Supply and Threshold Voltage Scaling for Low Power CMOS

Ricardo Gonzalez, Benjamin M. Gordon, and Mark A. Horowitz

***Abstract*— This paper investigates the effect of lowering the supply and threshold voltages on the energy efficiency of CMOS circuits. Using a first-order model of the energy and delay of a CMOS circuit, we show that lowering the supply and threshold voltage is generally advantageous, especially when the transistors are velocity saturated and the nodes have a high activity factor. In fact, for modern submicron technologies, this simple analysis suggests optimal energy efficiency at supply voltages under 0.5 V. Other process and circuit parameters have almost no effect on this optimal operating point. If there is some uncertainty in the value of the threshold or supply voltage, however, the power advantage of this very low voltage operation diminishes. Therefore, unless active feedback is used to control the uncertainty, in the future the supply and threshold voltage will not decrease drastically, but rather will continue to scale down to maintain constant electric fields.**

***Index Terms*— Energy-delay product, low power CMOS circuits, threshold scaling.**

I. Introduction

REDUCING power dissipation has become an important objective in the design of digital circuits. One common technique for reducing power is to reduce the supply voltage. For CMOS circuits the cost of lower supply voltage is lower performance. Scaling the threshold voltage can limit this performance loss somewhat but results in increased static power dissipation. Burr *et al.* [1], [2] have shown that if one optimizes for minimum energy, then operating in the subthreshold region is advantageous. Since minimum energy solutions are generally low performance solutions, we look instead at both energy and delay during optimization and use the energy-delay product as a measure of the efficiency of the circuit. In this paper we examine the effects of lowering the supply and threshold voltages on the energy efficiency of CMOS circuits.

The next section presents a first-order model of the energy-delay product (EDP) of CMOS circuits. Using this model, one can find the optimal operating point, that is the value of supply and threshold voltage for which the EDP is minimum, as well as how this optimal point will change as circuit and process parameters change. For a modern 0.25-μm technology the optimal operating point is a supply voltage of 250 mV and a threshold voltage of 120 mV. The importance of operating near the minimum is set by how steep the curve, or surface, is near the minimum point. As the curve becomes steeper the benefits of being near the optimal point increase. The performance cost of operating at this point is the ratio of the gate speed at this point to the original gate speed. We numerically solved the model described in Section II as a function of both V and $V_{\rm th}$ to determine the shape of the energy, delay, and EDP surfaces. We show that when transistors are velocity saturated, the EDP surface is pretty steep, and thus one wants to operate near the minima, but gates at this point are significantly slower than current operating conditions.

Finally, in the last section we extend the model to take into account the uncertainty in the value of the supply and threshold voltage. The effect of this variability is quite pronounced. It moves the optimal EDP to a higher operating voltage and threshold voltage and makes the EDP surface flatter.

Manuscript received July 15, 1996; revised November 19, 1996. This work was supported by the Advanced Research Projects Agency under contract J-FBI-92-194.

R. Gonzalez and M. A. Horowitz are with the Computer Systems Laboratory, Stanford University, Stanford, CA 94305 USA.

B. M. Gordon is with the Department of Electrical Engineering, University of Washington, Seattle, WA 98195 USA.

Publisher Item Identifier S 0018-9200(97)05302-X.

II. Energy and Delay in CMOS Circuits

The two main sources of power dissipation in CMOS circuits are static current, which results from resistive paths between power supply and ground, and dynamic power, which results from switching capacitive loads between different voltage levels. There is a third source of power dissipation in CMOS circuits, short-circuit current, which results from both transistors in a CMOS inverter being on at the same time while the input switches. The short-circuit component is small [3], [13], therefore we ignore it throughout this paper. Static power is due to current sources and to leakage current when a transistor is nominally off.

For a CMOS gate, the dynamic power is

$$P = aCV^2 f \tag{1}$$

where a is the activity factor of the output node, C is the total capacitance of the output node, V is the supply voltage, and f is the operating frequency. If the circuit performs one operation per cycle, then the energy per operation is

$$E = aCV^2. \tag{2}$$

For a complex chip, the total dynamic power is simply the sum of the dynamic power of all the gates. The resulting equation has the same form as (1); the only difference is that C is now the total capacitance of all the loads, and the activity factor is the average activity factor.

The leakage current for a gate can be written as

$$I_l = W I_s e^{(V_{\rm th}/V_o)} \tag{3}$$

where W is the effective transistor width[1] of the cell, I_s is the zero-threshold leakage current, $V_{\rm th}$ is the threshold voltage, and V_o is the subthreshold slope. We ignore the dependence of I_l on drain voltage, and also the leakage current in the reverse biased diodes. The leakage current for a complete chip is simply the sum of the leakage currents of all the gates.

