Metrics to Quantify Steady and Transient Gate Leakage in Nanoscale Transistors: NMOS vs. PMOS Perspective

Elias Kougianos
Electrical Engineering Technology
University of North Texas
P. O. Box 310679, Denton, TX 76203.
Email-ID: eliask@unt.edu
http://www.etec.unt.edu/~eliask

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

Abstract

In this paper we explore the use of a set of novel design metrics for characterizing the impact of gate oxide tunneling current in nanometer CMOS devices and perform Monte Carlo simulations to analyze the effects of variations of T_{ox} and V_{DD} on the statistical distribution of these metrics. We concentrate on 3 different unique quantities: (i) Steady-State ON Current (I_{ON}), (ii) Steady-State OFF Current (I_{OFF}), and (iii) Effective Tunneling Capacitance during transitions (C^t_{eff}). We define C^t_{eff} as the change in tunneling current with respect to the rate of change of input voltage, which represents the capacitive load of the transistor due to tunneling. It concisely encapsulates information about the swing in tunneling current during state transitions while simultaneously accounting for the transition rate. We demonstrate that the effect can be very significant due to the exponential dependence of the metrics on process parameters and this dependence also translates into a lognormal distribution for the metrics themselves. We first consider NMOS and PMOS devices individually and subsequently their interaction in an inverter.

1 Introduction and Contributions

Scaling of CMOS devices has been an unavoidable trend for addressing the increasing market demand for smaller and application packed portable electronic devices. The accompanying shrinking of feature size has led to a drastic 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 I_{ON}/I_{OFF}, and band-to-band tunneling [7]. To overcome these SCEs the ITRS roadmap predicts that high performance CMOS circuits will require ultra thin gate oxides [1]. Such devices will be susceptible to a more profound leakage mechanism due to tunneling through the gate oxide [8]. The gate oxide tunneling current (I_{ox}) is therefore emerging as the major component of the static power consumption of a thin nano-CMOS device. There is a critical need for analysis and characterization of the various tunneling mechanisms, targeted towards process variation modeling.

The gate oxide tunneling current is strongly dependent on the supply voltage of the transistor V_{DD} and gate SiO_2 thickness T_{ox}. Accordingly, a small change in T_{ox} can have a tremendous impact on gate oxide current. It is also a fact that, when T_{ox} is ultra thin, it is extremely difficult to maintain a constant T_{ox} for devices on a chip. During the fabrication process a displacement of even a few SiO_2 molecules can cause a significant variation in T_{ox}. This leads to a difference between the desired value of T_{ox} and the actual T_{ox} value obtained after fabrication. Similarly, a change in power supply voltage can cause variation in the gate oxide tunneling current. This necessitates the study of the impact of both process as well as design parameter variation on the tunneling current of a CMOS device. Various characterization and modeling issues for direct tunneling are discussed in [4], [5], [9] and [6]. However, most of the current works do not consider the effect of ON, OFF and transient states simultaneously.

In this paper we will demonstrate that due to a low I_{ON}/I_{OFF} ratio in thin nano-CMOS, both components need equal attention. In addition, the previously cited works investigate only steady-state conditions and do not account for transient effects of tunneling current. We have modeled this effect by assigning to it a meaningful physical

representation of an effective tunneling capacitance (C^t_{eff}). *C^t_{eff} quantifies the intra-device loading effect of the tunneling current and also gives a qualitative idea of the driving capacity of the gate.* We have performed our analysis and characterization on the basic devices and their interaction in an inverter. The purpose of considering isolated transistor devices and the inverter is that it will allow for an accurate modeling and the data and corresponding analysis can be applied to designs at progressively higher levels of abstraction. The contributions of the research in this paper are as follows: We analyzed in detail the behavior of NMOS, PMOS and Inverter during an entire cycle of operation : i) steady states (ON/OFF) and ii) transient states (LH/HL). We define three novel metrics, I_{ON}, I_{OFF}, and C^t_{eff}, completely characterizing the gate oxide tunneling current and its impact on device operation concerning leakage and capacitive effects. C^t_{eff} can be viewed as a crucial metric in considering the intrinsic loading of the transistor due to tunneling. We studied the dependence of these metrics on process parameter T_{ox} and design parameter V_{DD} and provided experimental verification corroborating theoretically predicted results. Finally, we performed Monte Carlo simulations that clearly indicate that small process and power supply variations can have very severe repercussions on the tunneling metrics.

