# Design of Superbuffers in sub-100nm CMOS Technologies with Significant Gate Leakage


Ali Bastani
Columbia University
Department of Electrical Engineering
500W 120th St., New York, NY 10027
ab2001@columbia.edu

Charles A. Zukowski
Columbia University
Department of Electrical Engineering
500W 120th St., New York, NY 10027
caz@columbia.edu



## ABSTRACT

In this paper, we first study the behavior of gate leakage current in a simple inverter. We make some important observations about the gate leakage in both PMOS and NMOS devices. We then study the trade off between power and delay in a standard inverter within a particular buffer chain when reducing the size of the pull-down device. In the last part of this paper, we present an alternate leakage suppressed inverter for driving large loads. We finally compare the performance of our proposed circuit with those of standard and scaled-down inverters and show that we can significantly reduce the standby power in certain situations with a modest reduction in speed.



## Categories and Subject Descriptors

B.7.1 [**Hardware**]: Integrated Circuits– *VLSI (very large scale integration)*

## General Terms

Design

## Keywords

Low power design, Gate leakage reduction, Superbuffers


## 1. INTRODUCTION

One of the most challenging aspects of today's CMOS VLSI circuits is standby power dissipation. In fact, feature size reduction has made the effects of leakage mechanisms more pronounced than ever. This becomes more complicated in sub-100nm technologies where we are dealing with not only the subthreshold leakage but also gate leakage. In particular, gate leakage is projected to surpass subthreshold leakage at around a 65nm technology [9]. At this technology node, which is the focus of this paper, standby gate leakage current may have a significant impact on total standby power. In certain applications where the standby time is relatively long, e.g. SRAM, gate leakage might be the dominant factor in total standby power and hence, needs special attention.

For 65nm technology with an ultra-thin 10Å $SiO_2$ gate insulator with dielectric constant of 3.9, gate leakage current can be around 50nA/μm [9]. One way to reduce the gate leakage is using high-k dielectrics. There are several high-k materials proposed so far such as oxinitrides (k=4.1) [5], $Si_3N_4$ (k=7.8), and $HfO_2$ (k=25~50) [6]. However, there are many process integration problems with these materials. Although Intel recently announced the first fully functional SRAM in 65nm technology with an unspecified high-k material [10], it will not be mass produced until 2005. In addition, extra costs of these processes could be significant. As a result, circuit engineers have focused on design methods to alleviate the adverse effects of gate leakage. In [3], the authors propose using p-type domino instead of n-type. They also suggest using p-type sleep transistors for MTCMOS circuits. This is because of the lower leakage current of PMOS devices. Others have investigated the state and structure dependencies of gate leakage. Based on these dependencies, authors in [8] have recommended some guidelines for designing low power circuits. This work has been expanded in [4], where the interaction between subthreshold and gate leakage is investigated in depth, and a pin-reordering technique is proposed to minimize total leakage.

As mentioned earlier, gate leakage becomes particularly important in circuits with a long standby time. The effect of this leakage mechanism grows to be more pronounced in drivers of large loads because of the direct relation between gate leakage current and device gate width. In such cases, the simple solution to the power problem is to scale down the width of the drive transistors, but this can lead to large delays in a component that is often in critical paths. Alternately, the large devices and critical nature of large drivers can justify a certain amount of circuit overhead to address the problem.

In section 2, we first study gate leakage phenomena in a simple inverter and make a few important observations. The section concludes with an analysis of the effect of NMOS gate width on the performance of a simple inverter. We show the trade-off between average standby power and delay that comes from shrinking the width of the pull-down device, and design a reasonable reference scaled-down driver. In section 3, we propose a leakage suppressed inverter which is a type of superbuffer. A simple feedback mechanism turns off the major source of gate leakage during standby time, resulting in a significant standby power reduction. The sizing of the proposed circuit and its effect on delay is studied. Finally, the performance of this circuit for different loads is investigated. It is shown that the leakage suppressed inverter, when used as a load driver, is able to significantly reduce the static power when the circuit has long standby time.


Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.


*GLSVLSI'04*, April 26-28, 2004, Boston, Massachusetts, USA


Copyright 2004 ACM 1-58113-853-9/04/0004...$5.00.

