Ultralow-voltage, minimum-energy CMOS

S. Hanson
B. Zhai
K. Bernstein
D. Blaauw
A. Bryant
L. Chang
K. K. Das
W. Haensch
E. J. Nowak
D. M. Sylvester

Energy efficiency has become a ubiquitous design requirement for digital circuits. Aggressive supply-voltage scaling has emerged as the most effective way to reduce energy use. In this work, we review circuit behavior at low voltages, specifically in the subthreshold ($V_{dd} < V_{th}$) regime, and suggest new strategies for energy-efficient design. We begin with a study at the device level, and we show that extreme sensitivity to the supply and threshold voltages complicates subthreshold design. The effects of this sensitivity can be minimized through simple device modifications and new device geometries. At the circuit level, we review the energy characteristics of subthreshold logic and SRAM circuits, and demonstrate that energy efficiency relies on the balance between dynamic and leakage energies, with process variability playing a key role in both energy efficiency and robustness. We continue the study of energy-efficient design by broadening our scope to the architectural level. We discuss the energy benefits of techniques such as multiple-threshold CMOS (MTCMOS) and adaptive body biasing (ABB), and we also consider the performance benefits of multiprocessor design at ultralow supply voltages.

1. Introduction

Mobile battery-powered electronic devices have created a growing demand for energy-efficient circuit design. Cellular phones alone represent a large industry and create both an opportunity for innovation and the potential for profitability. Future progress in mobile electronics will depend on the development of inexpensive devices with complex functionality and long battery life. The aim of this paper is to show how devices, circuits, and architectures within this design space may be optimized for minimum energy consumption.

Even in the realm of high-performance microprocessors, power has become a limiting constraint. Traditional scaling of high-performance FETs uses a combination of supply-voltage (V_{dd}) and threshold-voltage (V_{th}) reduction to accommodate both performance and power requirements, but the rapid rise of subthreshold and gate leakage has placed limits on this scaling strategy. It is clear that new strategies are necessary to address the power concerns in high-performance designs.

Voltage scaling is the most effective solution to stringent power requirements and has been practically demonstrated in a number of designs. Reduction of the supply voltage (with a fixed threshold voltage) results in a quadratic reduction of dynamic energy at the expense of decreased performance. For many applications, this performance penalty is tolerable. In fact, for a wide range of applications, including sensors and medical devices, a significant performance penalty may be tolerated without compromising the usefulness of the device. High-performance designs may also take advantage of supply-voltage reduction during idle periods when the circuit is performing simple background routines, because performance requirements are relaxed or removed altogether. Regardless of the application, the use of aggressive voltage scaling can lead to considerable energy reductions whenever performance demands are low for a circuit.

This paper explores the limits of minimum-energy CMOS. We show that, for large classes of circuits, minimum energy consumption occurs when the voltage is scaled below the device threshold voltage. In this region, called the subthreshold (sub-V_{th}) regime, energy consumption can be reduced by 20x relative to standard superthreshold ($V_{\mathrm{dd}} > V_{\mathrm{th}}$) operation. We use a

0018-8646/06/$5.00 © 2006 IBM

hierarchical approach to the exploration of sub-V_{th} design. We begin by discussing the evolution of device behavior as supply voltage is reduced into the sub-V_{th} regime. We then develop intuition and a methodology for minimum-energy design by considering circuit behavior in the sub-V_{th} regime. Finally, we use this intuition to discuss how architectural techniques may be used to improve energy efficiency in designs dedicated to low-energy operation as well as designs with both performance and energy requirements.

At the device level, we compare FET sub-V_{th} and super-V_{th} characteristics and sensitivities. We find that sub-V_{th} FET currents are exponentially dependent on V_{th} and V_{dd} and that this presents the biggest challenge to device, circuit, and architecture design. In a design that must operate over a wide range of voltages and performance levels, minor tradeoffs, such as lengthening FET channels slightly to reduce V_{th} variation and making minor V_{th} adjustments to maintain nominal matching between FETs at low V_{dd}, should be considered. However, if energy minimization at low V_{dd} is the critical goal, significant device-optimization studies must be considered. Dual-gate FETs show much promise for future work in energy-optimal design. If combined with low-workfunction metal gates and an increased channel length, dual-gate FETs can help minimize V_{th} variations and achieve a steep sub-V_{th} slope, the key parameters for sub-V_{th} operation.

At the circuit level, we present a simple analytical model for the energy-optimal supply voltage, V_{min}. This simple model illustrates the tradeoff between leakage energy and dynamic energy that occurs in energy-optimal circuits. We find that energy efficiency is limited by the rise of leakage and that the designers of energy-efficient circuits should reduce V_{min} until it approaches $V_{dd,limit}$, the minimum functional voltage. Additionally, we find that heightened sensitivity to the threshold voltage in combination with a low I_{on}/I_{off} ratio results in serious circuit-level robustness concerns when process variation is being considered. Energy efficiency also exhibits a strong sensitivity to threshold variability.

We pay special attention to SRAM arrays. We find that large SRAM arrays have higher V_{min} and $V_{dd,limit}$ values than those used for standard logic. For a standard six-transistor SRAM (6T-SRAM) array, $V_{dd,limit}$ is shown to be higher than V_{min}, suggesting that significant redesign will be necessary to produce robust, energy-efficient SRAM design. The problem is further complicated by process variation, particularly V_{th} mismatch introduced by random dopant fluctuations. The eight-transistor SRAM (8T-SRAM) cell is presented as a feasible solution.

In the final section of the paper, we discuss energy-efficient architectural techniques. We suggest that techniques such as multi-threshold CMOS (MTCMOS), adaptive body biasing (ABB), and the use of voltage islands can help a design achieve energy optimality by shifting V_{min} toward $V_{dd,limit}$. Architectural techniques, specifically those that involve multiprocessor design, can ameliorate the performance penalty suffered as a result of low-voltage operation.

The paper is organized as follows. In Section 2 we discuss device-level behavior at sub-V_{th} and near-V_{th} voltages. Section 3 includes a discussion of the implications of device-level changes at the circuit level and a general and useful energy model for sub-V_{th} operation. The complications introduced by SRAM design are given special consideration. Finally, in Section 4 we discuss energy-efficient architectural techniques targeted at both dedicated minimum-energy operation and high-performance operation.

2. Device characteristics at ultralow-voltage operation

As V_{dd} is reduced to minimize energy per operation, FETs make the transition from superthreshold (super-V_{th}) operation in strong inversion with large gate overdrives, to near-V_{th} operation in weak inversion with very small overdrives, and finally into sub-V_{th} operation. Sub-V_{th} operation differs from super-V_{th} operation primarily because the sub-V_{th} on-current ($I_{on\text{-}sub}$) depends exponentially on threshold voltage (V_{th}) and power-supply voltage (V_{dd}), while the typical super-V_{th} operation on-current ($I_{on\text{-}super}$) depends roughly linearly on V_{th} and V_{dd}. The $I_{on\text{-}sub}$ exponential sensitivities to V_{th} and V_{dd} are captured in the following equation:

$$I_{on\text{–}sub} = \frac{W}{L_{eff}} \cdot \mu_{eff} \cdot C_{ox} \cdot (m-1) \cdot v_T^2 \cdot \exp\left(\frac{V_{gs}-V_{th}}{m \cdot v_T}\right) \cdot \left[1-\exp\left(-\frac{V_{ds}}{v_T}\right)\right], \qquad (1)$$

where $v_T = kT/q$. In these equations, T is temperature, v_T is the thermal voltage, k is Boltzmann's constant, q is the charge of an electron, L_{eff} is the effective gate length, μ_{eff} is the effective mobility, C_{ox} is the oxide capacitance, W is the gate width, and m is the subthreshold slope factor. On-current is defined in this paper as I_{ds}, when $V_{gs} = V_{ds} = V_{dd}$. It is important to highlight the implicit V_{th} dependence on L_{eff} in Equation (1) because $I_{on\text{-}sub}$ becomes very sensitive to L_{eff} due to the V_{th} term. V_{th} is also dependent on V_{ds} via drain-induced barrier lowering (DIBL), which plays a role in determining the effect V_{dd} has on $I_{on\text{-}sub}$. The linear sensitivity of $I_{on\text{-}super}$ to V_{th} and V_{dd} for short-channel FETs is captured in the equation

$$I_{on\text{–}super} = \frac{g_{msat}}{1 + R_s \cdot g_{msat}} \cdot (V_{dd} - V_{th} - V_{PO}), \qquad (2)$$

where R_s is the FET source resistance and g_{msat} is the saturated transconductance, which depends on L_{eff}, C_{ox}, and the carrier saturation velocity. V_{PO} is the pinch-off voltage. The near-V_{th} I_{on} sensitivity to V_{th} and V_{dd} is bounded by the sub-V_{th} and super-V_{th} sensitivities. **Figure 1** highlights the differences between super-V_{th} and sub-V_{th} current characteristics. **Tables 1** and **2** compare key parametric sensitivities of FETs in sub-V_{th}, near-V_{th}, and super-V_{th} operation.

