A Circuit-Level Perspective of the Optimum Gate Oxide Thickness

Keith A. Bowman, *Student Member, IEEE*, Lihui Wang, *Student Member, IEEE*, Xinghai Tang, *Member, IEEE*, and James D. Meindl, *Life Fellow, IEEE*

***Abstract*—A performance constrained minimum power-area optimization is introduced to project the physical gate oxide thickness ($t_{\rm OX}$) scaling limit from a circuit-level perspective. The circuit optimization is based on the recent *physical* alpha-power law MOSFET model that enables predictions of CMOS circuit performance for future generations of technology. The model is utilized to derive an equation for propagation delay including the transition time effect. A physical compact gate-tunneling current model is also derived to analyze ultrathin oxide layers. Results indicate that the gate-tunneling power is *substantially less* ($<5\%$) than the drain-to-source leakage power at the oxide thickness required for optimum CMOS logic circuit performance. *As $t_{\rm OX}$ is scaled below 3.0 nm, the MOSFET performance improvement resulting from $t_{\rm OX}$ scaling diminishes due to an increasing effect of the polysilicon gate depletion depth on the electrical effective oxide thickness. The gate-tunneling power, however, remains exponentially dependent on $t_{\rm OX}$, thus resulting in an optimal value of $t_{\rm OX}$ where the gate-tunneling power is negligible in comparison to the drain-to-source leakage power*. The scaling limit of $t_{\rm OX}$ is projected as 2.2, 1.9, and 1.4 nm for the 180, 150, and 100 nm technology generations, respectively.**

***Index Terms*—CMOS scaling, gate oxide thickness scaling, gate-tunneling current model, low power optimization, physical alpha-power law model, propagation delay model.**

I. Introduction

THE thickness of the silicon dioxide (SiO_2) layer between the gate electrode and the channel region of a MOSFET is the smallest dimension in present-day silicon integrated circuits [1]. Scaling of gate oxide thickness ($t_{\rm OX}$) imposes one of the most restrictive barriers to achieving future gigascale integration (GSI). The International Technology Roadmap for Semiconductors (ITRS) [2] projects the thickness of an effective $t_{\rm OX}$ to be in the range of 1.0–1.5 nm for the 100 nm technology generation, which is expected for the year 2005. The ITRS, however, emphasizes that there are currently "No Known Solutions" for this forecast. To achieve this aggressive scaling of effective $t_{\rm OX}$, the ITRS assumes the development of new innovations such as replacing SiO_2 with a high-κ gate dielectric. This may be a valid assumption [3] since the semiconductor industry has overcome many previous obstacles to maintain Moore's Law [4]. Since a replacement for SiO_2 is yet to be accepted [1], [2] and the time interval for integrating this replacement into a 2005 manufacturing process is short, the importance of understanding the scaling limit of $t_{\rm OX}$ is crucial to future integrated circuit designs.

Historically, the scaling limit of $t_{\rm OX}$ has been defined by equating the gate-tunneling current to the subthreshold drain-to-source leakage current [5]–[8]. For example, assuming a subthreshold leakage current of $\sim$1.0 nA/μm, the oxide thickness limit occurs for 0.1 μm gate lengths at a tunneling current density of $\sim$1.0 A/cm^2 [5]–[8]. This "rule of thumb," however, is strictly limited to a device perspective and does not consider the impact on circuit behavior. Using a previously derived *physical* alpha-power law MOSFET model along with a newly derived gate-tunneling current model, a performance constrained minimum power-area optimization is utilized at a worst-case temperature to quantify the $t_{\rm OX}$ scaling limit at a circuit level for future technology generations outlined by the ITRS [2].

The circuit-level optimization is discussed in Section II by first explaining the physical device and circuit models and then utilizing these models to analyze a performance constrained minimum power-area optimization. In Section III, the circuit optimization of Section II is employed to evaluate the gate oxide thickness required for optimum CMOS logic circuit performance. Finally, concluding statements are offered in Section IV.

II. Circuit-Level Optimization

Circuit- and system-level optimizations provide valuable insight into maximizing the performance of future integrated circuit designs based on a set of material, device, circuit and system constraints. Typically, these optimizations require many iterations because of the large number of variable combinations. Circuit simulators such as HSPICE [9] provide accurate results and are commonly utilized to verify a final design. Circuit simulators, however, are costly in computation time and overall efficiency. Moreover, circuit simulators require a technology process file that is based on a considerable number of empirical parameters. It is unclear how these empirical parameters scale with technology, significantly limiting the simulator's ability to project future circuit performance. Thus, physical device and circuit models are required to enable expedient calculations of circuit performance for future generations of technology.

The remainder of Section II is divided into Sections II-A–E as follows. Section II-A reviews the *physical* alpha-power law MOSFET model [10], which describes the MOSFET drain current in the subthreshold, triode, and saturation regions of operation. Then Section II-B utilizes the *physical* alpha-power law

Manuscript received December 20, 2000; revised February 27, 2001. This work was supported by the Semiconductor Research Corporation and the Defense Advanced Research Projects Agency. The review of this paper was arranged by Editor G. Baccarani.

K. A. Bowman, L. Wang, and J. D. Meindl are with the Georgia Institute of Technology, Atlanta, GA 30332-0269 USA (e-mail: kbowman@ece.gatech.edu).

X. Tang is with the Motorola, Inc., Austin, TX 78730 USA.

Publisher Item Identifier S 0018-9383(01)05737-9.

$$I_D = \begin{cases} I_{D_{SUB}} & (V_{GS} \le V_T + \eta/\beta\text{: subthreshold region}) \\ I_{D_{TRI}} & (V_{DS} < V_{DS_{SAT}}\text{: triode region}) \\ I_{D_{SAT}} & (V_{DS} \ge V_{DS_{SAT}}\text{: saturation region}) \end{cases}$$

$$I_{D_{SUB}} = (W/L)\mu_0 C_{OX}\frac{\eta}{\beta^2}\left[1-\exp(-\beta V_{DS})\right]\exp\left((\beta/\eta)\left[V_{GS}-V_T-\eta/\beta\right]\right)$$

$$I_{D_{TRI}} = (W/L)\frac{\mu_0}{\left[1+\theta(V_{GS}-V_T)\right]\left[1+V_{DS}/(E_C L)\right]}C_{OX}V_{DS}\left[V_{GS}-V_T-(\eta/2)V_{DS}\right]$$

$$I_{D_{SAT}} = I_{D0}\left(\frac{V_{GS}-V_T}{V_{DD}-V_T}\right)^{\alpha};\quad I_{D0} = (W/L)\frac{\mu_0}{\left[1+\theta(V_{GS}-V_T)\right]\left[1+V_{DS_{SAT}}/(E_C L)\right]}C_{OX}V_{D0}\left[V_{DD}-V_T-(\eta/2)V_{D0}\right]$$

$$\alpha = \frac{1}{\ln(2)}\ln\left(\frac{2V_{D0}\left[V_{DD}-V_T-(\eta/2)V_{D0}\right]}{V_{Da}\left[V_{DD}-V_T-\eta V_{Da}\right]}\right)$$

$$V_{DS_{SAT}} = E_C L\left\{\sqrt{1+\frac{2}{E_C L}\left(\frac{V_{GS}-V_T}{\eta}\right)}-1\right\};\quad V_{D0} = V_{DS_{SAT}}\big|_{V_{GS}=V_{DD}};\quad V_{Da} = V_{DS_{SAT}}\big|_{V_{GS}=(V_{DD}+V_T)/2}$$

$$V_T = V_{TL} + \Delta V_{TS};\quad V_{TL} = V_{FB} + 2|\phi_F| - Q_{B0}/C_{OX}$$

MOSFET Parameters

$$\beta = q/(kT);\quad \eta = 1 + C_{D0}/C_{OX};\quad C_{D0} = \sqrt{q\varepsilon_{Si}N_A/\left[2\left(2|\phi_F| - V_{BS}\right)\right]};\quad C_{OX} = \varepsilon_{OX}/t_{OX};\quad \theta = \mu_0/\left[2t_{OX}v_{norm}\right]$$

$$v_{norm} = 2.2\times 10^9\,(\text{cm/s});\quad E_C = (v_{sat}/\mu_0)\left[1+\theta(V_{GS}-V_T)\right];\quad Q_{B0} = -\sqrt{2q\varepsilon_{Si}N_A\left(2|\phi_F| - V_{BS}\right)}$$

Fig. 1. *Physical* Alpha-Power Law MOSFET model.

model to derive an inverter propagation delay model. Section II-C describes the physical gate-tunneling current model that enables projections of circuit performance for ultrathin oxide layers. Employing the models in Sections II-A–C, Section II-D discusses a generic critical path model including the necessary equations to evaluate logic gate delay, power, and area. Finally, Section II-E provides a thorough explanation of the performance constrained minimum power-area optimization.

A. Physical Alpha-Power Law MOSFET Model

The alpha-power law MOSFET model [11] is one of the most widely utilized compact drain current models due to its simple mathematical form and high degree of accuracy. The model has been used to derive many expressions for evaluating circuit performance. Due to its empirical nature, however, several key parameters of the model are measured values, which largely precludes projections of circuit performance for future generations of technology. Moreover, the model does not describe the subthreshold region, thus prohibiting a thorough analysis of on/off drain current tradeoffs. The low power transregional MOSFET model [12] describes all regions of operation (subthreshold, triode, and saturation). The drain current equations are rigorously derived and provide insight into the physical basis of MOSFET behavior. Therefore, the low power transregional model is an advantageous choice for predicting performance of future technology generations and, in particular, for analyzing on/off drain current tradeoffs. The disadvantage of the low power transregional model is its relatively complex drain current equations. Coupling the alpha-power law and low power transregional models enables a new compact physics-based alpha-power law MOSFET model [10]. Salient features of this new model include 1) extension into the subthreshold region of operation, 2) the effects of vertical [13] and lateral [14] high field mobility degradation and velocity saturation, and 3) threshold voltage roll-off [15]. The complete *physical* alpha-power law MOSFET model is provided in Fig. 1.