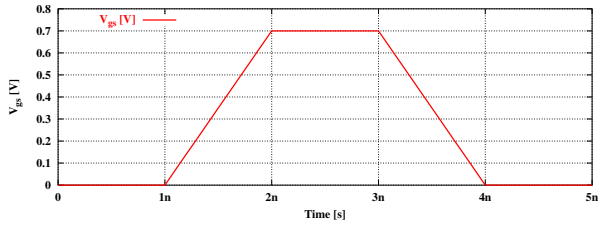
2 Dynamics of Gate Oxide Tunneling

In this section we discuss the physical mechanism of gate oxide tunneling with the help of NMOS and PMOS transistors and an inverter for nanoscale CMOS technology.

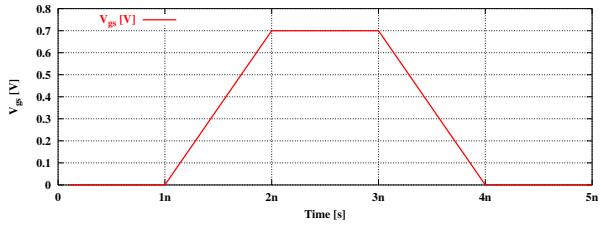
2.1 Tunneling Current SPICE Modeling

The BSIM 4 SPICE model identifies the various components of the gate tunneling current as gate-to-diffusion, gate-to-channel and gate-to-substrate contributions. Since the objective of this work is to analyze the effect of these currents in future nanoCMOS technologies, especially in the 65nm, 45nm ranges and below, no commercially available process data could be used. It is, however, expected that the widely used Berkeley Predictive Technology Model (BPTM) [3] accounts for these effects correctly and has been used throughout this work.

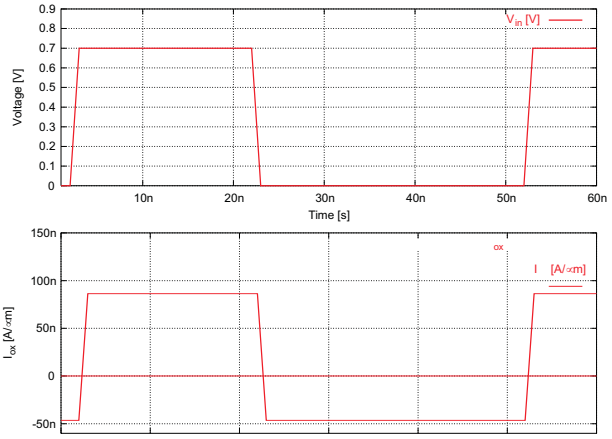
The BPTM model used in this work is for a 45nm device technology node with $T_{ox} = 1.4$nm and threshold voltage $V_{Th} = 0.22$V. The width of the device was chosen to be very large ($W = 1\mu$m), thus eliminating any narrow-width/width-modulation effects in the following analysis [2]. The supply voltage is initially held at $V_{DD} = 0.7$V. We characterized the gate direct tunneling current by evaluating all components (source, drain and bulk) from the BSIM 4 model during the switching operation of the transistor from



