A Compact Transregional Model for Digital CMOS Circuits Operating Near Threshold

Sean Keller, *Member, IEEE*, David Money Harris, *Member, IEEE*, and Alain J. Martin, *Member, IEEE*

***Abstract*— Power dissipation is currently one of the most important design constraints in digital systems. In order to reduce power and energy demands in the foremost technology, namely CMOS, it is necessary to reduce the supply voltage to near the device threshold voltage. Existing analytical models for MOS devices are either too complex, thus obscuring the basic physical relations between voltages and currents, or they are inaccurate and discontinuous around the region of interest, i.e., near threshold. This paper presents a simple transregional compact model for analyzing digital circuits around the threshold voltage. The model is continuous, physically derived (by way of a simplified inversion-charge approximation), and accurate over a wide operational range: from a few times the thermal voltage to approximately twice the threshold voltage in modern technologies.**

***Index Terms*— Circuit simulation, digital circuits, EKV, integrated circuit modeling, low-power electronics, minimum-energy point, near-threshold CMOS, subthreshold CMOS.**

I. Introduction

POWER and energy dissipation are critical design constraints in modern digital systems. Minimizing power and energy consumption in CMOS—the dominant digital circuit technology—requires supply voltage scaling below the process nominal supply voltage (V_{DD}). The minimum-energy operating point can occur below the device threshold voltage (V_t) or above it, and is a function of process parameters and environmental factors (such as activity factor) [1]–[4]. Even with additional constraints (e.g., performance, reliability, yield), the energy-optimal operating point typically occurs near the threshold voltage [5]–[8]. For these reasons, there is considerable interest in the analysis of circuits operating near threshold.

Modeling and analysis in this region of interest, around the threshold voltage, is complicated by the fact that even a rather narrow range of a few hundred millivolts around V_t spans three distinct MOSFET operating regimes: weak inversion, moderate inversion, and strong inversion. Conventional compact digital MOSFET models—the linear/quadratic strong-inversion model [9], the alpha-power-law model [10], [11], and the exponential weak-inversion model [9]—are discontinuous and inaccurate around V_t. Accurate continuous models exist [12], and some have been applied to digital circuit analysis. Nevertheless, it is apparent from [4] that even the simplest of continuous models are difficult to work with and yield complicated expressions for digital circuits (e.g., delay and energy) that somewhat obscure the relationship to supply voltage.[1] This relational complexity speaks to a clear need for MOS models that are simple enough to work with and reason about, while being sufficiently accurate to yield usable results. One of the goals of this paper is to address this problem; that is, to clarify the energy and delay relationship to the supply voltage (near threshold) by deriving a new simplified drain current model.

Compact MOS models are usually developed to be used in conjunction with numerical solvers and circuit simulators, as opposed to being designed for hand calculations. The most accurate of these models tend to have the greatest computational complexity and are the most difficult to work with by hand, while the simplest have reduced computational complexity at the expense of accuracy. Circuit simulation, along with the associated models, certainly plays an important role in digital system design; however, simple models and hand analysis can give the designer deeper insight into key tradeoffs, potential circuit problems, and optimizations than can be achieved by simulation alone. This paper presents a MOS device model designed specifically for hand calculations involving digital circuits.

Toward the goal of reducing model complexity, a number of simplifications are made throughout this paper. One such simplification reduces the drain-current (I_{ds}) model to a digital current model. In digital circuit design, first-order approximations for important characteristics (e.g., energy and delay) of large gate networks require only two MOSFET models: 1) the drain current of a logically "ON" transistor (I_{ON}) as a function of V_{DD} and 2) the drain current of a logically "OFF" transistor (I_{OFF}). Simple but accurate models for I_{OFF} exist, and those that include short channel effects are adequate. On the other hand, there is a need for new I_{ON} models that are accurate across all operating regimes. Using I_{ON} and I_{OFF} in lieu of a general I_{ds} model eliminates a number of variables and reduces model complexity but is only appropriate for digital applications.

This paper presents a simple physically derived inverted-charge MOS device model for I_{ON} (41) that is accurate for supply voltages ranging from a few times the thermal voltage to approximately twice the threshold voltage in modern

[1] Markovic *et al.*[4] acknowledge this complexity.

Manuscript received November 29, 2012; revised May 28, 2013; accepted August 20, 2013. Date of publication November 1, 2013; date of current version September 23, 2014. This work was supported by the National Science Foundation.

S. Keller and A. J. Martin are with the Department of Computer Science, California Institute of Technology, Pasadena, CA 91125 USA (e-mail: sean@async.caltech.edu; alain@async.caltech.edu).

D. M. Harris is with the Department of Engineering, Harvey Mudd College, Claremont, CA 91711 USA (e-mail: david_harris@hmc.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVLSI.2013.2282316

technologies; i.e., it is transregional. Since this model is approximately centered at V_t, it is referred to throughout as the near-threshold model. The model is continuous and also continuous in the first derivative; it makes use of three process-independent fitting parameters, and these parameters are stable. That is, these fitting parameters remain constant and the model remains accurate across different process technologies. Moreover, this paper shows the model to be accurate across four different commercial technologies from two different foundries ranging from 40 to 90 nm. The organization of the remainder of this paper is as follows. Section II gives the derivation of the near-threshold model. Section III applies the model derived in Section II to several problems. Section IV discusses related work, and Section V concludes this paper.

II. Compact Near-Threshold I_{ON} Model

It is tempting to avoid the considerable trouble of developing a physical model, and rather to use an empirical curve-fit as the foundation for a simplified transregional model. The problem with a purely empirical approach—even if the model is only intended for digital circuit analysis—is twofold. First, it is difficult to stabilize the model with respect to physical parameters that vary, such as the threshold voltage. Second, it is difficult to trust such a model; it is not clear how the fitting constants might change in new or different technologies. Fortunately, there are a number of established physical MOS models and approaches to compact modeling. One such approach, namely inversion-charge modeling, is used in this paper to generate the near-threshold model.

Inversion-charge models differ from the classic surface-potential-based models in that they make explicit the relation between MOS terminal voltages and the inversion-charge density (e.g., the charge due to electrons below the gate of an NFET). A continuous expression for drain current as a function of terminal voltages follows directly from this explicit relation when applied to the Pao–Sah [13] model. The inversion-charge density to terminal voltage relation is difficult to compactly model, and the choice of simplifying approximations is a key differentiating factor between inversion-charge models.

The goal of this section is to derive a new analytical expression for the NFET drain current of an "ON" transistor, I_{ON}, where I_{ON} is defined as the drain current when $V_{\mathrm{gb}} = V_{\mathrm{db}} = V_{\mathrm{DD}}$ and $V_{\mathrm{sb}} = 0V$. This expression for I_{ON} and the derivation are also applicable to a PFET; however, the corresponding derivation is not presented in this paper. The derivation begins with the quasi-static long-channel model for an NFET in terms of gate, source, and drain voltages, all relative to the bulk along the lines of the C. C. Enz, F. Krummenacher, and E. A. Vittoz (EKV) model derivation presented in [14]; as such, some of the content presented in Sections II-A and II-B is a review. It is a long/wide-channel inversion-charge model that makes use of the linearization of inversion charge to surface potential. The derivation starts with a well-accepted expression for drain current in terms of a diffusion component and a drift component, which is reduced to an expression where the drain current is proportional to a one-dimensional integral from the source potential to the drain potential of the mobile inversion charge (i.e., electrons) in the channel. A number of normalizations are applied to simplify the expression, and the integral is broken into an equivalent difference expression. Next, an expression is given for the mobile inversion charge in terms of the normalized gate, source, drain, and threshold voltages. This expression is directly solved for the mobile inversion charge without approximation—a task that previous works were unable to accomplish. Additionally, several approximations are explained, and a new approximation that yields the near-threshold model is presented. These approximations for mobile inversion charge can be directly applied to the integral expression for drain current to give drain current in terms of the terminal voltages. Finally, this drain-current expression is further simplified to give I_{ON} as a function of V_{DD}.

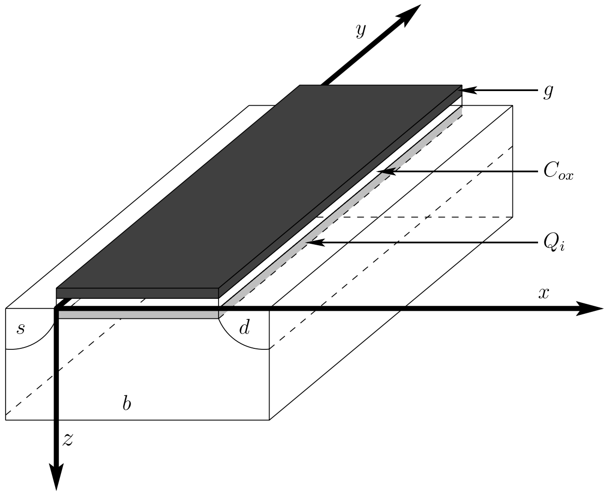


Fig. 1. NFET physical view.

A. Drain-Current Model

Consider an NFET labeled as in Fig. 1. The standard long/wide-channel expression for drain current is given by (1). See [12] and [15] for a full derivation and a discussion of the physical assumptions required for validity.

$$I_{\mathrm{ds}}(x) = \mu W \left(-Q_i' \frac{d\psi_s}{\mathrm{dx}} + \phi_t \frac{dQ_i'}{\mathrm{dx}} \right) \tag{1}$$

where μ is the effective electron mobility, W is the channel width, $Q_i'(x)$ is the mobile inversion charge per unit area as a function of position along the channel, ϕ_t is the thermal voltage,[2] and $\psi_s(x)$ is the surface potential (the potential drop from the semiconductor surface to deep into the body). The first term, $-Q_i'(d\psi_s/\mathrm{dx})$, models drift, and the second term, $\phi_t(dQ_i'/\mathrm{dx})$, models diffusion.

Assuming a constant channel width, constant electron mobility, the charge sheet approximation (the entire mobile inversion charge is at the surface potential), and the gradual channel approximation (the electric field along the z-axis is much larger than that along the x-axis), (1) reduces to (2); see

[2] Note that $\phi_t = (k_{BT}/q)$, where k_B is the Boltzmann constant, T is the absolute temperature, and q is the magnitude of the electrical charge on the electron.

