Steady and Transient State Analysis of Gate Leakage Current in Nanoscale CMOS Logic Gates

Saraju P. Mohanty
Dept. of Computer Science and Engineering
University of North Texas
P. O. Box 311366, Denton, TX 76203.
Email-ID: smohanty@cse.unt.edu

Elias Kougianos
Dept. of Engineering Technology
University of North Texas
P. O. Box 310679, Denton, TX 76203.
Email-ID: eliask@unt.edu

***Abstract*— Gate leakage (direct tunneling current for sub-$65nm$ CMOS) can severely affect both the transient and steady state behaviors of CMOS circuits. In this paper we quantify the transient and steady-state gate leakage effects as capacitances and state independent (equiprobable) average values, respectively. These metrics are characterized for two universal logic gates, 2-input NAND and NOR, and their sensitivity to variations in process and design parameters is studied. The effective tunneling capacitance of a logic gate is defined as the maximum change in tunneling current with respect to the rate of change of input voltage. It is an unique and novel metric and to our knowledge proposed here for the first time with respect to a logic gate. This metric concisely encapsulates both qualitative as well as quantitative information about the swing in tunneling current during state transitions while simultaneously accounting for the transition rate and represents the capacitive load of the logic gate due to transience in tunneling.**

I. Introduction

Scaling of CMOS devices has been an unavoidable trend for catering to the increasing market demand for ever smaller and application packed portable electronic devices. With the accompanying shrinking of feature size there has been exponential change in the leakage components of the device where each component of the total leakage has gained in relative importance. At this stage there are several short channel effects (SCE) such as drain induced barrier lowering (DIBL), large V_{th} roll-off, diminishing ratio of active and leakage currents, and band-to-band tunneling among others [1]. To overcome these SCEs the ITRS roadmap predicts that high performance CMOS circuits will require ultra-low gate oxide thicknesses [2]. However, such devices will be susceptible to a more profound leakage mechanism due to tunneling through the gate oxide. Thus, the gate oxide tunneling current is emerging as the major component of the static power consumption of a nanoscale CMOS devices.

The gate oxide tunneling current is strongly dependent on the supply voltage of the transistor V_{DD} and gate SiO_2 thickness T_{ox} [3]. During the fabrication process a displacement of even a few SiO_2 molecules can cause a significant variation in T_{ox} [4]. This leads to a difference between the desired value of T_{ox} and the actual T_{ox} obtained after fabrication. This indicates an urgent need for the study of impact of both process (gate dielectric thickness) as well as design parameter (supply voltage) variation on the tunneling current of a CMOS device targeted towards design for manufacturing (DFM) and process variation modeling.

Various characterization and modeling issues of ultra thin oxide MOS devices are discussed by Ghibaudo et al. [5] and Mukhopadhyay et al. [6]. Yang et al. presented methods for characterization and modeling of direct tunneling in [7]. Hertani et al. provide a detailed leakage current analysis of NAND, NOR, and XOR logic gates implemented in three circuit styles [8]. However, most of the available research works do not consider tunneling during both ON and OFF states. None of the above works examine the gate leakage when device or logic gate changes its states, say from ON to OFF, or vice versa. Moreover, no research recognizes the impact of transient gate leakage current on circuit load behavior.

Contributions of this paper: We first show that in nanoscale CMOS devices (and corresponding logic gates) both ON and OFF state gate oxide leakage currents are significant and one can not be ignored with respect to the other. Also, during the transition of states in a logic gate there is a considerable transient effect in the gate tunneling current. A new metric proposed in this paper, *the effective tunneling capacitance essentially quantifies the intra-device loading effect of the tunneling current and also gives a qualitative idea of the driving capacity of the logic gate.* We have performed our analysis and characterization on 2-input NAND, NOR, and other basic gates for a $45-nm$ CMOS process of Berkeley Predictive Transistor Model (BPTM) [9] following an identical procedure for all of them. The purpose of considering isolated logic gates is that it allows for accurate modeling at the most fundamental level of design and the data and corresponding analysis can be used for designs at higher levels of abstraction.

