A Transregional Model for Near-Threshold Circuits with Application to Minimum-Energy Operation

David Money Harris, Ben Keller, Julia Karl
Department of Engineering
Harvey Mudd College
Claremont, CA 91711
{David_Harris, Ben_Keller, Julia_Karl}@hmc.edu

Sean Keller
Department of Computer Science
California Institute of Technology
Pasadena, CA 91125
sean@async.caltech.edu

Abstract—The most energy-efficient operating point for CMOS circuits is near the threshold voltage. Conventional models are difficult to use in this region because they are piecewise and/or discontinuous around threshold. This paper proposes a simple new model for I_{on} that is valid in the near-threshold region. Based on the ON-current, a propagation delay model is derived. The model is applied to determine the minimum energy point for inverter chains. The transregional model matches simulated data within 15 mV, while the conventional exponential subthreshold model underestimates the supply voltage by up to 80 mV.

Index Terms—**low power, delay, minimum energy, subthreshold circuits, near-threshold circuits**

I. Introduction

To meet stringent energy budgets for systems such as implantable medical devices, wireless sensor networks, and RFID tags, researchers are increasingly interested in operating CMOS circuits below or near the device threshold voltage (V_t) [1]. Furthermore, throughput-oriented high-performance systems may improve energy efficiency by operating near 0.5 V and recouping performance through parallelism [2, 3]. Unfortunately, conventional MOSFET models are expressed in a piecewise fashion with breakpoints separating regions of operation; one such breakpoint occurs at or near V_t, making analysis of near-threshold circuits (NTC) difficult. This paper presents and validates a simple empirical transregional model for ON-current and delay in the NTC region.

The delay of a gate can be approximated as

$$t_{pd} = k_1 \frac{C_{load} V_{DD}}{I_{on}} \tag{1}$$

where C_{load} is the load capacitance, V_{DD} is the supply voltage, I_{on} is the current at $V_{gs} = V_{ds} = V_{DD}$, and k_1 is a small constant. ON-current depends on $V_{DD} - |V_t|$, which we will abbreviate as V_{DT}. In the saturation region with $V_{DT} > 0$, I_{on} is often approximated with the α-power law model [4]

$$I_{on} = k_2 W V_{DT}^{\alpha} \tag{2}$$

in which W is the transistor width, α is the velocity saturation index (typically around 1.3 – 1.5 in a nanometer process), and k_2 captures other process parameters.

In the subthreshold region with $V_{DT} < 0$, I_{on} is approximated with an exponential model [5]

$$I_{on} = I_0 W e^{\frac{V_{DT}}{n v_T}} \tag{3}$$

in which I_0 is the current per unit width at V_{DT} =0, v_T is the thermal voltage (kT/q, 29.6 mV at 70 ºC), and n is the subthreshold slope factor, $1 + C_d / C_{ox}$. This model is only valid for $V_{DD} > 3v_T$ so that the influence of drain voltage can be disregarded [12].

The discontinuity between these models is problematic in the analysis of circuits operating near V_t. Calhoun derived a reasonably simple formula [6] for the minimum energy point using a model valid for $V_{DD} < V_t$. Researchers have proposed empirical [7] and physical [8] piecewise transregional models that are continuous but use separate equations above and below threshold. The EKV model [9] is continuous and continuously differentiable. Marković proposes a near-threshold current model based on EKV [10]

$$I_{on} = \frac{2n\mu C_{ox}}{k_{fit}} \frac{W}{L} v_T^2 \left[\ln\left(e^{\frac{(1+\eta)V_{DD}-V_t}{2nv_T}} + 1 \right) \right]^2 \tag{4}$$

However, this model is rather difficult to work with analytically and is too complex to provide back-of-the-envelope insights.

In this work, we propose a simple empirical nonpiecewise modification to (3) that closely fits industry-standard BSIM models over the near-threshold region. We show that the model accurately predicts the voltage dependence of gate delay. We then apply the model to derive an equation for the minimum energy operating point, which is not necessarily in the subthreshold region when the activity factor is low.

II. Transregional ON-current Model

Fig. 1 plots simulated I_{on} vs. V_{DD} on a semilogarithmic scale for a 1 μm wide nMOS transistor with a threshold of approximately 0.3 V in a 65 nm process at 70 ºC [11, 12]. Observe that the curve is nearly straight for $V_{DD} < V_t$, corresponding to the exponential I-V relationship. Above V_t, the curve rolls off. The ON-current is closely fit by

$$I_{on} = I_0 W e^{\frac{V_{DT} - aV_{DT}^2}{n v_T}} \tag{5}$$

where a is an empirical fitting parameter. The model is expressed with a quadratic dependence on V_{DT} rather than V_{DD} so that a change in V_t caused by process variation or body bias does not require fitting a new value for a, n, or I_0.