[14] and [15] for a full discussion.

$$I_{\mathrm{ds}} = \mu C'_{\mathrm{ox}} \frac{W}{L} \int_{V_s}^{V_d} \frac{-Q'_i}{C'_{\mathrm{ox}}} dV_c \tag{2}$$

where C'_{ox} is the oxide capacitance per unit area, L is the channel length, and $V_c(x)$ is the channel potential which represents the quasi-Fermi potential of electrons in the channel as a function of position; to the first order, it varies monotonically from the source to the drain, i.e., from V_s to V_d.[3]

For a fixed V_g and V_s, as V_d increases, the device eventually enters the saturation region. This is due to the drain end of the channel pinching off as it enters weak inversion and the mobile inversion charge becomes negligible. Intuitively, this happens anywhere along the channel where the channel voltage is sufficiently large. In general, this property can be stated as the assumption that

$$\lim_{V_c \to \infty} Q'_i = 0. \tag{3}$$

As such, (2) can be broken into two parts: a forward current I_f which is independent of V_d, and a reverse current I_r which is independent of V_s. That is

$$I_{\mathrm{ds}} = \underbrace{\mu C'_{\mathrm{ox}} \frac{W}{L} \int_{V_s}^{\infty} \frac{-Q'_i}{C'_{\mathrm{ox}}} dV_c}_{I_f} - \underbrace{\mu C'_{\mathrm{ox}} \frac{W}{L} \int_{V_d}^{\infty} \frac{-Q'_i}{C'_{\mathrm{ox}}} dV_c}_{I_r}. \tag{4}$$

In order to solve this integral, the relationship between the channel potential and the mobile inversion charge needs to be established; however, the precise relation is quite complicated. One simple approach is to assume a linear relationship between the mobile inversion charge and the surface potential. This greatly reduces the complexity of the problem and yields a constant of proportionality, n, which is the slope factor. From [16]

$$n = \frac{Q'_i}{C'_{\mathrm{ox}}(\psi_s - \psi_p)} \tag{5}$$

where ψ_p is the pinch-off surface potential, i.e., the surface potential at which the inversion charge becomes zero.

Before solving (4), it is convenient to normalize the terms to unitless quantities using

$$\begin{aligned} q_i &= \frac{-Q'_i}{2n\phi_t C'_{\mathrm{ox}}} \\ I_0 &= 2n\mu C'_{\mathrm{ox}} \frac{W}{L} \phi_t^2 \\ v_c &= \frac{V_c}{\phi_t} \\ i &= \frac{I}{I_0} \\ \frac{V_s}{v_s} &= \frac{V_d}{v_d} = \frac{V_g}{v_g} = \phi_t \\ \frac{I_{\mathrm{ds}}}{i_{\mathrm{ds}}} &= \frac{I_f}{i_f} = \frac{I_r}{i_r} = I_0. \end{aligned} \tag{6}$$

Equation (2) now simplifies to

$$i_{\mathrm{ds}} = \int_{v_s}^{v_d} q_i dv_c \tag{7}$$

and (4) becomes

$$i_{\mathrm{ds}} = \underbrace{\int_{v_s}^{\infty} q_i dv_c}_{i_f} - \underbrace{\int_{v_d}^{\infty} q_i dv_c}_{i_r}. \tag{8}$$

Since the model is symmetric with respect to the source and drain, the forward and reverse component are of the same form; it is convenient to use a combined notation so as to work with both expressions (i_f and i_r) simultaneously. That is

$$i_{f,r} = \int_{v_{s,d}}^{\infty} q_i dv_c. \tag{9}$$

Finally, all that is needed to solve (9) is an expression for q_i, thus yielding an expression for drain current in terms of the three transistor terminal voltages—a goal of this section.

In normalized terms, the relation between mobile inversion-charge density and channel potential can be expressed as (see [14] for details)

$$2q_i + \ln q_i = v_p - v_c \tag{10}$$

where $v_p = (V_p/\phi_t)$ is the pinch-off voltage, defined in [17] and [18] as

$$v_p = \psi_p - \psi_0 \tag{11}$$

where ψ_0 is a process-dependent term with various approximations used in the literature. Conveniently, v_p can be approximated with common terms as

$$v_p \approx \frac{V_g - V_t}{n\phi_t} \tag{12}$$

where V_t is the threshold voltage [16], [17].

Equations (10) and (12) give the relation between the gate and channel potential and the mobile inversion charge with the process-dependent component compacted into the definition of v_p. Different [15] and more accurate [18] relations exist, but (10) is simple, practical, and differentiable

$$dv_c = -dq_i \left(2 + \frac{1}{q_i}\right). \tag{13}$$

Substituting this expression for dv_c into (9) and integrating results in

$$i_{f,r} = q_{s,d}^2 + q_{s,d} \tag{14}$$

where q_s is the normalized mobile inversion charge at the source end of the channel, and similarly for q_d at the drain end. Applying (10) to the source and drain ends of the channel yields

$$v_p - v_{s,d} = 2q_{s,d} + \ln q_{s,d}. \tag{15}$$

Prior work (see [14] and [18]) assumed that (15) [and (10)] is not invertible, but it actually can be inverted by using the principal branch of the Lambert W function. The Lambert W function is defined as the root of

$$\mathsf{W}(z)e^{\mathsf{W}(z)} = z \tag{16}$$

[3] Terminal voltages are body-referenced unless otherwise specified.

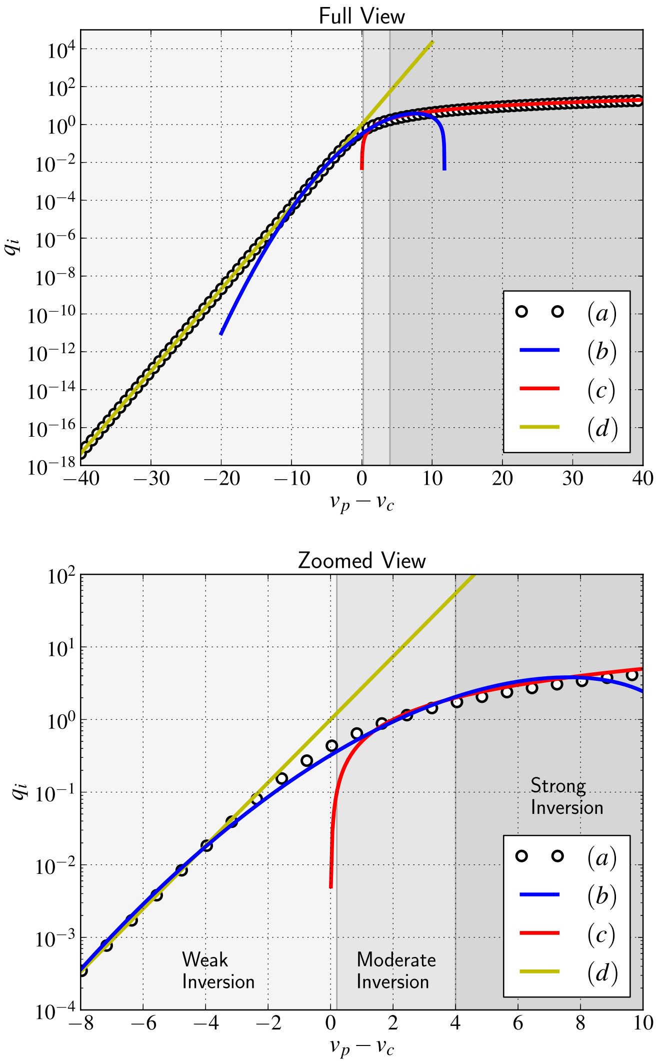


Fig. 2. (a) Equation (19). (b) Equation (35) (near-threshold Model). (c) Equation (26) (strong-inversion approximation). (d) Equation (22) (weak-inversion approximation).

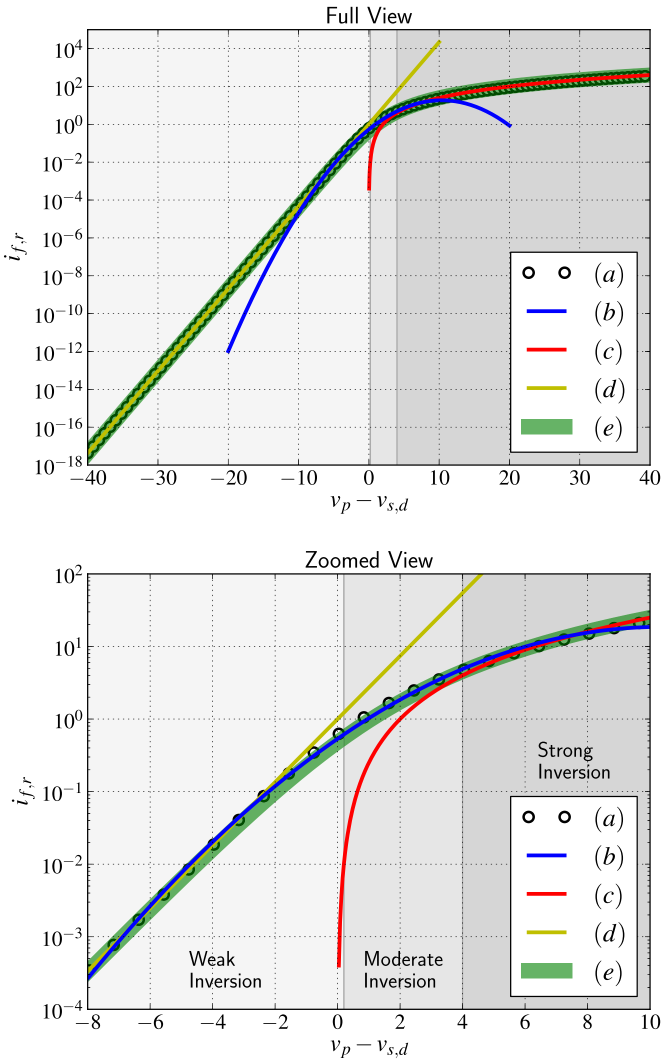


Fig. 3. (a) Equation (20). (b) Equation (36) (near-threshold approximation). (c) Equation (27) (strong-inversion approximation). (d) Equation (23) (weak-inversion approximation). (e) Equation (28) (EKV continuous approximation).

for any complex number z, (see [19] for details). The function dates back to the days of Euler and has been recently used in several related works (see Section IV).

