Circuit-Level Techniques to Control Gate Leakage for sub-100nm CMOS

Fatih Hamzaoglu, Mircea R. Stan
High-Performance Low-Power (HPLP) Lab
ECE Department, University of Virginia
Charlottesville, VA 22904-4743
{fatih,mircea}@virginia.edu

ABSTRACT
Although still negligible for state-of-the-art CMOS, gate leakage will become significant in the future for sub-100nm technologies, due to the scaling of oxide thickness. We propose several circuit techniques to control gate leakage based on the fact that PMOS transistors with SiO_2 gate oxide have an order of magnitude smaller gate leakage than NMOS transistors in the same technology. First, we compare n-type domino with p-type domino circuits in terms of performance, leakage and switching power, and explore the different tradeoffs between performance and power. Second, we compare n-type with p-type gating for MTCMOS to control the leakage during sleep. The proposed circuits are simulated for a predictive 70nm CMOS technology with 10Å gate oxide thickness and 1.2V supply voltage.

Categories and Subject Descriptors
B.7.1 [**Hardware**]: Integrated Circuits – *types and design styles.*

General Terms
Algorithms, Performance, Design, Reliability.

Keywords
Gate leakage, low power, domino circuits, MTCMOS.

1. INTRODUCTION
Device dimensions are scaled down with each technology generation in order to increase the complexity and performance of VLSI ICs. Extrapolating the present scaling trends, the standby power will become a major limit for sub-100nm technologies. The MOS devices will no longer be totally turned-off anymore, which will result in a non-zero "off-current" even for idle circuits. There are several sources for this off-current: (i) sub-threshold leakage current due to very low threshold voltage (V_t), (ii) gate leakage current due to very thin gate oxide (T_{ox}), (iii) band-to-band tunneling leakage current due to heavily-doped halo.

ISLPED '02, August 12-14, 2002, Monterey, CA, USA.

As a result of an exponential dependency on oxide thickness, gate leakage has the potential to become the dominant factor for sub-100nm generations [1],[2]. Process-level techniques to control the gate leakage involve using higher-k gate dielectrics than SiO_2 (k=3.9), such as Si_3N_4 (k=7.8), Zr and Hf oxides and silicates. Transistors with higher-k gate dielectrics can use a thicker oxide for a given technology node and thus reduce gate leakage. Inukai et al. proposed a circuit technique (BGMOS) similar to MTCMOS to cut off the standby current [1]. In this method, low V_t MOS transistors with ultra-thin T_{ox} are used for the circuit, and an NMOS switch with high V_t and thick T_{ox} is used to cut off both sub-threshold and gate leakages during sleep. The switch is driven by a boosted gate voltage in the active mode to decrease the area penalty. The off-current is suppressed with the cost of area and delay increase. The design also needs dual supply voltage and more complicated fabrication process due to the dual T_{ox}.

In this paper, we propose several circuit techniques to control the off-current, using a single V_{dd} and T_{ox}. The techniques are based on the fact that, under inversion bias, gate leakage through SiO_2 for the PMOS transistors is an order of magnitude lower than for the NMOS [3]. The reason for this difference is that electron tunneling from conduction band (ECB) is the dominant component of gate leakage for the NMOS, whereas it is the hole tunneling from valence band (HVB) for the PMOS. As the barrier height for HVB (4.5 eV) is significantly larger than for ECB (3.1 eV), this results in the much lower gate leakage for the PMOS [3].

Our first circuit-level method explores using p-type domino instead of n-type domino in order to control the gate leakage, which will be the dominant part of the leakage for regular V_t, sub-100nm circuits. The second circuit-level method explores the use of PMOS instead of NMOS as high V_t gating transistors in MTCMOS, in order to control the gate leakage of the transistors that cut off the total leakage of the low V_t logic circuit.

The rest of the paper is organized as follows: Section 2 presents an analysis of the gate current (I_g) for NMOS and PMOS transistors. Section 3 explores the first circuit-level method for domino circuits and Section 4 explores the second circuit-level method for MTCMOS. Finally, Section 5 concludes the paper.