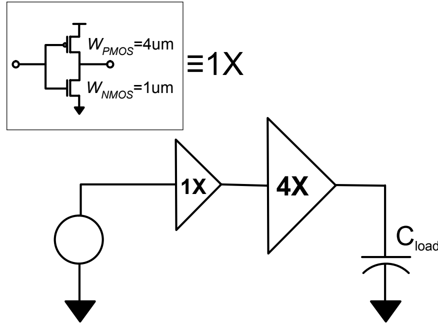



**Figure 1 A simple inverter chain. $C_{load}$ can be any large capacitve load (e.g. a bus or another inverter stage).**

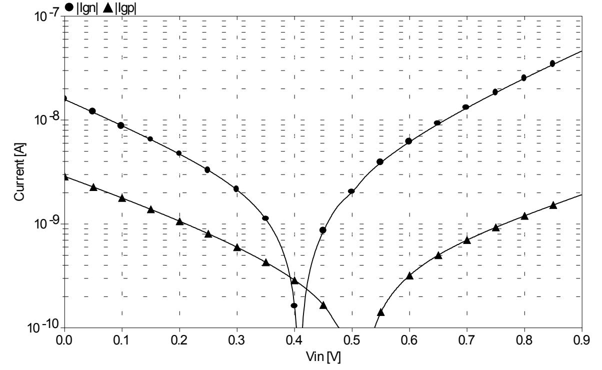



**Figure 2 NMOS ($I_{gn}$) and PMOS ($I_{gp}$) gate currents vs. input voltage for an inverter ($W_{PMOS}$=4×$W_{NMOS}$=16µm, $L_{PMOS}$=$L_{NMOS}$=65nm).**

## 2. GATE-LEAKAGE IN INVERTERS

In this paper, we have used BSIM4.2 [1] device models, which take into account different components of gate tunneling currents. We consider 65nm technology node transistors with ultra-thin 10Å $T_{ox}$, which have considerably large gate leakage. BPTM [1] BSIM4 model cards are used for MOS transistors in 65nm technology. BPTM is a customizable, detailed, predictive SPICE model of CMOS and interconnect technology for the next decade. The circuit simulator used in this research is AIM-Spice [2], which supports BSIM4.2.

At the 65nm technology node, gate leakage current in the NMOS devices surpasses the subthreshold current. This makes the 65nm node an interesting choice for studying the gate leakage. For the 45nm technology node, gate leakage is greater than subthreshold leakage for PMOS devices as well. Simulation results for both technology nodes are shown in Table 1. In this table, $I_{gate}$ and $I_{sub}$ are computed when the device is off and gate leakage is not at its maximum.

We first consider the inverter shown in the top left corner of Figure 1, where we have used minimum length transistors ($L_{PMOS}$=$L_{NMOS}$=65nm). By sweeping the input from 0 to $V_{DD}$, we can make three observations: i) the gate currents of both transistors first decrease to a certain point and then increase gradually (Figure 2). It should be noted that while the total gate current goes to zero at around 0.4V for the NMOS transistor and 0.5V for the PMOS transistor, the actual gate leakage current is not zero. In fact, gate current consists of two components ($I_{GD}$ and $I_{GS}$) which have equal values and different directions at that point, leading to the total gate current of zero ($I_G$=$I_{GD}$+$I_{GS}$) [3]; ii) when $V_{GS}$=$V_{DD}$ , the gate current is much larger than when $V_{GS}$=0 . This suggests we should focus on this region to get the most benefit; and iii) it is clear from Figure 2 that $I_{G,NMOS}$ is the dominant component of the total gate leakage current in an inverter. In fact, $I_{gate}$ for a PMOS device is typically one order of magnitude smaller than an NMOS device with identical $T_{ox}$ and $V_{DD}$ when using $SiO_2$ [9]. This is due to the much higher energy required for hole tunneling in $SiO_2$. However, in alternate dielectric materials the energy required for electron and hole tunneling can be completely different. In the case of $Si_3N_4$, $I_{G,PMOS}$ can actually exceed $I_{G,NMOS}$ for higher nitrogen concentrations [7].