Fig. 2 compares the model against HSPICE simulations [9] for a 0.25 μm ($L = 0.20$ μm) technology generation in the (a) superthreshold region (I_D versus $V_{\rm DS}$) and (b) subthreshold region (I_D versus $V_{\rm GS}$). Excellent agreement is demonstrated between the *physical* alpha-power law model and HSPICE simulations. Fig. 3 demonstrates that the model is in good agreement with measured data for submicron technology generations: (a) $L = 0.38$ μm [16] and (b) $L = 0.18$ μm [17]. The *physical* alpha-power law model retains the simplicity of the original alpha-power law model while providing a physical basis for the model parameters that enables circuit performance projections for future generations of technology including on/off current interdependence for low power GSI.

B. Inverter Propagation Delay Model

The CMOS inverter propagation delay ($T_{\rm PD}$) model is an extension of two previous model derivations [11], [18] that used the square-law model [19] and the original alpha-power law model [11], both of which include the input transition time (T_T) effect. During a rising ($0\ {\rm V} \rightarrow V_{\rm DD}$) or falling ($V_{\rm DD} \rightarrow 0$ V) input voltage ($V_{\rm IN}$) with a delay of T_T, $T_{\rm PD}$ is the time required from $V_{\rm IN} = 0.5V_{\rm DD}$ to $V_{\rm OUT} = 0.5V_{\rm DD}$, where $V_{\rm OUT}$ is the output voltage. The model analyzes the NFET drain current while neglecting the PFET drain current for a rising input

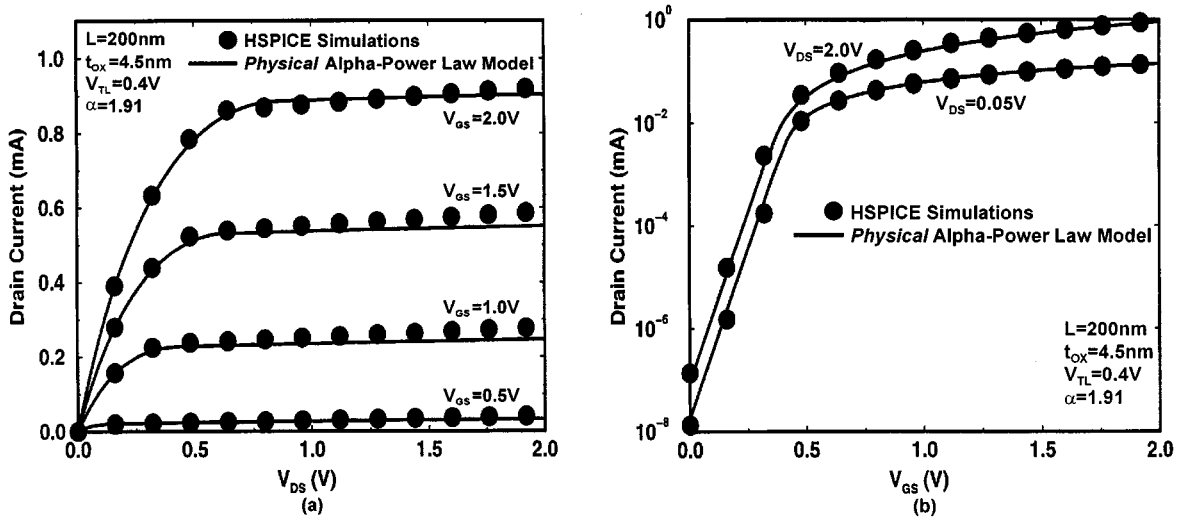

Fig. 2. Verification of the *Physical* Alpha-Power Law Model against HSPICE simulations for (a) I_D versus $V_{\rm DS}$ and (b) I_D versus $V_{\rm GS}$.

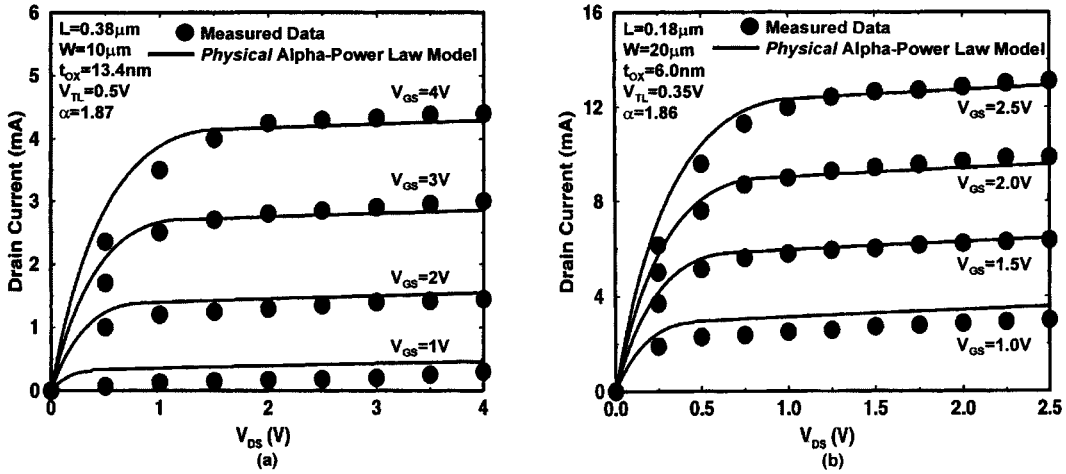

Fig. 3. Verification of the *Physical* Alpha-Power Law Model against measured data for (a) $L = 0.38\ \mu$m and (b) $L = 0.18\ \mu$m.

transition and vice versa for a falling input transition; these are valid assumptions for fast switching circuits [11], [18]. An important simplification results from analysis of short channel devices where the saturation voltage ($V_{\rm DS_{SAT}}$) is limited by carrier velocity saturation. In this regime, the saturation voltage is typically less than $0.5V_{\rm DD}$ [20], thus allowing the derivation of $T_{\rm PD}$ to evaluate only the saturation drain current of the driving MOSFET. $T_{\rm PD}$ is derived for a rising $V_{\rm IN}$ by equating the NFET drain current to the discharging current at the output node

$$I_{D_{\rm SATn}}|_{V_{\rm GS}=V_{\rm IN}(t)} = -C_L \frac{dV_{\rm OUT}(t)}{dt} \tag{1}$$

where C_L is the total output load capacitance, $I_{D_{\rm SATn}}|_{V_{\rm GS}=V_{\rm IN}(t)}$ is the NFET saturation drain current [10], given in Fig. 1, with $V_{\rm GS} = V_{\rm IN}(t)$, and t is the variable delay. The input voltage is assumed to increase linearly during the rising transition time [11], [18] as

$$V_{\rm IN}(t) = \begin{cases} (V_{\rm DD}/T_T)t & \text{for } 0 \le t \le T_T \\ V_{\rm DD} & \text{for } t > T_T \end{cases}. \tag{2}$$

Assuming worst-case mobility degradation models during the input transition, the propagation delay model as a function of T_T is derived as

$$T_{\rm PD} = T_T\left(\frac{1}{2} - \frac{1}{\alpha_n + 1}\frac{V_{\rm DD} - V_{\rm Tn}}{V_{\rm DD}}\right) + \frac{C_L V_{\rm DD}}{2I_{D_{\rm SATn}}|_{V_{\rm GS}=V_{\rm DD}}} \tag{3}$$

where α_n is a physical device parameter that models a portion of the NFET carrier velocity saturation [10], given in Fig. 1, and $V_{\rm Tn}$ is the NFET effective threshold voltage including threshold voltage roll-off [15] and temperature degradation effects.

A realistic input waveform is approximated by calculating the normalized delay required for $V_{\rm OUT}$ to transition from $0.9V_{\rm DD}$ to $0.1V_{\rm DD}$ with a step-input voltage [11],

$$\begin{aligned} T_T &= \frac{V_{\rm DD}}{0.9V_{\rm DD} - 0.1V_{\rm DD}} \times \left(T_{V_{\rm OUT}=0.1V_{\rm DD}} - T_{V_{\rm OUT}=0.9V_{\rm DD}}\right) \\ &= \frac{1}{0.8}\left(T_{V_{\rm OUT}=0.1V_{\rm DD}} - T_{V_{\rm OUT}=0.9V_{\rm DD}}\right). \end{aligned} \tag{4}$$

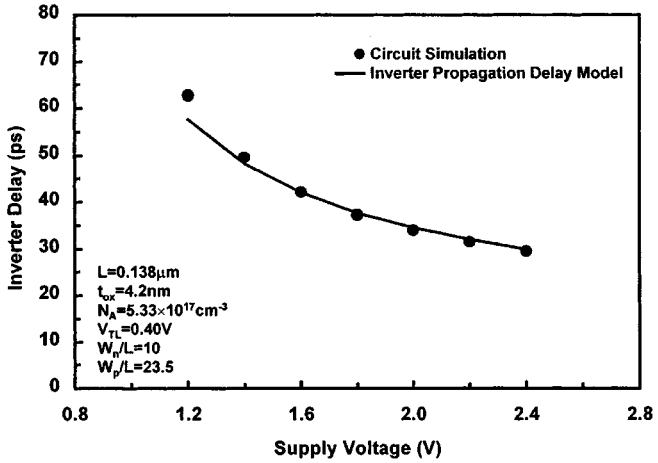

Fig. 4. Comparison of the inverter propagation delay model with a SPICE-equivalent circuit simulator.