After exponentiation, (15) can be rearranged as

$$2q_{s,d}e^{2q_{s,d}} = 2e^{v_p - v_{s,d}}. \tag{17}$$

Applying the Lambert W function[4] to (17) gives the closed-form expression

$$q_{s,d} = \frac{\mathrm{W}_0\left(2e^{v_p - v_{s,d}}\right)}{2}. \tag{18}$$

Analogously, applying the Lambert W function to (10) gives the closed-form expression

$$q_i = \frac{\mathrm{W}_0\left(2e^{v_p - v_c}\right)}{2} \tag{19}$$

[depicted in Fig. 2(a)]. This expression for q_i proves useful for making approximations in Section II-C.

Finally, (18) can be directly applied to (14), giving a new closed-form expression for normalized drain current

$$i_{f,r} = \left(\frac{\mathrm{W}_0\left(2e^{v_p - v_{s,d}}\right)}{2}\right)^2 + \frac{\mathrm{W}_0\left(2e^{v_p - v_{s,d}}\right)}{2}. \tag{20}$$

This expression for $i_{f,r}$ is exact, while the EKV approximation [14], discussed in Section II-B and given by (28), has a maximum absolute error of 21%. Using a more accurate approximation for inversion charge, e.g., [18] and then using the Lambert W function to give an exact expression for inversion charge may further improve total model accuracy; however, this analysis falls outside of the scope of this paper and is left as future work.

Fig. 3(a) depicts (20), and it makes clear the nonlinear nature of drain current as a function of the terminal voltages. It also helps relate $i_{f,r}$ to the standard operating regimes: weak, moderate, and strong inversion. The model presented in this paper is symmetric with respect to the source and drain; however, the ultimate goal of this derivation is to generate a model for I_{ON} wherein the drain end of the channel is tied

[4] When the domain of the Lambert W function is restricted to the nonnegative reals, the co-domain reduces to that of the reals, and $\mathrm{W}(z)$ has a single value denoted by the principal branch $\mathrm{W}_0(z)$.

TABLE I
OPERATING REGIME BOUNDS FOR I_{ON}

Operating Regime	Current Bounds	Potential Bounds
Weak Inversion	$i_f < -1.4$	$v_p - v_s < 0.20$
Moderate Inversion	$-1.4 \leq i_f < 3.6$	$0.20 \leq v_p - v_s < 4.0$
Strong Inversion	$3.6 \leq i_f$	$4.0 \leq v_p - v_s$

to V_{DD} and the source end to the body. From this and (15), it follows that $q_s > q_d$,[5] i.e., the inversion charge density at the source end of the channel always exceeds that of the drain end. From (14), it follows that $i_f > i_r$, and so the operating regime is determined exclusively by i_f and correspondingly $v_p - v_s$. The drain end of the channel and the drain dependent current i_r are pinned in the weak-inversion regime. The boundaries between the operating regimes are approximated in Table I.[6] It should be noted that, in the general case, the operating regime can be determined by the larger of i_f or i_r.

B. Existing Drain-Current Approximations

There are three well-accepted approximations for $i_{f,r}$: a simple weak-inversion approximation, a simple strong-inversion approximation, and a continuous approximation which is valid in all operating regions. The weak- and strong-inversion approximations, along with the new near-threshold model, are generated by modeling the mobile inversion charge as a function of the terminal voltages; Figs. 2 and 3 graphically depict these charge and current approximations, respectively.

In weak inversion, $v_p - v_c \ll 0$, and from (10) it follows that $2q_i + \ln q_i \ll 0$. The logarithmic term dominates, so

$$v_p - v_c \approx \ln q_i \tag{21}$$

$$q_i \approx e^{v_p - v_c}. \tag{22}$$

[See Fig. 2(d).] Integrating (9) with this approximation gives

$$i_{f,r} \approx e^{v_p - v_{s,d}} \tag{23}$$

which is depicted in Fig. 3(d). Removing the normalization, letting the approximation become an equality, and combining the forward and reverse components yields a well-known equation for subthreshold drain current

$$I_{\mathrm{ds}} = I_0 e^{\frac{V_g - V_t}{n\phi_t}} \left(e^{\frac{-V_s}{\phi_t}} - e^{\frac{-V_d}{\phi_t}} \right). \tag{24}$$

In strong inversion, $v_p - v_c \gg 0$, so the logarithmic term in (10) is negligible. That is

$$v_p - v_c \approx 2q_i \tag{25}$$

$$q_i \approx \frac{v_p - v_c}{2}. \tag{26}$$

[See Fig. 2(c).] With (9)

$$i_{f,r} \approx \left(\frac{v_p - v_{s,d}}{2} \right)^2 \tag{27}$$

as depicted in Fig. 3(c).

Finally, the continuous EKV approximation [14] given by

$$i_{f,r} \approx \ln^2 \left[1 + e^{\frac{v_p - v_{s,d}}{2}} \right] \tag{28}$$

as depicted in Fig. 3(e), is accurate over all operating regimes (at the expense of increased complexity).

C. Transregional Near-Threshold Drain-Current Approximation

This subsection presents a new inversion-charge approximation and the corresponding drain-current approximation for digital circuits. The model is simpler than the EKV model and continuously models digital devices operating across weak, moderate, and strong inversion. Consider (10); in weak inversion, the logarithmic terms dominates and in strong inversion the linear term dominates. In moderate inversion, neither term dominates, so a simple approximation is not possible. Fortunately, the exact expression for charge, i.e., (19), can be simplified for a narrow range of $v_p - v_c$.

First, assuming that $q_i > 0$, taking the logarithm of both sides of (19) gives

$$\ln q_i = \ln \mathsf{W}_0 \left(2e^{v_p - v_c} \right) - \ln 2. \tag{29}$$

Next, from [20], for $x > 0$ and $\mathsf{W}_0(x) > 0$

$$\ln \mathsf{W}_0(x) = \ln x - \mathsf{W}_0(x). \tag{30}$$

As such, (29) can be expressed as

$$\ln q_i = v_p - v_c - \mathsf{W}_0(2e^{v_p - v_c}). \tag{31}$$

Next, [19] shows that $\mathsf{W}(e^x)$ can be approximated by a Taylor series expansion. As such, (31) can be written as

$$\ln q_i \approx v_p - v_c - P(v_p - v_c) \tag{32}$$

for some polynomial P, wherein the coefficients and validity range are a function of $v_p - v_c$. The optimal polynomial approximation for a particular range of interest can be calculated to a high degree, but this does not aid in the simplification of the problem at hand. The approach taken in this paper is to use a degree-2 polynomial and to curve-fit the entire expression. That is

$$\ln q_i \approx k_a + k_b(v_p - v_c) + k_c(v_p - v_c)^2 \tag{33}$$

where k_a, k_b, and k_c are fitting constants.

Finally, letting $k_f = e^{k_a}$, $v_\omega = v_p - v_c$, and exponentiating both sides of (33) gives

$$q_i \approx k_f e^{k_b v_\omega + k_c v_\omega^2}. \tag{34}$$

In order to calculate $i_{f,r}$, integration is necessary, so it is helpful to approximate (34) as

$$q_i \approx k_0(k_1 + 2k_2 v_\omega) e^{k_1 v_\omega + k_2 v_\omega^2} \tag{35}$$

where k_0, k_1, and k_2 are new fitting constants.[7] (See Fig. 2(b) for a graphical depiction.)

[5] This requires that V_{DD} is a positive value relative to the body.

[6] Analytical bounds on operating regimes can be found in [12].

[7] This is valid because taking the logarithm of both sides of (35) gives $\ln q_i \approx \ln k_0(k_1 + 2k_2 v_\omega) + k_1 v_\omega + k_2 v_\omega^2$. Using the first few terms of the Taylor series for the ln term on the RHS reduces the entire RHS to a polynomial with new coefficients, i.e., $\ln q_i \approx P(v_\omega)$. As such, removing high-order terms and exponentiating both sides gives (34).

TABLE II

NEAR-THRESHOLD MODEL FITTING CONSTANTS—ERROR REPORTED FOR $i_{f,r}$ [i.e., (36) COMPARED TO (20)]

	Value
k_0	5.4e-1
k_1	6.9e-1
k_2	-3.3e-2
Mean Absolute Error	8.1%
Maximum Absolute Error	21%

After substituting (35) into (9) and integrating, the resulting expression for normalized drain current can be expressed as

$$\mathbf{i_{f,r} \approx k_0 e^{k_1 v_\varpi + k_2 v_\varpi^2}} \tag{36}$$

where $\upsilon_\varpi = \upsilon_p - \upsilon_{s,d}$. Fitting this expression in the near-threshold region, $-8 < \upsilon_p - \upsilon_{s,d} < 10$, (see Section II-D for boundary definition) gives the fitting constants given in Table II (used throughout this paper). (See Fig. 3(b) for a graphical depiction.)

Note that, because of the definition of the pinch-off voltage and the use of normalized variables, the fitting constants (k_0, k_1, k_2) are process-independent.

Finally, combining this expression for $i_{f,r}$ [i.e., (36)] with (8) gives

$$i_{\mathrm{ds}} = k_0 e^{k_1(\upsilon_p-\upsilon_s)+k_2(\upsilon_p-\upsilon_s)^2} - k_0 e^{k_1(\upsilon_p-\upsilon_d)+k_2(\upsilon_p-\upsilon_d)^2}. \tag{37}$$

Removing the normalization, and using (12) to approximate υ_p yields

$$\begin{aligned} I_{\mathrm{ds}} &= I_0 k_0 e^{k_1\left(\frac{V_g-V_t}{n\phi_t}-\frac{V_s}{\phi_t}\right)+k_2\left(\frac{V_g-V_t}{n\phi_t}-\frac{V_s}{\phi_t}\right)^2} \\ &\quad - I_0 k_0 e^{k_1\left(\frac{V_g-V_t}{n\phi_t}-\frac{V_d}{\phi_t}\right)+k_2\left(\frac{V_g-V_t}{n\phi_t}-\frac{V_d}{\phi_t}\right)^2}. \end{aligned} \tag{38}$$

Now, referencing all voltages to the source instead of the body and assuming that $V_{\mathrm{sb}} = 0V$ gives

$$\begin{aligned} I_{\mathrm{ds}} &= I_0 k_0 e^{k_1\frac{V_{\mathrm{gs}}-V_t}{n\phi_t}+k_2(\frac{V_{\mathrm{gs}}-V_t}{n\phi_t})^2} \\ &\quad \times \left(1 - e^{k_1\frac{-V_{\mathrm{ds}}}{\phi_t}+k_2\frac{n^2V_{\mathrm{ds}}^2-2nV_{\mathrm{ds}}V_{\mathrm{gs}}+2nV_{\mathrm{ds}}V_t}{n^2\phi_t^2}}\right). \end{aligned} \tag{39}$$