The exponential sub-V_{th} I_{on} sensitivity to V_{th} drastically affects circuit behavior. First, the circuit delay and power now also depend exponentially on V_{th} and V_{dd}. More significantly, current matching between two FETs is exponentially dependent on any difference in V_{th}. For example, while a reasonable 6σ 100-mV V_{th} mismatch disturbs the FET current ratios by only approximately 1.17x in super-V_{th} operation, a similar 100-mV V_{th} mismatch upsets the current matching by greater than 10x in sub-V_{th} operation. (We use "x" throughout this paper to indicate "times," so that, for example, "10x" means a factor of 10 times.) This extreme sensitivity to V_{dd} and V_{th} presents the most significant challenge to sub-V_{th} and near-V_{th} circuit functionality, and is discussed in later circuit sections.

Product requirements dictate how device optimizations may be used to increase energy efficiency. One product application may have performance restrictions and will therefore be required to operate at high V_{dd}. In this case, it is likely that the process will be similar to a typical super-V_{th} process. A multiple-core microprocessor may be an example of such an application, in which the V_{dd} of each core is varied according to performance needs and power constraints during operation. In this scenario, a few key circuits may require modification to enable low-voltage operation, but the potential to minimize energy is ultimately limited by the high-performance requirements.

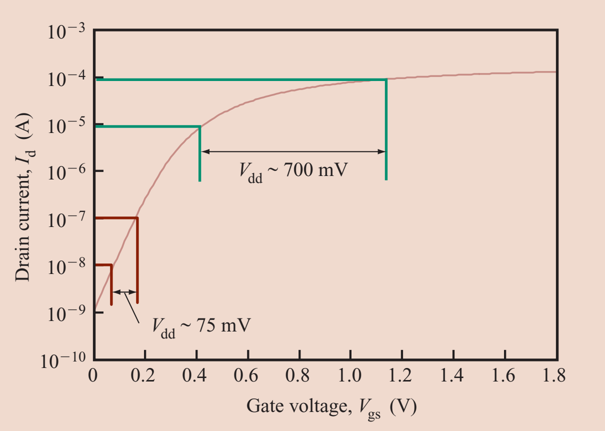

Figure 1

Comparison of super-V_{th} and sub-V_{th} current characteristics. In the sub-V_{th} region (red), the current increases at an exponential rate of ~85 mV/decade. Above V_{th} (green), current increases at a much slower non-exponential rate. Thus, the I_{on}/I_{off} ratio is maximized by setting $V_{th} > V_{dd}$ and operating in the sub-V_{th} regime.

There may also be some small technology modifications (e.g., small V_{th} adjustments) that enable low-V_{dd} operation without a significant high-V_{dd} performance impact. On the other hand, if the application is aimed solely at low-V_{dd} operation, the technology and circuits can be optimized to minimize total energy consumption. A number of techniques for addressing these two very different scenarios at the architectural level are discussed in Section 4. In this section, we first consider FET low-V_{dd} characteristics and sensitivities that have implications for both of these scenarios.

The exponential sensitivity to V_{th} in sub-V_{th} and near-V_{th} operation changes the impact that key device

Table 1 Comparison of key sub-V_{th}, near-V_{th}, and super-V_{th} n-FET sensitivities [65-nm technology, room temperature. Effective FET channel length (L_{gate}) is approximately 35 nm.].

	Sub-V_{th}	*Near-V_{th}*	*Super-V_{th}*
V_{dd}	200 mV	400 mV	1 V
$V_{th,sat}$	270 mV	250 mV	180 mV
I_{on}	~20 μA/μm	~80 μA/μm	~1 mA/μm
Sensitivity of I_{on} to 100-mV V_{dd} reduction	18x	4.6x	1.20x
Sensitivity of I_{on} to 100-mV V_{th} increase	11x	3.7x	1.17x
Sensitivity of I_{off} to 100-mV V_{th} increase	16x	15x	12x
Sensitivity of $I_{on,n\text{-}FET}/I_{on,p\text{-}FET}$ ratio to 100-mV V_{th} mismatch	10x	3.7x	1.17x
I_{on}/I_{off} ratio	160x	3,150x	7,000x
I_{on}/I_{off} ratio vs. 100-mV V_{th} increase	1.44x	4.2x	11x

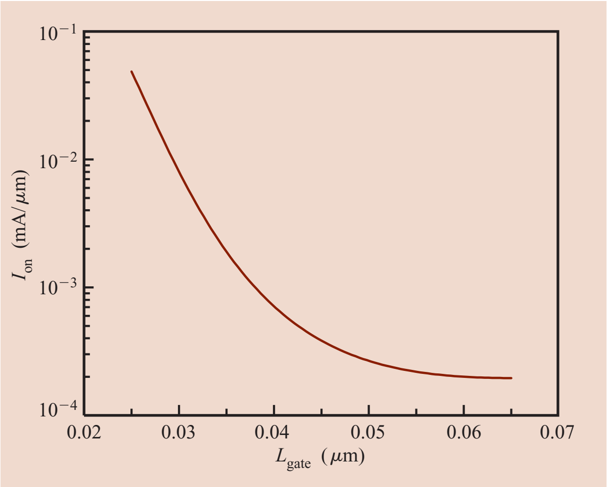

Figure 2

Sub-V_{th} I_{on} as a function of L_{gate} (SOI 65-nm technology with V_{dd} = 200 mV).

Table 2 Comparison of additional key sub-V_{th} and super-V_{th} n-FET sensitivities [65-nm technology, room temperature. Effective FET channel length (L_{gate}) is approximately 35 nm.].

	Sub-V_{th}	*Super-V_{th}*
Sensitivity of I_{on} to 0.9x inverse sub-V_{th} slope reduction (at constant I_{off})	~1.7x	~1.03x
Sensitivity of I_{on} to 1.3x decrease in T_{ox} (at constant I_{off})	~1.7x	~1.23x
Sensitivity of I_{on} to 1.3x increase in L (at constant V_{th} and slope)	~0.77x	~0.94x
Sensitivity of I_{on}/I_{off} ratio to 1.3x increase in L (at constant V_{th} and slope)	~1x	~1.22x
Sensitivity of I_{on} to 1.3x increase in mobility	~1.3x	~1.05x
Sensitivity of I_{on}/I_{off} ratio to 1.3x mobility change (at constant I_{off})	~1x	~1.04x

parameters have on FET currents (see Tables 1 and 2). For example, a 10% reduction in inverse sub-V_{th} slope increases sub-V_{th} I_{on} by 1.7x and super-V_{th} I_{on} by only 3% (sub-V_{th} slope measures the slope of the drain current with respect to gate voltage and is commonly quoted in its inverse form in mV/decade). As a result, sub-V_{th} I_{on} is much more sensitive to FET gate insulator thickness, t_{ox}, because t_{ox} plays a critical role in determining the sub-V_{th} slope. In typical high-performance technologies, FET channels are made as short as possible, and as a consequence sub-V_{th} slope is suboptimal. Reducing t_{ox} improves the sub-V_{th} slope and significantly increases sub-V_{th} I_{on}. The impact of the sub-V_{th} slope improvement in super-V_{th} I_{on} is considerably less. In reality, the observed super-V_{th} I_{on} increase results from the sublinear saturated transconductance dependence on t_{ox}. (*Transconductance* is an expression of the current-carrying ability of a FET. In general, the larger the transconductance value for a device, the greater the gain it is capable of delivering.) The example in Table 2 shows that a 1.3x reduction in t_{ox} improves the sub-V_{th} I_{on} by 1.7x and improves the super-V_{th} I_{on} by only 1.23x. The authors of [1] show that improved sub-V_{th} slope makes a significant contribution to the energy savings observed in devices optimized for sub-V_{th} operation. As we see in Section 3, the leakage reduction resulting from sub-V_{th} slope improvement provides an attractive strategy for energy minimization.