The total energy per operation of a chip thus can be written as

$$E = \sum_i a_i C_i V^2 + \sum_i W_i I_s e^{(V_{\rm th}/V_o)} V T_c \tag{4}$$

where T_c is the cycle time and i is an index that runs over all gates in the circuit. The circuit dissipates static current throughout the cycle, but each gate dissipates dynamic energy for a short period of time while it switches.

Notice that this equation is very similar to the energy consumed by a simple inverter (with the "correct" average activity a and load C and assuming W is the total transistor width of the gate), so optimizing the energy of this average inverter will yield an optimal operating point for the chip. In fact, the optimal point remains unchanged if we further normalize this equation by the width of this average inverter, yielding the average energy consumed per micron of transistor width

$$E = C_{\rm eff} V^2 + I_s e^{(V_{\rm th}/V_o)} V T_c \tag{5}$$

where $C_{\rm eff}$ is the average capacitance switched every cycle per micron of transistor width. This parameter is different for every design, depending on the types of circuits used. For the **StrongArm-110** processor from DEC $C_{\rm eff} = 0.2$ fF [4]. Since caches—which have very low activity factors—occupy about 50% of the area of this chip, we expect other designs to have larger values of $C_{\rm eff}$. Later on we show the location of the optimal point is highly insensitive to the value of $C_{\rm eff}$. Leakage power is more important when the effective switched capacitance is small. Thus, we use a value of 1 fF, which is relatively high. This will make lower voltages seem more attractive.

We use a similar technique to model the minimum operating cycle time, or critical path, of the chip. The critical path normally goes through a variety of gates, each with a different delay. Luckily, changes in supply voltage, temperature, and threshold voltage affect all gates in the same way so delay of any gate remains roughly proportional to the delay of an inverter, as is shown in Fig. 1. This figure shows the delay of different circuit elements normalized to the delay of an inverter. Solid lines show the delay at high temperature (125°C), dashed and doted lines show the delay at lower temperatures (25°C and −25°C, respectively). Thus, we can normalize the critical path by dividing the cycle-time by the delay of the average inverter (T_g) described above. We call this quantity the logic depth L_d, since it represents how many inverters are in a ring oscillator which has the same frequency as the maximum operating frequency of the chip. For modern microprocessors the logic depth is usually around 30 equivalent inverters. The cycle time is then just $L_d T_g$.

[1]Transistor width that contributes to the leakage current.

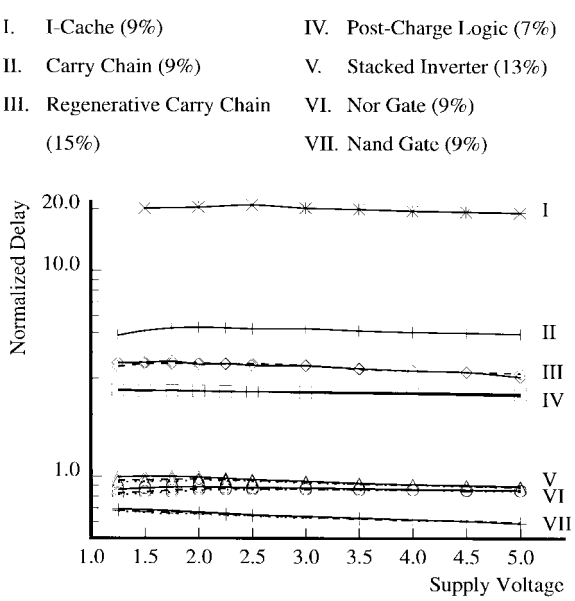

Fig. 1. Normalized delay of CMOS circuits.

To determine how the delay of an inverter varies with operating conditions we use a simple α power model for MOS current[2] [10]

$$T_g = K \frac{V}{(V - V_{\rm th})^\alpha} \tag{6}$$

where K is a proportionality constant specific to a given technology. The α power accounts for the fact that the transistors may be velocity saturated. It can be anywhere between one, complete velocity saturation, and two, no velocity saturation. For a 0.25-μm technology, α is likely to be 1.3–1.5.