For I_{ON}, $V_{\mathrm{DD}} = V_{\mathrm{gs}} = V_{\mathrm{ds}}$, so

$$\begin{aligned} I_{\mathrm{ON}} &= I_0 k_0 e^{k_1\frac{V_{\mathrm{DD}}-V_t}{n\phi_t}+k_2\left(\frac{V_{\mathrm{DD}}-V_t}{n\phi_t}\right)^2} \\ &\quad \times \left(1 - e^{k_1\frac{-V_{\mathrm{DD}}}{\phi_t}+k_2\frac{n^2V_{\mathrm{DD}}^2-2nV_{\mathrm{DD}}^2+2nV_{\mathrm{DD}}V_t}{n^2\phi_t^2}}\right). \end{aligned} \tag{40}$$

Assuming that V_{DD} is both a few times larger than ϕ_t and less than twice the threshold voltage allows the terms within parentheses to be approximated as unity, and letting $V_{\mathrm{DT}} = V_{\mathrm{DD}} - V_t$, yields

$$\mathbf{I_{\mathrm{ON}} = I_0 k_0 e^{k_1\frac{V_{DT}}{n\phi_t}+k_2\left(\frac{V_{DT}}{n\phi_t}\right)^2}}. \tag{41}$$

Equation (41) gives the drain current of a logically "ON" transistor as a function of the supply voltage—the goal of this section and one of the main goals of this paper. Within this expression, the constants k_0, k_1, and k_2 are process-independent, and the process-dependent terms are contained in the definitions of I_0, n, and V_t. The definition of I_0 [i.e., (6)] also contains the sizing ratio (W/L). This ratio is intentionally kept within the definition of I_0 throughout, because, in short/narrow-channel devices, modifying gate dimensions can affect some or all of the process dependent terms. As with most compact models, short/narrow-channel effects can be included in the near-threshold model as needed.[8] Additionally, regions of validity for both W and L can be established before using the near-threshold model (or any compact model) to calculate drain current as a function of either term.

D. Near-Threshold Model Validation

Fig. 3 depicts the different approximations for normalized drain current as a function of the transistor terminal voltages. It is clear that each approximation has a particular region of validity. The region boundaries are difficult to determine analytically but can be readily defined in terms of a maximum error. The original EKV approximation, i.e., (28), has a maximum absolute error of 21% compared to the analytical drain-current expression given by (20). The EKV approximation is a useful and well-accepted model, so the corresponding maximum absolute error of 21% against (20) can also be used as a validity bound for the other drain-current approximations. Table III provides the region boundaries in terms of both normalized voltages and currents, along with the mean absolute error.

The near-threshold model is further validated by application to four commercial bulk CMOS processes from two different foundries. Nominal devices, high-threshold transistors (HVT), and low-threshold transistors (LVT) are modeled in a 40-nm low-power (LP) technology, a 65-nm LP technology, a 65-nm general-purpose (GP) technology, and a 90-nm GP technology.[9] The foundry-provided BSIM4 models for each technology node are used as the basis for comparison and for parameter extraction. Parameter extraction is performed by way of a least-squares fit. This method of parameter extraction is common and convenient, but it has shortcomings. In simplified models, such as those presented in this paper, the extracted parameters may not correspond to the physical parameters that they are intended to represent. This is especially true of parameters that are greatly impacted by short-channel effects, e.g., V_t which is affected by drain-induced barrier lowering (DIBL) [21].

Figs. 4 and 5 each overlay the near-threshold model on top of the corresponding BSIM4 simulation of the 40-nm LP and 65-nm GP technologies, respectively. In these figures, the near-threshold model is plotted for the entire V_{DD} range to make clear how the model deviates outside of its range of applicability. Table IV gives the lower and upper bounds on model applicability along with error rates (relative to BSIM4 simulation) and extracted parameter values. Table IV also specifies the device dimensions and provides the data for

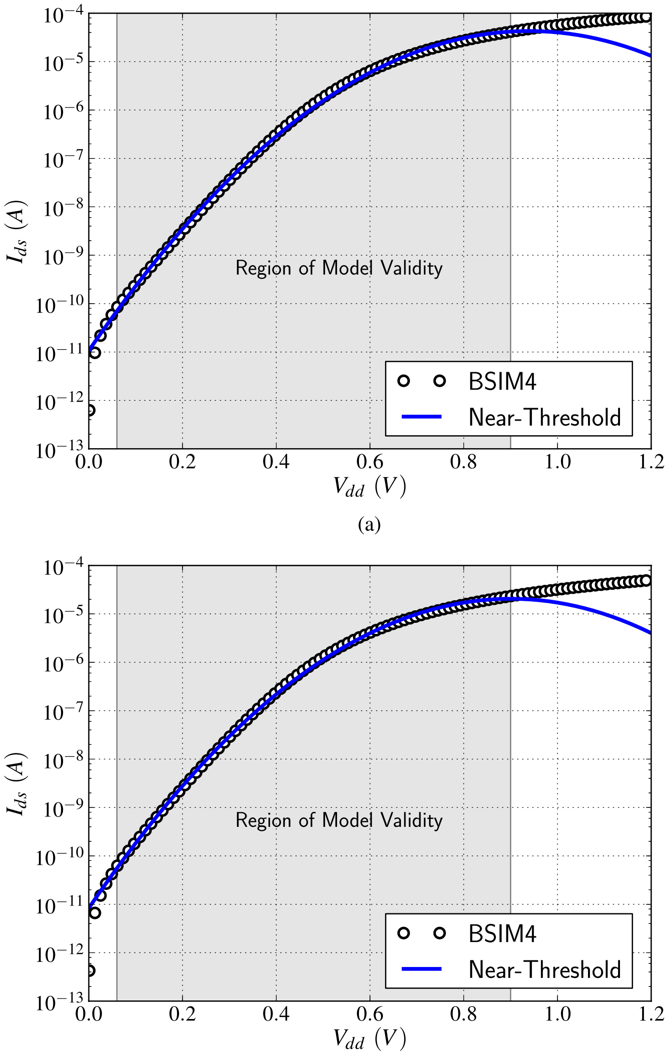
[8] See [12] for an example of incorporating short-channel effects into a strong-inversion model.

[9] The 90-nm technology only includes nominal and LVT devices, so 90-nm HVT devices are not modeled.

TABLE III

MODEL VALIDITY REGIONS (BOUNDED BY A MAXIMUM ABSOLUTE ERROR OF 21%)

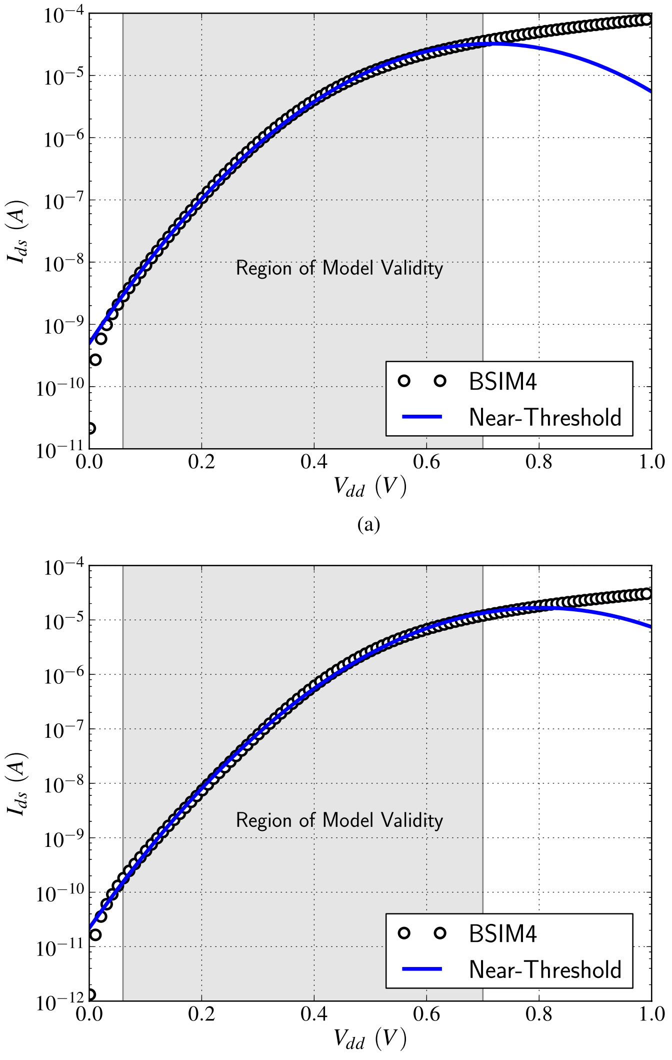
Approximation	$\min(v_p - v_{s,d})$	$\max(v_p - v_{s,d})$	$\min(i_f)$	$\max(i_f)$	Mean Absolute Error
EKV Continuous	< -40	> 40	$< 4.2e{-}18$	$> 3.6e2$	6.9%
Weak Inversion	< -40	-1.4	$< 4.2e{-}18$	0.20	0.59%
Strong Inversion	3.6	> 40	4.0	$> 3.6e2$	11%
Near-Threshold	-8.0	10	$3.4e{-}4$	23	8.1%



(a)

(b)

Fig. 4. Equation (36) (near-threshold model) plotted for entire V_{DD} range against SPICE simulation of BSIM4 model of a 40-nm LP process with minimum-size devices. (a) NFET. (b) PFET.



(a)

(b)

Fig. 5. Equation (36) (near-threshold model) plotted for entire V_{DD} range against SPICE simulation of BSIM4 model of a 65-nm GP process with minimum-size devices. (a) NFET. (b) PFET.

LVT and HVT devices (where applicable). Note that the error associated with HVT devices tends to be greater than that of the corresponding regular devices. This can be attributed to modeling error at the low end of the V_{DD} range; that is, with HVT devices, the quantity $v_p - v_{s,d}$ can be less than the near-threshold model lower bound given in Table III.

III. NEAR-THRESHOLD MODEL APPLICATIONS

The goal of this section is to demonstrate the applicability of the near-threshold model to digital circuit analysis in a modern technology. The model is first used to generate a closed-form analytical expression for delay, which is then used to give a closed-form equation for energy, and this is used to determine the minimum-energy operating point as a function of activity factor and frequency. Finally, parameter variation is incorporated into the model, and closed-form expressions for the stochastic path-delay are derived. All of these analyses, which yield closed-form expressions, leverage the simplicity of a digital I_{ON} model designed for hand calculations.