The impact of FET channel length (L_{gate}) on sub-V_{th} I_{on} is due predominantly to the dependence of V_{th} and the sub-V_{th} slope on L_{gate}. At short channels, V_{th} decreases and sub-V_{th} slope degrades as the value of L_{gate} is reduced because of drain barrier lowering and other short-channel effects. As a consequence, the sub-V_{th} current increases exponentially at short L_{gate} values, as shown in **Figure 2** for an n-FET in a typical 65-nm technology. Typically, high-performance FETs use L_{gate} in the region where these parameters vary strongly with length. This creates a considerable challenge for sub-V_{th} operation because small variations in L_{gate} values have an enormous impact on I_{on}. In particular, L_{gate} linewidth variation leads to significant mismatch in FET drive strengths. This effect can be reduced by increasing values of L_{gate} (**Figure 3** shows I_{on} variation resulting from linewidth variation as a function of L_{gate}), but increased gate length degrades super-V_{th} performance. (The term δL refers to a 3σ variation in linewidth.) On the other hand, sub-V_{th} performance is not affected as severely as super-V_{th} performance, because current can be regained with a small reduction of V_{th}, with no impact on the I_{on}/I_{off} ratio. More significantly, the additional capacitive loading associated with increasing L_{gate} is significantly smaller for sub-V_{th} than it is for super-V_{th}, as shown in **Table 3**. Sub-V_{th} operation at longer L_{gate} values gives the added advantage of a steeper sub-V_{th} slope. Similar tradeoffs must also be considered with respect to narrow FET channel widths (W). However, the choice of L_{gate} and W are greatly affected by the circuit application requirements. Gate dimensions have less flexibility in the extreme case in which high performance and high V_{dd} are at a premium. In this case, the designer can account for

Table 3 Sensitivity of sub-V_{th} vs. super-V_{th} inverter-chain node capacitance to channel length (130-nm technology).

	Sub-V_{th}	*Super-V_{th}*
V_{dd}	200 mV	1.2 V
Total node capacitance $L = 120$ nm (fF)	2.9	3.3
Total node capacitance at $L = 240$ nm (fF)	3.1	4.9
Ratio	1.1x	1.5x

Table 4 Sources of random V_{th} mismatch in 65-nm SOI technology. (ACLV: across-chip channel length variation; RTA: across-chip rapid thermal anneal.)

	ACLV	*RTA*	*Doping fluctuation ($L = 35$ nm, $W = 500$ nm)*	*Doping fluctuation ($L = 35$ nm, $W = 140$ nm)*
3σ V_{th} mismatch (mV)	25	28	30	58

the impact of the V_{th} and slope variations correctly only when predicting sub-V_{th} and near-V_{th} circuit behaviors.

Another point to consider is that p-FET and n-FET thresholds can decrease at different rates as V_{dd} is reduced. This implies that the n–p I_{on} matching can change drastically, as indicated in **Figure 4**. Small V_{th} adjustments can be made to provide the optimal matching for sub-V_{th} and near-V_{th} operation; but again, super-V_{th} operation may deteriorate.

Random channel dopant fluctuation (RDF) is another source of threshold variations that results in FET current mismatch. The 3σ V_{th} variation (δV_{th}) induced by RDF is inversely proportional to the square root of the channel area ($\delta V_{th} \sim A/(W \times L)^{1/2}$, where A is a constant of 4 in units of mV × μm, W is the FET channel width in μm, and L is the FET channel length in μm) [2]. **Table 4** compares V_{th} mismatch induced by RDF, across-chip channel-length variations (ACLVs), and across-chip rapid thermal anneal (RTA) variations in a 65-nm technology ($L_{gate} \sim 35$ nm). At 65 nm, RDF-induced V_{th} mismatch is comparable to other sources of V_{th} variability and will dominate in future technologies as channel areas are scaled down. Although the 30–60 mV of V_{th} variation has a significant impact on super-V_{th} matching, a similar variation in the sub-V_{th} region results in a 2–3x variation in I_{on}. Some relief from current variation can be gained by increasing FET dimensions, but the reduction is less significant than observed when channels are lengthened to reduce the impact of short-channel effects on V_{th} variation. One possibly useful approach is to provide feedback, at the circuit level, to FET back-gates or wells in order to match thresholds [3]; however, this adds circuit overhead and is impractical for any circuit that depends on the relative strengths of FETs (i.e., "ratioed circuits"). Nonetheless, back-gate or well feedback may enable lower V_{dd} in a large design if used only with highly sensitive circuits.

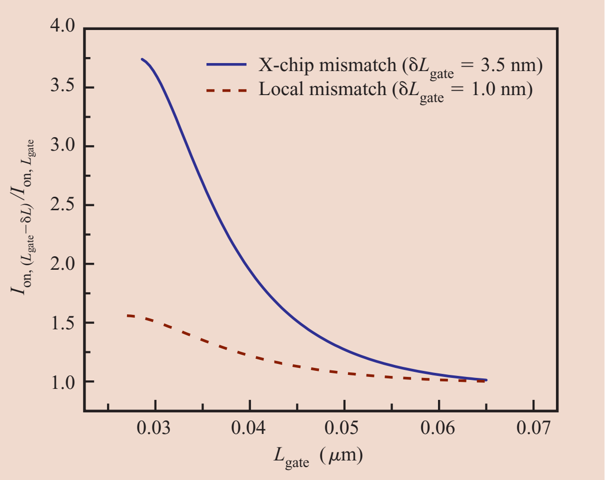

Figure 3

Gate-length dependence of sub-V_{th} I_{on} variation (3σ) due to line-width variation. Both full-chip and local within-circuit variations are considered (SOI 65-nm technology).

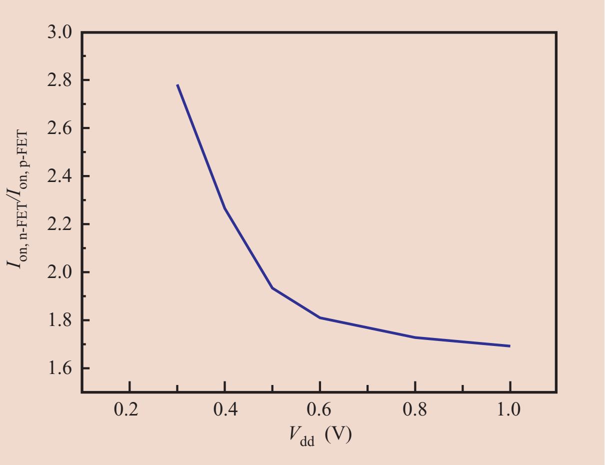

Figure 4

n-FET/p-FET current mismatch as a function of supply voltage.

The important lesson is that V_{th} and sub-V_{th} slope variations are the key challenges to sub-V_{th} and near-V_{th} circuit designs. If a circuit must operate at both high V_{dd} and very low V_{dd} (with a strong emphasis on high-voltage operation), minor tradeoffs should be considered, such as lengthening FET channels slightly to reduce V_{th} variation and making small V_{th} adjustments to maintain nominal matching between FETs at low V_{dd}. However, if energy minimization at low V_{dd} is of paramount importance, significant device optimization studies must be considered. Significant increases in channel length may be advantageous in order to achieve steep sub-V_{th} slopes and to minimize V_{th} and sub-V_{th} slope variations, keeping in mind that channel capacitance plays a smaller role in sub-V_{th} operation than in super-V_{th} operation. The use of mid-gap and quarter-gap metal gates in conjunction with the longer channels should be reconsidered [4]. At short channels, mid-gap metal gates can result in poor sub-V_{th} slopes, but steep slopes can be achieved at longer channels even with low body-doping concentrations. In this case, the V_{th} is controlled primarily by the metal workfunction and not body doping; hence, V_{th} variation due to random doping fluctuations is reduced. Furthermore, mobility increases as the doping concentration is reduced. In the future, advanced dual-gate FINFET [4] and back-gated FET [4] structures will become available. Each of these has different benefits for sub-V_{th} and near-V_{th} operation. Recent work has shown that dual-gated FETs are ideal sub-V_{th} devices because they offer the steepest sub-V_{th} slope [5]. If combined with mid-gap metal gates at long channels, dual-gated FETs may offer sufficient reduction in V_{th} variations. Back-gated FETs trade off sub-V_{th} slope for the possibility of further reduction in V_{th} variations via back-gate feedback. The ultimate choice in device type will depend on how well V_{th} can be controlled.

