Design Techniques for Gate-Leakage Reduction in CMOS Circuits

Rafik S. Guindi and Farid N. Najm
Department of Electrical and Computer Engineering
University of Toronto
Toronto, ON, Canada, M5S 3G4
rguindi@eecg.utoronto.ca – f.najm@utoronto.ca

Abstract

Oxide tunneling current in MOS transistors is fast becoming a non-negligible component of power consumption as gate oxides get thinner, and could become in the future the dominant leakage mechanism in sub-100nm CMOS circuits. In this paper, we present an analysis of static CMOS circuits from a gate-leakage point of view. We first consider the dependence of the gate current on various conditions for a single transistor, and identify 3 main regions in which a MOS transistor will operate between clock transitions. The amount of gate-current differs by several orders of magnitude from one region to another. Whether a transistor will leak significantly or not is determined by its position in relation to other transistors within a structure. By comparing logically equivalent but structurally different CMOS circuits, we find that the gate current exhibits a 'structure dependence'. Also, the total gate-leakage in a given structure varies significantly for different combinations of inputs, from which we derive "state-dependent gate-leakage tables" that can be used to estimate the total amount of gate-current for a large circuit. Finally, we suggest guidelines aimed at reducing the amount of oxide-leakage current based on the presented structure and state dependencies.

1 Introduction

As MOS transistors get smaller and the gate oxide gets thinner, tunneling current through the insulator becomes a non-negligible component with a potential impact on circuit operation and performance. With CMOS technology approaching the 100-nm regime, the current leaking through the gate oxide is becoming an important component of power consumption [1, 2]. Until a few years ago, oxide tunneling current was not considered to be an issue, however it was predicted that it could surpass weak inversion and DIBL as a dominant leakage mechanism in the future as oxides get thinner [3].

In this paper we look at the oxide tunneling current as a static leakage current component in CMOS logic circuits. We restrict the analysis to static CMOS circuits, however the results can be extended to other design styles, such as dynamic CMOS or domino logic. We first consider the dependence of the gate current on various conditions for a single transistor. We then consider how the gate of a transistor leaks in the context of a given logic structure. Also, given a specific structure, we present state-dependent gate-leakage tables that can be used to estimate the total amount of gate-current for a large circuit. Finally, we suggest guidelines aimed at reducing the amount of oxide-leakage current based on the presented structure and state dependencies.

2 Gate-Leakage Current in a Single Transistor

There are many published studies concerning gate-currents in MOS transistors [1, 4, 5, 9]. In this work we consider only the parameters that affect the magnitude of gate-current in a transistor as it operates in relation to other transistors at the circuit level. We are assuming that V_{dd}, V_t, and the oxide thickness are fixed and depend on the technology used. Variables considered are the *applied bias*, the transistor *type* (NMOS or PMOS), and the transistor *size*.

1. Applied bias:
The magnitude of the gate-current is a strong function of the applied bias [6]. Since the gate-current is a static leakage current, we will consider the transistors at steady-state, i.e. in-between transitions only and not during transient switching. In general, a transistor in a static CMOS logic gate will be operating in one of 3 regions of operation, each with a significantly different amount of gate leakage:

– *Strong inversion*, with $|V_{GS}| = V_{dd}$.
Gate-leakage current density for an NMOS in strong inversion can be as high as $10^3\ A/cm^2$ for an oxide thickness of

0-7695-1881-8/03 $17.00 © 2003 IEEE

$1.5\ nm$ at $V_{dd} = 3\ V$ [7]. For such a thin oxide, a more realistic value for V_{dd} is $1.2\ V$ [2], in which case the gate-leakage current density is around $20\ A/cm^2$ [7].

– *Threshold*, with $|V_{GS}| = V_t$.
A transistor operating at the threshold will leak significantly less that in strong inversion, typically 3 to 6 orders of magnitude less, depending on the value of V_{dd} and the oxide thickness [8].

– *Off*, with $|V_{GS}| = 0$.
For an NMOS device, there is no leakage if $V_{drain} = 0\ V$. However, if the drain is pulled up to V_{dd}, a small leakage component in the reverse direction (from drain to gate) may be present, due to the gate-drain overlap area. This current depends of course on the transistor geometry, and is around 10 orders of magnitude smaller than the gate-leakage current in the strong inversion case [6].