$T_{V_{\rm OUT}=0.9V_{\rm DD}}$ and $T_{V_{\rm OUT}=0.1V_{\rm DD}}$ are the delays required for $V_{\rm OUT}$ to equal 90% and 10% of $V_{\rm DD}$, respectively. This method of calculating T_T simplifies the nonlinear discharging ranges of $V_{\rm DD} \rightarrow 0.9V_{\rm DD}$ and $0.1V_{\rm DD} \rightarrow 0$ V. The transition time model is derived in both the saturation and triode regions of operation as

$$T_T = \frac{C_L V_{\rm DD}}{I_{D_{\rm SATn}}|_{V_{\rm GS}=V_{\rm DD}}} \left\{ \frac{0.9}{0.8} + \frac{V_{D0n}}{0.8V_{\rm DD}} \times \left[\frac{V_{\rm DD} - V_{\rm Tn} - (\eta/2)V_{D0n}}{V_{\rm DD} - V_{\rm Tn}} \times \ln\left(\frac{10V_{D0n}[V_{\rm DD} - V_{\rm Tn}]}{V_{\rm DD}[V_{\rm DD} - V_{\rm Tn} - (\eta/2)V_{D0n}]} \right) - 1 \right] \right\} \quad (5)$$

where V_{D0n} is the NFET saturation drain voltage with $V_{\rm GS} = V_{\rm DD}$ [10] and η is the subthreshold slope factor, both provided in Fig. 1.

The expressions for $T_{\rm PD}$, (3), and (5), simplify to the previously derived equations that used the original alpha-power law model [11] if $V_{\rm DD} - V_{\rm Tn} \gg (\eta/2)V_{D0n}$. Since the *physical* alpha-power law model [10] demonstrates better agreement in the triode region of operation than the original alpha-power law model [11], the $T_{\rm PD}$ model in (3) and (5) should provide improved accuracy in comparison to the $T_{\rm PD}$ model using the original alpha-power law model [11]. Moreover, the newly derived model is based on physical device and circuit parameters that allow projections of circuit delay for future generations of technology. Fig. 4 compares the $T_{\rm PD}$ model, (3) and (5), with a SPICE-equivalent circuit simulator [9] for a 0.25 μm technology process file. In simulating the circuit, a chain of symmetrical (equal rise and fall time) inverters are cascaded with a step-input voltage applied to the first inverter. The propagation delay through the first inverter is much faster than the other inverter stages, however, the propagation delay converges to a relatively constant value by the second inverter. The average propagation delay from the second inverter to the end of the chain is plotted in Fig. 4. Good agreement is achieved between the delay model and the circuit simulations for a large range of $V_{\rm DD}$.

C. Gate-Tunneling Model

Several physics-based gate-tunneling current models [7], [21], [22] have been proposed to examine the impact of ultrathin oxide layers on MOSFET designs. These models demonstrate a high degree of accuracy by numerically calculating the electron distribution at the Si/SiO_2 interface to determine the gate-tunneling current. The numerical computation, however, is too time-consuming for a circuit-level optimization. To alleviate this problem, compact gate-tunneling current models [23], [24] have been developed that also exhibit good agreement with measured data. These models, however, are based on empirical parameters, limiting their ability to predict gate-tunneling current for future technology generations. In this section, a compact gate-tunneling current model is derived from first principles of physics to enable device and circuit optimizations for future generations of technology.

The gate-tunneling current model is derived by calculating the electron distribution at the Si/SiO_2 interface and the corresponding probability of those electrons to tunnel through an oxide barrier. The probability of an electron with energy (E_x) tunneling through a barrier of height (E_B) and width $(t_{\rm OX})$ is approximated from the Wentzel-Kramers-Brillouin (WKB) method [25], [26] as

$$D(E_x) = \exp\{-\gamma\sqrt{E_B - E_x}\}. \quad (6)$$

The parameter γ is defined as

$$\gamma = \frac{4\pi t_{\rm OX}\sqrt{2m_{\rm OX}}}{h} \quad (7)$$

where $m_{\rm OX}(=0.32m_0)$ [22], [27] is the effective electron mass in the oxide, m_0 is the electron rest mass, and h is Planck's constant. The actual oxide barrier is approximated by an average rectangular barrier height calculated as

$$E_B = q\left(\chi - \frac{1}{2}V_{\rm OX}\right) \quad (8)$$

where

q electron charge;
χ modified electron affinity in Si;
$V_{\rm OX}$ voltage drop across the oxide layer.

The direct tunneling current density is then given by [28]

$$J_{\rm Tunnel} = \frac{4\pi m^* q}{h^3} \int_0^{E_B} \left\{ \int_0^{\infty} \left[f_{\rm Si/SiO_2}(E) - f_{\rm Gate}(E) \right] dE_t \right\} D(E_x)\, dE_x \quad (9)$$

where $m^*(= 0.19m_0)$ is the electron transverse mass and $f_{\rm Si/SiO_2}(E)$ and $f_{\rm Gate}(E)$ are the electron distributions at the Si/SiO_2 interface and gate, respectively, calculated by Fermi–Dirac statistics

$$f_{\rm Si/SiO_2}(E) = \frac{1}{1 + \exp\left(\frac{E - E_{F0,{\rm Si/SiO_2}}}{kT}\right)} \quad (10)$$

and

$$f_{\rm Gate}(E) = \frac{1}{1 + \exp\left(\frac{E - E_{F0,{\rm Gate}}}{kT}\right)}. \quad (11)$$

$E_{F0,\mathrm{Si/SiO_2}}$ and $E_{F0,\mathrm{Gate}}$ are the Fermi levels at the Si/SiO_2 interface and gate, respectively, k is Boltzmann's constant and T is temperature in °K. The total energy (E) is calculated as the sum of the energy in the direction of tunneling (E_x) and the energy transverse to the direction of tunneling (E_t),

$$E = E_x + E_t. \tag{12}$$

Assuming $E - E_{F0,\mathrm{Si/SiO_2}}$ is sufficiently larger than kT, the Maxwell–Boltzmann distribution can be used to simplify (10) while substituting (12) as

$$f_{\mathrm{Si/SiO_2}}(E) \approx \exp\left(-\frac{E_x + E_t - E_{F0,\mathrm{Si/SiO_2}}}{kT}\right). \tag{13}$$

For a positive gate-to-body potential for the NFET, $E_{F0,\mathrm{Si/SiO_2}}$ is sufficiently larger than $E_{F0,\mathrm{Gate}}$ leading to $f_{\mathrm{Si/SiO_2}}(E) \gg f_{\mathrm{Gate}}(E)$. Performing a two-term Taylor series expansion of $D(E_x)$ (6) around $E_x = 0$, gives

$$\begin{aligned} D(E_x) &= D(0) + E_x \frac{1}{1!}\left.\frac{dD(E_x)}{dE_x}\right|_{E_x=0} + \cdots \\ &= \exp\{-\gamma\sqrt{E_B}\} + E_x \frac{\gamma}{2\sqrt{E_B}} \exp\{-\gamma\sqrt{E_B}\} + \cdots \end{aligned} \tag{14}$$

Substituting (13) and (14) into (9) and utilizing the bias condition of $f_{\mathrm{Si/SiO_2}}(E) \gg f_{\mathrm{Gate}}(E)$, the gate-tunneling current density is simplified as

$$\begin{aligned} J_{\mathrm{Tunnel}} = \frac{4\pi m^* q}{h^3} kT \exp\left\{\frac{E_{F0,\mathrm{Si/SiO_2}}}{kT}\right\} \exp\{-\gamma\sqrt{E_B}\} \\ \times \int_0^{E_B} \exp\left\{-\frac{E_x}{kT}\right\} \left(1 + \frac{\gamma}{2\sqrt{E_B}} E_x\right) dE_x. \end{aligned} \tag{15}$$

The Fermi level at the Si/SiO_2 interface is given by

$$E_{F0,\mathrm{Si/SiO_2}} = q\phi_S - q\phi_F - E_G/2 \tag{16}$$

where ϕ_S is the surface potential and E_G is the Si band gap energy. The parameter $q\phi_F$ is the Fermi energy level either in the Si substrate for the gate-tunneling current through the channel or in the source/drain region for gate-tunneling current through the source/drain overlap. Since the electron distribution is approximately zero for large E_x, the integral in (15) is simplified by evaluating from 0 to ∞. Substituting (16) into (15), the integral in (15) is solved from 0 to ∞ to derive the gate-tunneling current density as

$$\begin{aligned} J_{\mathrm{Tunnel}} = \frac{4\pi m^* q}{h^3} (kT)^2 \left(1 + \frac{\gamma kT}{2\sqrt{E_B}}\right) \\ \times \exp\left(\frac{q\phi_S - q\phi_F - E_G/2}{kT}\right) \exp(-\gamma\sqrt{E_B}). \end{aligned} \tag{17}$$

Fig. 5 compares the gate-tunneling model (17) with measured data (2.9–3.6 nm) and numerical simulations (1.5–2.5 nm) [7]. The model demonstrates excellent agreement with both the measured data and numerical simulations for a large range of V_{GS} and t_{OX}. Fig. 5 demonstrates the validity of calculating the gate-tunneling current without including the effect of quantization of electron energy levels. Modeling the gate-tunneling current using a quantum mechanical quantization predicts more band bending and less electron density at the Si/SiO_2 interface than the classical electron distribution model. These two effects, however, produce opposite contributions to the tunneling current, thus rendering the quantization effect negligible in the gate-tunneling current [22], [29].