(a) I_{gate} in NMOS for a pulse input

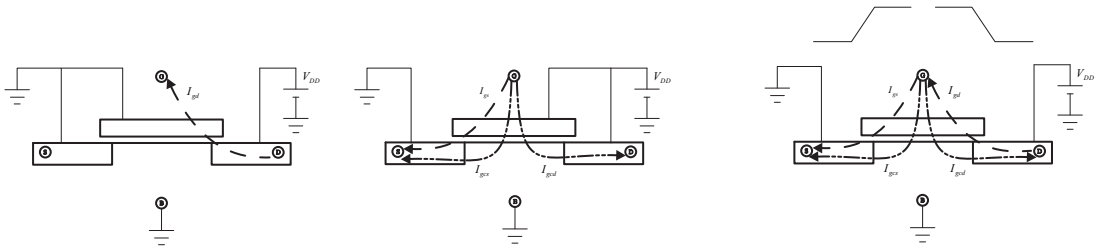


(b) I_{gate} in PMOS for a pulse input



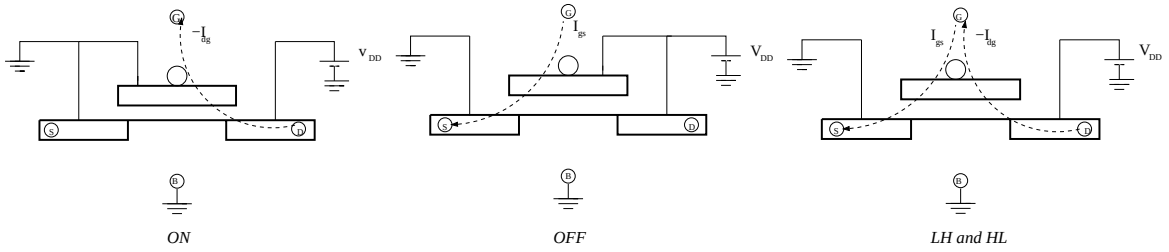
(c) I_{gate} in Inverter for a pulse input

Figure 1. Gate oxide tunneling current (I_{ox}) predicted by the BSIM4.4.0 model for a test input pulse. Indicated are the steady-state on and off currents (I_{ON}, I_{OFF}). The rise and fall times of the input pulse are $t_r = t_f = 1$ ns.



(a) (b) (c)

Figure 3. Gate tunneling current component flow in the various regions of operation of an NMOS. In this case (I_{gb}) is negligible and not shown.



(a) (b) (c)

Figure 4. Gate tunneling current component flow in the various regions of operation of a PMOS. The components (I_{gb}), I_{gcd} and I_{gcs} are neglected here.

the OFF to the ON state and vice versa. The results for NMOS, PMOS and inverter are shown in Fig. 1.

2.2 Physical Mechanism in Steady and Transient States

From Fig. 1 we can identify two distinct regions of operation and two sub-regions of each region of input and output transition of the transistor during a typical switching cycle: (i) Steady-state region [ON/OFF] and (ii) Transient region [Low-to-High(LH)/High-to-Low(HL)]. In order to obtain a better picture of the different tunneling current components and their relative contributions to the overall current, we used the BSIM 4 model to evaluate them. The results are shown in Fig. 2.

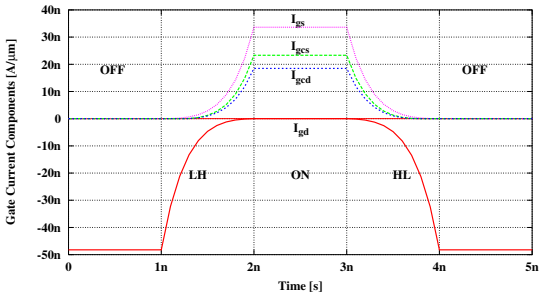
The gate to bulk component (I_{gb}) is negligible throughout all regions of operation and hence we ignore it for the remainder of this discussion. It is also clear that different mechanisms contribute to the overall current during different phases of the switching cycle. In the following discussion we refer to Fig. 3 which identifies the components of tunneling current in an NMOS which are active during each region of operation. The same methodology of analysis is extended to other devices such as PMOS (Fig. 4) and inverter (Fig. 5) in order to account for their characteristics, as shown in Fig. 2.

2.2.1 Steady State [ON/OFF]

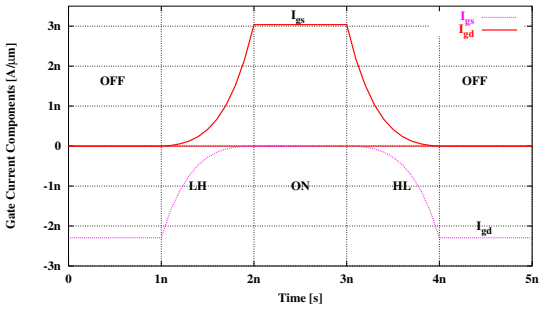
In the steady-state OFF region (Fig. 3(a)), both gate and source are at ground while the drain is at high (V_{DD}) voltage. Since no channel is formed in this condition, the only active component is I_{gd}. The direction of the current flow is from diffusion regions to gate. In the steady-state ON region (Fig. 3(b)), both the gate and drain of the device are held at high with the source being grounded. In this state a well-formed channel exists and three separate components of the gate tunneling current I_{gs}, I_{gcs} and I_{gcd} are active. 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.