The above 3 regions represent 3 distinct conditions or states for the channel of a MOS transistor. Whether an "on" transistor will operate in strong inversion or at the threshold is determined by its position inside a structure, as well as by the states of other transistors in the structure. These factors are discussed in Section 3.

2. Transistor type:
PMOS transistors in a static CMOS pull-up structure will also be operating in one of the same three modes as their NMOS counterparts. However, the main tunneling component in a PMOS device in inversion is hole tunneling from the valence band, as opposed to electron tunneling from the conduction band in NMOS devices, which results in PMOS gate currents being roughly 10 times smaller than NMOS currents [9]. This fact has important implications in assessing a static CMOS structure from a gate leakage point of view.

3. Transistor size:
Since gate leakage currents are measured as current densities, it follows that the leakage current will be directly proportional to the gate area $(W.L)$. Transistor sizing therefore has a direct impact on the amount of gate leakage in a CMOS circuit.

In static CMOS structures, the signals at transistor gates will always be strong (forcing), coming from either a signal source, or a previous output. In estimating the amount of gate-current in a given CMOS structure, we can therefore assume that a minimum-size transistor in strong-inversion will leak a specific amount that can be measured for a given technology, at a specific oxide thickness, with $|V_{GS}| = V_{dd}$. This is the basis for the values shown in Table 1, where gate-currents are normalized relative to the amount measured in a minimum-size NMOS transistor in the technology at hand, for a given V_{dd}. Without loss of generality, we will consider that a PMOS transistor leaks exactly one tenth the amount in an NMOS.

Table 1. Normalized gate currents in NMOS and PMOS transistors.

Channel condition	NMOS	PMOS
Off	0	0
Threshold	≈ 0	≈ 0
Strong inversion	$1\ \times\ \frac{W.L)_n}{W.L)_{min}}$	$0.1\ \times\ \frac{W.L)_p}{W.L)_{min}}$

3 Structure Dependence

The position of a specific transistor in relation to other transistors in a given structure will determine whether it will operate in strong inversion or at the threshold when switched on. In the following analysis, we will assume that the gate-leakage is significant only if a given transistor is operating in strong inversion.

3.1 Structure-dependent channel states

An NMOS transistor will be in *strong inversion* when its source is pulled to ground either through a direct connection or through another NMOS device, as illustrated in Figure 1 (a) and (b). This is the default state of an "on" transistor. Alternatively, the drain may be pulled low (through another device), as in Figure 1 (c). Also, the NMOS transistor will be in strong inversion if the gate is high, with both the source and drain floating low and the substrated grounded (Fig. 1 (d)). In general, an NMOS will be in strong inversion when switched on unless the channel is actively pulled towards V_{dd}, as in the following.

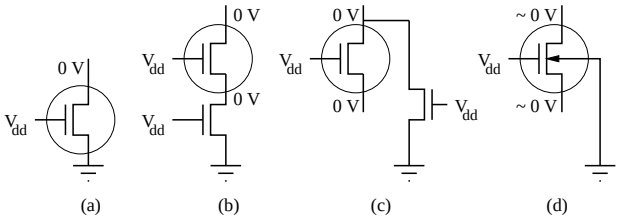

Figure 1. NMOS in strong inversion with $\mathbf{V_{gate} = V_{dd}}$, and (a) source tied to ground, (b) source brought down through another transistor, (c) drain brought down, and (d) source and drain floating low.

An NMOS transistor will be operating at the *threshold* when the drain is pulled high through the PMOS pull-up

network. This usually occurs when the transistor is in a stack, with one or more transistors in lower positions turned off. Depending on the transistor's position in the stack, the voltage at the drain may be either V_{dd} or $V_{dd} - V_t$, as shown in Fig. 2 (a) and (b). Alternatively, in more complex structures, the source of the transistor may be pulled to $V_{dd} - V_t$ through another NMOS device, as in Figure 2 (c).