2. NMOS/PMOS GATE CURRENT
The simulations in this paper use BSIM4 [4] device models, which explicitly account for gate-current effects. Unfortunately very few simulators from the major EDA companies support BSIM4 yet, furthermore most available device model cards are only available for BSIM3v3, not BSIM4. As circuit simulator we

used AIM-Spice [5], which includes BSIM4 among the supported models. For the BSIM4 device models we adapted available Berkeley Predictive Technology (BPTM) [6] BSIM3v3 model cards (for a 70nm technology) to BSIM4, by modifying and adding several parameters to account for the gate leakage. Gate current parameters have been adjusted to target 100 A/cm^2 (70 nA/μm) gate leakage for NMOS and 10 A/cm^2 (7 nA/μm) for PMOS in 70nm technology at 10Å oxide thickness and 1.2V supply voltage as predicted from device measurements [3],[7].

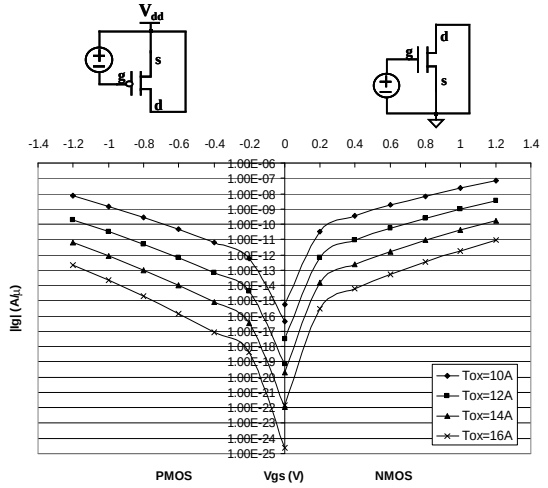

Figure 1. Dependence of gate current on Vdd and Tox for NMOS and PMOS transistors (at Vds= 0V).

Figure 1 shows IV-curve simulation results for the T_{ox} and V_{dd} dependencies of gate current for both NMOS and PMOS transistors. Gate current increases by an order of magnitude for each 2Å decrease in T_{ox}. Gate current also increases by an order of magnitude for each 0.3V increase in V_{dd}. The difference between NMOS and PMOS transistor gate currents, which can be observed in all the figures, is used in the next sections to investigate ways of decreasing the leakage power through circuit techniques.

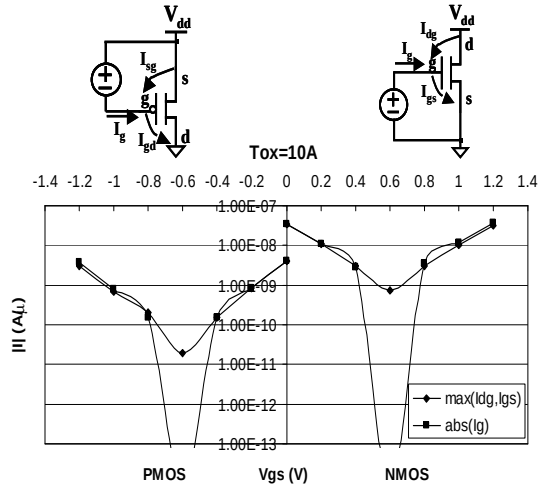

Figure 2. Dependence of gate current (Ig) and gate leakage (max(Idg,Igs)) on Vgs.

Figure 2 shows a subtle difference between gate *current* and gate *leakage*. In the case of NMOS, for example, the gate *current* (I_g) is zero for V_g around $V_{dd}/2$ because the gate-to-drain current I_{dg} and the gate-to-source current I_{gs} cancel each other. The gate *leakage* on the other hand, which can be defined as the actual current that flows from the power supply due to oxide tunneling, is not zero. This gate leakage is equal to the maximum of I_{dg} and I_{gs} and is nonzero even when the gate current is zero. Unfortunately, from the point of view of power consumption, the gate leakage is the one that counts, not the gate current.

3. GATE LEAKAGE CONTROL FOR DOMINO CIRCUITS

In this section we consider transistors with regular V_t and ultra-thin 10Å T_{ox}, which have relatively low sub-threshold leakage but large gate leakage in a 70nm CMOS technology. Since the gate leakage is the dominant part of the total leakage under these conditions, controlling the gate current is the most important task for reducing the total leakage. As the gate leakage for NMOS is larger than for PMOS, any method that decreases the total width of the NMOS in the circuit will also decrease the total leakage.