2.2.2 Transient State [LH and HL]

Finally, during the LH and HL transitions, all four components become active as shown quantitatively in Fig. 2 and qualitatively in Fig. 3(c). In this case the source is at ground,



(a) Components of I_{gate} in NMOS



(b) Components of I_{gate} in PMOS

Figure 2. Gate tunneling current components calculated in BSIM4.4.0 model. The gate-to-source diffusion (I_{gs}) and gate-to-channel (I_{gcs} and I_{gcd}) components are positive i. e. from the gate. The gate-to-drain diffusion component I_{gd} is negative i. e. towards the gate. The gate to bulk tunneling current (I_{gb}) is negligible and is not shown. Also note, as the values of (I_{gcs} and I_{gcd}) for a PMOS is very negligible, it has not been shown in the figure.

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 zero.

2.3 Three Metrics for Tunneling Current

Based on the results presented in 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 ON and OFF states but also during the transitions.* In particular, 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 situation is more complex

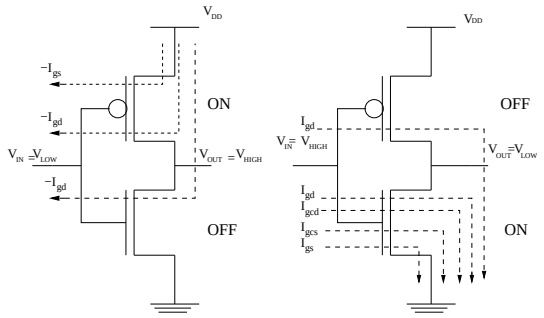


Figure 5. Gate tunneling current flow in an inverter during ON and OFF states.

during the LH and HL transitions due to the introduction of time-varying components which can be considered as an effective capacitive load defined by: $C^t_{eff} = \left|\frac{I_{ON}-I_{OFF}}{dV_g/dt}\right|$, where V_g is the voltage applied on the gate. For simplicity we assume that the rise (t_r) and fall (t_f) times of the gate input voltage are equal making the two transition regions symmetric with respect to their behavior during switching. We also 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 also gives the loading effect of tunneling current in the device during input transitions. When t_r and t_f are identical, this simplifies to: $C^t_{eff} = \frac{|I_{ON}-I_{OFF}|}{V_{DD}} t_r$.

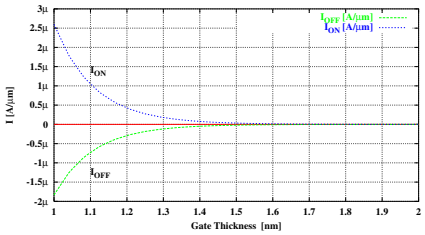
The three metrics presented here (I_{ON}, I_{OFF}, and C^t_{eff}) provide a concise and complete mechanism for characterizing the gate tunneling leakage during the entire operational cycle of an NMOS. Similar metrics are also defined for the PMOS and inverter considered here and can be extended to more complex logic gates.

3 Impact of Process and Supply Variation

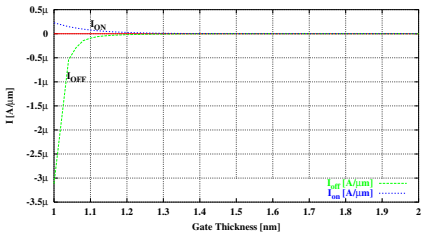
In this section we study the effect of T_{ox} and V_{DD} variation on the gate tunneling current components.