We analyzed in detail the behavior of the gates during an entire cycle of operation (i.e. all possible combinations of inputs.) Guided by the results of the analysis and using additional simulations, we defined the novel metrics: the transient gate leakage current as capacitive effects and the steady-state gate leakage current as state independent (equiprobable) average values. These metrics are appropriate for a complete characterization of the gate oxide tunneling current and its impact on operation of a logic gate as far as leakage and loading effects due to leakage are concerned. The effective

tunneling capacitance can be viewed as a crucial metric for considering the intrinsic loading of the logic gate due to tunneling and the corresponding fanout capacity. We studied the dependence of these metrics on process parameters (gate oxide thickness) and one design parameter (power supply) and provided experimental verification to the theoretically predicted results.

The rest of the paper is organized as follows. In Section II we distinguish between intrinsic gate capacitance and capacitance introduced due to gate tunneling. The steady state and transient gate leakage for a NMOS is discussed in Section III. Section IV introduce the equivalent tunneling capacitance of a logic gate. We present the experimental study and conclusions in Section V, and Section VI, respectively.

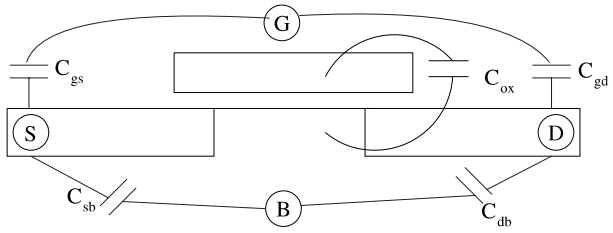
II. Gate Capacitances: Intrinsic Versus Tunneling

In this section we discuss the mechanism of gate oxide tunneling in a nanoscale MOS transistor. We study the behavior of the tunneling current and illustrate the qualitative aspect of capacitive effects of tunneling currents during transition of input states. A short-channel NMOS device is considered for illustration purposes.

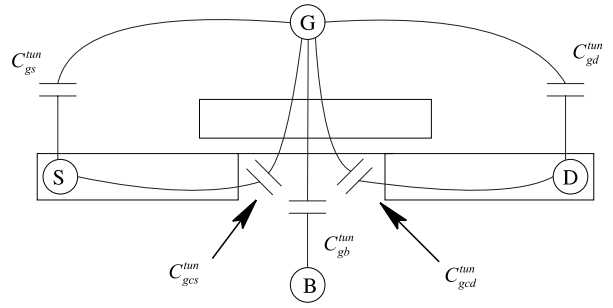
A number of gate and diffusion parasitic capacitances formed in a nominal MOS device. The various intrinsic capacitances of a MOS are shown in Fig. 1(a). It can be seen that during steady state all the terminals of the device have a capacitive effect on the neighboring terminal. These are modeled as capacitive loads. We propose that in addition to these intrinsic parasitic capacitances there is a group of transient capacitances that are formed due to transient gate tunneling current components that flow across the gate-drain, gate-source directly and through the channel as in Fig. 1(b).

We demonstrate that the contribution of gate leakage to power loss can be manifested in different mechanisms. In a short-channel device it is a persistent event that occurs in all states of the device. It is prominent not only during the dynamic state of device but also in static mode. In the static state of the device a steady gate tunneling current flows. However during the transient state, a number of time-varying components are introduced and need to be accounted for. What makes the tunneling effect in each state distinct is the identity of the component that it is comprised of. This effect of tunneling current can be identified as a system of tunneling gate capacitances which contribute to an increase in the overall gate capacitance of the device. These transient capacitances are shown in Fig. 1(b).

The various paths in the ON and OFF states in a NMOS are shown in Fig. 2. The gate oxide tunneling current (I_{ox}) predicted by the BSIM 4.4.0 model [9] for a test input pulse is shown in Fig. 3(a). In the same figure we indicate the steady-state on and off currents (I_{ON}, I_{OFF}). The rise time t_r and fall time t_f of the input pulse are kept equal to 1ns. In the Fig. 3(b) various gate tunneling components are shown gate-to-source diffusion (I_{gs}) and gate-to-channel (I_{gcs} and I_{gcd}). The gate-to-drain diffusion component I_{gd} is negative i. e. towards the gate. The gate to bulk tunneling current



(a) Various gate and diffusion parasitic capacitances in a MOS transistor as modeled in BSIM4. These capacitances comprise the intrinsic loading of the transistor during a steady state.



(b) Various tunneling capacitances in a MOS transistor as proposed in this paper. These capacitances comprise the intrinsic loading of the transistor during a state transition.