Consider the inverter chain shown in Figure 1, in which the first stage is the same as the one shown in the top left corner. In this circuit, gate leakage of the second stage, which drives a large capacitive load, is dominant because of its sizing. We can reduce the gate leakage current of this stage by reducing the pull-down gate width. This will not change the switching power much since that is mainly determined by the load. However, reducing the pull-down gate width has several important effects on the final driver stage as shown in Table 2, assuming a load '*C*' equivalent of a 16X inverter: i) average static power (with a duty cycle of 50%) decreases, partially as a result of gate current reduction; ii) rise time, defined as the time required the output to change from 10% to 90% of its final value, remains almost the same because the input capacitance of the pull-up device is always dominant; iii) fall time, defined as the time required for the output to change from 90% to 10% of its final value, increases dramatically. This is a direct adverse result of NMOS gate width reduction; iv) propagation delay defined as $t_p$=($t_{pLH}$+$t_{pLH}$)/2 where $t_{pLH}$ and $t_{pHL}$ are the time differences between the output and input of the middle-stage reaching 50% of the output final value for low-to-high transition and vice versa, increases as expected; and v) it can be seen that for $W_{NMOS}$=2µm, the average static power is reduced by 16%, compared to a standard inverter, while the delay is increased by a similar ratio. Beyond this point of scaling, static power continues to improve, but at a large cost in speed reduction. Thus, we use $W_{NMOS}$=2µm as reasonable choice for a reference "scaled-down" driver design.

**Table 1 Comparison between $I_{gate}$ and $I_{sub}$ for 65nm and 45nm technologies for $V_{ds}$=$V_{DD}$, $V_{gs,nmos}$=0V, and $V_{gs,pmos}$=$V_{DD}$.**

| Device | [nA/µm] | 65nm ($V_{DD}$=0.9V) | 45nm ($V_{DD}$=0.6V) |
|---|---|---|---|
| NMOS | $I_{sub}$ | 14 | 2.9 |
| | $I_{gate}$ | 16 | 27 |
| PMOS | $I_{sub}$ | 2.5 | 1.4 |
| | $I_{gate}$ | 0.5 | 1.9 |

**Table 2 Simulation results for a scaled-down inverter ($W_{PMOS}$=16µm, $C_{load}$=16X, Duty cycle=50%).**

| $W_{NMOS}$ (µm) | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $P_{static,avg}$ (µW) | 7.9 | 8.2 | 9.0 | 9.8 |
| $t_{rise}$ (psec) | 55 | 56 | 56 | 56 |
| $t_{fall}$ (psec) | 270 | 138 | 96 | 75 |
| $t_{delay}$ (psec) | 113 | 80 | 71 | 68 |

In many applications we would like to occasionally send a short enable pulse to a large load. In memory cells, we need to activate only one row at a time while the rest of the rows stay low. For these applications where the output stays at zero for a long time, standby power due to gate current will be significant. This becomes especially important when we recall that the gate leakage current in a transistor is much larger when the input is HIGH and the output is LOW. Solving this problem is the focus of the next section.

# 3. LEAKAGE SUPPRESSED INVERTER

An alternative to reducing gate width is replacing the second stage with a "leakage suppressed inverter" (Figure 3). The leakage suppressed buffer (LS-inverter) is a type of superbuffer [11]. Since the gate leakage current of the pull-down device is dominant, we would like to turn it off when $V_{in}$='1' and $V_{out}$='0'. In fact, we can split the pull-down transistor into two in parallel and shut down one part (M2) in order to reduce leakage power. We should, however, keep the other part (M4) on so that the output stays low after completion of the transition. The feedback mechanism of our proposed circuit consists of transistor M3 and delay elements INV1 and INV2. More precisely, when the output goes to zero, M3 turns off. This disconnects the gates of M1 and M2, and $V_{gn}$ discharges through the gate of M2. Although $V_{gn}$ cannot discharge completely in a reasonable time and settles close to zero as a result of gate leakage currents (Figure 4), this does not interfere with proper operation. However, if technology parameters lead to significant subthreshold conduction in M2, a minimum sized transistor M5 can be added to discharge $V_{gn}$ completely without significantly impacting delay. Finally, M3 has a critical effect on delay. It should be wide enough to supply the current M2 needs to turn on as fast as possible. Through simulation, one can find that $W_{M3}$=1µm is a reasonable choice for the 65nm 4X driver. An LS-inverter can also benefit from a dual-$V_{th}$ capability as a low-$V_{th}$ M3 improves its speed. In fact, our example 65nm technology has dual-$V_{th}$ devices ($V_{th,high}$=0.3V, $V_{th,low}$=0.17V), and low-$V_{th}$ is used for M3.

If we replace the second stage of the inverter chain (Figure 1) by the LS-inverter, we can reduce the static power. Since M2 is ON for a short period of time (Figure 4), M2 should be wide enough to make the high-to-low transition as fast as possible. We assume for comparison that $W_{M2}$+$W_{M4}$=4µm, matching the drive of the original standard inverter. $W_{M4}$ does not need to be large because it is only needed to keep the output low after M2 turns off, so we keep M4 minimum size ($W_{M2}$=3.9µm, and $W_{M4}$=0.1µm).