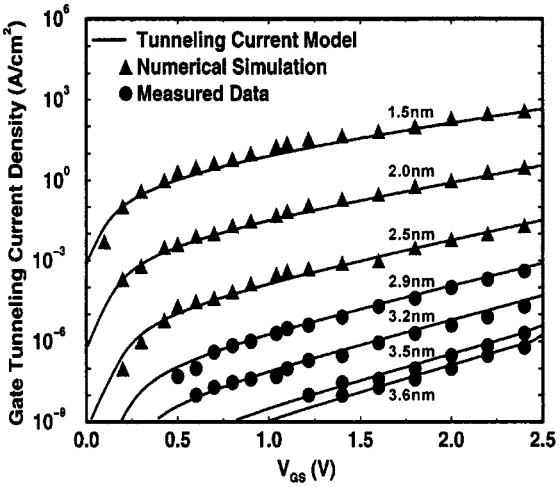

Fig. 5. Gate-tunneling current density (J_{Tunnel}) versus gate-to-source voltage (V_{GS}) for various values of gate oxide thickness (t_{OX}). Compares the gate-tunneling model with measured data (2.9–3.6 nm) and numerical simulations (1.5–2.5 nm).

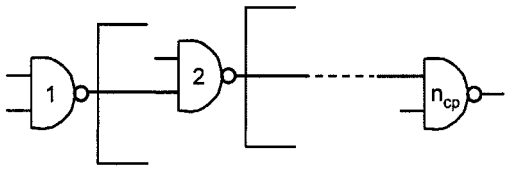

Fig. 6. Critical path model, where n_{cp} is the number of average gate delays.

D. Generic Critical Path Model

The device and circuit optimization is performed on a generic logic critical path with a number (n_{cp}) of identical two-input static CMOS NAND gates with a fan-out of three [30] as illustrated in Fig. 6. Each logic gate drives an average wiring capacitance calculated from a stochastic interconnect distribution [31], which utilizes the ITRS projections [2] of chip area and number of transistors per chip. The static CMOS logic gate was chosen for its low standby power drain, large operating margins, scalability, and flexibility of logic functions [32]. The average propagation delay through a two-input NAND gate is modeled by averaging the delay through two series-connected NFETs and the delay through one PFET given as

$$T_{\mathrm{PD,NAND}} = \frac{f_{\mathrm{ineff}} T_{\mathrm{PDn}} + T_{\mathrm{PDp}}}{2} \tag{18}$$

where f_{ineff} is the effective fan-in factor [33], [20] for series connected MOSFETs and T_{PDn} is the NFET propagation delay, (3) and (5). The PFET propagation delay (T_{PDp}) is calculated through (3) and (5) by substituting the corresponding PFET parameters for $I_{D_{\mathrm{SATn}}}|_{V_{\mathrm{GS}}=V_{\mathrm{DD}}}, \alpha_n, V_{\mathrm{Tn}}$ and V_{D0n}. The cycle

time for the critical path in Fig. 6, which is defined as the reciprocal of the clock frequency ($f_{\rm CLK}$), is described as

$$T_{\rm Cycle} = \frac{1}{f_{\rm CLK}} = \frac{n_{\rm cp} T_{\rm PD,NAND}}{b} \tag{19}$$

where b is the clock skew factor ($=0.9$).

The four major sources of total power consumption consist of 1) dynamic load power ($P_{\rm Dynamic}$), 2) dynamic short-circuit power ($P_{\rm Short-Circuit}$), 3) static drain-to-source leakage power ($P_{\rm Leakage}$), and 4) static gate-tunneling power ($P_{\rm Tunnel}$). For the analysis of high-performance logic critical paths in which the logic gate is sized to maintain equivalent rise and fall delays and the output load capacitance is substantially large, short-circuit power is assumed negligible [34] in comparison to the total power dissipation. The total power is then represented as

$$P_{\rm Total} \approx P_{\rm Dynamic} + P_{\rm Leakage} + P_{\rm Tunnel}. \tag{20}$$

The dynamic load power is given as

$$P_{\rm Dynamic} = a(1/2) C_L V_{\rm DD}^2 f_{\rm CLK} \tag{21}$$

where (a) is the activity factor ($=0.1$). The static drain-to-source leakage power is given as

$$P_{\rm Leakage} = V_{\rm DD} I_{D_{\rm OFF,NAND}} = V_{\rm DD} \frac{I_{D_{{\rm OFF}n}} + 2 I_{D_{{\rm OFF}p}}}{2} \tag{22}$$

where $I_{D_{\rm OFF,NAND}}$ is the average NAND gate leakage current for a chain of logic gates as given in Fig. 6. $I_{D_{{\rm OFF}n}}$ and $I_{D_{{\rm OFF}p}}$ are the NFET and PFET subthreshold drain currents [10] with $|V_{\rm GS}| = 0V$ and $|V_{\rm DS}| = V_{\rm DD}$ as provided in Fig. 1. Since the critical path is modeled as a series of cascaded NAND gates, the probability of a binary "1" or "0" at the output node approaches 50% for the entire critical path. Thus, $I_{D_{\rm OFF,NAND}}$ is calculated as the average leakage current through one NFET plus the leakage current through two PFETs. Neglecting the PFET contribution to gate-tunneling power due to the larger effective mass and barrier height for holes compared to electrons at the SiO_2/Si interface [35], the gate-tunneling power is calculated by averaging the total NFET gate-tunneling power using the same method to calculate (22),

$$\begin{aligned} P_{\rm Tunnel} &= V_{\rm DD} I_{\rm Tunnel,NAND} \\ &= V_{\rm DD} \left(\frac{1}{2}\right) \left\{ \begin{array}{c} [(5/3) I_{\rm Tunnel-S/D} + (2/3) I_{\rm Tunnel-C}] \\ +[4 I_{\rm Tunnel-S/D} + 2 I_{\rm Tunnel-C}] \end{array} \right\}. \end{aligned} \tag{23}$$

$I_{\rm Tunnel,NAND}$ is the average NAND gate-tunneling current, $I_{\rm Tunnel-S/D}$ is the NFET gate-tunneling current through the source/drain overlap regions and $I_{\rm Tunnel-C}$ is the NFET gate-tunneling current through the channel. The fractions $5/3$ and $2/3$ represent the percentage of $I_{\rm Tunnel-S/D}$ and $I_{\rm Tunnel-C}$, respectively, corresponding to the input combinations required for a two-input static CMOS NAND gate to produce a binary "1" at the output node; the factors 4 and 2 correspond to a binary "0." Both $I_{\rm Tunnel-S/D}$ and $I_{\rm Tunnel-C}$ are calculated from (17) using the corresponding doping concentrations and gate-tunneling area.

The equation for describing the logic gate area is given as [36]

$$\begin{aligned} A_{\rm Gate} = k_I &\left(1 + 4\sqrt{\frac{G_{\rm ar}}{k_I}}(f_{\rm in} - 1)\right) \\ &\times \left(1 + \frac{(1+\beta_G)}{\sqrt{k_I G_{\rm ar}}} \left(\frac{W_n}{F} - 1\right)\right) F^2 \end{aligned} \tag{24}$$

where

W_n	NFET width;
F	minimum feature size for a technology;
$k_I\ (=102)$	area of a minimum size inverter with respect to F^2;
$G_{\rm ar}\ (=17/6)$	aspect ratio of the logic gate;
$f_{\rm in}\ (=2)$	number of inputs;
β_G	ratio of the PFET width to the NFET width.

E. Performance Constrained Minimum Power-Area Optimization

The circuit-level methodology minimizes the product of total power consumption and logic gate area for a generic critical path by simultaneously optimizing the supply voltage ($V_{\rm DD}$), the long channel threshold voltage ($V_{\rm TL}$), and the NFET and PFET channel width-to-length ratios, W_n/L and W_p/L, respectively, while satisfying a desired clock frequency. This optimization is an extension of a previous technique [20], [30], which is performed at a worst-case temperature of 400 °K to evaluate the worst-case delay and power constraints. The technology parameters are determined by using the ITRS [2] as a guideline.

Figs. 7 and 8 elucidate the simultaneous optimization of $V_{\rm DD}, V_{\rm TL}, W_n/L$, and W_p/L to minimize the $P_{\rm Total}$ and $A_{\rm Gate}$ product. Fig. 7(a) illustrates the optimum in $V_{\rm TL}$ by plotting the contributions of $P_{\rm Total}, P_{\rm Dynamic}, P_{\rm Leakage}$, and $P_{\rm Tunnel}$ for constant values of W_n/L and W_p/L, thus logic gate area, in a 180 nm technology generation. As $V_{\rm TL}$ decreases through a reduction in channel doping concentration, $P_{\rm Leakage}$ increases resulting from the leakage current's exponential dependence on V_T, the effective threshold voltage including roll-off [15] and temperature degradation. Accompanying the reduction of $V_{\rm TL}, V_{\rm DD}$ decreases to maintain a constant $f_{\rm CLK}$, as shown in Fig. 7(b), leading to a decrease in dynamic power. The optimal long channel threshold voltage ($V_{\rm TL,opt}$) is defined as the value of $V_{\rm TL}$ at which $P_{\rm Total}$ is minimized.

The optimum NFET width-to-length ratio is evaluated in Fig. 8(a), where the left axis plots $P_{\rm Total}$ versus W_n/L for a constant value of $V_{\rm TL}$. W_p/L increases along with an increasing W_n/L to satisfy equal worst-case two-input NAND gate rise and fall times. As W_n/L increases, $V_{\rm DD}$ decreases to maintain the desired $f_{\rm CLK}$, as shown in Fig. 8(b). Initially, increasing W_n/L leads to tremendous power savings. As indicated in Fig. 8(a), however, the power reduction resulting from an increasing W_n/L starts to "saturate." As W_n/L continues to increase from this saturation area, the total power dissipation reduces at a relatively slow rate. The optimum W_n/L for minimum total power is approximately 50% greater than the W_n/L at which $P_{\rm Total}$ "saturates." To achieve the value of W_n/L for a minimum $P_{\rm Total}$, a significant increase in logic gate area is required with a negligible reduction in $P_{\rm Total}$.