FET-channel resistances are very large for sub-V_{th} and near-V_{th} operation, providing optimization possibilities in applications in which minimization of energy is the key goal. Because the channel resistances are high, FET series resistance (R_{s}) and interconnect resistances can be larger without having an impact on performance. For example, the increase in R_{s} required to reduce super-V_{th} I_{on} by 10% is approximately 100 Ω-μm, while a 10% sub-V_{th} reduction requires an increase in R_{s} of approximately 6 kΩ-μm. When considering challenges at the FET level, the high R_{s} values that can occur with the simplest dual-gate process options may not be a concern. Decreasing the gate overlap of source and drain diffusions in order to reduce Miller capacitance at the cost of larger R_{s} should be considered. Decreasing the dimensions of interconnect is attractive because capacitances can be reduced at the cost of increased resistance. For a net loaded by wire capacitance, this latter tradeoff could be quite significant. A simple sizing suggests that a 1.25x reduction in active power results when a 4x increase in interconnect resistance is accompanied by a 2x reduction in interconnect capacitance. Of course, all of these options compromise performance at high V_{dd}.

Reliability and susceptibility to wear-out mechanisms in low-V_{dd} and high-V_{dd} operation also differ. Hot-carrier degradation is greatly reduced at low V_{dd}. However, if circuits must operate at both high and low V_{dd}, degradation during high-V_{dd} operation can have a large impact on low-V_{dd} operation. Negative bias temperature instability (NBTI) and channel hot carrier (CHC) effects that cause V_{th} shifts will be the major concern. Even if circuits operate only at low voltage, where the NBTI effect is greatly reduced, NBTI is made worse by long standby periods. SRAMs, for example, can have very long standby periods and may be susceptible to NBTI even at low voltage. Thus, it is important to evaluate the impact of these reliability effects at low V_{dd}. On the other hand, electromigration, which increases interconnect resistance, is less of a concern. Susceptibility to radiation needs consideration, and careful analysis of soft-error rates in sub-V_{th} logic must be conducted.

In summary, the major challenge in sub-V_{th} FET design is V_{th} control. In circuits with high V_{dd} performance requirements, designers must make small compromises to reduce V_{th} variation and maintain current matching between FETs. However, if energy minimization at low V_{dd} is the critical goal, significant device optimization studies must be considered. For minimum-energy CMOS, dual-gate FETs show much promise. If combined with low-workfunction metal gates at long channel lengths, dual-gate FETs will help minimize V_{th} variation and improve sub-V_{th} slope, the key parameters for sub-V_{th} operation.

3. Circuit characteristics at ultralow-voltage operation

The previous section described significant changes in device-level behavior as supply voltage is lowered toward the sub-V_{th} regime. In particular, the $I_{\text{on}}/I_{\text{off}}$ ratio is reduced significantly in the sub-V_{th} region, and devices show an increased sensitivity to the threshold voltage, supply voltage, and sub-V_{th} slope. At the circuit level, this leads to changes in three areas of concern: noise margins, energy optimality, and sensitivity to process variations. Each of these topics is discussed thoroughly for general CMOS logic, and special consideration is given to the design of SRAM arrays.

CMOS characteristics at the voltage-scaling limit

Most useful designs have maintained a “safe” difference between the values of supply voltage and threshold voltage to guarantee robustness and performance. However, as designers have known for many years,

Table 5 Lowest functional supply voltage for several common CMOS gates in a 130-nm technology. $V_{\mathrm{dd,limit}}$ increases with the number of inputs because of the imbalance between pull-up and pull-down networks.

Gate	$V_{dd,limit}$ (mV)
INV	52
Two-input NAND	72
Three-input NAND	87
Four-input NAND	97
Two-input NOR	65
Three-input NOR	74
Four-input NOR	80

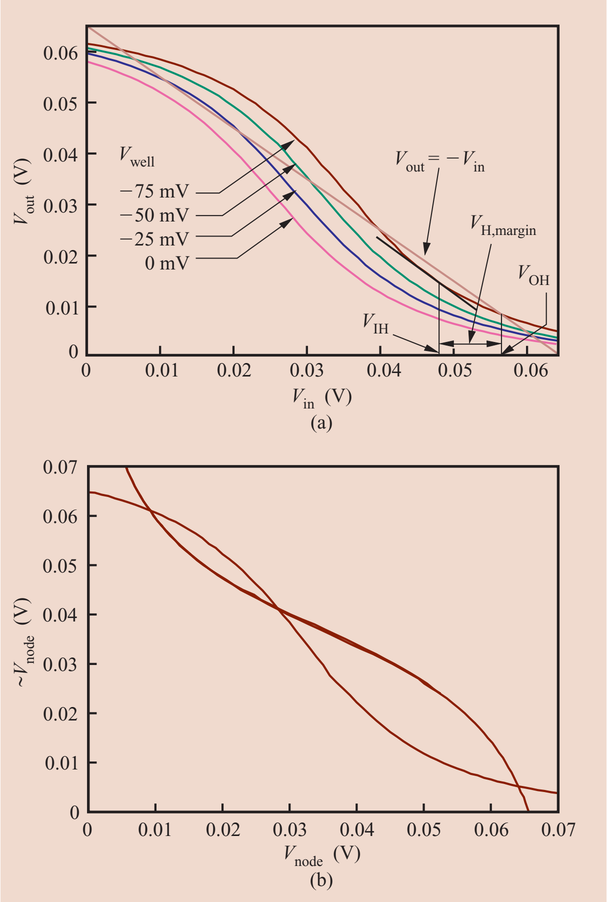

Figure 5

(a) Inverter VTC at $V_{\mathrm{dd}} = 65$ mV and various well biases (V_{well}). (b) A standard "butterfly" curve demonstrates 70-mV sub-V_{th} SRAM cell bi-stability. Wordline and bitline voltages are set to V_{dd}. Voltage is forced via bitlines on one node and measured on the other. Well-biasing is used for optimum n–p matching. V_{node} and $\sim V_{\mathrm{node}}$ correspond to the voltages of the SRAM internal data nodes.

CMOS is a very robust logic family, and in the absence of variability it is unnecessary to maintain a margin between the supply and threshold voltages to guarantee functionality. The supply voltage of a design can therefore be dramatically lowered to limit the dynamic energy consumed. Once the supply voltage drops below the threshold voltage, current takes the form of weak inversion sub-V_{th} current, which is modeled by Equation (1). Using sub-V_{th} current to charge and discharge nodal capacitances, a circuit may function at very low voltages. The theoretical lower limit on voltage scaling was first established in [6, 7] as

$$V_{\mathrm{dd,limit}} = 2 \cdot \frac{kT}{q} \cdot \left(1 + \frac{C_{\mathrm{fs}}}{C_{\mathrm{ox}} + C_{\mathrm{d}}}\right) \cdot \ln\left(2 + \frac{C_{\mathrm{d}}}{C_{\mathrm{ox}}}\right)$$