Combining (5) with (6), the energy-delay product can be written as

$$\text{EDP} = \frac{K^2 I_s L_d V^3}{(V - V_{\rm th})^\alpha} \left(K_2 + \frac{e^{V_{\rm th}/V_o)}}{(V - V_{\rm th})^\alpha} \right) \tag{7}$$

where K_2 is a constant for the given technology and is given by

$$K_2 = \frac{C_{\rm eff}}{I_s K L_d}. \tag{8}$$

To find the optimal supply and threshold voltage we differentiate (7) with respect to V and $V_{\rm th}$ and set the equations to zero. Solving for V and $V_{\rm th}$, one gets

$$V = \frac{3V_{\rm th}}{3 - \alpha} + \frac{3\alpha}{3 - \alpha} V_o \tag{9}$$

and

$$e^{-n} = \frac{K_2 \left[\frac{\alpha}{3 - \alpha} (n + 3) V_o \right]^\alpha (3 - \alpha)}{(n + 2\alpha - 3)} \tag{10}$$

or

$$n = -\alpha \ln \left(\frac{\alpha}{3 - \alpha} (n + 3) V_o \right) = \ln(K_2) + \ln(n + 2\alpha - 3) \tag{11}$$

[2]This ignores subthreshold current and assumes transistors are always in the current saturation mode.

TABLE I
PROCESS AND CIRCUIT PARAMETERS FOR 0.25-μm TECHNOLOGY

Variable	Value
C_{eff}	1.0fF
K	155E-6
I_s	1μA
α	1.3
L_d	30
V_o	1.3KT/q

where

$$n = \frac{V_{\rm th}}{V_o}. \tag{12}$$

Although it is not possible to find a closed-form solution for n, we can numerically solve (9) and (11). For the parameters given in Table I, the optimal operating voltage and threshold voltage are quite low, $V = 254$ mV and $V_{\rm th} = 119$ mV, and only weakly depend on most of the technology parameters. The strongest dependence is on the velocity saturation parameter, α, since the optimal $V_{\rm th}$ is proportional to it. As transistors become more velocity saturated, $V_{\rm th}$ at the optimal point decreases, and the optimal operating voltages decreases as well. The threshold depends logarithmically on the other technology parameters $(C_{\rm eff}, L_d)$ present in the K_2 constant. For every order of magnitude change in the effective switched capacitance, due to a change in activity or capacitance, the optimal $V_{\rm th}$ changes by about 70 mV. The logic depth is not likely to change by more than an order of magnitude so its influence on the optimal $V_{\rm th}$ is small. Since order-of-magnitude changes in the other technology parameters are not likely, their effect on the threshold voltage is likely to be small.

This simple analysis indicated that for advanced technologies, there might be a potential energy saving by reducing both V and $V_{\rm th}$ to relatively small values. To determine the magnitude of the savings, we plot contours of constant EDP versus V and $V_{\rm th}$ by finding numerical solutions for the EDP equation. The equations solved are more complete versions of (5) and (6) which include subthreshold currents and are given in the Appendix.

If the transistors are not velocity saturated $(\alpha = 2)$, the EDP surface is relatively flat, as is shown in Fig. 2. This figure shows contours of the inverse of the relative EDP. The relative EDP can be found by normalizing to the value of the EDP at the optimal point. Thus, any point on the curve labeled 0.5 has an EDP value twice that of the minimum. The optimal point is at $V = 533$ mV and $V_{\rm th} = 127$ mV. But at $V = 1$ V and $V_{\rm th} = 300$ mV, the EDP of the circuit has increased by only a factor of 1.5×. Thus, the benefit of operating at the optimal point is small. The small "kinks" in the curves near the border of the subthreshold region are artifacts of the model and can be ignored. The current is slightly nonmonotonic as the transistors switch from the subthreshold to the active region.

When transistors are velocity saturated, however, the EDP surface is much steeper. Fig. 3 shows contours similar to those of Fig. 2 but with $\alpha = 1.3$. In this case the contour

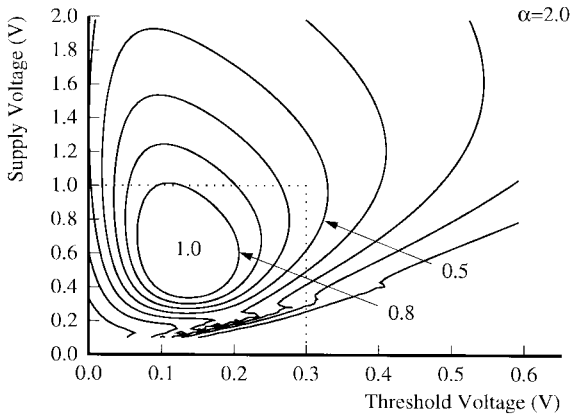

Fig. 2. EDP contours without velocity saturation.