Model validity is determined by comparing the analytical expressions against corresponding BSIM4 SPICE simulations,

TABLE IV

NEAR-THRESHOLD MODEL COMPARED TO SPICE SIMULATION OF BSIM4 MODEL FOR COMMERCIAL TECHNOLOGIES (AT 70 °C). THE CIRCUIT PARAMETERS V_t, I_0, AND n ARE EXTRACTED FROM A LEAST-SQUARES FIT AGAINST THE CORRESPONDING BSIM4 SIMULATION

Technology	Device	V_t (mV)	L (nm)	W (nm)	$I_0(A)$	n	Maximum Absolute Error	Mean Absolute Error	Lower Bound (mV)	Upper Bound (mV)
40-nm Low Power	NFET	485	36	108	$2.30e-6$	1.48	15%	8.5%	60	900
40-nm Low Power	PFET	458	36	108	$1.10e-6$	1.24	13%	7.4%	60	900
40-nm Low Power	HVT NFET	586	36	108	$2.30e-6$	1.58	20%	12%	60	1100
40-nm Low Power	HVT PFET	572	36	108	$1.23e-6$	1.56	21%	13%	60	1100
40-nm Low Power	LVT NFET	415	36	108	$1.34e-6$	1.49	17%	8.3%	60	900
40-nm Low Power	LVT PFET	425	36	108	$2.67e-6$	1.45	16%	8.4%	60	900
65-nm General Purpose	NFET	310	50	100	$1.74e-6$	1.32	9.8%	4.9%	60	700
65-nm General Purpose	PFET	392	50	100	$8.99e-7$	1.34	12%	7.6%	60	700
65-nm General Purpose	HVT NFET	371	50	100	$1.83e-6$	1.34	12%	7.7%	60	700
65-nm General Purpose	HVT PFET	475	50	100	$1.11e-6$	1.50	11%	5.0%	60	700
65-nm General Purpose	LVT NFET	273	50	100	$2.22e-6$	1.33	13%	6.4%	60	700
65-nm General Purpose	LVT PFET	348	50	100	$8.76e-7$	1.32	9.3%	5.2%	60	700
65-nm Low Power	NFET	504	60	120	$1.86e-6$	1.46	19%	11%	60	900
65-nm Low Power	PFET	504	60	120	$1.12e-6$	1.50	16%	9.1%	60	900
65-nm Low Power	HVT NFET	610	60	120	$1.86e-6$	1.50	22%	13%	60	900
65-nm Low Power	HVT PFET	602	60	120	$1.25e-6$	1.59	21%	13%	60	900
65-nm Low Power	LVT NFET	373	60	120	$1.44e-6$	1.25	12%	6.3%	60	750
65-nm Low Power	LVT PFET	460	60	120	$1.28e-6$	1.52	11%	5.0%	60	750
90-nm General Purpose	NFET	304	80	120	$1.52e-6$	1.25	14%	7.0%	60	600
90-nm General Purpose	PFET	395	80	120	$1.24e-6$	1.30	15%	9.3%	60	600
90-nm General Purpose	LVT NFET	218	80	120	$1.24e-6$	1.18	21%	6.0%	60	600
90-nm General Purpose	LVT PFET	271	80	120	$3.62e-7$	1.20	6.2%	3.3%	60	600

and the errors are reported. For simplicity, chains of minimum-size inverters are used as a basis throughout; minimum-size devices are typical of circuits designed to minimize energy in the near-threshold region. Chains of other gates can be normalized to this basis; the delays of more complex digital circuits, e.g., a ripple-carry adder, track that of the inverter over a wide supply voltage range [6]. Similar analyses also make use of inverters as a canonical basis for analytical evaluation, see [3] and [22]. Furthermore, Table V (in Section III-A) reports the error (as compared to SPICE) when the closed-form delay model is applied to combinational gates other than minimum-size inverters.

A. Delay Model

Numerous delay models of varying accuracy and complexity have been used to model the switching delay of gates operating at superthreshold, see [1], [23], and [24]. For circuits operating subthreshold and near threshold, [2]–[4], [6] use and validate a simple linear RC-delay model. That is, the delay of a gate can be approximated as

$$t_{\text{pd}} = k_f C_{\text{load}} \frac{V_{\text{DD}}}{I_{\text{ON}}} \tag{42}$$

where C_{load} is the load capacitance, and k_f is a small fitting constant. This fitting constant serves to normalize the RC time constant and is necessary because propagation delay more closely tracks the drain current of devices that are only partially "ON" [23].

Using the near-threshold model for I_{ON} [i.e., (41)], t_{pd} from (42) can be expressed as

$$t_{\text{pd}} = \frac{k_f C_{\text{load}}}{k_0 I_0} V_{\text{DD}} e^{-k_1 \frac{V_{\text{DT}}}{n\phi_t} - k_2 \left(\frac{V_{\text{DT}}}{n\phi_t}\right)^2}. \tag{43}$$

TABLE V

FO4 DELAY OF COMBINATIONAL GATES DETERMINED USING (44) AND COMPARED TO BSIM4 SPICE SIMULATION OF 65-nm GP PROCESS AT 70 °C. FIT FROM 170 TO 750 mV WITH $V_t = 386$ mV AND $n = 1.43$

Gate	$t_{pd}(ns)$ at $V_{DD} = V_t$	$\frac{C_{\text{load}}}{I_F}\left(\frac{ns}{V}\right)$	Maximum Absolute Error	Mean Absolute Error
INV_1X	0.57	1.42	13%	8.3%
INV_4X	0.49	1.29	11%	5.7%
INV_8X	0.48	1.28	10%	5.6%
NAND2	1.2	3.03	19%	11%
NOR2	1.2	3.02	14%	7.8%
AOI21	2.4	5.42	29%	15%

Since I_0 is typically treated as a fitting constant, (43) can be simplified by combining the constants k_f, k_0, and I_0 into a single term I_F. This gives

$$t_{\text{pd}} = \frac{C_{\text{load}}}{I_F} V_{\text{DD}} e^{-k_1 \frac{V_{\text{DT}}}{n\phi_t} - k_2 \left(\frac{V_{\text{DT}}}{n\phi_t}\right)^2}. \tag{44}$$

In order to apply (44) to an inverter, separate delays for the PFET and NFET can be calculated. A simpler approach used in this paper is to calculate an average propagation delay that simultaneously models the delay of both types of devices, but this requires refitting I_0 and V_{DT}. Fig. 6 plots the FO4 delay of a minimum-size inverter in the 65-nm GP process. The near-threshold model is plotted against the BSIM4 model with the near-threshold model fit from 135 to 700 mV (the inverters do not function below 135 mV). The mean absolute error is 8.0%, and the maximum absolute error is 13% with $V_t = 386$ mV, $n = 1.43$, and $(C_{\text{load}}/I_F) = 1.42$ (ns/V).

Equation (44) can be applied to a variety of gates by fitting (C_{load}/I_F); Table V gives the corresponding error (as compared to the BSIM4 model) for the FO4 delay of

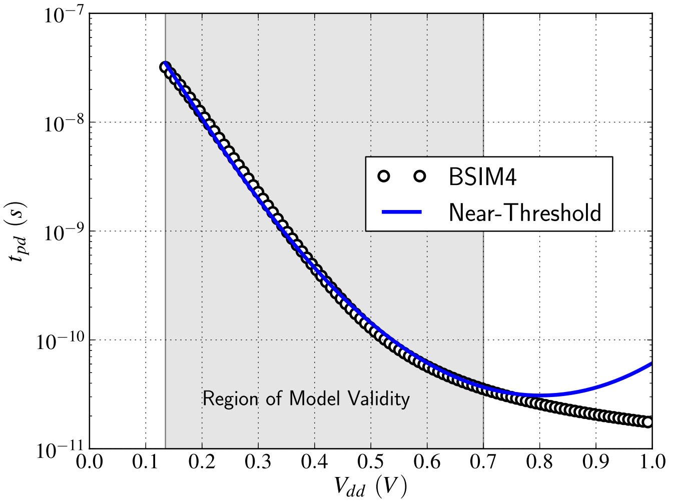


Fig. 6. Inverter FO4 delay, i.e., (44), plotted for entire V_{DD} range against a BSIM4 SPICE simulation of 65-nm GP process at 70 °C for a minimum-size inverter driving an FO4 load. Fit from 135 to 700 mV, yielding $V_t = 386$ mV, $n = 1.43$, and $(C_{\mathrm{load}}/I_F) = 1.42$ (ns/V).

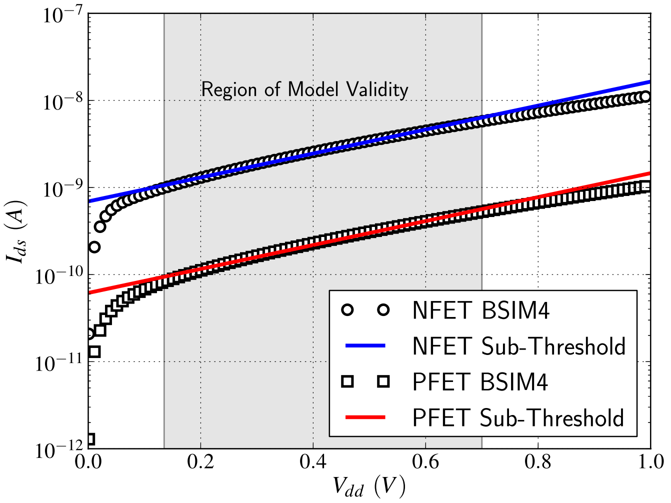


Fig. 7. Off-current, i.e., (51), plotted for the entire V_{DD} range against a BSIM4 SPICE simulation of 65-nm GP process at 70 °C for minimum-size devices with $V_t = 386$ mV, $n = 1.43$. Fit from 135 to 700 mV, resulting in $\eta = 0.134$, NFET $I_0 = 6.34$ μA, and PFET $I_0 = 0.564$ μA.

several combinational gates: a minimum-size inverter, a four-times minimum-width inverter, an eight-times minimum-width inverter, a NAND2 gate, a NOR2 gate, and an AOI21 gate.