$$\cong 2 \cdot v_{\mathrm{T}} \cdot \ln(2)$$

$$\cong 36\,\mathrm{mV}, \qquad (3)$$

where C_{fs} is the fast surface state capacitance per area, C_{ox} is the gate-oxide capacitance per unit area, C_{d} is the depletion capacitance per unit area, and v_{T} is the thermal voltage kT/q. The second form of Equation (3) is an approximation of the first assuming an ideal MOSFET (a sub-V_{th} swing of 60 mV/dec at 300 K) with $C_{\mathrm{fs}} \ll C_{\mathrm{ox}}$ and $C_{\mathrm{d}} \ll C_{\mathrm{ox}}$ [6]. An ideal MOSFET can therefore theoretically operate at voltages as low as 36 mV. Sub-V_{th} swing is generally much higher than 60 mV/dec, so an inverter based on realistic MOSFETs will cease to function at a voltage higher than 36 mV. Furthermore, the result in Equation (3) depends on matching between p-FET and n-FET currents. Proper balancing of pull-up and pull-down networks becomes very difficult when gates have transistor stacks (i.e., series-connected transistors). The relative strengths of the pull-up and pull-down networks are dependent on the values of the inputs to the gate, so the use of stacks raises $V_{\mathrm{dd,limit}}$ well above that of a simple inverter. **Table 5** shows simulated $V_{\mathrm{dd,limit}}$ values for several common CMOS gates. $V_{\mathrm{dd,limit}}$ is the lowest voltage for which the voltage transfer characteristic (VTC) has a gain greater than magnitude one when the input voltage V_{in} equals the output voltage V_{out} (i.e., $|\mathrm{gain}| \geq 1$ when $V_{\mathrm{in}} = V_{\mathrm{out}}$) [6]. **Figure 5(a)** shows the measured voltage transfer characteristic (VTC) of an inverter at 65 mV with various well biases and confirms that sub-V_{th} logic functions at room temperature below the previously reported low value of 70 mV [8]. In **Figure 5(b)**, a butterfly curve for a 6T-SRAM cell is shown at a supply voltage of 70 mV, proving that sequential elements also maintain functionality well into the sub-V_{th} regime. The robustness and energy efficiency of SRAM receives special attention

Table 6 Delays for wire-dominated and logic-dominated representative microprocessor paths with voltage scaling. A delay value indicates that output switching occurred; "fail" indicates that the circuit did not work at the specified voltage. (R and F: rising and falling transitions.)

V_{dd} (V)	*1.0*	*0.4*	*0.3*	*0.2*	*0.15*	*0.14*	*0.12*	*0.1*	*0.06*
Wire (R)	1	6	27	204	319	503	632	2,034	5,130
Wire (F)	1	6	26	192	300	469	587	1,750	4,433
Logic (R)	1	11	49	333	876	1,032	1,258	fail	fail
Logic (F)	1	8	37	289	925	1,189	2,110	fail	fail

Table 7 Delays for a logic-dominated microprocessor path with voltage scaling. 100 mV of n-FET/p-FET mismatch is introduced. The rows labeled "n-FET increase" indicate that the n-FET threshold has been increased by 50 mV. (R and F: rising and falling transitions.)

V_{dd} (V)	*1.0*	*0.4*	*0.3*	*0.2*	*0.18*	*0.16*	*0.15*	*0.14*	*0.12*
n-FET increase (R)	1	11	58	381	fail	fail	fail	fail	fail
n-FET increase (F)	1	10	54	517	1,107	fail	fail	fail	fail
p-FET increase (R)	1	13	63	434	646	961	1,199	1,721	fail
p-FET increase (F)	1	9	42	307	460	656	734	fail	fail

in the subsection below on sub-$V_{\rm th}$ SRAM design issues.

Voltage scaling is further limited when complex gates are placed in series. Circuit simulation has been conducted to examine the voltage-scaling limits of typical logic using circuit models of an IBM 65-nm PD-SOI process. For this study, we had to redefine $V_{\rm dd,limit}$ because we were not considering static isolated gates. Here, we define $V_{\rm dd,limit}$ as the lowest voltage at which an input signal switching event results in a correct output switching event (the switching threshold is defined as $0.5V_{\rm dd}$). All delay values have been normalized with respect to the delays of the respective gates at $V_{\rm dd} = 1.0$ V. **Table 6** illustrates $V_{\rm dd,limit}$ for a wire-load-dominated, representative microprocessor critical path. Such critical paths contain long metal interconnects and are interspaced with optimally sized inverting repeaters. Because of the absence of complex static gates (with stacked devices) in this type of critical path, $V_{\rm dd,limit}$ is very low. As Table 6 shows, this circuit functions properly at supply voltages as low as 60 mV. Table 6 also shows the $V_{\rm dd,limit}$ for a logic-dominated, representative microprocessor critical path. This path includes a wide variety of gates including NAND3s, AND-OR-INVERT (AOI) gates, inverters, and transmission-gate-based multiplexers. Multiple instances of each of the mentioned static gates are present in this circuit with varying device sizes and p/n ratios. It is important to mention that this circuit does not have any NOR gates and includes parasitic wire capacitances. Because of the presence of stacked n-FETs and p-FETs in various static gates, this circuit fails to operate below 120 mV.

Increasing $V_{\rm th}$ mismatch between n-FETs and p-FETs further increases $V_{\rm dd,limit}$, as is illustrated in **Table 7**. Two of the entries in the table are labeled "n-FET increase," and the other two are labeled "p-FET increase," indicating that the n-FET and p-FET thresholds have been increased by 100 mV. When the n-FET threshold is increased, circuit failure occurs below 200 mV, while when the p-FET threshold is increased, the circuit is operational at values as low as 150 mV. The simulated circuit has large n-FET stacks and no p-FET stacks (because of the presence of NAND3s and the absence of NORs), so a higher-p-FET $V_{\rm th}$ is more acceptable than a high-n-FET $V_{\rm th}$. This explains the lower $V_{\rm dd,limit}$ for the case in which the p-FET threshold is increased. Although not illustrated in these tables, $V_{\rm dd,limit}$ is also influenced by the p/n sizing ratios in the various gates of the design.

CMOS is clearly functional at very low voltages. There is a performance price, however, for low-voltage operation. The sensitivity of delay to both supply voltage and threshold voltage in sub-$V_{\rm th}$ operation has been alluded to in both Section 2 and Tables 6 and 7. **Figure 6** shows inverter delay in a 130-nm process as a function of supply voltage. Simulations of the same inverter also show that a threshold shift of only 50 mV results in a 4x change in delay. These general trends become important in discussions of leakage energy and variability later in Section 3.

Noise susceptibility in sub-V_{th} logic

Unwanted signals, such as noise, must be addressed in any system of logic, particularly in ultralow-power CMOS. For convenience, we partition the problem into two parts, the circuit noise margin and the noise level in the system. In the first part of this section, we compare the quantitative behavior of the noise margin in sub-V_{th} logic with that of conventional CMOS. We follow this analysis with a view of noise generation in sub-V_{th} logic, again comparing the behavioral issues to those found for conventional CMOS.

The noise margin is the difference between a valid output logic level and an input level at which the data of a "victim" circuit will be corrupted. (A victim circuit is one that is subject to noise from an external source.) Thus, for a "high" logic level, the high-state margin is given by $V_{\mathrm{H,margin}} = V_{\mathrm{OH}} - V_{\mathrm{IH}}$, where V_{OH} is the least positive guaranteed logic output voltage for a valid high state, and V_{IH} is the least positive input voltage required to disturb the logic state of the receiving circuit [see Figure 5(a)]. $V_{\mathrm{L,margin}}$ is similarly defined, such that a positive value for $V_{\mathrm{L,margin}}$ is sufficient for valid transmission of a "low" logic level. In all following discussions, for convenience we refer to these noise margins in fractional values of the V_{dd}.

The input levels V_{IH} and V_{IL} can be approximated by the unity-gain points of the receiver in question. For this approximation, the relative input levels increase as V_{dd} is decreased in the extreme sub-V_{th} region ($V_{\mathrm{dd}} < 100$ mV) as long as n-FET- and p-FET-drive levels are carefully balanced. Both V_{IH} and V_{IL} necessarily approach $0.5V_{\mathrm{dd}}$ as V_{dd} is decreased toward the limiting low-voltage limit for bi-stability. **Figure 7** shows the increase in $(V_{\mathrm{dd}} - V_{\mathrm{OH}})/V_{\mathrm{dd}}$ and $V_{\mathrm{OL}}/V_{\mathrm{dd}}$ for a simulated 90-nm-generation CMOS sub-V_{th} inverter. The reason for the increase is clear; the $I_{\mathrm{on}}/I_{\mathrm{off}}$ ratio is decreasing exponentially with V_{dd}, roughly as $\exp(V_{\mathrm{dd}}/mv_{\mathrm{T}})$, where m is the "ideality factor," that is, the subthreshold slope factor described in Section 2. Hence $V_{\mathrm{OL}}/V_{\mathrm{dd}}$, for example, is expected to increase similarly, as $\exp(-V_{\mathrm{dd}}/mv_{\mathrm{T}})$. The net effect is that the fractional noise margin in sub-V_{th} logic is fairly constant as V_{dd} is reduced from a few hundred mV to 100 mV, below which the increasing values of $(V_{\mathrm{dd}} - V_{\mathrm{OH}})/V_{\mathrm{dd}}$ and $V_{\mathrm{OL}}/V_{\mathrm{dd}}$ result in a decreasing fractional noise margin. **Figure 8** illustrates how the fractional output and input levels behave from $V_{\mathrm{dd}} = 200$ mV to $V_{\mathrm{dd}} = 45$ mV, where operation becomes unstable.