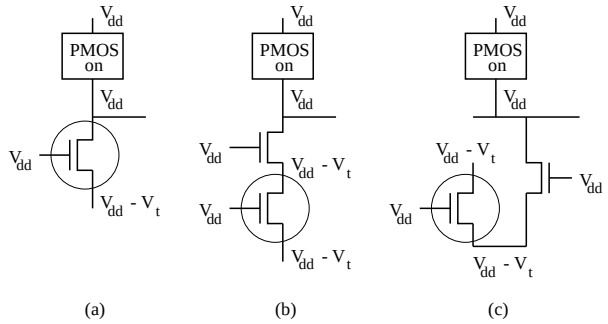

Figure 2. NMOS at the threshold with $\mathbf{V_{gate} = V_{dd}}$, and $\mathbf{V_{drain} = V_{dd}}$ (a) or $\mathbf{V_{dd} - V_t}$ (b,c).

The above structure-dependent transistor states can be re-iterated as follows:

• An NMOS transistor which is turned on (its gate is high) and whose drain or source is connected to ground (either directly or through a path of "on" transistors) will be in *strong inversion* and will leak.

• In a static CMOS gate whose output is high, any NMOS transistor which is turned on (its gate is high) and whose drain or source is connected to the output of the logic gate (either directly or through a path of "on" transistors) will *not* leak. Such transistors will be in the *threshold* mode.

• The higher up an NMOS is in a stack, the lower the chances that it would leak (*strong inversion*), because there are fewer transistors which, if turned on, would cause it to be in the *threshold* mode.

3.2 Comparing NOR and NAND structures

The above observations are clearly illustrated in NMOS NOR and NAND structures (see Fig. 3). In a NOR-gate, each NMOS transistor will leak when turned on, independently from others. In that sense, the NOR structure is a 'worst-case' structure. In the NAND structure however, we find that transistor B is at the threshold if A is off and C is on. Note that when A is off the output node is high. Transistor C leaks in only one case, when all three transistors are on. The different leakage states in the NOR and NAND pull-down structures are given in Table 2 for all input combinations. It is evident that a NAND structure (stacked NMOSes) will leak less that a NOR structure (parallel NMOSes) in 3 out of 8 possible cases, corresponding to the inputs (A,B,C) = (0,0,1), (0,1,1), and (1,0,1).

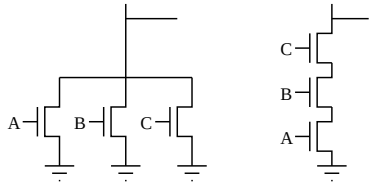

Figure 3. NOR and NAND pull-down structures.

Table 2. Gate leakage in NOR and NAND pull-down networks.

			NOR			NAND		
A	B	C	T_A	T_B	T_C	T_A	T_B	T_C
0	0	0	0	0	0	0	0	0
0	0	1	0	0	Yes	0	0	≈ 0
0	1	0	0	Yes	0	0	Yes	0
0	1	1	0	Yes	Yes	0	≈ 0	≈ 0
1	0	0	Yes	0	0	Yes	0	0
1	0	1	Yes	0	Yes	Yes	0	≈ 0
1	1	0	Yes	Yes	0	Yes	Yes	0
1	1	1	Yes	Yes	Yes	Yes	Yes	Yes

The insight gathered from the pull-down structure of the NAND-gate is applicable to the pull-up of the NOR-gate, and vice-versa. However, since PMOS transistors leak much less than NMOS transistors, it follows that the overall leakage in a NOR-gate (stacked PMOSes, parallel NMOSes) will be larger than in a NAND-gate (stacked NMOSes, parallel PMOSes) if equal-size transistors are used. Transistor sizing will play a role here, as well as the probabilities of the input signals. These issues will be addressed in section 4.

3.3 Logically equivalent structures

Structure-dependent channel states are useful for comparing different implementations of the same logic function. Considering the two logically equivalent structures shown in Fig. 4,it is clear that the first structure leaks more than the second structure, because of the parallel transistors connected to ground, which will always leak when switched on. In the second structure, any or all of the parallel transistors may be on, however there will be practically no gate-leakage through all these devices when the bottom transistor is off. Note also that both pull-up structures for this circuit are identical, making the second structure overall better from the point of view of gate-leakage.

