.

subthreshold leakage paths for gate1, which mean more subthreshold leakage. For gate leakage current, gate1 has three sources which are permanently connected to ground, this definitely means more possibility of electron tunneling from gate to source than gate1 equivalence which only has one source permanently connected to ground. As a joint effect of subthreshold leakage and gate leakage, gate1 is more leaky than its equivalence. What's more, when biased in their best input states, the total leakage power of gate1 equivalence is about one half of gate1's leakage power. In summary, gate1 equivalence is superior than gate1 not only from the view of average total leakage power but also from the view of best input state leakage power. The above analysis is also applicable to gate2 and gate2 equivalence except that the leakage power under best input state of these two gates is same.

IV. CONCLUSIONS

This paper embarks on a comprehensive quantitative approach to leakage power analysis. For the first time, we comprehensively analyze the novel stacking effects on gate leakage current, including the long-time ignored reverse tunneling current from drain/source to gate, which has strong effects on the stacking analysis in sub-100nm nodes. Based on the novel stacking effect analysis, a new best input vector, which is totally different from previous works, is deduced to reduce the total leakage power; an efficient and accurate leakage power estimation macromodeling methodology is put forward, which is tested to generate mean error as small as 3.1% compared with HPSICE results and can be integrated into a leakage power estimator in the future work; novel stacking effect on logically equivalent circuit structures is discussed and optimal structure for leakage reduction is selected for circuit synthesis. We believe that this detailed analysis will facilitate leakage reduction techniques

REFERENCES

- [1] "International technology roadmap for semiconductors 2003 edition executive summary." [Online]. Available: <http://public.itrs.net>
- [2] Y. Cao and H. Yasuura, "Reducing Dynamic Power and Leakage Power for Embedded Systems," in *15th Annual IEEE International ASIC/SOC Conference*, Sept. 2002, pp. 25–28.
- [3] S. Yang, W. Wolf, and N. Vijaykrishnan, "Search Speed and Power Driven Integrated Software and Hardware Optimizations for Motion Estimation Algorithms," in *Proc. Int. Conf. Multimedia and Expo*, June 2004.
- [4] N. S. Kim, et al., "Leakage Current: Moore's Law Meets Static Power," *IEEE Computer*, vol. 36, no. 12, pp. 68–75, Dec. 2003.
- [5] F. Hamzaoglu and M. R. Stan, "Circuit-level Techniques to Control Gate Leakage for sub-100nm CMOS," in *Proc. Int. Symp. Low Power Electronics and Design*, Aug. 2003, pp. 60–63.