Domino gates with NMOS networks in the dynamic stage, n-type domino, are widely used for high-speed applications (Figure 3(a)). Since only the zero-to-one transition at the output is critical for n-type domino, the static stage, generally represented by an inverter, is typically *skewed* in the PMOS direction. An alternative to n-type domino can be obtained by using a PMOS network in the dynamic stage and a static stage skewed in the NMOS direction as shown in Figure 3(b). Such a "p-type" domino has the advantage of requiring a smaller total NMOS width than the regular n-type domino. In order to have a fair comparison between n-type and p-type domino we compare gates that have the same topology for the corresponding NMOS and PMOS networks. The justification for doing this is that a fair comparison should not be affected by the fact that the network for one type has a transistor stack and the other one does not. In general an n-type OR gate (no stack) will always be faster than a p-type OR gate (stack), and a p-type AND gate (no stack) will always be faster than an n-type AND gate (stack). The interesting comparison then is between an n-type OR gate and a p-type AND gate (both with no stack), and between an n-type AND gate and a p-type OR gate (both with stack).

There is also the issue of transistor sizing. The typical method for sizing transistors in a high-performance domino circuit is to up-size all transistors on the critical path until the reduction in delay with sizing saturates. The main idea is to counteract the net effect of wire loads on the overall delay, without unnecessarily increasing the total transistor area. The final result of such sizing is that there will be a fixed optimal ratio of wire load to transistor gate capacitance, this ratio being the one that results in the "knee" of the delay to total-transistor-size curve. In other words the optimal transistor size at the input is given once the wire load is known. For this reason a fair comparison needs to use the same transistor sizes in the dynamic part of both n-type and p-type domino. For similar reasons the inverter stage skew, as well as the transistor-strength to dynamic-stage network-strength ratios, for both clock and keeper, need to be also kept the same for both n-type and p-type domino.

Assuming no parasitic capacitances, the p-type and the n-type domino gates could have similar performance for a first-order

analysis. Although the PMOS network in Figure 3(b) has less drive current than the NMOS network in Figure 3(a) due to the lower mobility of PMOS, it is also driving less capacitance since the inverter is skewed in the NMOS direction and is smaller. Unfortunately p-type domino is affected by parasitic capacitances more than n-type domino, which will result in slightly slower performance. However the gate leakage for p-type domino will also be smaller since the total NMOS transistor area is smaller than for n-type domino.

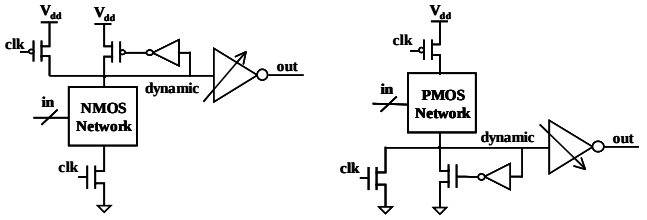

Figure 3. Domino gates. (a) N-type. (b) P-type.

The comparison of n-type and p-type domino gates in terms of delay, standby leakage, energy, and energy-delay product is given in Table 1 for two different dynamic stage circuit topologies; 2-input parallel (OR gate n-type and AND gate p-type domino) and 2-input series (AND gate n-type and OR gate p-type domino). The delay results are for a fanout of four, while the power and the energy results are for a single domino gate. The larger delay of p-type domino can be explained as follows: although the static stage gate capacitance is smaller in this case, the parasitic capacitances at the dynamic node are similar for both cases. Hence, the total capacitance at the dynamic node for p-type domino is not as small as could be expected from a first-order analysis. In other words, the drive current degradation due to the lower mobility of the PMOS network is larger than the corresponding capacitance decrease at the dynamic node. As expected, this delay degradation is more significant for the series topology. The reason is that the inverter size for the optimal domino gate delay in the series case is smaller than for the parallel one, thus the parasitic capacitance at the dynamic node is more significant in this case.

Table 1. P-type vs. n-type domino simulation results

	Delay Increase (%)	Standby Leakage Reduction (%)	Switching Energy Reduction (%)	Energy-Delay Product Reduction (%)
2-input parallel	10.3	48.6	45.0	39.3
2-input series	27.0	61.0	51.3	38.1

Even if the delay is increased, the standby leakage, mainly gate leakage, for the p-type domino circuits is 48.6% to 61% less than for n-type domino circuits as a result of decreasing the total NMOS size. Besides, the switching energy of p-type domino is also 45% to 51.3% less than for n-type domino, because, as explained above, the total transistor size is also decreased. This is due to the previously mentioned method of sizing the dynamic networks at the "knee" of the delay to total size curve, by which the ratio of wire to gate capacitance at he input is kept the same for both n-type and p-type domino.