Over the near-threshold operating range of 0.06 – 0.7 V, the least squares fit has an average error of 2.4% and a maximum error of 6.1%. Below the bottom end of the range, current abruptly drops to zero because of the dependence on the lateral electric field. Above the top end of the range, the transistor is well into the saturation region and current is better described with the α-power law model. Note that *n* is now an empirical parameter that lacks the physical meaning it had in the subthreshold current model of (3). If *a* is set to 0, (5) reduces to (3) and the fit is only valid in the subthreshold region. Over a range of 0.06 – 0.24 V, (3) has an average error of 4.8% and a worst-case error of 10.0%. The transregional model is a better match to data over a much broader range than the standard subthreshold model. A pMOS transistor fits the same model with average and worst-case errors of 5.8% and 12%. Fits for 90 and 130 nm models use similar values of *a* and *n* and show average errors of 9%.

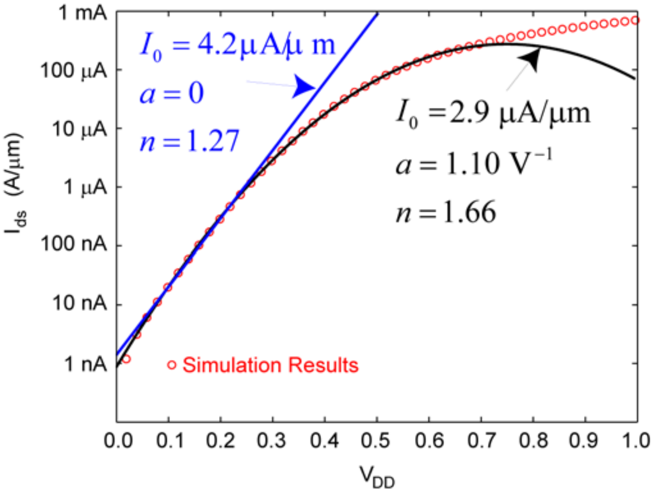

Figure 1. I_{on} vs. V_{DD} for an nMOS transistor in a 65 nm process showing good fit to the proposed transregional model in the near-threshold region.

III. Transregional Delay Model

Combining (5) and (1) gives a simple model for gate delay dependence on V_{DD} in the near-threshold region

$$t_{pd} = k\frac{C_{load}}{W}V_{DD}e^{-\frac{V_{DT}-aV_{DT}^2}{nv_T}} \qquad (6)$$

where *k*, *n*, and *a* are fitting parameters. The propagation delay is sensitive to current when transistors are partially ON as the input rises [13]. Although it is reasonable to use the same values of *a* and *n* that are obtained from I_{on} simulations, the delay fits measured data better when these parameters are allowed to vary; however, the difference between the parameters is small.

Fig. 2 plots the propagation delay of a fanout-of-4 (FO4) inverter as a function of supply voltage. Minimum-width nMOS and pMOS transistors are used in the inverter to achieve low energy. The inverters function down to a minimum of 140 mV. The fit to (6) over the range of V_{DD} = 0.2 – 0.7 V has an average error of 2.8% and a worst-case error of 8.7%. If *a* is forced to 0, (6) reduces to the conventional subthreshold delay model and the fit is only good below threshold. The delay of a ring oscillator also fits (6) with average and worst-case errors of 2.0% and 6.9%.

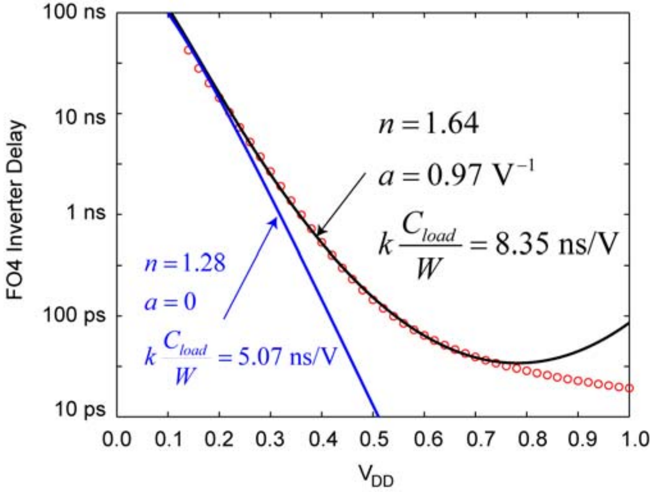

Figure 2. FO4 inverter delay vs. V_{DD}

Fig. 3 plots the worst-case delay of an 8-bit ripple carry adder (using minimum-width devices) in terms of FO4 inverter delays as a function of supply voltage. Notice that the normalized delay varies by only 8% over the range of 0.2 – 0.7 V. This indicates that the delay of complex gates tracks well with that of inverters even in near-threshold operation. Hence, EQ (6) is a reasonable model for predicting how circuit delay scales with supply and threshold voltage.