Fig. 1. Illustration of the various tunneling capacitances that increase the capacitive loading effect at the gate of a MOS transistor.

(I_{gb}) is negligible. The ON and OFF currents contribute to the intrinsic gate capacitances of the transistor, whereas the transient components of the tunneling current during switching from one state to other contribute to the tunneling capacitance in the device.

III. Steady-state and Transient Tunneling Currents in CMOS Transistors

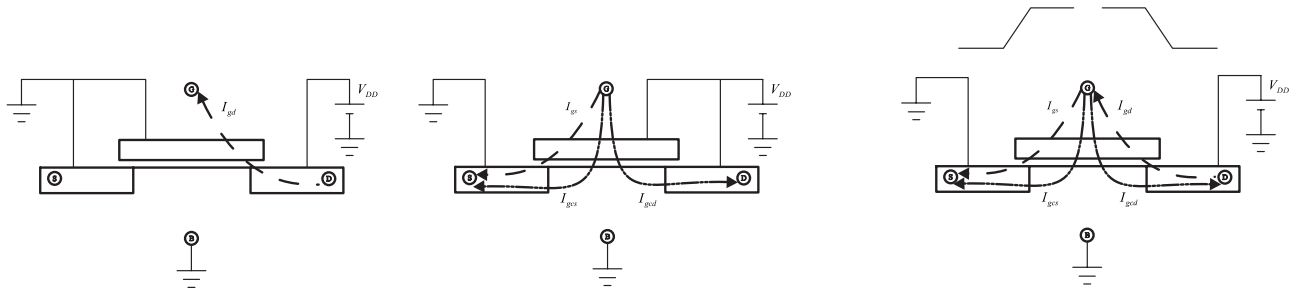
In this section we study the physical mechanism of the gate oxide tunneling in a MOS transistor. We perform a case study using a NMOS transistor; similar analogy will hold good for a PMOS transistor. We identify the regions of operation of the device distinguishing its transition and steady states. We also provide physical explanation of the device behavior during those states. Then we propose a quantitative interpretation of the tunneling current in a nanoscale NMOS device during the steady and transient state. It is observed that the gate to bulk component (I_{gb}) is negligible throughout all regions of operation and hence we ignore it for the remainder of this discussion.

A. Mechanism of Gate Oxide Tunneling

The physical mechanism of the tunneling current can be studied in separate regions: (i) Steady-state (ON or OFF) and (ii) Transient state corresponding to ON to OFF or OFF to ON state transitions as observed in Fig. 2 and Fig. 3.

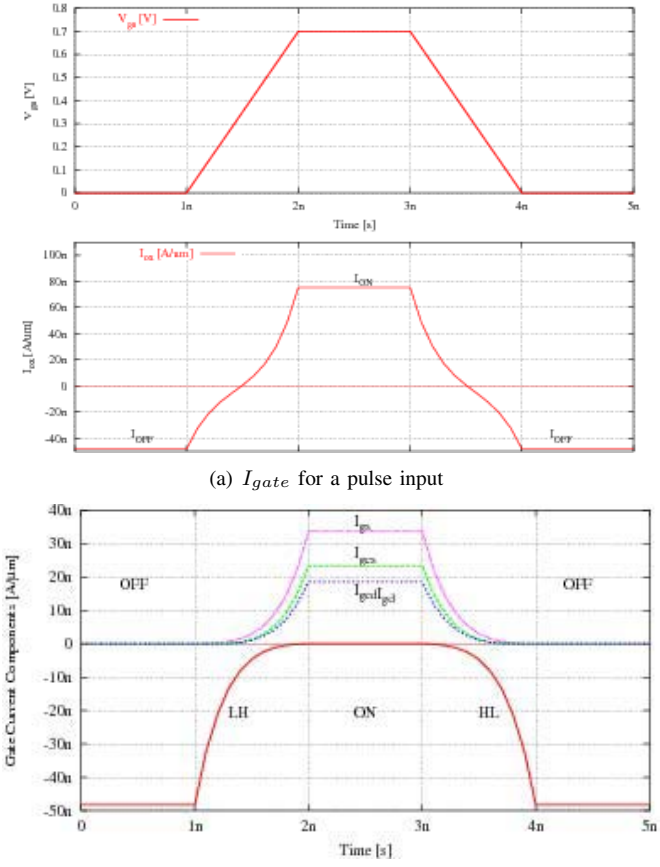
(i) Steady State: This can be either ON or OFF region.