4. GATE LEAKAGE CONTROL FOR MTCMOS

In the previous section we assumed that only regular V_t transistors are used in the logic. However, usage of low V_t transistors in speed-critical paths of current microprocessors becomes unavoidable in order to decrease the delay further. Low V_t transistors lead to an exponential increase in sub-threshold leakage current with reduced V_t. For applications that are sometimes idle, this large "off" current in sleep mode becomes a significant portion of the total power, and therefore unacceptable.

Multi-threshold CMOS (MTCMOS) [8],[9], shown in Figure 4, has been proposed to reduce the standby current. MTCMOS uses low V_t transistors in the main logic for fast operation in active mode, and high V_t gating transistors for reducing the "off" current in sleep mode. In active mode the high V_t gating transistors are turned on, but the *virtual* ground for the entire circuit becomes slightly higher than zero (NMOS gating), or the *virtual* power line becomes slightly less than V_{dd} (PMOS gating), due to the *IR* drop across the gating transistors. This degrading of the power supply levels leads to an increase in gate delay. The gating transistors need to be sized properly in order to minimize this effect of *IR* drop on circuit speed.

In sleep mode, the gating transistors are turned off and all the node voltages in the circuit become close to V_{dd} or ground, including the *virtual ground (vgnd)* or *virtual supply (vsup)*, in the case of PMOS or NMOS gating, respectively. As a result, the overall leakage is determined by the gating transistor sub-threshold and gate current, with gate leakage being dominant because of high V_t but large gate-to-drain voltage in sleep mode.

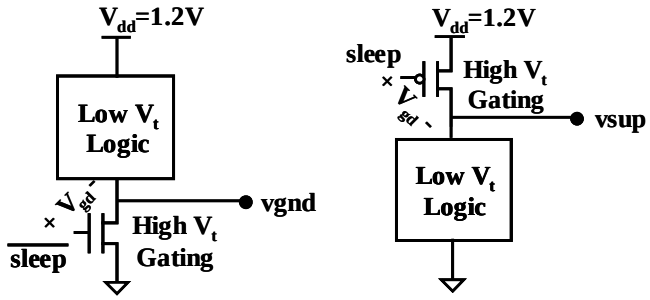

Figure 4. MTCMOS for cutting off the leakage in sleep mode. (a) NMOS gating. (b) PMOS gating.

Although the MTCMOS technique can be applied to any circuit family, it is less applicable to domino circuits. Similar reduction of leakage as with MTCMOS can be obtained without any gating transistors, by appropriately using dual V_t transistors in domino circuits [9]. Because of this we compare NMOS gating with PMOS gating for a more generic circuit, a nine-stage ring oscillator using static CMOS inverters with a fanout of four. In active mode, alternating stage inverters switch in the same direction and there is no overlap between these alternating stage currents; because of that the optimal gating transistor size is independent of the number of stages for such a cicuit. In an n-well process, widely used in current technology, all NMOS transistor bodies are connected to ground through the substrate, whereas the

PMOS transistor bodies can be connected either directly to V_{dd} or to the virtual supply *vsup*.

Table 2. Normalized leakage, area penalty, and the gating transistor gate-to-drain voltage in sleep mode, compared to the original case without gating

	Normalized Leakage	Area Penalty (%)	vsup (V)	vgnd (V)	|Vgd| (V)
Original (No Gating)	1	0	NA	NA	NA
NMOS Gating	0.027	4.6	NA	1.063	1.063
PMOS Gating (pbody=vsup)	0.016	11.1	0.009	NA	1.191
PMOS Gating (pbody=V_{dd})	0.010	19	0.139	NA	1.061