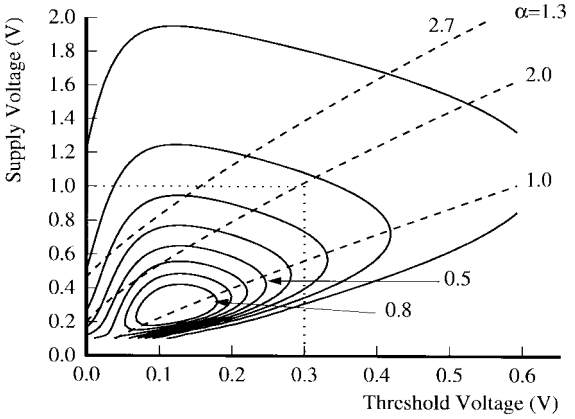

Fig. 3. Contours of constant energy delay product.

lines are much closer together, indicating the surface is much steeper. At the same operating point as before, $V = 1$ V and $V_{\rm th} = 300$ mV, the EDP of the circuit is four times that of the optimal. Thus, the model as described so far predicts that very low supply and threshold voltages would be beneficial for reducing power.

Most circuits must meet specific performance targets, so it is important to look at the actual performance in addition to the energy-delay product. The dashed lines in Fig. 2 show contours of constant performance. Performance was normalized to that of the performance contour that runs approximately through the optimal point. If the circuit must operate somewhere along the topmost performance contour, then the designer could reduce power by more than a factor of three without changing performance by moving from using a 2-V supply and 0.5-V $V_{\rm th}$ to using a 0.8-V supply and 0.1-V $V_{\rm th}$. But this point is still not at the optimal energy-delay product. To further improve the energy efficiency requires reducing the supply with constant threshold voltage, which will make the gates slower. The gates at the optimal point are 2.7× slower than the highest performance curve. If the logic can be changed to reduce the levels of logic in the design, there is about another factor of three reduction in power that is possible.

While the reduced gate performance is one factor that is keeping supply voltages from dropping too quickly, this

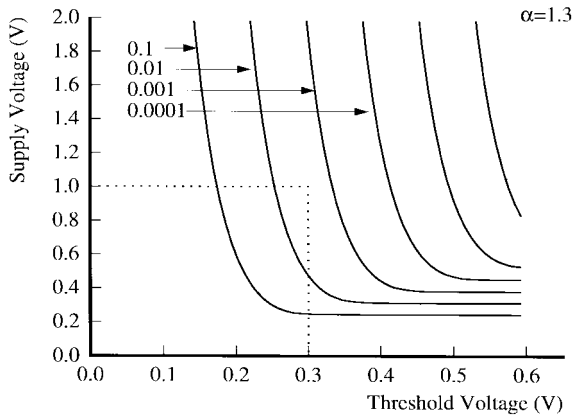

Fig. 4. Ratio of leakage to total power.

simple analysis indicates that submicron technologies with low threshold voltages should be very attractive for low power applications. By simply moving to a low $V_{\rm th}$ process, a designer could reduce the supply voltage and power without requiring a major change of the design, since the gate speed would remain constant. If the design could be changed to use slower basic gates, further power savings would be possible. Unfortunately, the next two sections show that much of this advantage is illusory; when sleep modes and variation in voltages are taken into account, the advantage of operating at these very low voltages is greatly diminished.

III. Sleep Mode

There are some applications where the circuit must be idle for extended periods of time. During this time it is not always possible to shut-off the power supply since the circuit must maintain state. One can reduce the dynamic power by simply reducing or completely stopping the clock signal. However, during this period leakage power remains constant. In recent processor implementations [5], [7] the idle power has been limited by the leakage current of the transistors. One can estimate the ratio of active to idle power by finding the ratio of leakage to total power in the circuit. Fig. 4 shows the ratio of leakage power to total power. When the ratio of active to idle power must be a factor of 1000, then the threshold voltage will be limited to be at least 300 mV or so. Thus, the minimum EDP of the circuit will be limited. In order to get around this constraint, the circuit would need a mechanism to change the effective $V_{\rm th}$ of its transistors. It would use the lower value during normal operation and use the higher values during sleep modes. Some papers have proposed using high threshold power switch devices for this function [8] while others have proposed direct control of $V_{\rm th}$ [11]. While this adaptive threshold control requires overhead in design time, area, and energy, it might be needed to deal with the more significant problem caused by process and operating point variation, which is described in the next section.

IV. Process and Operating Point Variation

The previous analysis assumed that the supply and threshold voltages were fixed, although in reality, both have small variations. In this section we derive new energy-delay curves which take into account the effect of these variations. We first look at how to define the energy-delay product when variations are present and then derive energy-delay curves for the same technology used in the previous section. We also look at the results from a real 0.25-μm technology by using HSPICE to generate the power and delay numbers for the average inverter.