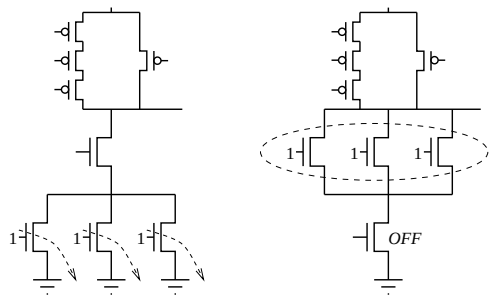

Figure 4. Two logically equivalent circuits, with different pull-down structures.

4 State Dependence

In assessing a structure, or when comparing two logically equivalent implementations, we need to consider all transistors together instead of individually as in Section 3 above. For a given structure, we can compute the total amount of gate-leakage current in the circuit for every "state of the logic gate", as determined by the set of applied inputs. Such *state-dependent* leakage figures can be used to compute the total leakage due to gate-currents in a large network.

4.1 State-dependent leakage tables

Considering all the possible states of a 2-input NAND-gate (see Fig. 5) we find that the leakiest state corresponds to the inputs (1,1) (2 NMOSes leaking, case 4 in the figure). The following state corresponds to the inputs (0,1) (1 NMOS and 1 PMOS leaking), followed by the state corresponding to the inputs (0,0) (2 PMOSes leaking). The least leaky state occurs when the inputs (1,0) are applied, in which case only one PMOS transistor is leaking (case 3 in the figure).

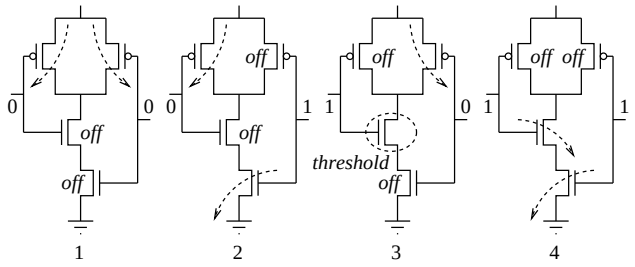

Figure 5. Gate-leakage for the different states in a 2-input NAND-gate.

For comparison purposes, we show in Table 3 the normalized total gate leakage per state for a 2-input NAND-gate and a 2-input NOR-gate, assuming that a minimum-size NMOS device leaks one unit. As in Table 1, we are assuming that a PMOS leaks exactly one tenth the amount in a similar size NMOS. Unsized gates are assumed to have equal minimum-sized NMOS and PMOS transistors, whereas sized gates are designed for equal rise and fall times with respect to a reference inverter, with $\mu_n = 3\mu_p$. Such state-dependent leakage tables can be constructed for a logic gate of any complexity, and be used to identify the input conditions that produce more or less leaky states. It is interesting to note that symmetrical inputs do not produce the same amount of leakage (see rows corresponding to inputs (0,1) and (1,0) in Table 3). If we know in advance that one input is more likely to be high than the other (i.e. signal probabilities are known), it makes sense to assign the input signals in a way that favors the least leaky state.

Table 3. Normalized state-dependent gate currents in NAND and NOR-gates

	NAND		NOR	
	unsized	sized	unsized	sized
	$W_n = 1\times$	$W_n = 2\times$	$W_n = 1\times$	$W_n = 1\times$
input	$W_p = 1\times$	$W_p = 3\times$	$W_p = 1\times$	$W_p = 6\times$
0 0	0.2	0.6	0.2	1.2
0 1	1.1	2.3	1	1
1 0	0.1	0.3	1.1	1.6
1 1	2	4	2	2

4.2 Low-leakage standby state

It is a common approach in low-power designs to put parts of the circuit in a 'standby' mode when not in use, by holding the clock fixed to reduce dynamic power dissipation. It is important when doing so to place the circuit in a static standby state which exhibits low gate-leakage. To find such a state for a large circuit, we must search for an input vector that produces a small total leakage current, which can be computed using state-dependent leakage tables corresponding to the specific logic gates making up the large circuit. Forcing a large multi-gate circuit into a low-leakage state during an idle period can be easily done using minimal additional circuitry as in [10], where a methodology for finding a low-leakage state from the point of view of *subthreshold* leakage was presented. A similar algorithm can be used for placing the circuit in a low-leakage state from the point of view of *gate*-leakage, or ultimately for finding a low-leakage state which takes into consideration *both* varieties of leakage currents.