Noise generation can be approached from a simplified model, shown in **Figure 9**. Noise is generated by an "offending" path driven by R_{driver} (R_{driver} may be thought of as the equivalent impedance of a CMOS output stage) with a load consisting of a path directly to ac ground, and a second path with "bad" coupling capacitance to the input of a victim circuit, which in turn has some "good" input capacitance to ac ground. We refer to coupling capacitance as "bad" because it allows noise from one wire to affect another wire. In contrast, we refer to grounded capacitance as "good" because it helps a wire to resist noise. In addition, the gate of the "victim" is driven by R_{good}, which, like R_{driver}, is the equivalent impedance of a CMOS output stage. Typically, the "bad" coupling capacitance arises from adjacent wires that are parallel to each other. To simplify analysis, we consider two limiting cases: 1) R_{good} is much greater than the impedance of

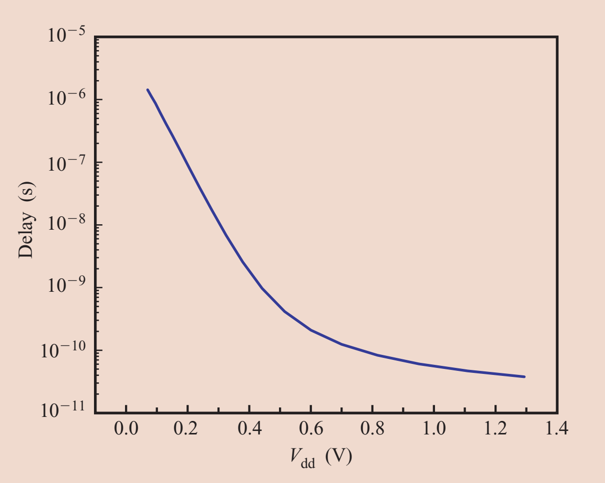

Figure 6

Inverter delay in seconds as a function of supply voltage (130-nm technology).

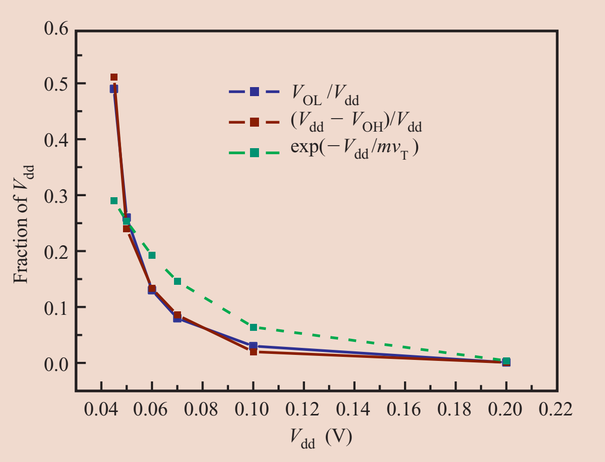

Figure 7

Simulated V_{OH} and V_{OL} for an inverter. Both output levels increase roughly as $\exp(-V_{\mathrm{dd}}/mv_{\mathrm{T}})$ (90-nm technology).

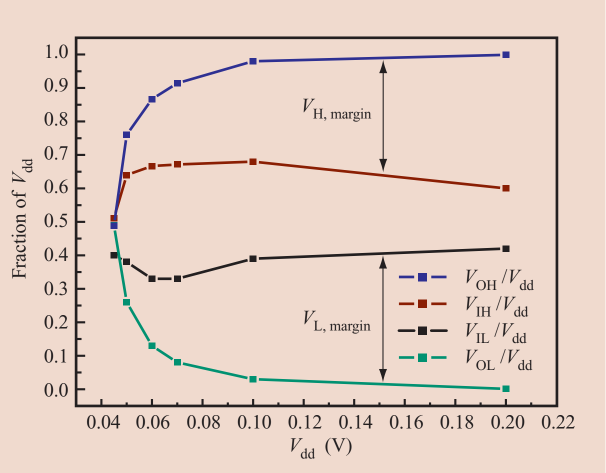

Figure 8

Relative noise margins for an inverter for a range of supply voltages. Above 100 mV, relative noise margins stay fairly constant. Below 100 mV, relative noise margins degrade significantly (90-nm technology).

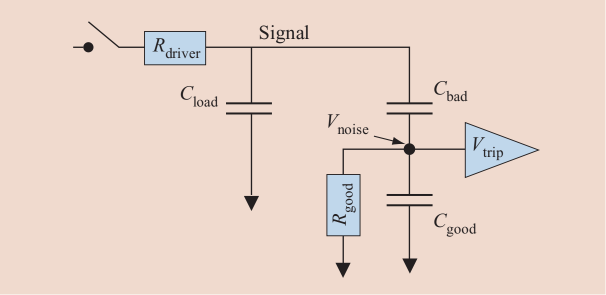

Figure 9

Noise diagram showing both aggressor line (labeled “signal”) and “victim.” The victim circuit may be thought of as the combination of the circuit elements corresponding to R_{good}, C_{good}, and V_{trip}.

C_{good}, and 2) R_{good} is much less than the impedance of C_{good}.

In the first case, noise generation is effectively given by the ratio $2C_{bad}/(C_{good} + C_{bad})$, where the factor of 2 results from a worst-case scenario in which the offending wire and the victim wire switch in opposite directions. This factor of increase is due to the so-called Miller effect. While this ratio is insensitive to the drive characteristics of the transistors, a significant fraction of C_{good} can be due to gate capacitance to (ac) ground; this gate capacitance is shown to decrease in sub-V_{th} operation. Thus, in the case in which C_{gate} may comprise 30% of the total wire load to ground, the noise coupling may increase by as much as 15% in sub-V_{th} operation. This provides further impetus to customize the interconnect technology toward using narrow, thin wires, providing lower capacitance at the expense of higher interconnect resistance. In particular, noise “scaling” can be preserved, provided that the interconnect capacitances are reduced in proportion to the effective reduction in gate capacitance.

In the second case, in which R_{good} is much less than the impedance of C_{good}, the noise coupling is given by $(\zeta C_{bad})^{-1}/[R_{good} + (\zeta C_{bad})^{-1}]$. Note that no factor of 2 is needed because R_{good} can be dominant only in a static case. Here we must consider the scaling of ζ, which is given by the inverse of the characteristic rise or fall time of the offending signal. This rise or fall time is in turn driven by R_{driver}, which is increasing at the same rate as R_{good}, and pushes the design into sub-V_{th} operation. Thus, the noise coupling is expected to be unchanged in this limit. Consequently, sub-V_{th} noise coupling may actually be smaller than super-V_{th} noise coupling if sub-V_{th} interconnect is optimized for lower capacitance and higher resistance.

In conclusion for this section, note that noise margins are largely unchanged in sub-V_{th} CMOS (to as low as 100 mV), provided that n-FET/p-FET matching can be adequately constrained. This may be non-trivial, particularly in light of stochastic mechanisms such as random dopant fluctuations (RDFs). Below 100 mV, even with perfect matching, the noise margin intrinsically decays until it collapses at the stability limit. Interconnect optimization of capacitance at the expense of resistance is desirable to compensate for reduced gate capacitance as a noise shunt. Otherwise, some extra allowance for noise coupling may be required.

Finding the energy minimum

Optimal power and energy analysis

Though absolute noise margins and delay both degrade in the sub-V_{th} region, CMOS logic continues to function at very low voltages. To justify these delay and robustness penalties, we now consider the power and energy efficiency when operating in the sub-V_{th} regime. Power consumption has two components: dynamic and leakage. Thus, the expression for power is

$$P = P_{dyn} + P_{leak}\,,$$

$$P = \frac{1}{2}\cdot C_s\cdot V_{dd}^2\cdot \alpha\cdot f + I_{leak}\cdot V_{dd}\,, \tag{4}$$

where C_s is the switched capacitance of a single inverter, α is an activity factor, f is the clock frequency, and I_{leak} is the leakage current. The “activity factor” is the average number of transitions on a node per clock cycle.