B. Energy Model

The total energy dissipated by a CMOS circuit, E_{tot}, can be expressed as the summation of a dynamic component, E_{dyn}, corresponding to the charging and discharging of capacitance and a leakage component, E_{leak}, attributed to parasitic leakage current. Assuming periodic operation, e.g., clocking, the total energy can be defined in terms of energy per cycle. That is

$$E_{\mathrm{tot}} = \alpha E_{\mathrm{dyn}} + E_{\mathrm{leak}} \tag{45}$$

where α is the switching activity factor; it models the common case in which only a fraction (alpha) of the logic gates is in the circuit switch. There are numerous physical mechanisms behind leakage currents in modern MOSFETs [25], but in current technology nodes operating in the near-threshold region, sub-threshold drain-to-source current dominates [26]. The leakage energy is thus defined as

$$E_{\mathrm{leak}} = N_l I_{\mathrm{OFF}} V_{\mathrm{DD}} T_c \tag{46}$$

where T_c is the cycle time (typically the critical path delay), and N_l is the number of representative gates that are leaking in a cycle. The dynamic energy of the circuit, E_{dyn}, is defined as

$$E_{\mathrm{dyn}} = C_{\mathrm{dyn}} V_{\mathrm{DD}}^2 \tag{47}$$

where C_{dyn} represents the entire switching capacitance, i.e., it includes glitching and crowbar current. The cycle time can be defined in terms of a sequence of representative gates and corresponding delays as

$$T_c = t_d \qquad t_d = t_{\mathrm{pd}} L_{dp} \tag{48}$$

where t_d is the path delay, and L_{dp} is the number of gates on the path each with a delay of t_{pd}.

The off-current, I_{OFF}, for a single gate can be defined in terms of the subthreshold drain current from (24); letting $V_g = V_s = 0$ and $V_d = V_{\mathrm{DD}}$ gives

$$I_{\mathrm{OFF}} = I_0 e^{\frac{-V_t}{n\phi_t}} \left(1 - e^{\frac{-V_{\mathrm{DD}}}{\phi_t}}\right). \tag{49}$$

Assuming V_{DD} is a few times larger than the thermal voltage allows the terms in parentheses to be approximated as unity; that is

$$I_{\mathrm{OFF}} = I_0 e^{\frac{-V_t}{n\phi_t}}. \tag{50}$$

In modern technologies, the inclusion of short-channel effects in the off-current model can significantly improve model accuracy. For example, ignoring the effects of DIBL can result in an order of magnitude of error [6]. The effects of DIBL can be included by explicitly making V_t a function of V_{ds} [12]. That is, the effective threshold voltage becomes $V_t - \eta V_{\mathrm{DD}}$, where η is the DIBL factor. Substituting this value into (50) (in place of V_t) gives

$$I_{\mathrm{OFF}} = I_0 e^{\frac{\eta V_{\mathrm{DD}} - V_t}{n\phi_t}}. \tag{51}$$

Fig. 7 shows the application of the off-current equation to an NFET and PFET in the 65-nm GP process. The values of fitting bounds, n, and V_t are taken from the t_{pd} model detailed in Fig. 6. The least-squares fit value for η is 0.134, NFET $I_0 = 6.34$ μA, and PFET $I_0 = 0.564$ μA. For the NFET, there is a mean absolute error is 3.4% and a maximum absolute error of 10%; for the PFET there is a mean absolute error is 3.7% and a maximum absolute error of 15%.

The off-current equation (51) can be substituted into the expression for leakage energy (46). That is

$$E_{\mathrm{leak}} = I_0 V_{\mathrm{DD}} N_l T_c e^{\frac{\eta V_{\mathrm{DD}} - V_t}{n\phi_t}}. \tag{52}$$

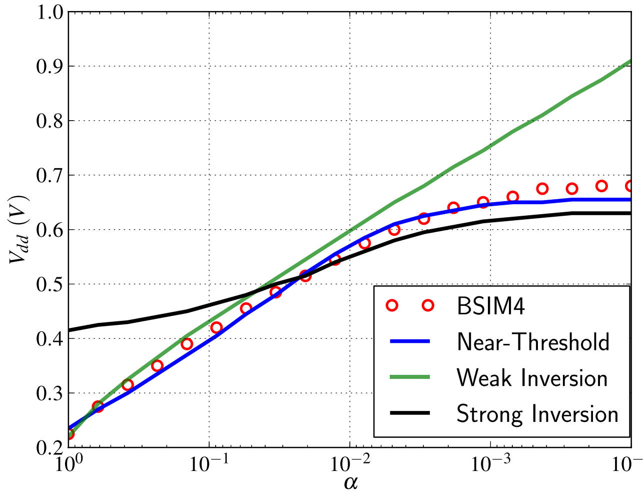


Fig. 8. Minimum-energy operating voltage versus activity factor (α). The circuit consists of a linear chain of 20 minimum-size inverters with FO4 loads in a 65-nm GP process at 70 °C.

TABLE VI

MINIMUM-ENERGY OPERATING VOLTAGE ERROR RELATIVE TO SPICE SIMULATION OF BSIM4 MODEL. THE CORRESPONDING PLOTS ARE DEPICTED IN FIG. 8

Model	Maximum Absolute Error	Mean Absolute Error
Near-Threshold	5.1%	2.6%
Weak Inversion	34%	11%
Strong Inversion	84%	15%

Combining this with the dynamic-energy equation (47) by way of (45) yields in an expanded expression for energy

$$E_{\text{tot}} = \alpha C_{\text{dyn}} V_{\text{DD}}^2 + I_0 V_{\text{DD}} N_l T_c e^{\frac{\eta V_{\text{DD}} - V_t}{n\phi_t}}. \tag{53}$$

Finally, expanding the term T_c with (48) and (44) results in the full expression for energy per cycle

$$\begin{aligned} E_{\text{tot}} &= \alpha C_{\text{dyn}} V_{\text{DD}}^2 \\ &\quad + N_l L_{\text{dp}} \frac{I_0}{I_F} V_{\text{DD}}^2 C_{\text{load}} e^{-k_1 \frac{V_{\text{DT}}}{n\phi_t} - k_2 \left(\frac{V_{\text{DT}}}{n\phi_t}\right)^2 + \frac{\eta V_{\text{DD}} - V_t}{n\phi_t}}. \end{aligned} \tag{54}$$

Equation (54) is continuously differentiable, and can be used to solve traditional and sensitivity-based optimization problems. For example, in the 65-nm GP process, for a chain of FO4 inverters, $C_{\text{dyn}} \approx 1.8 \text{ fF} * L_{\text{dp}}$. Using this value for C_{dyn}, with $L_{\text{dp}} = N_l = 20$, and the parameters from Figs. 6 and 7, Fig. 8 gives the minimum-energy operating voltage as a function of activity factor for the 65-nm GP process using (54). Fig. 8 also shows the minimum-energy operating voltage when the weak-inversion approximation [i.e., (23)] and the strong-inversion approximation [i.e., (27)] are used as models for I_{ON}. The errors for these approximations relative to SPICE simulation of the BSIM4 model are listed in Table VI. The strong-inversion approximation is a poor model for high activity factors, and the weak-inversion approximation is a poor model for low activity factors. The near-threshold model proves to be accurate for a wide range of activity factors.

C. Statistical Delay Model

Timing and delay play a critical role in digital circuit optimization, and in modern technologies the effects of parameter variation on path delay cannot be ignored. In order to account for parameter variation, static timing analysis (STA)—the most prominent method of delay analysis in digital circuit design—must incorporate statistical methods [e.g., by way statistical static timing analysis (SSTA)] [27]–[29]. Parameter variation can be modeled in a number of different ways, and a global corner model with local random variation is accurate but slightly pessimistic [30]. In this model, global variation affects all devices in the same way (e.g., the TT, FS, SF, SS corners), and local variation is truly random: i.e., neighboring identically drawn devices may behave differently. With local variation, the physical effects that dominate parameter variation depend on the operating region. In the subthreshold and near-threshold regions, parameter variation is dominated by random uncorrelated normally distributed V_t variation [31]. That is, when modeling the delay of circuits operating subthreshold or near-threshold region, for any particular global corner, the effects of parameter variation can be modeled by considering the V_t of each device as an independent normal random variable (RV). The goal of this section is to generate a closed-form stochastic delay model by way of the near-threshold delay model [i.e., (44)] with the new assumption that V_t is an RV.

If X is a normally distributed RV with mean denoted as μ_X, and variance denoted as σ_X^2, then the corresponding probability density function $f(X)$ is given by

$$f(X) = \frac{1}{\sigma_X \sqrt{2\pi}} e^{-\frac{(X-\mu_X)^2}{2\sigma_X^2}}. \tag{55}$$

For $g(X)$, a function of X, the expected value E can be calculated as

$$E[g(X)] = \int_{-\infty}^{\infty} g(X) f(X) dx \tag{56}$$

and the variance Var as

$$\text{Var}[g(X)] = E[(g(X)^2] - (E[g(X)])^2. \tag{57}$$

Treating V_t as a normally distributed RV with mean μ_{V_t} and standard deviation σ_{V_t}, the expected value of t_{pd} can be calculated by applying t_{pd} from (44) to (56). That is

$$\begin{aligned} &E[t_{\text{pd}}(V_t)] \\ &= \int_{-\infty}^{\infty} \frac{C_{\text{load}}}{I_F} \frac{V_{\text{DD}}}{\sigma_{V_t}\sqrt{2\pi}} e^{-k_1 \frac{V_{\text{DD}} - V_t}{n\phi_t} - k_2 \left(\frac{V_{\text{DD}} - V_t}{n\phi_t}\right)^2 - \frac{(V_t - \mu_{V_t})^2}{2\sigma_{V_t}^2}} dV_t. \end{aligned} \tag{58}$$

Similarly, applying t_{pd} from (44) to (57) gives the variance as

$$\begin{aligned} &\text{Var}[t_{\text{pd}}(V_t)] \\ &= \int_{-\infty}^{\infty} \frac{C_{\text{load}}^2}{I_F^2} \frac{V_{\text{DD}}^2}{\sigma_{V_t}\sqrt{2\pi}} e^{-2k_1 \frac{V_{\text{DD}} - V_t}{n\phi_t} - 2k_2 \left(\frac{V_{\text{DD}} - V_t}{n\phi_t}\right)^2 - \frac{(V_t - \mu_{V_t})^2}{2\sigma_{V_t}^2}} dV_t \\ &\quad - (E[t_{\text{pd}}(V_t)])^2. \end{aligned} \tag{59}$$