3.1 Process Parameter (T_{ox}) Variation

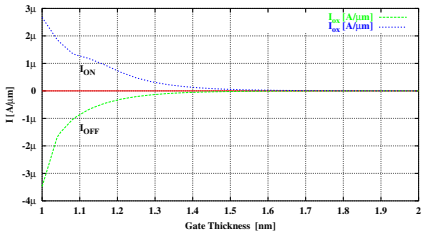
Initially, we held the power supply fixed at $V_{DD} = 0.7$V and varied the oxide thickness from $T_{ox} = 1$nm to $T_{ox} = 2$nm. The BSIM 4 based simulation results are shown in Figs. 6. We notice the very strong exponential dependence of all three proposed metrics on T_{ox}. Moreover the I_{ON} of NMOS when compared to PMOS is larger while the reverse is true for I_{OFF}. This can be observed from the figures in Figs 6(a) and 6(b). For the inverter this is a combination of the effects of NMOS and PMOS as expected



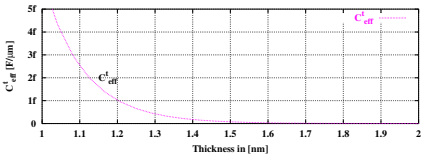
(a) NMOS: I_{ON} or I_{OFF} Vs T_{ox}



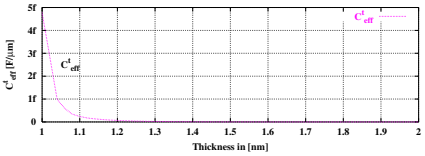
(b) PMOS: I_{ON} or I_{OFF} Vs T_{ox}



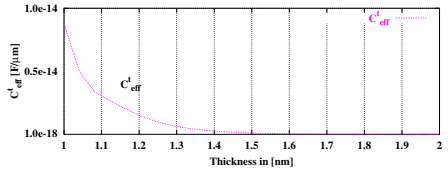
(c) Inverter: I_{ON} or I_{OFF} Vs T_{ox}



(d) NMOS: C^t_{eff} Versus T_{ox}

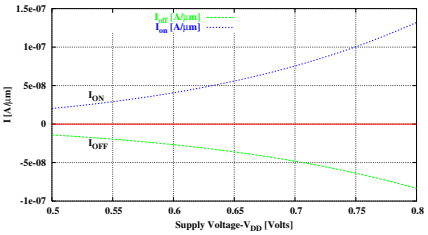


(e) PMOS: C^t_{eff} Versus T_{ox}

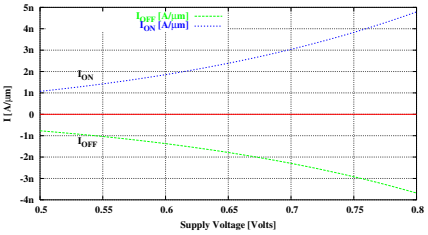


(f) Inverter: C^t_{eff} Versus T_{ox}

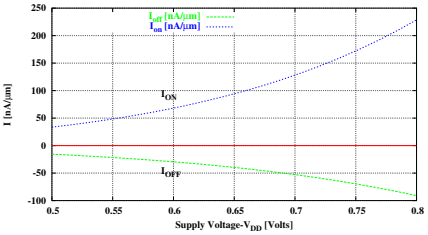
Figure 6. Dependence of Steady-States' I_{ON} and I_{OFF} and Transient States' Effective Tunneling Capacitance C^t_{eff} on Gate Oxide Thickness (T_{ox}). V_{DD} is maintained constant at 0.7V.



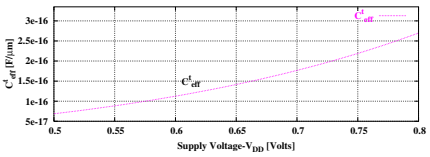
(a) NMOS: I_{ON} or I_{OFF} Vs V_{DD}



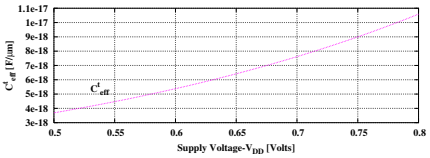
(b) PMOS: I_{ON} or I_{OFF} Vs V_{DD}



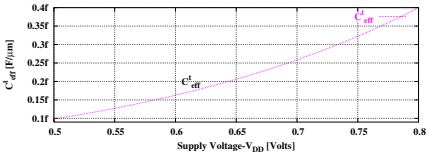
(c) Inverter: I_{ON} or I_{OFF} Vs V_{DD}



(d) NMOS: C^t_{eff} Versus V_{DD}



(e) PMOS: C^t_{eff} Versus V_{DD}



(f) Inverter: C^t_{eff} Versus V_{DD}

Figure 7. Dependence of Steady-States' I_{ON} and I_{OFF} and Transient States' Effective Tunneling Capacitance C^t_{eff} on power supply (V_{DD}). T_{ox} is maintained constant at 1.4nm.