We have so far assumed that the supply and threshold voltages can be controlled perfectly. For real circuits there is always uncertainty in the value of V and $V_{\rm th}$. There are two main sources of uncertainty. The first is that circuits must work over a range of operating conditions. The supply voltage is normally specified to be within $\pm$10% of the nominal value. The operating temperature can be anywhere between 25°C and 100°C. Since the threshold voltage changes by $-$0.8 mV/°C, the variation in T introduces approximately 60 mV of uncertainty in $V_{\rm th}$. The second source of uncertainty is the variability in $V_{\rm th}$ due to the manufacturing process. The threshold voltage of transistors within a chip, or a single transistor across chips, varies randomly. We model $V_{\rm th}$ as a random variable with Gaussian probability density function (PDF) and a standard deviation of 35 mV. This gives a 3 σ of about $\pm$100 mV.

Introducing uncertainty into V and $V_{\rm th}$ causes the delay and energy to be spread out over a range. The solid line in Fig. 5 shows how the energy and delay vary as $V_{\rm th}$ varies over a range. In this example supply voltage and temperature are fixed. Every design is specified to have a maximum energy and delay. That is, the design is guaranteed not to dissipat more energy than the maximum and guaranteed to meet or exceed the minimum performance. These limits correspond to the vertical and horizontal lines in Fig. 5. The design is guaranteed to lie somewhere below and to the left of these two lines. If $V_{\rm th}$ is a fixed quantity, then setting the design specifications is easy. All parts will have the same energy and the same delay. In Fig. 5 the star represents this point for some value of $V_{\rm th}$. If $V_{\rm th}$ varies over a range, then the specifications need to be relaxed, or else few parts will meet the targets. This corresponds to shifting the target lines up and right in Fig. 5, as shown by the dotted lines. Reducing the area below and to the left of the target or cutoff lines improves the energy and delay specifications, but reduces the number of parts that meet the specifications. Where to draw the cutoff lines is somewhat arbitrary. The important point is that even though the parts will fall somewhere along the solid line, they will be sold as if they operated at the intersection of the energy and delay cutoff lines, marked with a filled triangle. Thus, we use the product of the cutoff numbers as the EDP of the circuit, even though no part can have both the worst-case power and worst-case delay simultaneously.

Since we model $V_{\rm th}$ as a normal random variable, both E and D will follow some kind of distribution. One possible cutoff line is at the mean of the energy and the mean of delay. However, at this point a relatively small number of chips would meet both spec limits. We use instead the mean plus one standard deviation. In addition to the uncertainty due to $V_{\rm th}$ variations, we must also account for the variations due to V and T. We therefore solve the equations at four process

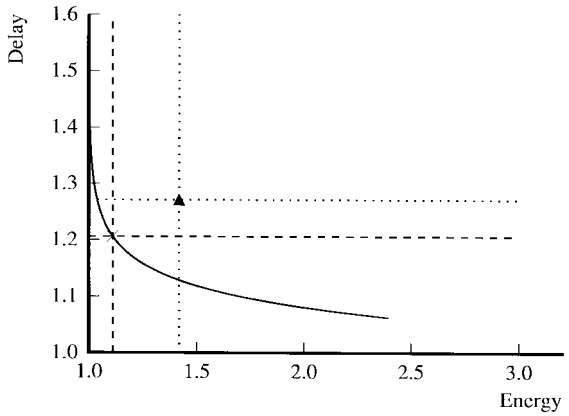

Fig. 5. Variation in energy and delay.

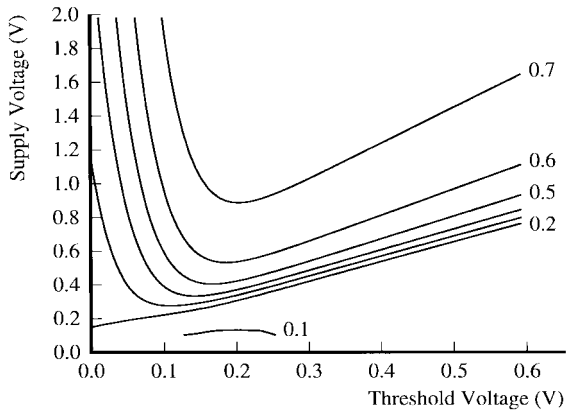

Fig. 7. Ratio of EDP without and with uncertainty.

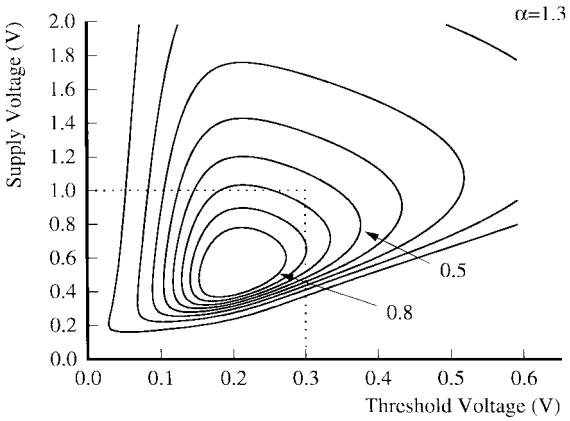

Fig. 6. EDP contours with uncertainty.