Owing to the form of (44), $\log(t_{\text{pd}}(V_t))$ is an RV with a noncentral χ^2 distribution, and $t_{\text{pd}}(V_t)$ can be approximated

TABLE VII

NEAR-THRESHOLD STATISTICAL DELAY MODEL (60) AND (61) COMPARED TO MC SPICE SIMULATIONS OF BSIM4 STATISTICAL MODEL FOR 65-nm GP CMOS FROM 300 TO 700 mV AT 100-mV INTERVALS (AT TT-CORNER, 70 °C, AND WITH 10 K MC TRIALS PER V_{DD} ACCOUNTING FOR LOCAL PARAMETER VARIATION). PATH DELAYS CORRESPONDING TO CHAINS OF 2, 10, AND 20 INVERTERS (WITH F04 LOADS AT EACH INVERTER) ARE CONSIDERED

Measurement	Path Length (gates)	Maximum Absolute Error	Mean Absolute Error
$E[t_d]$	2	13%	7.8%
$Var[t_d]$	2	32%	18%
$E[t_d]$	10	13%	8.9%
$Var[t_d]$	10	16%	12%
$E[t_d]$	20	20%	12%
$Var[t_d]$	20	17%	16%

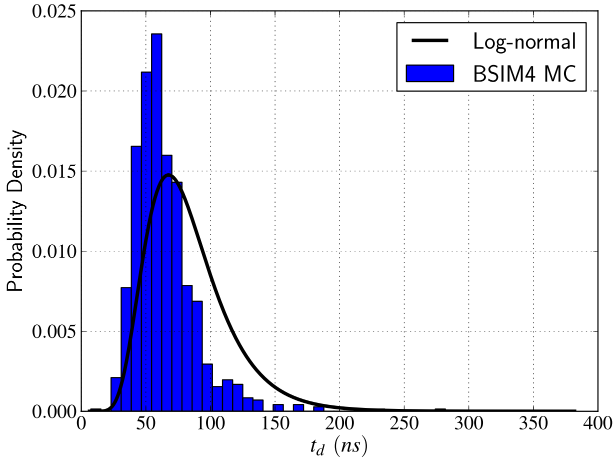


Fig. 9. Log-normal distribution for path delay, using an expected value and variance calculated with the near-threshold statistical delay model (60) and (61) compared to MC SPICE simulations of BSIM4 statistical model for a chain of 20 minimum-size inverters (with FO4 loads at the output of each inverter) in 65-nm GP CMOS with $V_{\mathrm{DD}} = 300$ mV (at TT-corner, 70 °C, and with 10K MC trials accounting for local parameter variation).

as log-normal with expected value and variance given by (58) and (59), respectively.[10] The sum of log-normal RVs can be approximated as log-normal [32], [33], giving closed-form equations for the path delay t_d of a sequence of gates with L_{dp} gates on the path [from(48)]

$$E[t_d(V_t;\, L_{dp})] = \sum_{i \in \{1,2,\ldots,L_{dp}\}} E[t^i_{\mathrm{pd}}(V_t)] \tag{60}$$

$$\mathrm{Var}[t_d(V_t;\, L_{dp})] = \sum_{i \in \{1,2,\ldots,L_{dp}\}} \mathrm{Var}[t^i_{\mathrm{pd}}(V_t)] \tag{61}$$

where t^i_{pd} is the delay of the ith gate in the path.

With these statistical delay models, short-channel effects cannot be completely ignored. As with the I_{OFF} model [i.e., (51)], DIBL can be easily incorporated by using an effective threshold voltage of $V_t - \eta V_{\mathrm{DD}}$ in lieu of V_t. In the

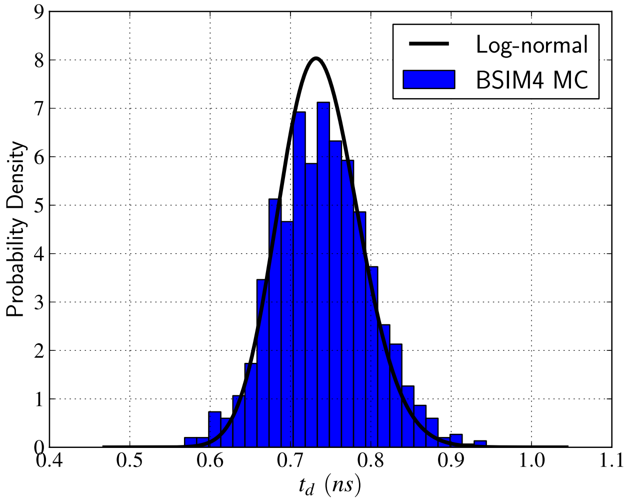


Fig. 10. Log-normal distribution for path delay using an expected value and variance calculated with the near-threshold statistical delay model [i.e., (60) and (61)] compared to MC SPICE simulations of BSIM4 statistical model for a chain of 20 minimum-size inverters (with FO4 loads at the output of each inverter) in 65-nm GP CMOS with $V_{\mathrm{DD}} = 700$ mV (at TT-corner, 70 °C, and with 10K MC trials accounting for local parameter variation).

65-nm GP process, incorporating the effects of DIBL into (44) yields new parameters: $\eta = 0.134, n = 1.61, (C_{\mathrm{load}}/I_F) = 1.23$ (ns/V). V_t is normally distributed with mean $\mu_{V_t} =$ 449 mV and standard deviation, $\sigma_{V_t} = 56.9$ mV (computed at the TT-corner from statistical BSIM4 models using the methods from [31]). In order to measure model accuracy, (60) and (61) are compared to Monte Carlo (MC) simulations using SPICE and foundry-provided statistical BSIM4 models. Path lengths of 2, 10, and 20 inverters are considered from 300 to 700 mV (at 100-mV intervals) with 10 K MC trials at each V_{DD}. The error in both the expected value and the standard deviation are reported in Table VII. Figs. 9 and 10 depict the histograms generated from 10 K MC trials at 300 and 700 mV, respectively, with a path length of 20 inverters; the corresponding log-normal distributions with expected value and variance calculated from (60) and (61), respectively, overlay each histogram.

Approximately 1.3 core-hours of computation time is needed to perform each set of 10 000 MC BSIM4 transient simulations on modern hardware with modern commercial SPICE software. In practice, fewer trials per set may be necessary; however, the computation cost of a broad analysis (e.g., of a large gate set over a wide range of supply voltages at multiple temperatures and multiple global process corners) is significant. The computation cost associated with solving (60) and (61) is comparatively negligible; the only significant computation expense is incurred when calculating the fitting constants in (44): 1.1e–2 core-hours when V_{DD} is swept from 60 mV to 1 V with a 10-mV step size.

IV. RELATED WORK

MOS modeling dates back many decades, so the set of works that present and discuss various approaches to it are numerous (see [34] or [35] for a historical discussion). The

[10] X is a log-normal RV iff $\log(X)$ is normally distributed.

models used in this paper are based on existing inversion-charge models. The EKV [14], [16], [36] model and the ACM [15] model are examples of accurate and mature continuous compact inversion-charge models. The forthcoming BSIM6 [37] model is a new and purportedly extremely accurate inversion-charge model that is still under development. The work in this paper differs in that the models are reduced and simplified to the point of limiting applicability to that of digital circuit modeling.

The Lambert W function is currently supported in numerous mathematical computation frameworks, e.g., Maple, MATLAB, Mathematica, SciPy. Calhoun used it to define a closed-form approximation for the minimum-energy operating point of CMOS circuits in [3], and Ortiz-Conde used it to model diode current [38] and surface potential in an undoped-body MOSFET [39].

One of the primary goals of this paper is to give a simple and continuous digital MOSFET model that can be used for hand calculations of circuits operating near threshold. A number of works, see [2] and [3], perform digital circuit analysis in this region ([40] includes variability and derives a sophisticated sub-threshold statistical delay model), but these works rely on the weak-inversion approximation. The weak-inversion model is inaccurate at and above the device threshold voltage, which makes it difficult to perform analysis or establish trends for circuits operating near threshold. The authors of [4] and [41] address this shortcoming by using the EKV approximation, but this makes hand analysis nearly impossible. Simple continuous models such as [6] and [42] exist but are purely empirical, so they lack the rigor and fitting-constant stability associated with the analytical model presented in this paper.

V. Conclusion

This paper presented the near-threshold model [i.e., (41)], which is a simplified transregional MOS drain-current model designed specifically for digital circuit analysis of near-threshold circuits. The near-threshold model is continuous and continuous in the first derivative, and it accurately models I_{ON} over a wide supply voltage range. The model derivation follows that of previous inversion-charge-based models with the addition of a new exact expression for inversion charge (19) and a new simplified inversion-charge approximation, i.e., (34)–(36). The exact expression for inversion charge may improve the accuracy of certain analyses, e.g., small signal; verification of this is left as future work. The near-threshold model was validated in four modern CMOS technologies against BSIM4 SPICE simulations, and it was used to solve a circuit analysis problem: finding the minimum-energy operating point of a digital circuit.

As with all models, the near-threshold model has limitations. In a technology with extremely high or low nominal threshold voltages, the model accuracy may degrade. In the technologies examined in this paper, the HVT devices tended to have higher error rates than those of regular devices (see Table IV). Explicitly including short-channel effects within the model may somewhat mitigate this problem; however, this is left as future work. Similarly, model accuracy may degrade when modeling I_{ON} as a function of transistor length or width unless short-channel effects are explicitly included (as discussed in Section II-C). This problem is apparent in most compact models, but examining it in the context of the near-threshold model is left as future work.

Acknowledgment

The authors would like to thank S. S. Bhargav, C. Moore, and X. Chang for their excellent comments and helpful discussions.

References

[1] N. H. E. Weste and D. M. Harris, *CMOS VLSI Design*, 4th ed. Reading, MA, USA: Addison-Wesley, 2010.

[2] B. Zhai, D. Blaauw, D. Sylvester, and K. Flautner, "Theoretical and practical limits of dynamic voltage scaling," in *Proc. 41st Design Autom. Conf.*, Jul. 2004, pp. 868–873.

[3] B. Calhoun, A. Wang, and A. Chandrakasan, "Modeling and sizing for minimum energy operation in subthreshold circuits," *IEEE J. Solid-State Circuits*, vol. 40, no. 9, pp. 1778–1786, Sep. 2005.