- *OFF region:* In the steady-state OFF region (Fig. 2(a)), both gate and source are at ground while the drain is at high (V_{DD}) voltage. Since no channel is formed in this



OFF (a) *ON* (b) *LH and HL* (c)

Fig. 2. Gate tunneling current component flow in the various regions of operation of a NMOS transistor. In this case (I_{gb}) is negligible and not shown.



(a) I_{gate} for a pulse input

(b) Components of I_{gate} during transient state

Fig. 3. Study of gate tunneling current components in response to a pulse input in a NMOS. The input voltage V_{gs} is given for a complete cycle to make a NMOS OFF, OFF-to-ON, ON, and ON-to-OFF.

condition, the only active component is I_{gd}. The direction of the current flow is from diffusion to gate.

- *ON region:* In this region (Fig. 2(b)), both gate and drain of the device are held at high with the source being grounded. There exists a well-formed channel with three separate components of the gate tunneling current I_{gs}, I_{gcs} and I_{gcd}. The component from gate to drain overlap (I_{gd}) has been extinguished due to the almost zero electric field in that region of the oxide. The overall current flow is from gate to source and channel, opposite to the flow in the OFF state.

(ii) Transient State: This state becomes important when the device switches from ON to OFF or OFF to ON. This change is not an instantaneous process; both the intrinsic gate capacitance and the gate tunneling capacitance will inhibit it. During Low-to-High (LH) and High-to-Low (HL) input transitions all four components of the tunneling current become active as shown quantitatively in Fig. 3(b) and qualitatively in Fig. 2(c). In this case the source is at ground, the drain is at V_{DD} and the gate is switched from low to high or high to low. In the LH transition, the channel gradually originates at the source and extends to the drain and the components I_{gs}, I_{gcs} and I_{gcd} start becoming significant, in that order. Conversely, as the field across the oxide region over the drain is reduced, I_{gd} decreases to almost zero.

B. Quantification of Tunneling Currents

During the ON and OFF steady-states of a MOS device, I_{ON} and I_{OFF}, respectively provide sufficient information for gate leakage characterization. The situation is more complex during the LH and HL input transitions due to the introduction of time variant components. In order to provide a meaningful metric during this transition period, we need to account for the change in magnitude and direction of the total gate tunneling current and consider the time in which this transition takes effect. This presents the loading effect of tunneling current in the device during input (HL or LH) and state (OFF to ON or ON to OFF) transitions. An effective capacitive load due to gate oxide tunneling is defined by:

$$C_{eff}^{tun} = \left| \frac{I_{ON} - I_{OFF}}{dV_g/dt} \right|, \tag{1}$$

where V_g is the voltage applied on the gate. When the rise time (t_r) and fall time (t_f) are identical, this simplifies to:

$$C_{eff}^{tun} = \frac{|I_{ON} - I_{OFF}|}{V_{DD}} t_r. \tag{2}$$

The three metrics presented here (I_{ON}, I_{OFF}, and C_{eff}^{tun}) provide a concise and complete mechanism for characterizing

the gate tunneling leakage during the entire operational cycle of a CMOS device (NMOS or PMOS). The C_{eff}^{tun} quantifies the effective tunneling capacitance at the gate of a transistor during its transient state which is contributed by C_{gs}^{tun}, C_{gd}^{tun}, C_{gcs}^{tun}, C_{gcd}^{tun} and C_{gb}^{tun}, as shown in Fig. 1(b).

IV. Transient Gate Oxide Tunneling Capacitative Effects in Logic Gates

Based on the discussion from the previous sections, it is apparent that *the behavior of the device in terms of gate tunneling leakage current must be characterized not only during the steady state (ON and OFF) but also during transient states*. The OFF state current is comparable in magnitude to the ON state current and hence forms a major source of leakage which needs to be accounted for in any characterization effort. The transient gate leakage current produces an effective tunneling capacitance which has an additive effect with the intrinsic gate capacitance of the device.

This observation can be extended to the case of a logic gate as a higher level of abstraction. However, the interaction of various ON and OFF devices in a logic gate complicates the analysis. The various switching states in a 2-input NAND are shown in Fig. 4. The corresponding transitions in the output and the various components of the tunneling current are also shown in Fig. 5(a). A similar state transition analysis done for a 2-input NOR logic gate is shown in Fig. 5(b). We performed similar analysis for all basic types of logic gates, but presented a comparative perspective of the two universal logic gates (NAND and NOR) for brevity.