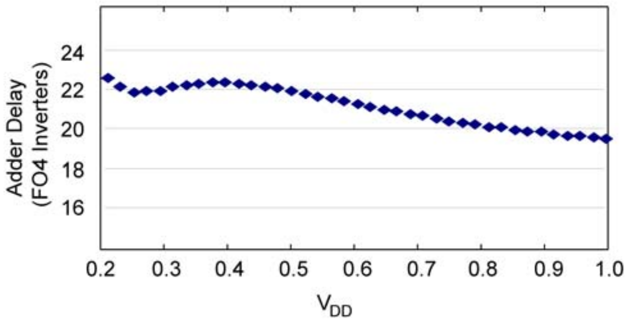

Figure 3. Normalized 8-bit ripple carry adder delay

IV. Minimum Energy Point

This section derives the minimum energy operating point using the proposed transregional model. The analysis is similar to that of Zhai [14] and Calhoun [6] but no longer presupposes subthreshold operation and uses a better model for OFF-current. The total energy of a circuit is

$$E_{tot} = E_{dyn} + E_{leak} \qquad (7)$$

The dynamic energy of a circuit is

$$E_{dyn} = C_{dyn}V_{DD}^2 \qquad (8)$$

where C_{dyn} is the effective switching capacitance of the entire circuit accounting for activity factor, glitching, and short circuit current. The leakage energy is

$$E_{leak} = I_{off} V_{DD} T_c \tag{9}$$

I_{off} is the leakage current at V_{gs} = 0 and V_{ds} = V_{DD}. Calhoun approximated I_{off} as I_{on} at V_{DD} = 0 [6]. This approximation does not account for drain-induced barrier lowering (DIBL) and thus may be inaccurate by as much as an order of magnitude. A better model for OFF-current for $V_{ds} >> v_T$ is

$$I_{off} = I_1 W_{eff} e^{\frac{\eta V_{DT}}{n v_T}} \tag{10}$$

where I_1 is the OFF-current per micron at V_{gs} = 0 and V_{DT} = 0, η is the DIBL coefficient and W_{eff} is the effective width of all the OFF transistors contributing leakage. Note that I_1 is a function of V_t, so this model is inappropriate for evaluating the impact of V_t variation on the minimum energy point. Fig. 4 shows the simulated leakage currents for minimum-width devices in the 65 nm process along with the curve fit. Over the range of 0.2 – 0.7 V, the model describes OFF-current very well. The fit uses the values of n from (5). Note that this process demonstrates a strong narrow-width effect so that I_1 must be adjusted for non-minimum devices. However, the minimum energy point is obtained using minimum-sized transistors. The leakage of the pMOS transistors is an order of magnitude lower than that of the nMOS.

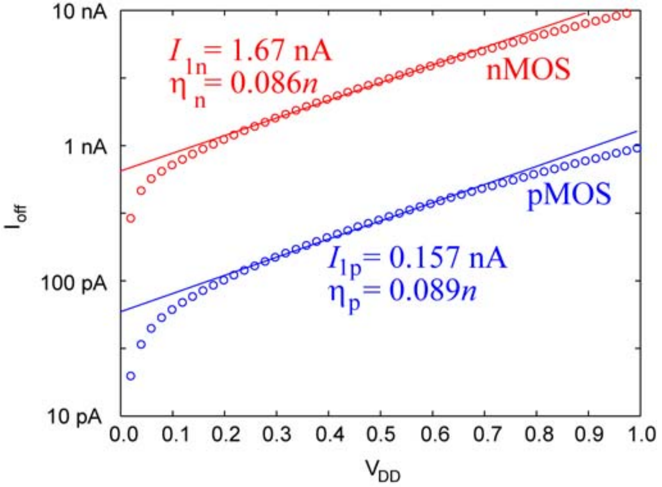

Figure 4. OFF-current of minimum-width transistors, illustrating DIBL

The cycle time for a path with a logic depth of L_{dp} gates on the critical path is

$$T_c = t_{pd} L_{dp} \tag{11}$$

Putting these all together gives

$$E_{leak} = I_1 W_{eff} L_{dp} k \frac{C_{load}}{W} V_{DD}^2 e^{-\frac{V_{DD}(1-\eta)+\eta V_t - a V_{DT}^2}{n v_T}} \tag{12}$$