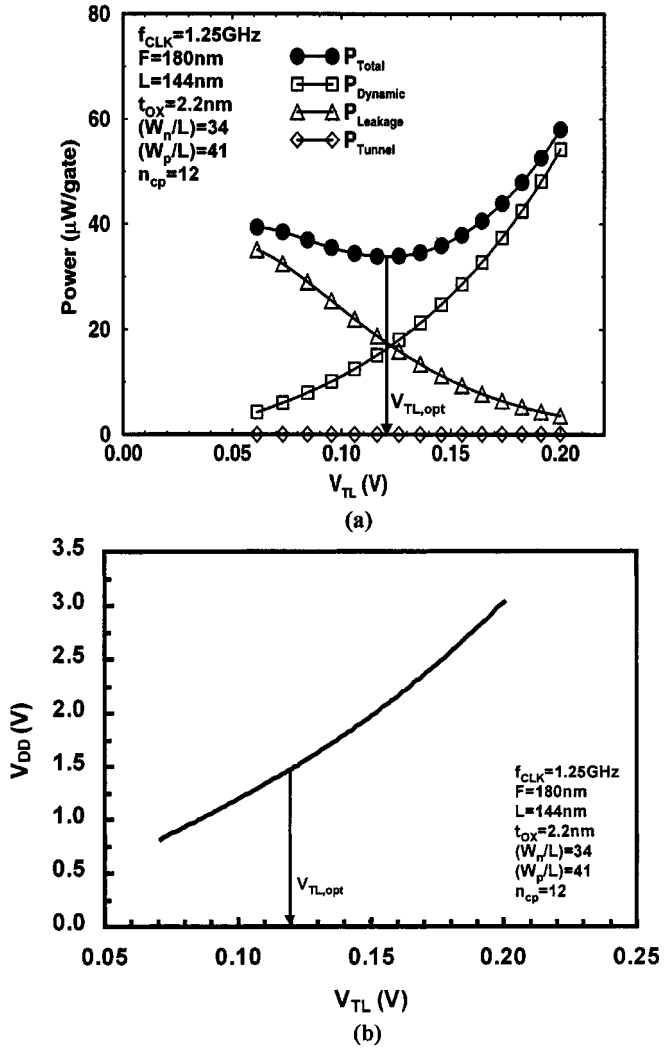

Fig. 7. (a) $P_{\rm Total}, P_{\rm Dynamic}, P_{\rm Leakage}$, and $P_{\rm Tunnel}$ versus $V_{\rm TL}$. (b) $V_{\rm DD}$ versus $V_{\rm TL}$.

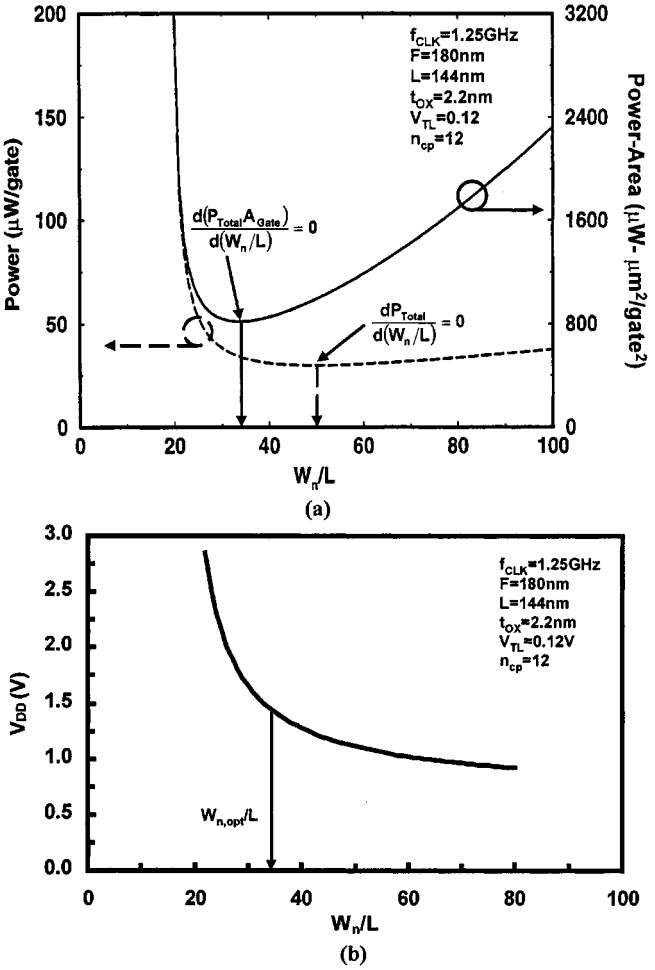

Fig. 8. (a) $P_{\rm Total}$ and $P_{\rm Total}A_{\rm Gate}$ versus W_n/L. (b) $V_{\rm DD}$ versus W_n/L.

The physical reason that $P_{\rm Total}$ "saturates" is understood through Figs. 8 and 9 by describing the critical path cycle time (19), the reciprocal of $f_{\rm CLK}$, as

$$T_{\rm Cycle} = \frac{1}{f_{\rm CLK}} \propto \frac{C_L}{(W_n/L)} = \frac{(W_n/L)C_{\rm Dev0} + C_W}{(W_n/L)} = C_{\rm Dev0} + \frac{C_W}{(W_n/L)}. \quad (25)$$

$C_{\rm Dev0}$ is the logic gate overlap, junction, and fan-out capacitance per W_n/L and C_W is the wiring capacitance. In Figs. 8 and 9, as W_n/L is increased initially, C_L is dominated by C_W and a small increase in W_n/L provides a significant decrease in $V_{\rm DD}$, and consequently $P_{\rm Total}$, to achieve a constant $f_{\rm CLK}$. As the W_n/L ratio continues to increase, however, C_W becomes a smaller percentage of C_L and further increase of W_n/L gives only a small reduction in $V_{\rm DD}$.

Fig. 8(a) also plots on the right axis the product of $P_{\rm Total}$ and $A_{\rm Gate}$ ($P_{\rm Total}A_{\rm Gate}$) versus W_n/L. The value of W_n/L at which $P_{\rm Total}$ "saturates" is approximately equal to the value of W_n/L at which the minimum in $P_{\rm Total}A_{\rm Gate}$ occurs. The optimum W_n/L is defined as the value of W_n/L at which $P_{\rm Total}A_{\rm Gate}$ is minimized. As illustrated in Fig. 8(a), minimizing $P_{\rm Total}A_{\rm Gate}$ results in a significant reduction in logic gate area with a negligible cost in $P_{\rm Total}$ in comparison to just minimizing $P_{\rm Total}$, thus resulting in a more efficient design. Moreover, constraining the logic gate area enables additional silicon real estate to be utilized for other design opportunities such as repeater placement.

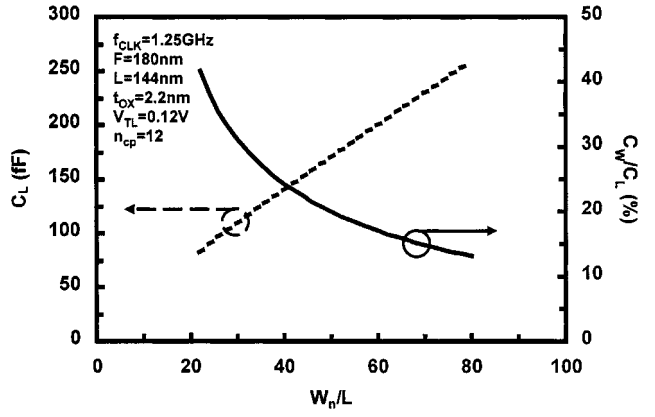

Fig. 9. Load capacitance (C_L) and ratio of wiring-to-load capacitance (C_W/C_L) versus W_n/L.