In any steady state the gate tunneling can be represented by the state independent (equiprobable) average gate tunneling current which can be defined as:

$$I_{tun} \equiv \text{State independent average gate tunneling current.} \tag{3}$$

The situation is more complex during the transitions due to the introduction of new time variant components which are consolidated as a capacitive load at the fanin of the device and can be defined as:

$$C_{tun} \equiv \text{Effective tunneling capacitance at the input of a logic gate.} \tag{4}$$

We formulate a quantitative definition of the effective tunneling capacitance by considering the worst case among the transitions in the input switching states of the logic gate. Therefore the variation of the total tunneling current at this transition from I_{max}^{logic} to I_{min}^{logic} over a switching period can be used to define the *worst case effective tunneling capacitance* (C_{tun}) of the logic gate as:

$$C_{tun} = \left| \frac{I_{max}^{logic} - I_{min}^{logic}}{dV_{in}/dt} \right|, \tag{5}$$

where V_{in} is the voltage applied at the input of a logic gate. When the rise and fall times are equal this simplifies to:

$$C_{tun} = \frac{\left| I_{max}^{logic} - I_{min}^{logic} \right|}{V_{DD}} t_r. \tag{6}$$

C_{tun} is a worst case estimate, however state-specific estimation can be done for more exhaustive characterization.

The metrics presented here, I_{tun} and C_{tun} provide a concise and complete mechanism for characterizing the gate tunneling leakage during the entire operational cycle of a logic gate. In addition, we believe that C_{tun} contains valuable information for reliability studies since it can be directly related to transient charge flow through the gate oxide. Thus, an early characterization which delivers information on these metrics provides invaluable information to the designers in order to accurately estimate power and delay losses due to gate tunneling leakage.

In summary, *we can define the total internal self capacitive loading of a logic gate as*

$$C_{logic} = C_{tun} + C_{intrinsic}, \tag{7}$$

i. e. sum of capacitances due to transient gate oxide tunneling current and intrinsic gate capacitance.

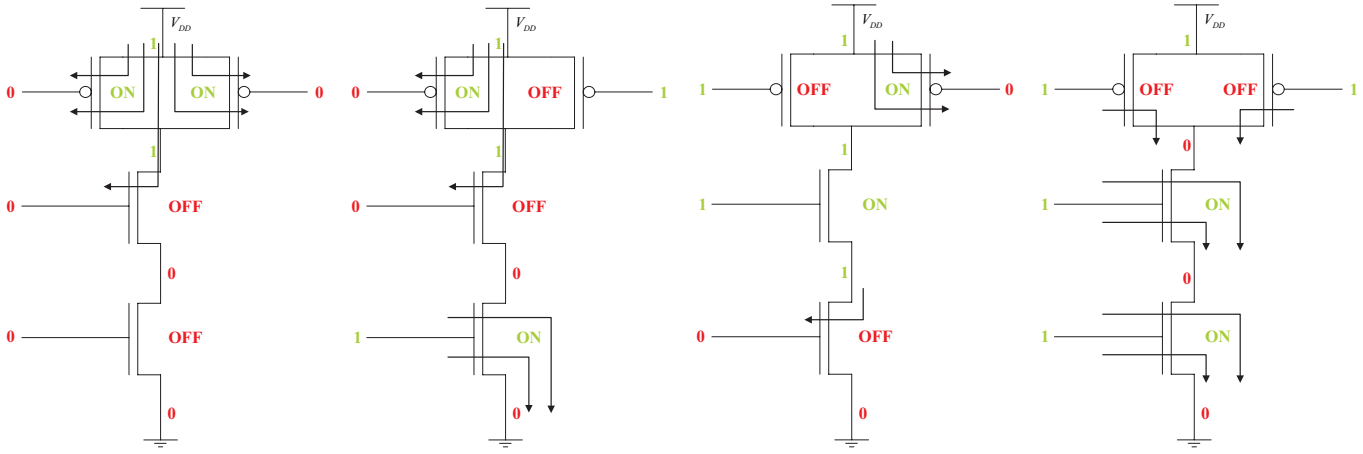
V. Experiments for study of effect of variation in Design and Process parameters

In this section we study the effects of variation in one process parameter T_{ox} and one design parameter V_{DD} on the proposed metrics I_{tun} and C_{tun}. We chose T_{ox} and V_{DD} as these two parameters have maximum impact on the value of the tunneling current as discussed in Section I.