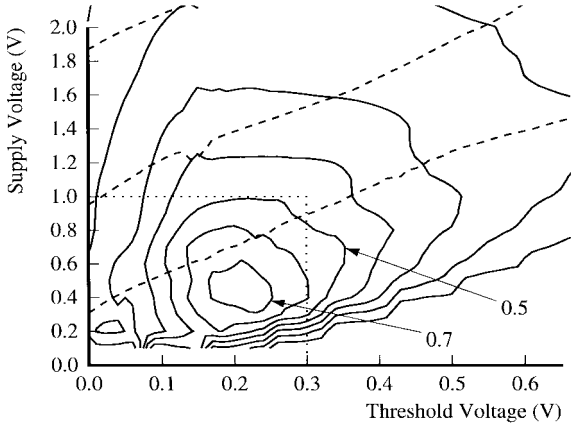

Fig. 8. EDP contours using HSPICE models.

corners (low V low T, high V low T, high V high T, low V high T). At each corner we find the mean and the standard deviation of E and D. The EDP then is the product of the worst case energy at any of the four corners times the worst case delay at any of the four corners.

Fig. 6 shows contours of the inverse of the relative EDP when there is uncertainty in $V, V_{\rm th}$, and T. In this case the EDP surface is again relatively flat even though transistors are velocity saturated ($\alpha = 1.3$). The optimal operating point has moved to approximately $V = 0.5$ V and $V_{\rm th} = 200$ mV. The surface becomes flat because at lower supply and threshold voltages the delay and the energy become more sensitive to variations in V and $V_{\rm th}$. Thus, operating at higher voltages is beneficial. The largest change in $V_{\rm th}$ is due to the temperature variation, which introduces approximately 60 mV of uncertainty in $V_{\rm th}$.

The effect of uncertainty then is to reduce the overall efficiency of the circuit. Although on average the circuit will operate at much higher efficiency, it can only be guaranteed to work for the worst-case conditions. The penalty for this worst-case operation becomes very severe for low-voltage, low-threshold operating conditions. This is shown in Fig. 7, which gives the ratio of the EDP without uncertainty to the EDP with uncertainty. At operating voltages above 1 V, and $V_{\rm th}$ above 0.2 V, the effect of the variations is modest, less than 40% reduction. But the cost of the variations is around a factor of four at the previously found optimal point, $V = 250$ mV, $V_{\rm th} = 120$ mV. This high cost at low operating voltages is what flattens out the curves.

So far we have used simple models to approximate the energy and delay of a CMOS gate. Using this model allows us to understand how the EDP depends on the different circuit and process parameters, but the model's simplicity raises questions about its accuracy. In order to ensure that our models are correct, we compared our model with the results from running HSPICE. Using HSPICE we simulated a chain of inverters and computed the delay and energy per μm of gate width, as we swept the supply and threshold voltage. We used the level 37 models of a next generation 0.25-μm process from Texas Instruments [9] for our simulations. The nominal threshold voltage is $V_{\rm th} = 0.4$ V. We adjusted $V_{\rm th}$ by modifying the $VT0$ parameter of the HSPICE models. The results of the simulations are shown in Fig. 8. The figure shows contours of the inverse of the relative EDP. The simulations consider variations in V and T.

The contour lines are strikingly similar to previous figures. The EDP surface is steeper, but this is not surprising since our simulations do not consider uncertainty in $V_{\rm th}$. We should note, however, that absolute values of $V_{\rm th}$ are hard to compare across processes, because there is no common definition of threshold voltage. We also simulated a chain of inverters using BSIM2 models of the HP CMOS14B process. This is a 0.6-μm

process with nominal $V_{\rm th} = 0.9$ V. We set $V_{\rm th}$ by adjusting the flat-band voltage parameter VFB. The most striking difference with the curves shown in Fig. 7 is that the optimal operating point occurs at a much higher threshold voltage (0.5 V versus 0.2 V). Most of this difference is a result of HSPICE using a different definition of $V_{\rm th}$, which accounts for about 250 mV of the shift. The remaining difference is because of drain-induced barrier lowering (DIBL) [12] which lowers the effective threshold voltage at higher supply voltages. Thus, for high supply voltages the effective $V_{\rm th}$ is lower than shown in the figure.