The parameters multiplying voltage in (12) have an overall unit of capacitance. Although they do not represent a physical capacitance, it is convenient to lump them into an effective leakage capacitance analogous to the effective switching capacitance

$$C_{leak} = I_1 W_{eff} L_{dp} k \frac{C_{load}}{W} \tag{13}$$

Finally, we arrive at the total energy of the circuit

$$E_{tot} = V_{DD}^2 C_{dyn} \left(1 + Re^{-\frac{V_{DT}(1-\eta) - a V_{DT}^2}{n v_T}} \right) \tag{14}$$

with $R = C_{leak} / C_{dyn}$. It is possible to differentiate energy with respect to V_{DD} to solve for the minimum energy point. However, the result lacks a closed-form solution, so we solve (14) numerically instead.

Calhoun showed that, in the subthreshold region, delay and leakage have inverse sensitivity to V_t and hence the minimum energy point is independent of V_t. However, in near-threshold operation, (14) shows that raising V_t reduces leakage more than it increases cycle time. Hence, at higher V_t, leakage is less important and the minimum energy point occurs at a lower supply voltage.

V. Application to Inverter Chains

This section examines the minimum energy point for inverter chains. It compares the conventional and transregional models to simulation results in a 65 nm process.

Consider the circuit in Fig. 5 consisting of M chains of N FO4 inverters. Only one of the chains switches in a given cycle. The circuit models an arbitrary block of logic with a logic depth of N and an activity factor of 1/M. Each FO4 inverter contains minimum-width (0.1 μm) nMOS and pMOS transistors. The transistors have a gate capacitance of 1.0 fF/μm, so each inverter has an input capacitance of 0.2 fF. Hence, with an FO4 load, C_{dyn} = 0.8N fF. Half the nMOS and half the pMOS transistors are leaking at any given time, but the pMOS leakage is negligible in comparison, so the effective leakage width is W_{eff} = 2MN minimum-sized nMOS transistors. Notice that in this process, the pMOS leakage is negligible; in other processes this may not be true, and W_{eff} could increase by up to a factor of 2. The logic depth is L_{dp} = N. We use the following parameters extracted from Figures 2 and 4:

TABLE I. Model Parameters

	Subthreshold	Transregional
A	0	0.97
N	1.28 V^{-1}	1.64 V^{-1}
$k\frac{C_{load}}{W}$	5.07 ns/V	8.35 ns/V
I_1	5.75 μA	
V_t	0.3 V	
η	0.086n V^{-1}	
R	0.021MN	0.034MN

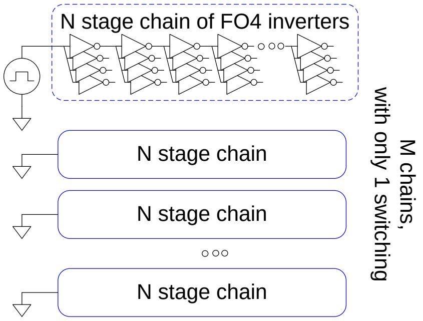
N stage chain of FO4 inverters
N stage chain
N stage chain
N stage chain
M chains, with only 1 switching

Figure 5. Test circuit

Fig. 6 shows the supply voltage, V_{DDopt}, providing minimum energy as a function of M for 12-stage inverter chains. The transregional model matches the results of HSPICE simulation within 15 mV. The subthreshold model diverges significantly from simulation because it is not valid at the voltages of interest. It underestimates the best supply voltage by as much as 80 mV for circuits with very low activity factors. This 80 mV error corresponds to underestimating the total energy by 25%, a significant deviation.