The Berkeley Predictive Transistor (BPTM) model used in this work is for a $45nm$ device technology node with $T_{ox} = 1.4nm$ and threshold voltage $V_{Th} = 0.22V$ [9]. The effect of varying T_{ox} was incorporated by varying $TOXE$ in the spice model deck directly. The width of the device was chosen to be very large ($W = 1\mu m$), thus eliminating any narrow-width effects in the analysis [10]. As a first step in the characterization we selected an appropriate capacitive load for the logic gate under test. A value of 10 times the total gate capacitance C_{gg} of the PMOS device was used for this purpose [11]. Cadence Design Systems' Analog Design Environment and Spectre circuit simulator were used [12]. The supply voltage is initially held at $V_{DD} = 0.7V$.

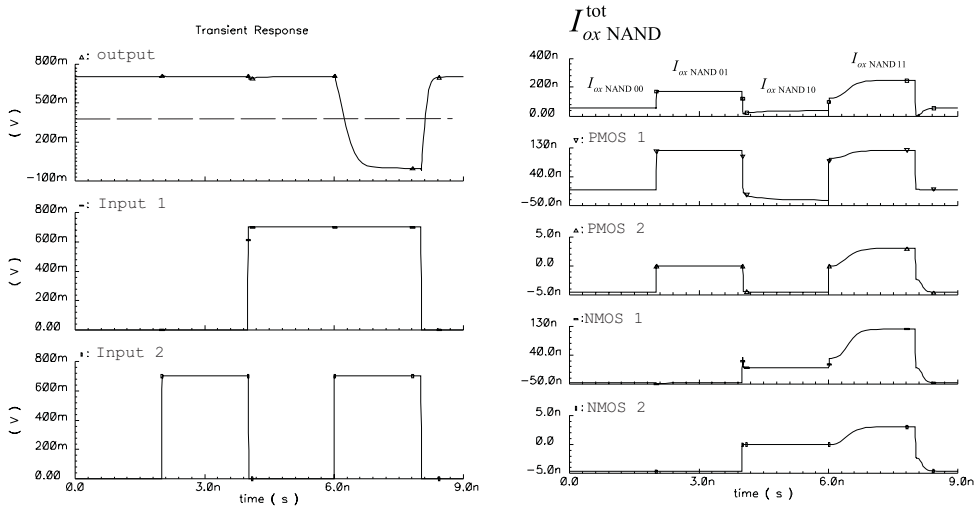
Gate thickness T_{ox}variation: Initially, we held the power supply fixed at $V_{DD} = 0.7V$ and varied the oxide thickness from $T_{ox} = 1.2nm$ to $T_{ox} = 2nm$. The BSIM 4 model based simulation results are shown in Figs. 6(a) and 6(b). It may be noted that the plots are shown in logscale and there is appreciable quantitative difference between the actual values. We notice the purely exponential dependence on T_{ox}. Moreover the I_{tun} of the NAND gate is lower than that of the NOR but the opposite is true for C_{tun}.

Power supply V_{DD} variation: We studied the effect of variation of power supply V_{DD} on the values of I_{tun} and C_{tun} while keeping the other parameters constant at their nominal values. Similarly, we held T_{ox} to a nominal value of $1.4nm$, appropriate for a $45nm$ CMOS technology and investigated the dependence of I_{tun} and C_{tun} on power supply variation. The results are shown in Figs. 6(c) and 6(d).