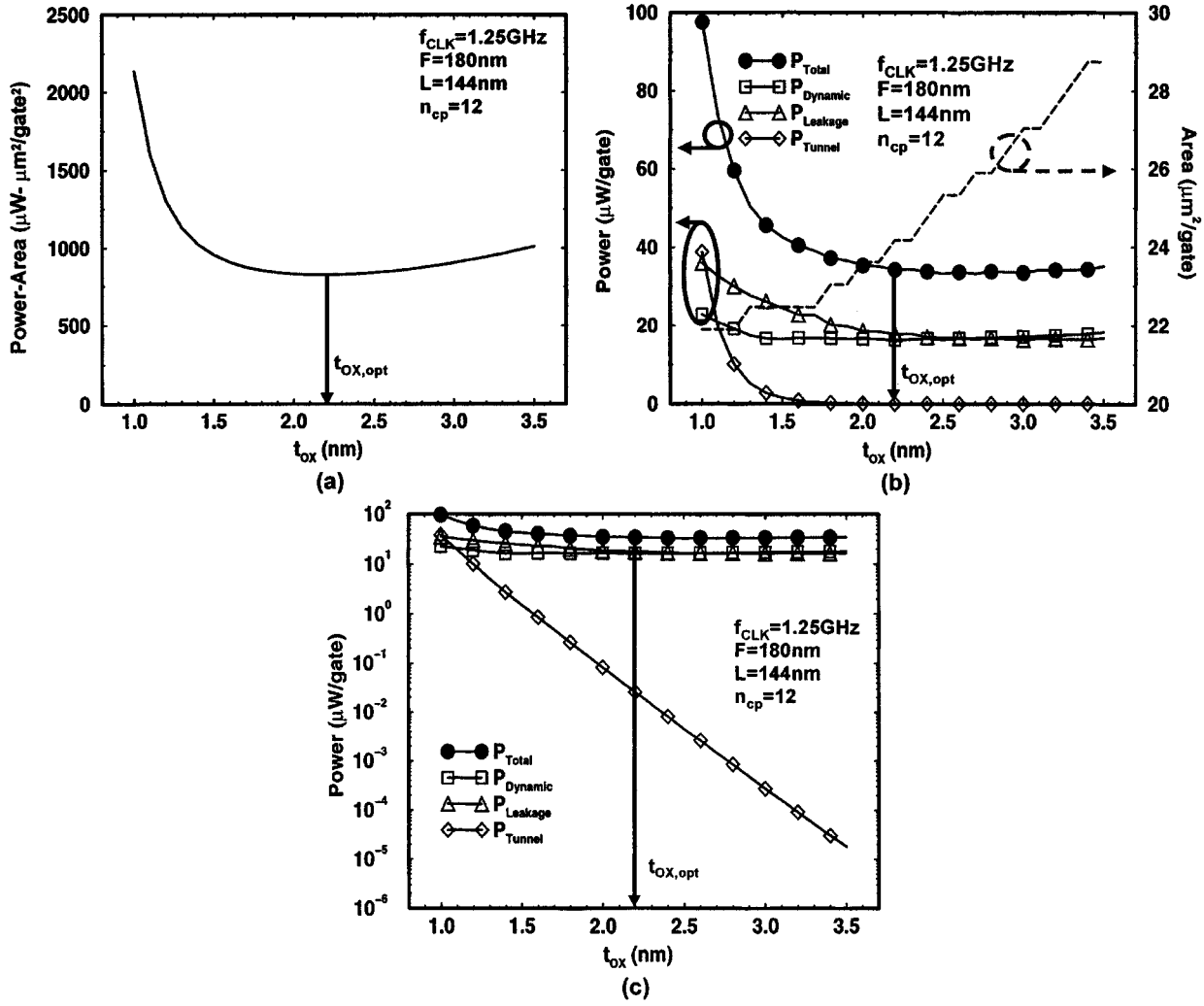

Fig. 10. (a) $P_{\rm Total}A_{\rm Gate}$ versus $t_{\rm OX}$. (b) $P_{\rm Total}$, $P_{\rm Dynamic}$, $P_{\rm Leakage}$, $P_{\rm Tunnel}$, and $A_{\rm Gate}$ versus $t_{\rm OX}$. (c) Log scale of $P_{\rm Total}$, $P_{\rm Dynamic}$, $P_{\rm Leakage}$, and $P_{\rm Tunnel}$ versus $t_{\rm OX}$.

III. Gate Oxide Thickness Scaling Limit for Optimum CMOS Logic Circuit Performance

Employing the circuit methodology in Section II, the optimum physical gate oxide thickness from a circuit-level perspective is determined. The gate-tunneling current model uses the physical oxide thickness $(t_{\rm OX})$ for the barrier width as mentioned in Section II-C while the electrical effective oxide thickness $(t_{\rm OXeff})$ is calculated by including the gate-depletion effect [37] given as

$$t_{\rm OXeff} = t_{\rm OX} + \frac{\varepsilon_{\rm OX}}{\varepsilon_{\rm Si}} X_{\rm Gate} \tag{26}$$

where $\varepsilon_{\rm OX}$ and $\varepsilon_{\rm Si}$ are the permittivity of oxide and silicon, respectively, and $X_{\rm Gate}$ is the gate polysilicon depletion depth. For $V_{\rm GS} > V_T$, $X_{\rm Gate}$ is derived [37] as

$$X_{\rm Gate} = \left(\frac{\varepsilon_{\rm Si}}{\varepsilon_{\rm OX}}\right) t_{\rm OX} \left\{ \sqrt{1 + \frac{2\varepsilon_{\rm OX}^2 (V_{\rm GS} - V_{\rm FB} - \phi_S)}{qN_{\rm Gate}\varepsilon_{\rm Si} t_{\rm OX}^2}} - 1 \right\} \tag{27}$$

where $V_{\rm FB}$ is the flatband voltage and $N_{\rm Gate}$ is the gate polysilicon doping concentration. The effect of inversion layer quantization is omitted in the $t_{\rm OXeff}$ expression (26) as previous research [6] has demonstrated that the polysilicon depletion has a significantly greater impact as $t_{\rm OX}$ is scaled to extremely small dimensions (1.5–2.0 nm). The voltage across the oxide in the gate-tunneling current model is also calculated by including the gate-depletion effect [37].

To quantify the physical oxide thickness for optimal CMOS logic circuit performance, Fig. 10(a) plots the power-area product resulting from the performance constrained minimum power-area optimization for each 1 nm interval of $t_{\rm OX}$ from 1.0 to 3.5 nm. The parameters of $V_{\rm DD}, V_{\rm TL}, W_n/L$, and W_p/L are simultaneously optimized for each value $t_{\rm OX}$. *Thus*, Fig. 10(a) *illustrates the minimum product of total power consumption and logic gate area for a generic critical path by optimizing* $V_{\rm DD}, V_{\rm TL}, W_n/L, W_p/L$, *and* $t_{\rm OX}$, *while maintaining a specified clock frequency at a worst-case temperature of* 400 °K. The optimal oxide thickness $(t_{\rm OX,opt})$ is defined as the value of $t_{\rm OX}$ at which the power-area product is minimized. Fig. 10(a) indicates that $t_{\rm OX,opt}$ is 2.2 nm for the 180 nm technology generation. Fig. 10(b) provides a physical interpretation of the optimum $t_{\rm OX}$ by plotting the individual components of power dissipation and logic gate area from Fig. 10(a). As $t_{\rm OX}$

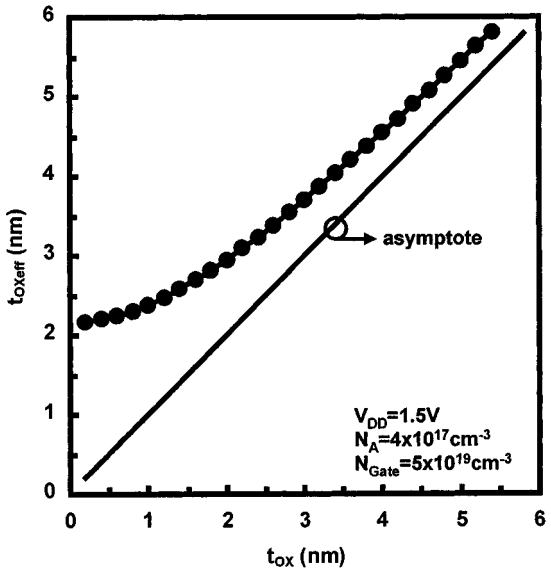

Fig. 11. Electrical effective oxide thickness (t_{OXeff}) versus physical oxide thickness (t_{OX}).

increases, A_{Gate} increases linearly corresponding to a required increase in W_n/L to maintain balance in the device-to-wiring capacitance ratio as discussed in Section II-E. As t_{OX} decreases, P_{Tunnel} increases exponentially, thus increasing P_{Total}. For t_{OX} greater than $t_{\mathrm{OX,opt}}$, the optimum values of V_{DD} and V_T remain relatively constant with respect to t_{OX} as indicated by P_{Dynamic} and P_{Leakage}, respectively. In this regime, V_{TL} and the magnitude of threshold voltage degradation due to roll-off and temperature both decrease resulting in a relatively constant value of V_T. As Fig. 10(b) confirms, P_{Tunnel} is ***significantly less*** ($<$0.2%) than P_{Dynamic} and P_{Leakage} at the optimum t_{OX} of 2.2 nm [38].

To explain this result, Fig. 10(c) plots the power curves from Fig. 10(b) on a logarithmic scale to demonstrate P_{Tunnel}'s exponential dependency on t_{OX}. The optimum t_{OX} occurs for the derivative of $P_{\mathrm{Total}}A_{\mathrm{Gate}}$ with respect to t_{OX} equal to zero,

$$\frac{d(P_{\mathrm{Dynamic}}A_{\mathrm{Gate}})}{dt_{\mathrm{OX}}}+\frac{d(P_{\mathrm{Leakage}}A_{\mathrm{Gate}})}{dt_{\mathrm{OX}}}+\frac{d(P_{\mathrm{Tunnel}}A_{\mathrm{Gate}})}{dt_{\mathrm{OX}}}=0. \tag{28}$$

Since the derivative of power (P) may be expressed as

$$\frac{dP}{dt_{\mathrm{OX}}}=P\frac{d\ln(P)}{dt_{\mathrm{OX}}} \tag{29}$$

and the rate of change of $\ln(P_{\mathrm{Tunnel}})$ is vastly larger than the rate of change of $\ln(P_{\mathrm{Dynamic}})$ and $\ln(P_{\mathrm{Leakage}})$ with respect to t_{OX} as illustrated in Fig. 10(c), P_{Tunnel} is ***considerably less*** ($<$0.2%) than P_{Dynamic} and P_{Leakage} at the optimum t_{OX}.

For an additional insight of this result, Fig. 11 plots t_{OXeff} versus t_{OX} to illustrate the increasing impact of X_{Gate} on t_{OXeff} as t_{OX} is reduced. As t_{OX} decreases below 3.0 nm for a constant V_{DD} and N_{Gate}, X_{Gate} (27) contributes a significant portion to t_{OXeff}. Since saturation drain current is inversely proportional to t_{OXeff}, the advantages of reducing t_{OX} for the logic gate delay starts to saturate in this ultrathin oxide regime. For a constrained performance, the reduction in logic gate area corresponding to a reduction in t_{OX} also starts to saturate in this regime as illustrated in Fig. 10(b). The gate-tunneling current, however, remains exponentially dependent on t_{OX} so that the gate-tunneling power continues to increase dramatically with t_{OX} scaling, as illustrated in Fig. 5. *As t_{OX} enters the ultrathin oxide layer regime* ($<$3.0 nm), *the MOSFET performance gain resulting from further scaling of t_{OX} diminishes due to an increasing impact of the polysilicon depletion, while the gate-tunneling power remains an exponential function of t_{OX}. Thus, the optimal value of t_{OX} results in a gate-tunneling power that is much less than the drain-to-source leakage power*. Note that even though the effect of inversion layer quantization was not considered in t_{OXeff} (26), the same conclusion would still be reached since this effect would add to t_{OXeff} as t_{OX} is scaled below 3.0 nm.