[4] D. Markovic, C. Wang, L. Alarcon, T.-T. Liu, and J. Rabaey, "Ultralow-power design in near-threshold region," *Proc. IEEE*, vol. 98, no. 2, pp. 237–252, Feb. 2010.

[5] S. Keller, S. S. Bhargav, C. Moore, and A. J. Martin, "Reliable minimum energy CMOS circuit design," in *Proc. 2nd Eur. Workshop CMOS Variabil.*, May 2011, pp. 1–6.

[6] D. Harris, B. Keller, J. Karl, and S. Keller, "A transregional model for near-threshold circuits with application to minimum-energy operation," in *Proc. ICM*, Dec. 2010, pp. 64–67.

[7] R. Dreslinski, M. Wieckowski, D. Blaauw, D. Sylvester, and T. Mudge, "Near-threshold computing: Reclaiming Moore's law through energy efficient integrated circuits," *Proc. IEEE*, vol. 98, no. 2, pp. 253–266, Feb. 2010.

[8] L. Chang, D. Frank, R. Montoye, S. Koester, B. Ji, P. Coteus, R. Dennard, and W. Haensch, "Practical strategies for power-efficient computing technologies," *Proc. IEEE*, vol. 98, no. 2, pp. 215–236, Feb. 2010.

[9] B. Sheu, D. Scharfetter, P.-K. Ko, and M.-C. Jeng, "BSIM: Berkeley short-channel IGFET model for MOS transistors," *IEEE J. Solid-State Circuits*, vol. 22, no. 4, pp. 558–566, Aug. 1987.

[10] T. Sakurai and A. Newton, "Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas," *IEEE J. Solid-State Circuits*, vol. 25, no. 2, pp. 584–594, Apr. 1990.

[11] K. Bowman, B. Austin, J. Eble, X. Tang, and J. Meindl, "A physical alpha-power law MOSFET model," *IEEE J. Solid-State Circuits*, vol. 34, no. 10, pp. 1410–1414, Oct. 1999.

[12] Y. Tsividis and C. McAndrew, *Operation and Modeling of the MOS Transistor*, 3rd ed. London, U.K.: Oxford Univ. Press, 2011.

[13] H. Pao and C. Sah, "Effects of diffusion current on characteristics of metal-oxide (insulator)-semiconductor transistors," *Solid-State Electron.*, vol. 9, no. 10, pp. 927–937, Oct. 1966.

[14] C. C. Enz and E. A. Vittoz, *Charge-Based MOS Transistor Modeling: The EKV Model for Low-Power and RF IC Design*. New York, NY, USA: Wiley, Sep. 2006.

[15] C. Galup-Montoro, M. Schneider, A. Cunha, F. de Sousa, H. Klimach, and O. Siebel, "The advanced compact MOSFET (ACM) model for circuit analysis and design," in *Proc. IEEE CICC*, Sep. 2007, pp. 519–526.

[16] C. C. Enz, F. Krummenacher, and E. A. Vittoz, "An analytical MOS transistor model valid in all regions of operation and dedicated to low-voltage and low-current applications," *Analog Integr. Circuits Signal Process.*, vol. 8, no. 1, pp. 83–114, Jul. 1995.

[17] J.-M. Sallese, M. Bucher, and C. Lallement, "Improved analytical modeling of polysilicon depletion in MOSFETs for circuit simulation," *Solid-State Electron.*, vol. 44, no. 6, pp. 905–912, Jul. 2000.

[18] J.-M. Sallese, M. Bucher, F. Krummenacher, and P. Fazan, "Inversion charge linearization in MOSFET modeling and rigorous derivation of the EKV compact model," *Solid-State Electron.*, vol. 47, no. 4, pp. 677–683, Apr. 2003.

[19] R. M. Corless, D. J. Jeffrey, and D. E. Knuth, "A sequence of series for the Lambert W function," in *Proc. ISSAC*, 1997, pp. 197–204.

[20] D. Veberic, *Having fun with Lambert W(x) Function*. Ithaca, NY, USA: Cornell Univ. Press, Mar. 2010.

[21] R. Troutman, "VLSI limitations from drain-induced barrier lowering," *IEEE Trans. Electron Devices*, vol. 26, no. 4, pp. 461–469, Apr. 1979.

[22] B. Zhai, D. Blaauw, D. Sylvester, and K. Flautner, "The limit of dynamic voltage scaling and insomniac dynamic voltage scaling," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 13, no. 11, pp. 1239–1252, Nov. 2005.

[23] M. Na, E. Nowak, W. Haensch, and J. Cai, "The effective drive current in CMOS inverters," in *Proc. IEDM*, 2002, pp. 121–124.

[24] J. Rabaey, A. Chandrakasan, and B. Nikolic, *Digital Integrated Circuits*, 2nd ed. Englewood Cliffs, NJ, USA: Prentice-Hall, 2003.

[25] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits," *Proc. IEEE*, vol. 91, no. 2, pp. 305–327, Feb. 2003.

[26] *International Technology Roadmap for Semiconductors*, Semiconductor Industry Association, Washington DC, USA, 2011.

[27] A. Devgan and C. Kashyap, "Block-based static timing analysis with uncertainty," in *Proc. ICCAD*, Nov. 2003, pp. 607–614.

[28] A. Agarwal, D. Blaauw, V. Zolotov, and S. Vrudhula, "Computation and refinement of statistical bounds on circuit delay," in *Proc. Design Autom. Conf.*, Jun. 2003, pp. 348–353.

[29] A. Agarwal, K. Chopra, D. Blaauw, and V. Zolotov, "Circuit optimization using statistical static timing analysis," in *Proc. 42nd Design Autom. Conf.*, Jun. 2005, pp. 321–324.

[30] P. Asenov, N. Kamsani, D. Reid, C. Millar, S. Roy, and A. Asenov, "Combining process and statistical variability in the evaluation of the effectiveness of corners in digital circuit parametric yield analysis," in *Proc. ESSDERC*, Sep. 2010, pp. 130–133.

[31] N. Drego, A. Chandrakasan, and D. Boning, "Lack of spatial correlation in MOSFET threshold voltage variation and implications for voltage scaling," *IEEE Trans. Semicond. Manuf.*, vol. 22, no. 2, pp. 245–255, May 2009.

[32] N. Beaulieu, A. Abu-Dayya, and P. McLane, "Comparison of methods of computing lognormal sum distributions and outages for digital wireless applications," in *Proc. IEEE ICC Serving Humanity Through Commun.*, vol. 3. May 1994, pp. 1270–1275.

[33] N. Beaulieu, A. Abu-Dayya, and P. McLane, "Estimating the distribution of a sum of independent lognormal random variables," *IEEE Trans. Commun.*, vol. 43, no. 12, pp. 2869–2873, Dec. 1995.

[34] S. Chih-Tang, "Evolution of the MOS transistor-from conception to VLSI," *Proc. IEEE*, vol. 76, no. 10, pp. 1280–1326, Oct. 1988.

[35] P. Balk, "40 years MOS technology–From empiricism to science," *Microelectron. Eng.*, vol. 48, nos. 1–4, pp. 3–6, Sep. 1999.

[36] C. Enz, "An MOS transistor model for RF IC design valid in all regions of operation," *IEEE Trans. Microw. Theory Tech.*, vol. 50, no. 1, pp. 342–359, Jan. 2002.

[37] Y. S. Chauhan, M. Karim, V. Sriram, P. Thakur, N. Paydovosi, A. Sachid, A. Niknejad, and C. Hu, "Transitioning from BSIM4 to BSIM6," in *Proc. MOS-AK Workshop*, Mar. 2012, pp. 1–29.

[38] A. Ortiz-Conde, F. J. G. Sánchez, and J. Muci, "Exact analytical solutions of the forward non-ideal diode equation with series and shunt parasitic resistances," *Solid-State Electron.*, vol. 44, no. 10, pp. 1861–1864, Oct. 2000.

[39] A. Ortiz-Conde, F. G. Sánchez, and M. Guzmán, "Exact analytical solution of channel surface potential as an explicit function of gate voltage in undoped-body MOSFETs using the Lambert W function and a threshold voltage definition therefrom," *Solid-State Electron.*, vol. 47, no. 11, pp. 2067–2074, Nov. 2003.

[40] B. Zhai, S. Hanson, D. Blaauw, and D. Sylvester, "Analysis and mitigation of variability in subthreshold design," in *Proc. ISLPED*, Aug. 2005, pp. 20–25.

[41] S. Fisher, R. Dagan, S. Blonder, and A. Fish, "An improved model for delay/energy estimation in near-threshold flip-flops," in *Proc. IEEE ISCAS*, May 2011, pp. 1065–1068.

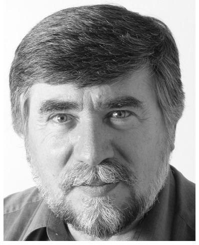
[42] K. Nose and T. Sakurai, "Optimization of VDD and VTH for low-power and high speed applications," in *Proc. Asia South Pacific Design Autom. Conf.*, 2000, pp. 469–474.

Sean Keller (S'07–M'13) received the B.S. and M.Eng. degrees from Cornell University, Ithaca, NY, USA, the M.S. degree from the University of Illinois at Urbana-Champaign, Urbana, IL, USA, and the Ph.D. degree from the California Institute of Technology, Pasadena, CA, USA.

He is a Researcher with Microsoft Corporation. He has designed and taped out chips at Situs Logic. His current research interests include low-power and reliable digital circuit design, asynchronous VLSI, and device modeling.

David Money Harris (S'95–M'99) received the S.B. and M.Eng. degrees from the Massachusetts Institute of Technology, Cambridge, MA, USA, and the Ph.D. degree from Stanford University, Stanford, CA, USA.

He is a Professor of engineering with Harvey Mudd College, Claremont, CA, USA. He has designed chips with Intel, Sun Microsystems, Hewlett Packard, Broadcom, and holds more than a dozen circuit patents. He is the author of *CMOS VLSI Design, Digital Design and Computer Architecture, Logical Effort*, and *Skew-Tolerant Circuit Design*.



Alain J. Martin (M'97) received the Degree from Institut National Polytechnique de Grenoble, Grenoble, France, in 1969.

He is a Professor of computer science with the California Institute of Technology, Pasadena, CA, USA. In 1988, he designed the world-first asynchronous microprocessor. His current research interests include concurrent computing, asynchronous VLSI design, formal methods for the design of hardware and software, and computer architecture.

Prof. Martin is a member of the Association for Computing Machinery.