The dashed lines in Fig. 8 show contours of constant performance. Thus, if a circuit is operating at the nominal point for this process ($V = 2.0$ V, $V_{\rm th} = 0.4$ V) the range of EDP one can reach is limited. In order to approach the minimum, one must be willing to give up a factor of $2\times$ in gate performance, but this gives a factor of $1.7\times$ reduction in the energy of the operation.

One way to reduce the effect of uncertainty is to use adaptive techniques to regulate the supply and threshold voltage. That is, dynamically adjust the supply and/or threshold voltage such that the circuit meets the required specifications [6], [11], [14], [15]. In Fig. 5 this corresponds to being able to guarantee that all chips will be in a small rectangle near the point marked with a star. As was stated earlier, using adaptive techniques, it is also possible to adjust the threshold voltage when the circuit is idle in order to reduce the leakage power [11]. These techniques seem most promising when applied to circuits attempting to operate at very low voltages, where the cost of variations is very high. The area overhead of the regulating circuits is usually negligible [6], [11]. The cost of these adaptive techniques is hard to determine, since it depends on how well the feedback controls the desired parameters, the overhead of the feedback control, and how frequently the circuit switches between active and idle modes.[3] These techniques look promising, but more research is needed to fully understand what dominates the cost and when using them is advantageous.

V. Conclusions

We found the supply and threshold voltages for optimal EDP using a first-order model of energy and delay in CMOS circuits that take into account leakage current. The location of this point and the shape of the EDP surface near the minimum are a strong function of how velocity-saturated the transistors are. If transistors are not velocity-saturated then the EDP surface is relatively flat. As transistors become more velocity-saturated, the EDP surface becomes steeper and the optimal point moves closer to the origin. For a 0.25-μm technology, this analysis yielded a supply of 250 mV and $V_{\rm th}$ of 120 mV. One difficulty with operating at this point is that the speed of each gate is modest, forcing the designers to reduce the levels of logic in their design to maintain performance. If this was the only issue, designers would use technologies with scaled thresholds and simply operate them at higher voltages (1 V versus 250 mV) to recoup the lost gate speed. The main difficulty with these low voltage operating points is dealing with the variations caused by changes in operating conditions and threshold voltages.

While low operating voltages looked very attractive for low power operation, they are very sensitive to both manufacturing variations and operating point changes. If you need to provide margins in your circuit to ensure it will meet certain speed and power requirements, the advantage of using technologies with very low threshold voltages disappears. When variations are considered, the optimal point moves to higher voltages, and the whole energy-delay surface becomes flatter. In order to achieve the large potential gains of operating at low voltages, the circuit needs to use some kind of adaptive control on both the threshold voltage and the supply, to reduce the effective variation that the circuit sees. Until energy efficient techniques are developed to accomplish this, the supply and threshold scaling is likely to be more modest, probably at a rate that maintains constant electric fields within the devices.

[3] If substrate bias is used to control $V_{\rm th}$, every time $V_{\rm th}$ changes it is necessary to charge or discharge the substrate capacitance, which is large.

Appendix

In order to find a mathematical solution for the optimal V and $V_{\rm th}$, the equations used in Section II are relatively simple. Accounting for second-order effects or different modes of operation makes the equations hard to manage. When numerically solving for the delay and energy, however, it is possible to account for second-order effects. Thus, the equations solved are not the ones presented in Section II, but rather those shown below. One of the most important differences is that transistor current is not zero in the subthreshold region. Thus, it is possible to find both the energy and the delay of circuits even when $V < V_{\rm th}$

$$I_l = I_s e^{(V_{gs}-V_{\rm th})/\gamma V_o}\left(1 - e^{-(V_{ds}/\gamma V_o)}\right). \tag{13}$$

If $V > V_{\rm th}$

$$I = K(V - V_{\rm th})^{\alpha} + I_s\left(1 - e^{-(V_{ds}/\gamma V_o)}\right) \tag{14}$$

else

$$I = I_l \tag{15}$$

$$T_g = \frac{CV}{I} \tag{16}$$

$$E = aCV^2 + I_s e^{-(V_{\rm th}/\gamma V_o)}\left(1 - e^{-(V_{ds}/\gamma V_o)}\right) V L_d T_g. \tag{17}$$

Acknowledgment

The authors would like to thank D. Ramsey of MIPS Technology Inc. for the data shown in Fig. 1.

References

[1] J. B. Burr, "Cryogenic ultra low power CMOS," in *IEEE Int. Symp. Low Power Electronics*, 1995, p. 9.4.
[2] J. B. Burr and A. M. Peterson, "Ultra low power CMOS technology," in *NASA VLSI Design Symp.*, 1991, pp. 4.2.1–4.2.13.
[3] A. Chatterjee, M. Nandakumar, and I. Chen, "An investigation of the impact of technology scaling on power wasted as short current in low voltage CMOS," in *IEEE Int. Symp. Low Power Electronics and Design*, Aug. 1996, pp. 145–150.
[4] D. Dobberphul, Personal communication.