F (nm)	180	150	100
L (nm)	144	120	80
f_{CLK} (MHz)	1250	1767	3500
n_{cp} (# gates)	12	10	8
$t_{OX,opt}$ (nm)	2.2	1.9	1.4
$t_{Oxeff,opt}$ (nm)	3.1	2.7	2.4
$V_{DD,opt}$ (V)	1.5	1.2	1.2
$V_{TL,opt}$ (V) @ 300°K	0.12	0.11	0.13
$W_{n,opt}/L$	34	32	42
$W_{p,opt}/L$	41	38	50
$P_{Dynamic}$ (μW/gate) @ 400°K	16.36	11.25	15.89
$P_{Leakage}$ (μW/gate) @ 400°K	17.83	18.25	20.06
P_{Tunnel} (μW/gate) @ 400°K	0.03	0.06	0.91
P_{Total} (μW/gate) @ 400°K	34.22	29.56	36.86
A_{Gate}/F^2	746	711	887
$I_{D_{OFFn}}/W_n$ (nA/μm) @ 300°K	289	518	666
$I_{D_{OFFn}}/W_n$ (nA/μm) @ 400°K	2,486	4,023	4,898
$I_{Tunnel\text{-}C}/W_n$ (nA/μm) @ 300°K	0.11	0.53	14.74
$I_{Tunnel\text{-}S/D}/W_n$ (nA/μm) @ 300°K	0.62	2.49	37.89

Fig. 12. Results of the minimum power-area oxide thickness optimization for the 180, 150, and 100 nm technology generations.

Fig. 12 tabulates the results of the minimum power-area oxide thickness optimization for the 180, 150, and 100 nm technology generations. Fig. 12 indicates that the drain-to-source leakage current is significantly larger than the channel gate-tunneling current when evaluated at a circuit-level perspective. Also, notice that the gate-tunneling current through the source/drain overlap region is larger than the gate-tunneling current through the channel [39] even though the overlap length (=0.1 F) is much smaller than the channel length (=0.8 F). Fig. 13 projects the scaling limit of t_{OX} as 2.2, 1.9, and 1.4 nm for the 180, 150, and 100 nm technology generations, respectively, while

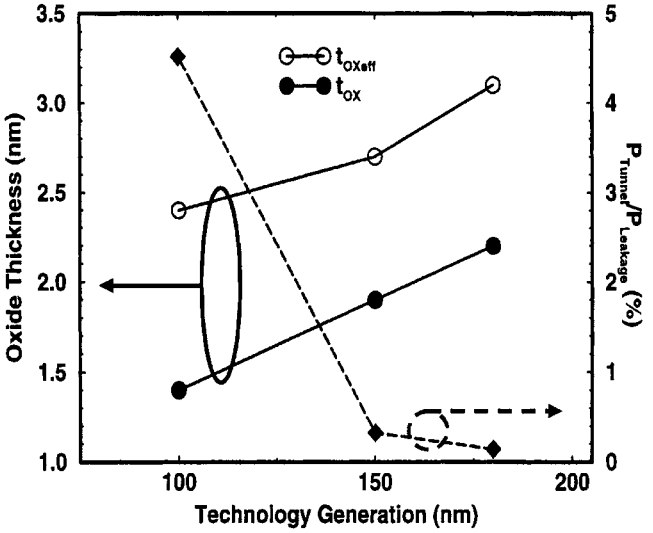

Fig. 13. Optimum physical ($t_{\rm OX}$) and electrical effective ($t_{\rm OXeff}$) oxide thicknesses and ratio of tunneling to leakage power versus technology generation.

the corresponding $t_{\rm OXeff}$ is calculated as 3.1, 2.7, and 2.4 nm. The ratio of $P_{\rm Tunnel}$ to $P_{\rm Leakage}$ is 0.14, 0.32, and 4.5% for the 180, 150, and 100 nm technology generations, respectively. These results indicate that aggressively scaling $t_{\rm OX}$ into this ultrathin oxide layer regime does not provide the typical device performance improvements as those demonstrated in past technology generations. A recently manufactured MOSFET [40] with a $t_{\rm OX}$ of 0.8 nm supports this conclusion with a corresponding $t_{\rm OXeff}$ of 1.8 nm at a $V_{\rm DD}$ of 0.85 V. Until new gate materials are available, this analysis suggests that the focus of future MOSFET designs should be invested in other areas of the device such as channel engineering or junction depth scaling to exploit more significant performance advantages. Previous research [41] has simulated a 25 nm effective channel length using a gate oxide thickness of 1.5 nm along with novel channel engineering techniques. This type of MOSFET design philosophy seems to have a greater potential for impacting future GSI circuits.

IV. Conclusion

A performance constrained minimum power-area optimization is introduced to predict the physical oxide thickness ($t_{\rm OX}$) scaling limit from a circuit-level perspective. The circuit optimization is based on the *physical* alpha-power law MOSFET model that enables projections of CMOS logic circuit performance for future generations of technology by linking the simple mathematical expression of the original alpha-power law model with their physical origins. Using the *physical* alpha-power law MOSFET model, an expression for propagation delay including the transition time effect is developed. Also, a physical compact gate-tunneling current model is derived to evaluate ultrathin oxide layers. Results indicate that the gate-tunneling power is ***significantly less*** ($<5\%$) than the drain-to-source leakage power at the oxide thickness required for optimum CMOS logic circuit performance. *As $t_{\rm OX}$ is scaled below 3.0* nm, *the MOSFET performance gain resulting from further reductions of $t_{\rm OX}$ diminishes due to an increasing impact of the polysilicon depletion depth on the electrical effective oxide thickness. The gate-tunneling power, however, remains an exponential function of $t_{\rm OX}$, thus resulting in an optimal value of $t_{\rm OX}$ where the gate-tunneling power is negligible in comparison to the drain-to-source leakage power*. The scaling limit of $t_{\rm OX}$ is projected as 2.2, 1.9, and 1.4 nm for the 180, 150, and 100 nm technology generations, respectively. Until new gate materials become acceptable, the key recommendation from this work is for future MOSFET designs to focus more attention on other areas of the device such as channel engineering or junction depth scaling to exploit more significant performance opportunities.

Acknowledgment

The authors would like to express their sincere appreciation to J. Joyner of Georgia Tech for his helpful suggestions regarding this work.

References

[1] D. A. Muller *et al.*, "The electronic structure at the atomic scale of ultrathin gate oxides," *Nature*, pp. 758–761, June 24, 1999.
[2] *ITRS*, 1999.
[3] S. Borkar, "Obeying Moore's law beyond 0.18 micron," in *Proc. 13th Annual IEEE Intl. ASIC/SOC Conf.*, Sept. 2000, pp. 26–31.
[4] G. E. Moore, "Progress in digital integrated electronics," in *IEDM Tech. Dig.*, Dec. 1975, pp. 11–13.
[5] S. Thompson, P. Packan, and M. Bohr, "MOS scaling: Transistor challenges for the 21st century," *Intel Tech. J.*, 3rd qtr. 1998. 3rd qtr..
[6] Y. Taur *et al.*, "CMOS scaling into the nanometer regime," *Proc. IEEE*, pp. 486–504, Apr. 1997.
[7] S. H. Lo, D. A. Buchanan, Y. Taur, and W. Wang, "Quantum-mechanical modeling of electron tunneling current from the inversion layer of ultra-thin oxide nMOSFET's," *IEEE Electron Device Lett.*, pp. 209–211, May 1997.
[8] T. Sorsch *et al.*, "Ultra-thin, 1.0–3.0 nm, gate oxides for high performance sub-100 nm technology," in *Symp. VLSI Tech. Dig.*, June 1998, pp. 222–223.
[9] *HSPICE User's Manual*, Mar. 1995.
[10] K. A. Bowman, B. L. Austin, J. C. Eble, X. Tang, and J. D. Meindl, "A physical alpha-power law MOSFET model," *IEEE J. Solid-State Circuits*, vol. 34, pp. 1410–1414, Oct. 1999.
[11] T. Sakurai and A. R. Newton, "Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas," *IEEE J. Solid State-Circuits*, vol. 25, pp. 584–594, Apr. 1990.
[12] B. L. Austin, K. A. Bowman, X. Tang, and J. D. Meindl, "A low power transregional MOSFET model for complete power-delay analysis of CMOS gigascale integration (GSI)," in *Proc. 11th Annu. IEEE Intl. ASIC Conf.*, Sept. 1998, pp. 125–129.
[13] S. L. Garverick and C. G. Sodini, "A simple model for scaled MOS transistors that includes field-dependent mobility," *IEEE J. Solid-State Circuits*, vol. 22, pp. 111–114, Feb. 1987.
[14] B. T. Murphy, "Unified field-effect transistor theory including velocity saturation," *IEEE J. Solid-State Circuits*, vol. SC-15, pp. 325–327, June 1980.
[15] B. Agrawal, V. K. De, and J. D. Meindl, "Opportunities for scaling FET's for gigascale integration (GSI)," in *Proc. 23rd ESSDERC*, Sept. 1993, pp. 919–926.
[16] R. A. Chapman, C. C. Wei, D. A. Bell, S. Aur, G. A. Brown, and R. A. Haken, "0.5 Micron CMOS for high performance at 3.3 V," in *IEDM Tech. Dig.*, Dec. 1988, pp. 52–55.
[17] M. Bohr *et al.*, "A high performance 0.35 μm logic technology for 3.3 V and 2.5 V operation," in *IEDM Tech. Dig.*, Dec. 1994, pp. 273–276.
[18] N. Hedenstierna and K. O. Jeppson, "CMOS circuit speed and buffer optimization," *IEEE Trans. Computer-Aided Design*, vol. CAD-6, pp. 270–281, Mar. 1987.
[19] W. Shockley, "A unipolar field effect transistor," *Proc. IRE*, pp. 1365–1376, Nov. 1952.
[20] A. J. Bhavnagarwala, B. L. Austin, K. A. Bowman, and J. D. Meindl, "A minimum total power methodology for projection limits on CMOS GSI," *IEEE Trans. VLSI Syst.*, vol. 8, pp. 235–251, June 2000.