Simulation results for three gating scenarios are given in Table 2 and Table 3. Gating transistors are sized in all cases to tolerate a maximum of 5% delay increase in active mode. The standby leakage and area penalties in Table 2 are normalized with respect to the case without gating whose standby leakage is assumed to be 1. The leakage is reduced from 1 to 0.027 for NMOS gating with a 4.6% area overhead, the remaining leakage being due to the gating transistor gate leakage because of its large (1.063V) gate-to-drain voltage in sleep mode. Since the PMOS has smaller gate leakage than NMOS, using PMOS gating instead of NMOS gating, and connecting inverter PMOS transistor bodies to *vsup* in order to eliminate the body effect, decreases the leakage further down to 0.016 with a 11.1% area overhead. Although the leakage is further reduced from 0.027 to 0.016, the reduction is not as large as expected because of two reasons: first, the PMOS gating transistor needs to be larger than the NMOS for the same 5% delay; second, the gate-to-drain voltage of PMOS gating transistor (1.191V) is larger than for the NMOS gating transistor (1.063V). The third scenario connects inverter PMOS transistor bodies to V_{dd} for the case of PMOS gating. This has a larger area overhead (19%) due to the body effect, but surprisingly has smaller leakage (0.01) as a result of a smaller gating-transistor gate-to-drain voltage (1.061V instead of 1.191V). The contribution of body effect on gating-transistor gate-to-drain voltage is due to the fact that, in the case of PMOS gating, the inverter PMOS will need a larger gate-to-source voltage to overcome the reverse body effect in order to be able to sink the gating transistor leakage current. This results in a larger voltage at node *vsup*, which leads to a smaller gate-to-drain voltage for the gating transistor.

Table 3. Standby leakage reduction and the area penalty of PMOS gating compared to NMOS gating

	Leakage Reduction (%)	Area Penalty (%)
PMOS Gating (pbody=vsup)	41	6.2
PMOS Gating (pbody=V_{dd})	60	13.3

Table 3 shows a comparison of both PMOS gating scenarios with the NMOS gating scenario, in terms of standby leakage reduction and area overhead. It is shown that using PMOS gating with the PMOS bodies connected to V_{dd} reduces the standby leakage by 60% with a 13.3% area overhead, for the same 5% delay penalty, compared to NMOS gating.

It is also worthwhile to mention that the area overheads of these three cases decreases as the number of stages increases, since the gating transistor size is independent of the number of stages in the ring oscillator.

5. CONCLUSION

In this paper we have presented several circuit techniques to control the gate leakage based on the fact that the PMOS transistors with SiO_2 gate oxide have much smaller gate leakage than NMOS transistors in the same technology. First, we analyzed NMOS and PMOS gate leakage and also observed an interesting difference between gate current and gate leakage. Next, we compared p-type domino circuits with n-type domino circuits, the results showing that although there is a 10.3% performance penalty, the standby leakage and the energy-delay product of a 2-input parallel domino gate can be reduced by 49% and 39%, respectively. Similarly, the standby leakage and the energy-delay product of a 2-input series domino circuit can be reduced by 61% and 38%, respectively, with a slightly larger 27% penalty in performance. Finally, we compared p-type gating with n-type gating for MTCMOS to reduce the leakage during sleep mode. The results show that using p-type gating reduces the leakage by 60% with a 13% area overhead compared to n-type gating.

6. REFERENCES

[1] T. Inukai et al., "Boosted Gate MOS (BGMOS): device/circuit cooperation scheme to achieve leakage-free giga-scale integration," Proc. CICC, pp. 409-412, May 2000.

[2] T. Ghani et al., "Scaling challenges and device design requirements for high performance sub-50 nm gate length planar CMOS transistors," Proc. Symp. VLSI Tech., Dig. Tech. Papers, pp. 174-175, June 2000.

[3] Y.C. Yeo et al., "Direct tunneling gate leakage current in transistors with ultrathin silicon nitride gate dielectric," IEEE Electron Device Letters, vol. 21, no. 11, pp. 540-542, Nov. 2000.

[4] K.M. Cao et al., "BSIM4 gate leakage model including source-drain partition," IEEE IEDM Tech. Dig., pp. 815-818, San Francisco, CA, Dec. 2000.

[5] http://www.acm.org/sigs/pubs/proceed/template.html

[6] http://www-device.eecs.berkeley.edu/~ptm/mosfet.html

[7] "The International Technology Roadmap for Semiconductors-Process Integration, Devices, & Structures", 1999.

[8] M.R. Stan, "Low threshold CMOS circuits with low standby current," in Proc. ISLPED, pp. 97-99, 1998.

[9] J.T. Kao, and A.P. Chandrakasan, "Dual-threshold voltage techniques for low-power digital circuits," IEEE JSSC, vol. 35, no. 7, pp. 1009-1018, July 2000.