Both dynamic and leakage power benefit from supply-voltage reduction, and dynamic power will continue to improve quadratically as voltage is scaled to the lowest value that guarantees functionality. Circuit designers have traditionally given power more attention than energy, but energy is generally a more suitable metric when battery life is the overriding priority. We therefore begin with a study of energy and show that, unlike power optimization, energy optimization relies upon a compromise between dynamic and leakage energies. Just as in the case for power, energy comprises dynamic and leakage components:

$$E = E_{\mathrm{dyn}} + E_{\mathrm{leak}} ,$$

$$E = \frac{1}{2} \cdot C_{\mathrm{s}} \cdot V_{\mathrm{dd}}^2 \cdot \alpha + I_{\mathrm{leak}} \cdot V_{\mathrm{dd}} \cdot t_{\mathrm{p}} . \qquad (5)$$

Note that we can safely ignore short-circuit energy in this analysis; the reason is explained below. Typical short-circuit current analysis assumes that direct-path current approaches zero when the supply voltage drops below $V_{\mathrm{th,n\text{-}FET}} + |V_{\mathrm{th,p\text{-}FET}}|$ because the direct-path current is entirely sub-V_{th} current below this voltage. Because sub-V_{th} logic is driven exclusively by sub-V_{th} current, we redefine short-circuit current as any direct-path current beyond the leakage current present in steady state. A first-order analysis shows that the total short-circuit energy is an approximately quadratic function of V_{dd} and may be combined with dynamic energy. Assuming a triangular short-circuit current distribution and that Q_{sc} is short-circuit charge, I_{sc} is the peak short-circuit current, and t_{sc} is the total time that short-circuit current exists:

$$Q_{\mathrm{sc}} \propto I_{\mathrm{sc}} \cdot t_{\mathrm{sc}} \propto I_{\mathrm{sc}} \cdot \frac{C \cdot V_{\mathrm{dd}}}{I_{\mathrm{on}}} \propto V_{\mathrm{dd}} . \qquad (6)$$

Note that I_{sc} and I_{on} are assumed to scale identically with V_{dd}, so their dependencies cancel. As a result, Q_{sc} is linear with V_{dd}, and short-circuit energy, $E_{\mathrm{sc}} = Q_{\mathrm{sc}} V_{\mathrm{dd}}$, is quadratic with V_{dd}. Simulations show that the quadratic relation fits very well in the super-V_{th} region, but in the sub-V_{th} region, E_{sc} actually decreases faster than predicted by the quadratic model. This change in behavior in the sub-V_{th} region is minimal, though, and can be ignored with only a small penalty. If we assume a quadratic dependence on V_{dd}, E_{sc} may be modeled using a multiplier in front of dynamic energy because dynamic energy is also quadratically dependent on V_{dd}. We therefore ignore short-circuit energy without invalidating our analysis.

A critical difference between energy and power exists as illustrated by Equations (4) and (5), namely that leakage energy (per operation) is dependent on circuit delay, t_{p}. Figure 6 shows that the delay increases rapidly as the supply voltage scales, particularly when the supply approaches the threshold voltage. Even though I_{leak} [Equation (1)] decreases with supply voltage, the increase in delay is so dramatic that leakage energy quickly overtakes dynamic energy. **Figure 10** shows the average power consumption and the energy consumed per operation for a chain of 50 inverters in an industrial 130-nm technology. Here, an "operation" is the work done in a single clock period. In this example, V_{dd} is scaled, and V_{th} is fixed at approximately 400 mV. Power decreases monotonically, while energy shows an inflection point caused by the rapid rise in leakage energy. For the circuit under consideration, the energy-optimal point occurs

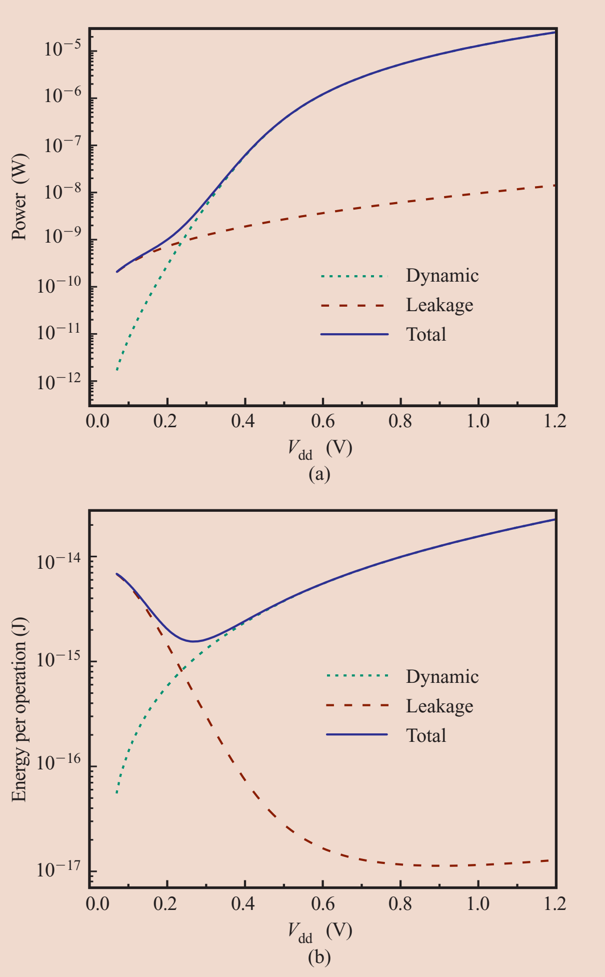

Figure 10

(a) Power consumption for a 50-stage inverter chain decreases monotonically. (b) Energy consumption shows a minimum with respect to V_{dd} (130-nm technology).

in the sub-V_{th} region and yields an energy reduction greater than 20x. This result is confirmed by the authors of [9] and [10], who observe sub-V_{th} voltages to be optimal for an inverter chain and an FIR filter, respectively, with V_{th} fixed. However, the threshold voltage does not have to be fixed when scaling the supply voltage. The authors of [11] study the energy benefits of simultaneous V_{dd} and V_{th} scaling and find that sub-V_{th} operation is generally more energy-efficient than super-V_{th} operation when performance requirements are low. Because we are primarily concerned with minimizing energy, we can accept low performance and can take advantage of sub-V_{th} operation. Given the results of [9–11], we first consider the optimization of circuits in the sub-V_{th} region. Because a number of applications are expected to have higher energy-optimal supply voltages, we also consider the implications of near-threshold and super-V_{th} operation. In both cases, the supply voltage is scaled relative to a fixed threshold voltage.

Inverter chain analysis

Simple analysis of an inverter chain helps build an understanding of the balance between dynamic and leakage energies that occurs at the energy-optimal supply voltage. In [9], an inverter chain with n identical stages and an activity factor of α is considered. The energy per switching event of this system is given by

$$E = \frac{1}{2} \cdot (n \cdot C_{\mathrm{s}}) \cdot V_{\mathrm{dd}}^2 \cdot \alpha + (n \cdot I_{\mathrm{leak}}) \cdot V_{\mathrm{dd}} \cdot (n \cdot t_{\mathrm{p}}) , \tag{7}$$

where n is the number of inverter stages, t_{p} is the delay of a single inverter, I_{leak} is the leakage current of a single inverter, and all other variables are as previously described for Equation (4). It is clear that sub-V_{th} operation is optimal for many circuits [9, 10], so this analysis assumes sub-V_{th} operation. In this analysis, the delay of an inverter with a step input voltage, $t_{\mathrm{p,step}}$, is modeled using

$$t_{\mathrm{p,step}} = \frac{C_{\mathrm{s}} \cdot V_{\mathrm{dd}}}{I_{\mathrm{on}}} . \tag{8}$$