[21] N. Yang, W. K. Henson, J. R. Hauser, and J. J. Wortman, "Modeling study of ultrathin gate oxides using direct tunneling current and capacitance-voltage measurements in MOS devices," *IEEE Trans. Electron Devices*, vol. 46, pp. 1464–1471, July 1999.
[22] E. M. Vogel *et al.*, "Modeled tunnel currents for high dielectric constant dielectrics," *IEEE Trans. Electron Devices*, vol. 45, pp. 1350–1354, June 1998.
[23] C. Choi, K. Oh, J. Goo, Z. Yu, and R. W. Dutton, "Direct tunneling current model for circuit simulation," in *IEDM Tech. Dig.*, Dec. 1999, pp. 735–738.
[24] P. O'Sullivan, A. Fox, K. G. McCarthy, and A. Mathewson, "Toward a compact model for MOSFETS with direct tunneling gate dielectrics," in *Proc. 29th ESSDERC*, Sept. 1999, pp. 488–491.
[25] J. J. Sakurai, *Modern Quantum Mechanics*. Norwell, MA: Addison-Wesley, 1995.
[26] R. Holm, "The electric tunnel effect across thin insulator films in contacts," *J. Appl. Phys.*, pp. 569–574, May 1951.
[27] M. Depas, B. Vermeire, P. W. Mertens, R. L. Meirhaehe, and M. M. Heyns, "Determination of tunneling parameters in ultra-thin oxide layer poly-Si/SiO_2/Si structures," *Solid-State Electron.*, pp. 1465–1471, Aug. 1995.
[28] S. M. Sze, *Physics of Semiconductor Devices*, 2nd ed. New York: Wiley, 1981.
[29] F. Ranna, S. Tiwari, and D. A. Buchanan, "Self-consistent modeling of accumulation layers and tunneling currents through very thin oxides," *Appl. Phys. Lett*, pp. 1104–1106, Aug. 1996.
[30] A. J. Bhavnagarwala, B. L. Austin, and J. D. Meindl, "Minimum supply voltage for bulk Si CMOS GSI," in *ISLPED*, Aug. 1998, pp. 100–102.
[31] J. A. Davis, V. K. De, and J. D. Meindl, "A stochastic wire-length distribution for gigascale integration (GSI)—Parts I and II," *IEEE Trans. Ele.ctron Devices*, vol. 45, pp. 580–597, Mar. 1998.
[32] J. D. Meindl, "Low power microelectronics: Retrospect and prospect," *Proc. IEEE*, vol. 83, pp. 619–635, Apr. 1995.
[33] T. Sakurai and A. R. Newton, "Delay analysis for series-connected MOSFET circuits," *IEEE J. Solid-State Circuits*, vol. 26, pp. 122–131, Feb. 1991.
[34] A. P. Chandrakasan, S. Sheng, and R. W. Brodersen, "Low-power CMOS digital design," *IEEE J. Solid-State Circuits*, vol. 27, pp. 473–484, Apr. 1992.
[35] K. Yang *et al.*, "A physical model for hole direct tunneling current in P^+ poly-gate PMOSFET's with ultrathin gate oxides," *IEEE Trans. Electron Devices*, vol. 47, pp. 2161–2166, Nov. 2000.
[36] J. C. Eble, III, "A Generic system simulator with novel on-chip cache and throughput models for gigascale integration," Ph.D. dissertation, Georgia Inst. Technol., Atlanta, Nov. 1998.
[37] B. Yu *et al.*, "Gate engineering for deep-submicron CMOS transistors," *IEEE Trans. Electron Devices*, vol. 45, pp. 1253–1262, June 1998.
[38] K. A. Bowman, L. Wang, X. Tang, and J. D. Meindl, "Oxide thickness scaling limit for optimum CMOS logic circuit performance," in *Proc. 30th ESSDERC*, Sept. 2000, pp. 300–303.
[39] N. Yang, W. K. Henson, and J. J. Wortman, "A comparative study of gate direct tunneling and drain leakage currents in N-MOSFET's with sub-2nm gate oxides," *IEEE Trans. Electron Devices*, vol. 47, pp. 1636–1644, Aug. 2000.
[40] R. Chau *et al.*, "30 nm physical gate length CMOS transistors with 1.0 ps n-MOS and 1.7 ps p-MOS gate delays," in *IEDM Tech. Dig.*, Dec. 2000, pp. 45–48.
[41] Y. Taur, C. H. Wann, and D. J. Frank, "25 nm CMOS design considerations," in *IEDM Tech. Dig.*, Dec. 1998, pp. 789–792.

Keith A. Bowman (S'97) received the B.S. degree in electrical engineering from North Carolina State University, Raleigh, in 1994 and the M.S. and Ph.D. degrees in electrical engineering from Georgia Institute of Technology, Atlanta, in 1995 and 2001, respectively. His doctoral research focused on the impact of power consumption and parameter fluctuations on future circuit performance to enable opportunities for further advancement of gigascale integration. He is currently performing post-doctoral research with Professor James D. Meindl at Georgia Tech.

In the summer of 2000, he performed research in modeling the impact of die-to-die and within-die parameter fluctuations on the maximum clock frequency distribution for a 0.25 μm microprocessor while interning with the Technology CAD Division at Intel Corporation, Santa Clara, CA.

Lihui Wang (S'01) was born in Hubei, China, on January 1, 1975. He received the B.S. degree in microelectronics from Peking University, Beijing, China, in 1997 and is currently pursuing the Ph.D. degree in electrical engineering at the Georgia Institute of Technology, Atlanta.

His research interests focus on the simulation and modeling of quantum effects in MOSFET's as well as device and circuit optimization for gigascale integration.

Xinghai Tang (S'96–M'99) was born in Heilongjiang, China. He received the B.S. and M.S. degrees in electrical engineering from Beijing Institute of Technology, China, in 1985 and 1988, respectively, and the Ph.D. degree in electrical engineering from the Georgia Institute of Technology, Atlanta, in 1999.

He is currently with the Somerset Design Center of Motorola, Inc., Austin, TX. Since joining Motorola in 1999, he has been analyzing circuit performance variations resulting from process and application variations for future generation Power PC microprocessors. His primary research interests are high performance/low power circuit techniques as embodied in more than 20 technical publications in refereed international conferences and journals.

James D. Meindl (M'56–SM'66–F'68–LF'97) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Carnegie Mellon University (CMU), Pittsburgh, PA, in 1955, 1956, and 1958, respectively.

He is presently the director of the Joseph M. Pettit Microelectronics Research Center and the Pettit Chair Professor of Microelectronics at the Georgia Institute of Technology, Atlanta. Previously. he served from 1986 to 1993 as Senior Vice President for Academic Affairs and Provost of Rensselaer Polytechnic Institute, Troy, NY. From 1967 through 1986, he was with Stanford University, Stanford, CA, where he was John M. Fluke Professor of Electrical Engineering, Associate Dean for Research in the School of Engineering, Director of the Center for Integrated Systems, Director of the Electronics Laboratories, and founding Director of the Integrated Circuits Laboratory. He is a co-founder of Telesensory Systems, Inc., the principal manufacturer of electronic reading aids for the blind, and served as a member of the Board from 1971 through 1984. From 1965 through 1967, he was founding Director of the Integrated Electronics Division at the Fort Monmouth, New Jersey, U.S. Army Electronics Laboratories. He is the author of the book *Micropower Circuits* and over 500 technical papers on ultra large scale integration, integrated electronics, and medical electronics; and editor of the book *Brief Lessons in High Technology*, which elucidates the most important economic event of our lives, the emergence of the information society. His major contributions have been new medical instruments enabled by custom integrated electronics, projections and codification of the hierarchy of physical limits on integrated electronics, and leadership in creation of academic environments promoting high quality teaching and research.

Dr. Meindl is a Life Fellow of the American Association for the Advancement of Science, and a member of the American Academy of Arts and Sciences and the National Academy of Engineering and its Academic Advisory Board. He most recently was awarded the Georgia Institute of Technology Distinguished Professor Award. He received the IEEE Third Millennium Medal, the 1999 SIA University Research Award, the 1997 Hamerschlag Distinguished Alumnus Award from CMU, and the 1991 Benjamin Garver Lamme Medal from ASEE. He was the recipient of the 1990 IEEE Education Medal "for establishment of a pioneering academic program for the fabrication and application of integrated circuits" and the recipient of the 1989 IEEE Solid-State Circuits Medal for contributions to solid-state circuits and solid-state circuit technology. At the 1988 IEEE International Solid-State Circuits Conference, he received the Beatrice K. Winner Award. In 1980, he was the recipient of the IEEE Electron Devices Society's J.J. Ebers Award for his contributions to the field of medical electronics and for his research and teaching in solid-state electronics. From 1970 through 1978, he and his students received five outstanding paper awards at IEEE International Solid-State Circuits Conferences, along with one received at the 1985 IEEE VLSI Multilevel Interconnections Conference.