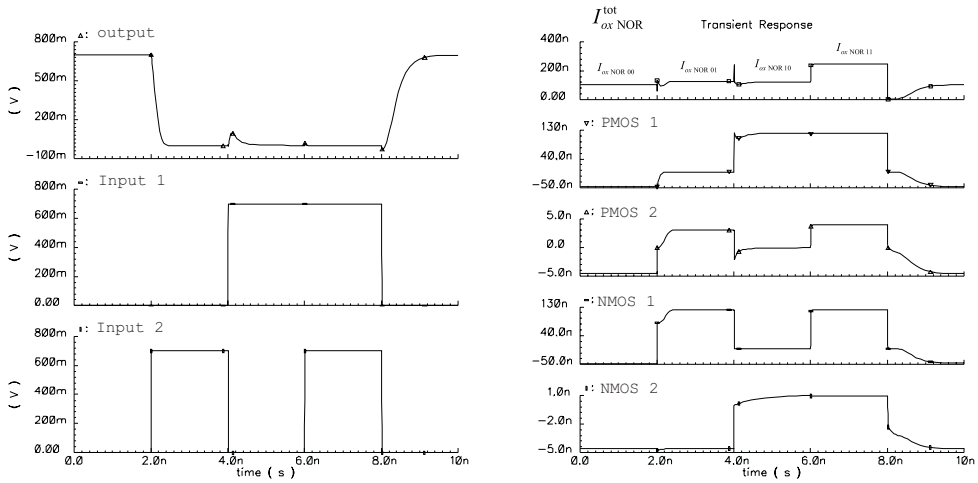


(a) With input 00 (b) With input 01 (c) With input 10 (d) With input 11

Fig. 4. Gate oxide tunneling current paths in various switching states of a 2-input NAND logic gate for different inputs. High logic level is indicated by "1" while low level is indicated by "0".

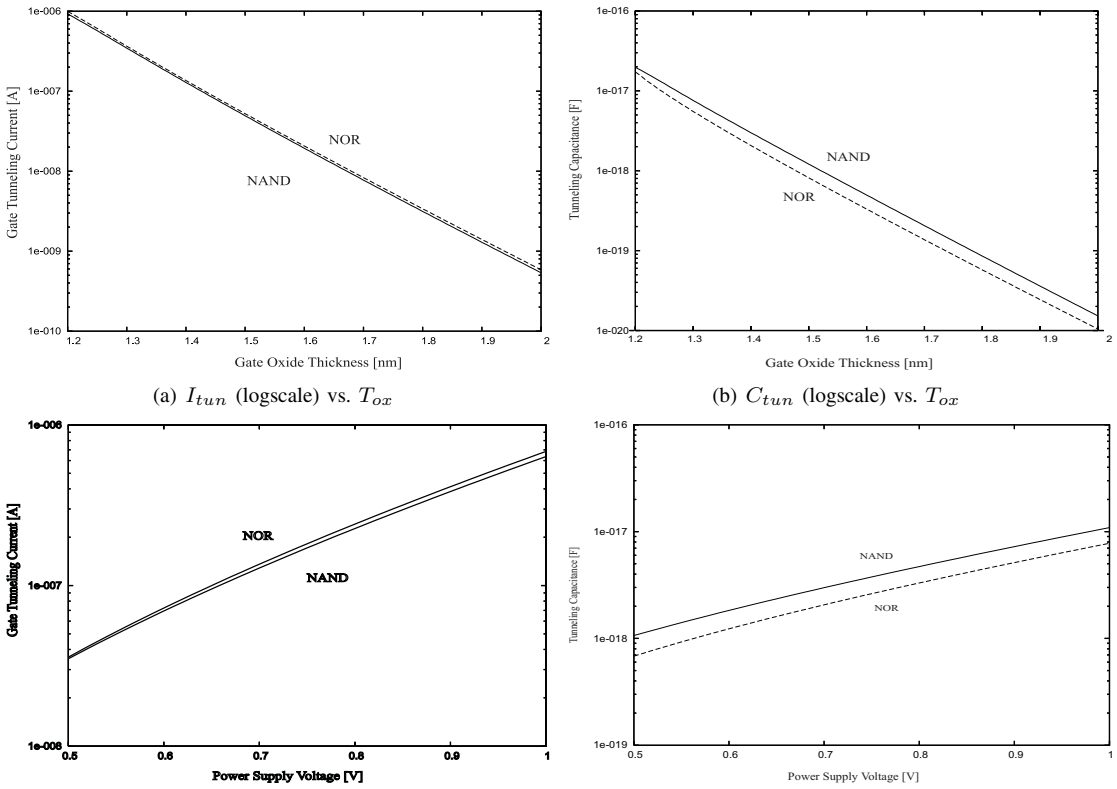


(a) For 2-input NAND gate



(b) For 2-input NOR gate

Fig. 5. Transient response of a 2-input NAND and NOR logic gates for a fixed load. The left figures show voltages for two inputs voltages and the resulted output voltage. The gate oxide tunneling current components in various individual transistors and the total value for the logic gate is presented in right side figure. The drops in threshold voltage V_{th} have not been taken into account.