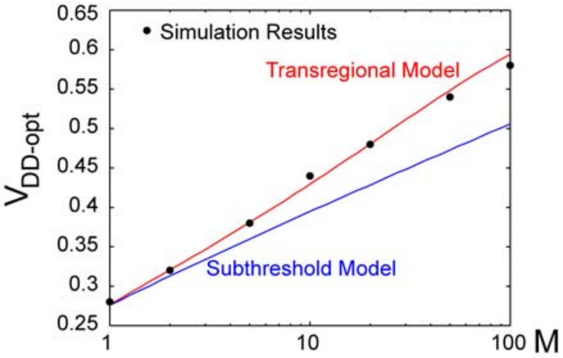

Figure 6. Supply voltage for minimum energy for 12-stage inverter chains with activity factor of $1/M$ ($V_t = 0.3$ V)

Observe that the curve fits are nearly straight lines on a semilogarithmic scale. This indicates that V_{DDopt} can be curve-fit from Figure 6 as a linear function of log M. Substituting R for M using the relationship from Table 1 puts the result in a form like that which has been reported in [14].

$$V_{DDopt} = \left(1.37 \ln R + 6.96\right) nv_T \tag{15}$$

Variability exacerbates the expected leakage and delay, effectively increasing R [15]. This pushes the minimum energy point to a higher voltage.

VI. Conclusion

Designers investigating energy-efficient operating points need models for drive current and delay that are valid both above and below threshold. This paper proposes a simple empirical transregional model for ON-current that matches simulated characteristics to within 10% over the range of 0.2 – 0.7 V. It also develops a model for delay proportional to CV/I and shows that the model fits simulations well. The model is applied to determine the minimum energy point for inverter chains. Circuits with high logic depths and/or low activity factors have a minimum energy point above threshold. The new model closely matches simulated data, while the conventional subthreshold model underestimates the minimum energy supply voltage by up to 80 mV at low activity factors, and thus underestimates the total energy by 25%.

Acknowledgment

This work is supported by the National Science Foundation under grant 910606 and by the Clay-Wolkin Fellowship.

References

[1] A. Wang, B. Calhoun, and A. Chandrakasan, *Sub-Threshold Design for Ultra Low-Power Systems*, New York: Springer, 2006.

[2] L. Chang *et al*., “Practical Strategies for Power-Efficient Computing Technologies,” *Proc. IEEE*, vol. 98, no. 2, pp. 215-236, Feb. 2010.

[3] R. Dreslinski, M. Wiekowski, 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.

[4] T. Sakurai and R. 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, April 1990.

[5] B. Sheu, D. Sharfetter, P, Ko, and M. Jeng, “BSIM: Berkeley Short-Channel IGFET Model for MOS Transistors,” *IEEE J. Solid-State Circuits*, vol. SC-22, no. 4, pp. 558-566, Aug. 1987.

[6] 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, Sept. 2005.

[7] K. Nose and T. Sakurai, “Optimization of VDD and VTH for Low-Power and High-Speed Applications,” *Proc. Asia South Pacific Design Automation Conf.*, pp. 469-474, 2000.

[8] A. Bhavnagarwala, B. Austin, K. Bowman, and J. Meindl, “A Minimum Total Power Methodology for Projecting Limits on CMOS GSI,” *IEEE Trans. VLSI*, vol. 8, no. 3, pp. 235-251, June 2000.

[9] C. Enz, F. Krummenacher, and E. Vittoz, “An Analyitical MOS Transistor Model Valid in All Regions of Operation and Dedicated to Low-Voltage and Low-Current Applications,” *Analog Integrated Circuits and Signal Processing*, vol. 8, no. 1, pp. 83-114, July 1995.

[10] D. Marković, C. Wang, L. Alarcón, T. Liu, and J. Rabaey, “Ultralow-Power Design in Near-Threshold Region,” *Proc. IEEE*, vol. 98, no. 2, pp. 237-252, Feb. 2010.

[11] Z. Luo, et al., “High Performance and Low Power Transistors Integrated in 65nm Bulk CMOS Technology,” *IEEE Intl. Electron Devices Meeting,* pp. 28.3.1-28.3.4, 2004.

[12] N. Weste and D. Harris, *CMOS VLSI Design*, 4th Ed., Boston: Addison-Wesley, 2010.

[13] M. Na, E. Nowak, W. Haensch, and J. Cai, “The Effective Drive Current in CMOS Inverters,” *Proc. Intl. Electron Dev. Meet,*, 2002, pp. 121-124.

[14] B. Zhai, D. Blaauw, D. Sylvester, and K. Flautner, “The Limit of Dynamic Voltage Scaling and Insomniac Dynamic Voltage Scaling,” *IEEE Trans. VLSI*, vol. 13, no. 11, pp. 1239-1252, Nov. 2005.

[15] B. Zhai, S. Hanson, D. Blaauw, and D. Sylvester, “Analysis and Mitigation of Variaiblity in Subthreshold Design,” *Intl. Symp. Low Power Electronics and Design*, 2005, pp. 20-25.