The data in Figure 5 covers a wide range of ratios between the final driver and the load ($C_{load}$ has been substituted by inverters equivalent to 16X, 32X, 48X, and 64X for a fixed driver chain of 1X-4X, as shown in Figure 1). Dynamic power is mostly determined by the size of the load, and hence does not depend on the driver. Since the final stage of a driver chain (4X in this case) dominates its power consumption, this data should be able to roughly scale for larger loads with the same final-stage ratio. The delay data roughly reflects the contribution of only the last stage in a longer chain with similar final stage ratio.

Figure 5(a) shows the growth of delay with load size. The delay of the LS-inverter is slightly higher than of the standard inverter, but not as large as the scaled-down inverter. Figure 5(b) compares the rise and fall times. Since the rise time depends on the pull-up device, the standard and the scaled-down inverters have the same values for different loads, while the rise time values of the LS-inverter are slightly higher due to the overhead. However, the LS-inverter closely follows the standard inverter in terms of fall time because both enjoy the same drive. The fall time values of a scaled-down inverter are considerably higher as a result of pull-down device shrinkage. Figure 5(c) compares the average static power of the circuits when the output of the load-driver is HIGH. All three follow the same path because this quantity mainly depends on the pull-up device in the last stage. Figure 5(d) shows the results for the average static power when the output of the load-driver is LOW. The LS-inverter has a smaller value compared to the other two circuits. This shows that our leakage suppression technique, for the case that the output is low, is very effective. It also shows that the overhead in the LS-inverter does not have much effect on the overall performance of the circuit. The shift reflects differences in driver power. For a 0.5pF capacitive load, the output low static power in the driver is reduced by 27 % at the expense of 7% increase in delay.

Finally, it should be noted that we could use a combination of the LS-inverter and scaled-down inverter to further reduce the static power when the output stays low for a considerable amount of time. The penalty, however, would be an increase in delay. The LS-inverter approach can expand the options available for driver designs to trade between power, speed, and area.

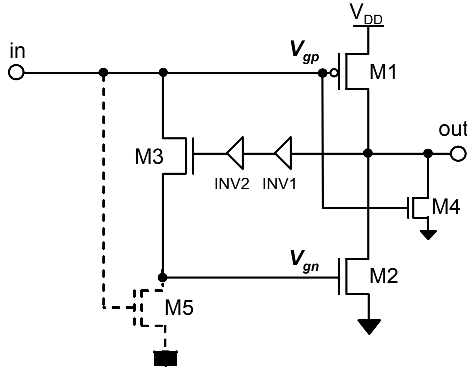



**Figure 3 Schematic of a leakage suppressed inverter. The load-driver (4X) in Figure 1 can be replaced by this circuit ($W_{M2}$+$W_{M4}$=4µm).**

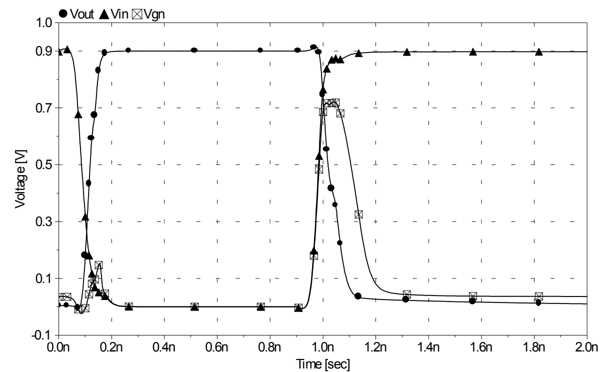
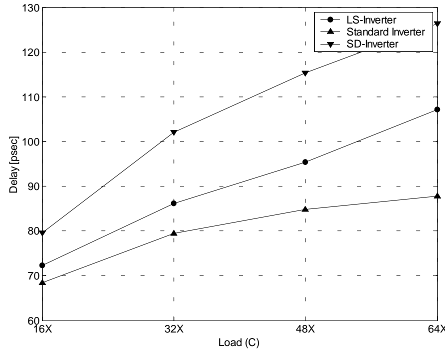
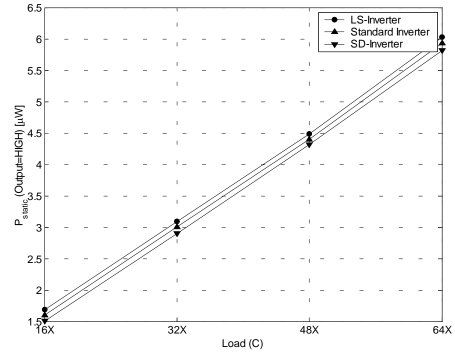