(Fig. 6(c)). The variation in C^t_{eff} with respect to T_{ox} are shown in Fig. 6(d), 6(e), and 6(f).

3.2 Design Parameter (V_{DD}) Variation

We held T_{ox} to a nominal value of 1.4nm, appropriate for a 45nm CMOS technology and investigated the dependence of the currents and C^t_{eff} on power supply variation. The results are shown in Figs. 7. With a variation in the value of supply voltage it can be seen that both I_{ON} and I_{OFF} of the NMOS are almost 100 times that of the PMOS (Figs 7(a) and 7(b). The effect on the inverter is thereby predominantly that of an NMOS (Fig. 7(c)). Similarly the value of C^t_{eff} with respect to V_{DD} is 10 times higher as shown in Fig. 7(d), 7(e), and 7(f). Again we note the strong exponential dependence of all 3 proposed metrics on V_{DD}.

4 Experimental Monte Carlo Results

Our ultimate objective in determining a functional relationship between the metrics and T_{ox} and V_{DD} is to translate statistical information for the distributions of T_{ox} and V_{DD} to statistical information about the metrics themselves.

In the Monte Carlo method followed here we assume that the statistical distribution of process (T_{ox}) and on-chip power supply factors (V_{DD}) is known. For both variables we use a normal distribution with standard deviation (σ) equal to 10% of the mean (μ). The mean value for T_{ox} was 1.4nm and for V_{DD} was 0.7V.

Using these distributions, a statistical sample of N (T_{ox}, V_{DD}) pairs was generated and N simulations were performed with each pair being used only once. These simulations resulted in N triplets of (I_{ON}, I_{OFF}, C^t_{eff}) metrics which were subsequently processed to generate frequency plots. These results indicate that even though approximately 65% of the metrics follow the mean very closely, a significant number of them fall within the range from 2σ to 3σ of the mean. In addition, the distribution is lognormal and the σ is almost 1.8 times the value of the mean.

In summary, a small (10%) variation in process and supply parameters can influence the gate oxide tunneling current metrics significantly. This influence can be manifested by metrics that are two or more times the mean value. Clearly this wide distribution must be taken into account in the design and synthesis of next generation ICs.

5 Conclusions

We presented a comprehensive analysis of the various gate tunneling current components present during the entire switching cycle of NMOS, PMOS and inverter for a realistic 45nm model and used this information to identify metrics for the characterization of the 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 for and towards this objective, we introduced the metrics I_{ON}, I_{OFF} and C^t_{eff}.*

We used directly the variations in T_{ox} and V_{DD} and performed Monte carlo simulations using a baseline BSIM 4 model to obtain the statistical distribution of the metrics.

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 nanoCMOS circuit design.

References

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

[2] 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*, pages 353 – 368, 2005.

[3] 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*, pages 201–204, 2000.

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

[5] R. S. Guindi. Gate-Leakage Estimation and Minimization in CMOS Combinatorial Circuits. In *Proceedings of the 15th International Conference on Microelectronics*, pages 85–88, 2003.

[6] S. Mukhopadhyay, K. Keunwoo, K. J. J. Kim, L. S. Hsien, R. V. Joshi, C. C. Te, and K. Roy. Modeling and Analysis of Gate Leakage in Ultra-Thin Oxide Sub-50nm Double Gate Devices and Circuits. In *International Symposium on Quality Electronic Design*, pages 410–415, 2005.

[7] K. Roy, S. Mukhopadhyay, and H. M. Meimand. Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits. *Proceedings of the IEEE*, 91(2):305–327, February 2003.

[8] A. K. Sultania, D. Sylvester, and S. S. Sapatnekar. Tradeoffs Between Gate Oxide Leakage and Delay for Dual T_{ox} Circuits. In *Proceedings of Design Automation Conference*, pages 761–766, 2004.

[9] N. Yang, W. K. Henson, J. R. Hauser, and J. J. Wortman. Modeling Study of Ultra-Thin Gate Oxides Using Direct Tunneling Current and Capacitance-Voltage Measurements in MOS Devices. *IEEE Transactions on Electron Devices*, 46(7):1464–1471, July 1999.