5 Recommendations and Design Guidelines

The design guidelines presented here derive from the structure and state dependencies discussed in previous sections. In general, these recommendations will result in slower circuits, and are applicable in situations where speed is not the most important issue. For ultra low power applications, we recommend the following:

1. Minimize the number ot NMOS transistors connected to the ground (PMOSes connected to V_{dd}), and maximized the number of NMOSes and PMOSes connected to the output node of a logic gate. The reduction in leakage will be however at the expense of an increased capacitance at the output node.

2. Use minimum-size devices as much as possible.

3. A sum-of-products mapping of a function (NAND-based implementation) leads to potentially less gate-leakage than a product-of-sums mapping (NOR-based implementation) assuming equal-sized transistors are used, and signal probabilities are not known.

4. If signal probabilities are known at the inputs of NAND or NOR gates, connect the signal with the highest probability to the top-most transistor in the stack (NMOS or PMOS) followed by the next-highest signal probability in descending order.

5. For complex gates, construct a state-dependent leakage table and use the signal probabilities to obtain the best assignment of signals for interchangeable inputs (transistors in series or in parallel).

6 Conclusion

In this paper, we have presented an analysis of static CMOS circuits from a gate-leakage point of view. We first identified the 3 main modes in which a transistor operates between clock transitions, and the large difference in the amount of gate-leakage in each mode. We then considered specific combinations of transistors to illustrate the structure dependence of the leakage current. Structure dependence is used as the basis for designing low gate-leakage gates. We then considered the state dependence of the leakage current through state-dependent leakage tables, which can be used for a proper pin assignment if the input signal probabilities are known, as well as to estimate the total power lost due to gate-leakage in a large multi-gate circuit. Finally, we presented recommendations and design guidelines that summarize the main points of the paper.

References

[1] B. Yu *et al.*, "Limits of Gate-Oxide Scaling in Nano-Transistors," *IEEE Symposium on VLSI Technology Digest of Technical Papers*, pages 90 – 91, 2000.

[2] S. Song *et al.*, "CMOS Device Scaling Beyond 100nm," *IEDM Technical Digest*, pages 235 – 237, IEEE, 2000.

[3] A. Keshavarzi, K. Roy, and C.F. Hawkins, "Intrinsic Leakage in Low Power Deep Submicron CMOS ICs," *IEEE International Test Conference*, pages 146 – 155, 1997.

[4] K.F. Schuegraf *et al.*, "Ultra-thin Silicon Dioxide Leakage Current and Scaling Limit," *IEEE Symposium on VLSI Technology Digest of Technical Papers*, pages 18 – 19, 1992.

[5] C-H. Choi *et al.*, "C-V and Gate Tunneling Current Characterization of Ultra-Thin Gate Oxide MOS (t_{ox}=1.3-1.8 nm)," *IEEE Symposium on VLSI Technology Digest of Technical Papers*, pages 63 – 64, 1999.

[6] W. Kirklen Henson *et al.*, "Analysis of Leakage Currents and Impact on Off-State Power Consumption for CMOS Technology is the 100-nm Regime," *IEEE Trans. Electron Devices*, vol. 47, p. 1393, 2000.

[7] S.-H. Lo *et al.*, "Quantum-Mechanical Modeling of Electron Tunneling Current from the Inversion Layer of Ultra-Thin-Oxide nMOSFET's," *IEEE Electron Device Letters*, vol. 18, p. 209, 1997.

[8] K.M. Cao *et al.*, "BSIM4 Gate Leakage Model Including Source-Drain Partition," *IEDM Technical Digest*, pages 815 – 818, IEEE, 2000.

[9] W.C. Lee and C. Hu, "Modeling Gate and Substrate Currents due to Conduction- and Valence-Band Electron and Hole Tunneling," *IEEE Symposium on VLSI Technology Digest of Technical Papers*, pages 198 – 199, IEEE, 2000.

[10] J.P. Halter and F.N. Najm, "A Gate-Level Leakage Power Reduction Method for Ultra-Low-Power CMOS Circuits," *Custom Integrated Circuits Conference*, pages 475 – 478, IEEE, 1997.