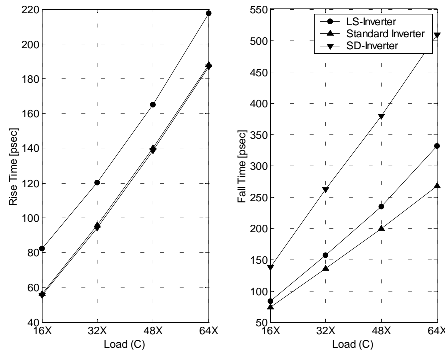

**Figure 4 Timing analysis of an LS-inverter ($W_{M1}$=16µm, $W_{M2}$=4µm, $W_{M3}$=1µm, $W_{M4}$= $W_{M5}$=0.1µm, INV1 and INV2 are minimum sized.).**



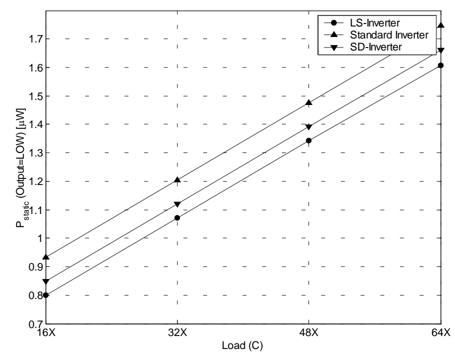

**(a) Delay vs. load**




**(c) Avg. static power (output=HIGH) vs. load**




**(b) Rise/fall time vs. load**




**(d) Avg. static power (output=LOW) vs. load**

**Figure 5 (a) Delay, (b) rise/fall time, avg. static power when the output is (c) HIGH and (d) when it is LOW of an LS-inverter ($W_{PMOS}$=16µm, $W_{M2}$=3.9µm, $W_{M3}$=1µm, $W_{M4}$=0.1µm) compared to a standard inverter ($W_{PMOS}$=16µm, $W_{NMOS}$=4µm) and a scaled-down inverter ($W_{PMOS}$=16µm, $W_{NMOS}$=2µm) vs. $C_{load}$ substituted by inverters equivalent to 16X, 32X, 48X, and 64X. Note that the static power is calculated for both driver and load inverter. The shift in (d) reflects differences in driver power.**

# 4. CONCLUSIONS

In this paper, we presented a leakage suppressed superbuffer for driving large loads in technologies with significant gate leakage currents. We showed that the circuit, which is especially effective when the output stays low for a long time, is able to achieve significant static power savings at the expense of a reasonable increase in propagation delay. We presented data showing the performance of the leakage suppressed driver and driver scaling over a range of driver/load ratios. The best design choice ultimately depends on the performance cost function, circuit context, and technology; but the leakage suppressed approach may become more useful as technologies scale.

# 5. ACKNOWLEDGEMENTS

The authors would like to give thanks for the support of the Microelectronic Design Center (MDC) of the New York State Office of Science, Technology, and Academic Research (NYSTAR).

# 6. REFERENCES

[1] http://www-devices.eecs.berkeley.edu/~ptm

[2] http://www.aimspice.com/

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

[4] D. Lee, W. Kwong, D. Blaauw, and D. Sylvester, "Analysis and Minimization Techniques for Total Leakage Considering Gate Oxide Leakage," Proc. DAC, 2003.

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

[6] I. Polishchuk, Gate Stacks for sub-50nm CMOS Devices: Materials, Engineering, and Modeling, Ph.D. Dissertation, UC Berkeley, 2002.

[7] 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.

[8] R. S. Guindi, F. N. Najm, "Design Techniques for Gate-Leakage Reduction in CMOS Circuits," Fourth International Symposium on Quality Electronic Design, pp.61-65, 2003.

[9] T. Inukai, et al., "Boosted- gate MOS (BGMOS): Device/circuit cooperation scheme to achieve leakage-free giga-scale integration," Custom Integrated Circuits Conf., pp.409-412, 2000.

[10] http://www.intel.com/pressroom/archive/releases/20031124tech.htm

[11] Carver Mead, Lynn Conway, *Introduction to VLSI Systems*, Reading, MA: Addison-Wesley, c1980.