(a) I_{tun} (logscale) vs. T_{ox}

(b) C_{tun} (logscale) vs. T_{ox}

(c) I_{tun} (logscale) vs. V_{DD}

(d) C_{tun} (logscale) vs. V_{DD}

Fig. 6. Change in tunneling current and capacitance due to oxide thickness and power supply variations. The NAND gate always exhibits lower tunneling current than the NOR gate. The NOR gate always exhibits lower tunneling capacitance than the NAND gate.

VI. Conclusions

We presented a comprehensive analysis of the various gate tunneling current components present during the entire switching cycle of 2-input NAND and NOR. This information was used to identify metrics for the quantitative definition of transient and steady state in the gate tunneling effect. *A study of these metrics reveals that not only the ON cycle but the OFF as well as switching cycles must be accounted. Towards this objective, we introduced the metrics I_{tun} and C_{tun}.* We also observed that while the NAND gate always exhibits lower leakage current, the NOR gate always exhibits lower tunneling capacitance.

Further analysis of the dependence of these metrics on gate oxide thickness, gate dielectric permittivity and power supply variation was presented. This methodology can provide valuable information and estimates for the effect of gate tunneling leakage on power consumption and delay which can then be used to characterize entire cells and libraries leading ultimately to optimized synthesis algorithms for nanoscale CMOS circuit design.

References

[1] K. Roy, S. Mukhopadhyay, and H. M. Meimand, "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits," *Proceedings of the IEEE*, vol. 91, no. 2, pp. 305–327, February 2003.

[2] "Semiconductor Industry Association, International Technology Roadmap for Semiconductors," `http://public.itrs.net`.

[3] S. P. Mohanty and E. Kougianos, "Modeling and Reduction of Gate Leakage during Behavioral Synthesis of NanoCMOS Circuits," in *Proceedings of the 19th IEEE International Conference on VLSI Design (VLSID)*, 2006, pp. 83–88.

[4] S. H. Lo, D. A. Buchanan, and Y. Taur, "Modeling and characterization of quantization, polysilicon depletion and direct tunneling effects in mosfets with ultrathin oxides," *IBM Journal on Research and Development*, vol. 43, no. 3, May 1999.

[5] G. Ghibaudo and R. Clerc, "Characterization and modeling issues in MOS structures with ultra thin oxide," in *Proceedings of the International Conference on Microelectronics*, 2004, pp. 103–113.

[6] S. Mukhopadhyay, C. Neau, R. T. Cakici, A. Agarwala, C. H. Kim, and K. Roy, "Gate Leakage Reduction for Scaled Devices using Transistor Stacking," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 11, no. 4, pp. 716–730, Apr 2003.

[7] N. Yang, W. K. Henson, J. R. Hauser, and J. J. Wortman, "Modeling Studey of Ultra-Thin Gate Oxides Using Direct Tunneling Current and Capacitance-Voltage Measirements in MOS Devices," *IEEE Transactions on Electron Devices*, vol. 46, no. 7, pp. 1464–1471, July 1999.

[8] H. Al-Hertani, D. Al-Khalili, and C. Rozon, "Leakage Power Dissipation in UDSM Logic Gates," in *Proceedings of the 3rd IASTED International Conference on Circuits, Signals, and Systems*, 2005.

[9] Y. Cao, T. Sato, D. Sylvester, M. Orshansky, and C. Hu, "New Paradigm of Predictive MOSFET and Interconnect Modeling for Early Circuit Design," in *Proceedings of the IEEE Custom Integrated Circuits Conference*, 2000, pp. 201–204.

[10] A. Agarwala, S. Mukhopadhyay, C. H. Kim, and K. Roy, "Leakage Power Analysis and Reduction: Models, Estimation and Tools," in *IEE Proceedings in Computers and Digital Techniques*, 2005, pp. 353 – 368.

[11] J. G. Hansen, "Design of CMOS Cell Libraries for Minimal Leakage Currents," Master's thesis, Dept. of Informatics and Mathematical Modelling, Computer Science and Engineering Technical University of Denmark, Fall, 2004.

[12] Cadence Design Systems Inc., *Spectre User's Guide*.