- [6] R. S. Guindi and F. N. Najm, "Design Techniques for Gate-Leakage Reduction in CMOS Circuits," in *Proceedings of Fourth International Symposium on Quality Electronic Design*, Mar. 2003, pp. 24–26.
- [7] K. M. Cao, et al., "BSIM4 Gate Leakage Model Including Source-Drain Partition," in *International Electron Devices Meeting, Technical Digest*, Dec. 2000, pp. 815–818.
- [8] B. Govoreanu, P. Blomme, K. Henson, J. V. Houtdt, and K. D. Meyer, "An Investigation of the Electron Tunneling Leakage Current through Ultrathin Oxides/High-k Gate Stacks at Inversion Conditions," in *Proc. Int. Conf. Simulation of Semiconductor Processes and Devices*, Sept. 2003, pp. 287–290.
- [9] I. Polishchuk, P. Ranade, T. J. King, and C. Hu, "Dual work function metal gate cmos transistors by ni-ti interdiffusion," *IEEE Electron Device Letters*, vol. 23, no. 4, pp. 200–202, Apr. 2002.
- [10] R. Kaushik, M. Saibal, and M. M. Hamid, "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuit," *Proc. IEEE*, vol. 91, no. 2, pp. 305–327, Feb. 2003.
- [11] S. Mukhopadhyay and K. Roy, "Accurate Modeling of Transistor Stacks to Effectively Reduce Total Standby Leakage in Nano-Scale CMOS Circuits," in *Proc. Symp. VLSI Circuit*, June 2003, pp. 53–56.
- [12] Y. K. Choi, et al., "Ultra-thin Body SOI MOSFET for Deep-sub-tenth Micron Era," in *International Electron Devices Meeting, Technical Digest*, Dec. 1999, pp. 919–921.
- [13] R. X. Gu and M. I. Elmasry, "Power Distribution Analysis and Optimization of Deep Submicron CMOS Digital Circuits," *IEEE J. Solid-State Circuits*, vol. 31, no. 5, pp. 707–713, May 1996.
- [14] Z. Chen, M. Johnson, L. Wei, and K. Roy, "Estimation of Standby Leakage Power in CMOS Circuits Considering Accurate Modeling of Transistor Stacks," in *Proc. Int. Symp. Low Power Electronics and Design*, Aug. 1998, pp. 239–244.
- [15] W. T. Shiu, "Leakage Power Estimation and Minimization in VLSI Circuits," in *Proc. IEEE Int. Symp. Circuits and Systems*, May 2001, pp. 178–181.
- [16] L. Wei, Z. Chen, M. Johnson, K. Roy, and V. De, "Design and Optimization of Low Voltage High Performance Dual Threshold CMOS Circuits," in *Proc. Design Automation Conf.*, June 1998, pp. 489–494.
- [17] F. Assaderaghi, D. Sinitzky, S. A. Bokor, and C. Hu, "Dynamic threshold-voltage mosfet(dtmos) for ultra-low voltage vlsi," *IEEE Trans. Electron Devices*, vol. 44, no. 3, pp. 414–422, Mar. 1997.
- [18] Y. F. Tsai, D. Duarte, N. Vijaykrishnan, and M. J. Irwin, "Implications of Technology Scaling on Leakage Reduction Techniques," in *Proc. Design Automation Conf.*, June 2003, pp. 187–190.
- [19] D. Duarte, Y. F. Tsai, N. Vijaykrishnan, and M. J. Irwin, "Evaluating Run-Time Techniques for Leakage Power Reduction," in *Proc. Asia and South Pacific Design Automation Conf. and Int. Conf. VLSI Design*, Jan. 2002, pp. 31–38.
- [20] S. Mukhopadhyay and K. Roy, "Modeling and Estimation of Total Leakage Current in Nano-Scaled CMOS Devices Considering the Effect of Parameter Variation," in *Proc. Int. Symp. Low Power Electronics and Design*, 2003, pp. 25–27.
- [21] S. Mukhopadhyay and K. Roy, "Accurate Estimation of Total Leakage Current in Scaled CMOS Logic Circuits Based on Compact Current Modeling," in *Proc. Design Automation Conf.*, June 2003, pp. 169–174.
- [22] D. Lee, W. Kwong, D. Blaauw, and D. Sylvester, "Analysis and Minimization Techniques for Total Leakage Considering Gate Oxide Leakage," in *Proc. Design Automation Conf.*, June 2003, pp. 175–180.
- [23] D. Lee, H. Deogun, D. Blaauw, and D. Sylvester, "Simultaneous State, Vt and Tox Assignment for Total Standby Power Minimization," in *Proc. Design & Test Europe Conf.*, Mar. 2004.
- [24] W. Jiang, V. Tiwari, E. D. L. Iglesia, and A. Sinha, "Topological Analysis for Leakage Prediction of Digital Circuits," in *Proc. Asia and South Pacific Design Automation Conf. and Int. Conf. VLSI Design*, Jan. 2002, pp. 39–44.
- [25] Y. Liu and Z. Gao, "Timing Analysis of Transistor Stack for Leakage Power Saving," in *Proc. Int. Conf. Electronics, Circuits and Systems*, Sept. 2002, pp. 41–44.