[5] V. Kaenel, D. Aebischer, C. Piquet, and E. Dijkstra, "A 320 MHz 1.5 mW at 1.35 V CMOS PLL for microprocessor clock generation," in *IEEE Int. Solid-State Circuits Conf.*, Feb. 1996, pp. 132–133.
[6] T. Kuroda, T. Fujita, S. Mita, *et al.* "A 0.9 V 150 MHz 10 mW 4 mm^2 2-D discrete cosine transform core processor with variable-threshold-voltage scheme," in *IEEE Int. Solid-State Circuits Conf.*, Feb. 1996, pp. 166–167.
[7] J. Montanaro, R. T. Witek, *et al.*, "A 160 MHz 32 b 0.5 W CMOS RISC microprocessor," in *IEEE Int. Solid-State Circuits Conf.*, Feb. 1996, pp. 215–215.
[8] S. Mutoh, S. Shigematsu, Y. Matsuya, *et al.*, "A 1 V multi-threshold voltage CMOS DSP with an efficient power management technique for mobile phone applications," in *IEEE Int. Solid-State Circuits Conf.*, Feb. 1996, vol. 39, pp. 168–169.
[9] M. Nandakumar, A. Chatterjee, M. Rodder, *et al.*, "A device design study of 0.25 μm gate length CMOS for 1 V low power applications," in *IEEE Int. Symp. Low Power Electronics*, Oct. 1995, pp. 80–81.
[10] T. Sakurai and A. R. Newton, "Alpha-power law MOSFET model and its application to CMOS inverter delay and other formulas," *IEEE J. Solid-State Circuits*, vol. 25, pp. 584–593, Apr. 1990.
[11] K. Seta, H. Hara, T. Kuroda, *et al.*, "50% active-power saving without speed degradation using standby power reduction (SPR) circuit," in *IEEE Int. Solid-State Circuits Conf.*, Feb. 1995, vol. 38, pp. 318–319.
[12] R. R. Troutman. "Drain-induced barrier lowering," *IEEE J. Solid-State Circuits*, vol. 14, p. 383, Apr. 1979.
[13] H. Veendrick, "Short-circuit dissipation of static CMOS circuitry and its impact on the design of buffer circuits," *IEEE J. Solid-State Circuits*, vol. 19, pp. 468–473, Aug. 1984.
[14] C. Vieri, I. Yang, A. Chandrakasan, *et al.*, "SOIAS: dynamically variable threshold SOI with active substrate," in *IEEE Int. Symp. Low Power Electronics*, Oct. 1995, pp. 86–87.
[15] G.-Y. Wei and M. Horowitz, "A low power switching power supply for self-clocked systems," in *IEEE Int. Symp. Low Power Electronics and Design*, Aug. 1996, pp. 313–317.

Ricardo Gonzalez received the B.S. and M.S. degrees in electrical engineering from Stanford University, Stanford, CA, in 1990 and 1992, respectively. He is currently pursuing the Ph.D. degree in the same field at Stanford.

His research interests are in the area of processor architecture and low-power circuits.

Benjamin M. Gordon was born in State College, PA, in 1965. He received the B.S. degree in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 1987 and the M.S. and Ph.D. degrees from Stanford University, Stanford, CA, in 1992 and 1996, respectively.

From 1987 to 1991 he worked as a Systems Engineer for Advanced Processing Labs, Inc., San Diego, CA, on real-time signal processing systems. In 1996 he joined the faculty at the University of Washington, Seattle, as an Assistant Professor of Electrical Engineering. His research interests include integrated circuits for low power and multimedia applications.

Mark A. Horowitz received the B.S. and M.S. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 1978 and the Ph.D. degree in the same field from Stanford University, Stanford, CA, in 1984.

Since September 1984, he has been working in the Computer Systems Laboratory at Stanford where he is currently an Associate Professor in electrical engineering. His research area is in digital system design. He has led a number of processor design projects at Stanford including MIPS-X, one of the first processors to include an on-chip instruction cache, and TORCH, a statically-scheduled, superscalar processor. He has also worked in a number of other chip design areas including high-speed memory design, high-bandwidth interfaces, and fast floating point. In 1990 he took leave from Stanford to help start Rambus, Inc., a company designing high-bandwidth memory interface technology. His current research includes multiprocessor design, low-power circuits, memory design, and processor architecture.

Dr. Horowitz is the recipient of a 1985 Presidential Young Investigator Award, an IBM Faculty development award, as well as the 1993 Best Paper Award at the 1994 International Solid State Circuits Conference.