This expression is valid for both super-V_{th} and sub-V_{th} operation, but in the latter case I_{on} is modeled using Equation (1). The authors of [9] show that t_{p} in the sub-V_{th} region can be approximated as

$$t_{\mathrm{p,actual}} = \eta \cdot t_{\mathrm{p,step}} , \tag{9}$$

where η is a technology-dependent parameter that represents delay degradation due to the slope of the input signal. Substituting Equations (8) and (9) into Equation (7) results in the following equation:

$$\begin{aligned} E &= \frac{1}{2} \cdot (n \cdot C_{\mathrm{s}}) \cdot V_{\mathrm{dd}}^2 \cdot \alpha + (n \cdot I_{\mathrm{leak}}) \cdot V_{\mathrm{dd}} \cdot \left(n \cdot \frac{\eta C_{\mathrm{s}} V_{\mathrm{dd}}}{2 I_{\mathrm{on}}}\right) \\ &= \frac{1}{2} \cdot (n \cdot C_{\mathrm{s}}) \cdot V_{\mathrm{dd}}^2 \cdot \left(\alpha + \eta \cdot n \cdot \frac{I_{\mathrm{leak}}}{I_{\mathrm{on}}}\right) . \end{aligned} \tag{10}$$

The variables I_{leak} and I_{on} both take the form of Equation (1) because we are assuming operation in the sub-V_{th} regime. Consequently, all terms in the I_{leak} and I_{on} expressions cancel except for the exponential V_{g} dependence, and Equation (10) may be further simplified to

$$E = \frac{1}{2} \cdot (n \cdot C_{\mathrm{s}}) \cdot V_{\mathrm{dd}}^2 \cdot \left[\alpha + \eta \cdot n \cdot \exp -\left(\frac{V_{\mathrm{dd}}}{m \cdot v_{\mathrm{T}}}\right)\right] . \tag{11}$$

Equation (11) reveals a great deal about the energy dependencies in the sub-V_{th} voltage regime. For example, the strong dependence of energy on supply voltage (V_{dd}) is evident. The quadratic voltage term is initially the dominant term in the expression, but as voltage is reduced far into the sub-V_{th} regime, the exponential dependence on supply voltage (which reflects circuit delay) begins to dominate. Figure 10(b) illustrates the fact that there is a distinct energy minimum with respect to voltage. We can easily find the supply voltage at the minimum by determining the derivative of Equation (11) and setting it equal to zero. The resulting equation is nonlinear and must be solved using numerical methods. The final expression for the supply voltage at the energy minimum, which we denote V_{min}, is shown in the following equation:

$$V_{\mathrm{min}} = \left[1.587 \cdot \ln\left(\eta \cdot \frac{n}{\alpha}\right) - 2.355\right] \cdot m \cdot v_{\mathrm{T}} . \tag{12}$$

V_{min} does not necessarily correspond to $V_{\mathrm{dd,limit}}$, described in the subsection on CMOS characteristics at the voltage-scaling limit. In fact, as the subsequent discussion shows, V_{min} is usually well above $V_{\mathrm{dd,limit}}$ because of the dominance of leakage.

In Equation (12), V_{min} depends only on the number of device stages, the activity factor, and two process-related parameters, η and m. This simple model has great value because switching between technologies requires only the determination of η and m. The accuracy of the model is confirmed in [9].

The importance of logic depth n and activity factor α in Equation (12) is obvious. To understand the relationship between these two parameters, we replace the ratio of n to α with a single parameter n_{eff}. This substitution is valid because logic depth and activity factor affect the energy characteristics of a circuit in very similar ways. A circuit with many stages (large n) will be leaky because the leakage time for each stage is increased. Similarly, a circuit with a low activity factor is more likely to be leakage-dominated because dynamic energy is proportional to the activity factor. In both cases, the circuit will exhibit a higher V_{min}.

To properly understand the effect of n_{eff} and V_{min} on a circuit, it is important to understand the notion of *transistor utility* [12], which embraces the idea that all transistors in an energy-efficient design should spend as much time as possible doing useful computation because the circuit consumes wasted energy during idle time. We can nominally assume that dynamic energy is a measure of useful computation (because switching transistors consume dynamic energy) and that leakage energy is the penalty paid for idle time. The goal of an energy-efficient design is therefore to optimize the circuit structure such that the ratio of dynamic energy to leakage energy is maximized. Maximizing the dynamic-to-leakage ratio enables a designer to effectively use supply voltage as a "lever" to decrease dynamic energy consumption and consequently total energy consumption. In other words, a design with high transistor utility generally has a lower V_{min} than a similar design with low transistor utility. In the ideal case, V_{min} approaches $V_{dd,limit}$, the lowest functional voltage.

If we adopt transistor-utility maximization as the goal of a design, it is obvious that a large n_{eff} is undesirable. Long paths or paths with low activity increase effective idle time and increase V_{min}. Logic depth and activity factor tend to be a function of high-level architectural decisions, so they may not be characteristics that a circuit designer can easily exploit. However, the circuit designer does have the ability to decrease the penalty for idle time. Leakage-reduction techniques such as multiple-threshold CMOS (MTCMOS) [13], input vector control [14], and threshold control via adaptive body biasing (ABB) [15] are all tools that have the potential to lower V_{min} and consequently the total energy consumed by the circuit. The leakage problem may also be addressed at the device level by improving the sub-V_{th} slope. Section 2 pointed to the importance of sub-V_{th} slope in low-voltage design, and our simple analysis clearly shows that minimizing sub-V_{th} slope should be a key goal of energy-optimal design.

We now consider whether the previous analysis is valid for a more complex design. Silicon measurements of a simple 8-bit microprocessor with 2-Kb memory in a 130-nm technology are shown in **Figure 11**. A more detailed description of the architecture may be found in [12]. The form of the curve is identical to that of the inverter chain (thus confirming the validity of the inverter chain analysis), but V_{min} is higher than that of the inverter chain. The memory, which has a very low activity factor compared with typical logic, is largely responsible for the higher V_{min}. This example suggests that very low circuit activity could push V_{min} into the super-V_{th} regime. The next section investigates this topic.

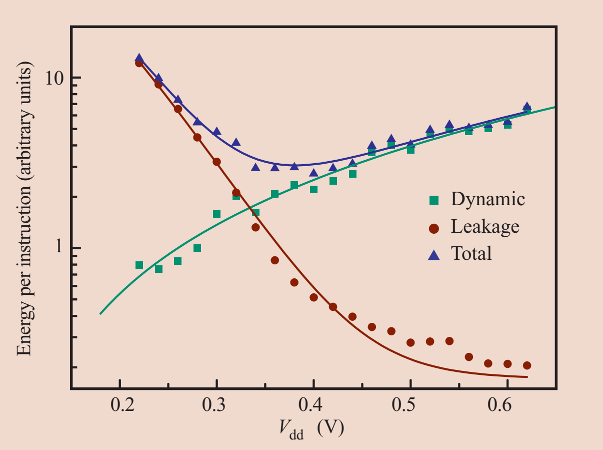

Figure 11

Measured energy per operation for an 8-bit microprocessor with a 2-Kb memory (130-nm technology).

Energy optimality in the near-V_{th} and super-V_{th} regions

As Figure 11 shows, circuit structures with very low activity factors, such as memories, face serious leakage penalties for extremely low-voltage operation. As the leakage energy of a design increases relative to dynamic energy, the optimal supply voltage tends toward higher values. If the threshold voltage is held constant, the optimal supply voltage will likely be near or above the threshold voltage.

We can extend the inverter chain analysis example from the previous section to make some simple but powerful conclusions about operation in the near-V_{th} and super-V_{th} regions. Consider an inverter chain with variable activity factor. As the switching activity decreases, the leakage energy of the circuit is unchanged, but the dynamic energy shifts lower. This downward shifting of the dynamic energy curve is illustrated in **Figure 12** for a range of α values. As the dynamic-energy curve moves relative to the leakage curve, the location of the minimum-energy voltage, V_{min}, shifts. When V_{min} approaches the threshold voltage, the leakage-energy curve flattens because delay (and consequently leakage) is less sensitive to supply-voltage changes above the threshold voltage. As a consequence, the energy minimum is flattened, so that the choice of supply voltage can deviate slightly from the optimum with only a small energy penalty.

The core problem, however, remains unchanged. The dynamic- and leakage-energy curves still cross over one another and create an energy minimum. This is a key conclusion that is independent of the region of operation: The location of the energy minimum is entirely determined by the way in which the leakage-energy and

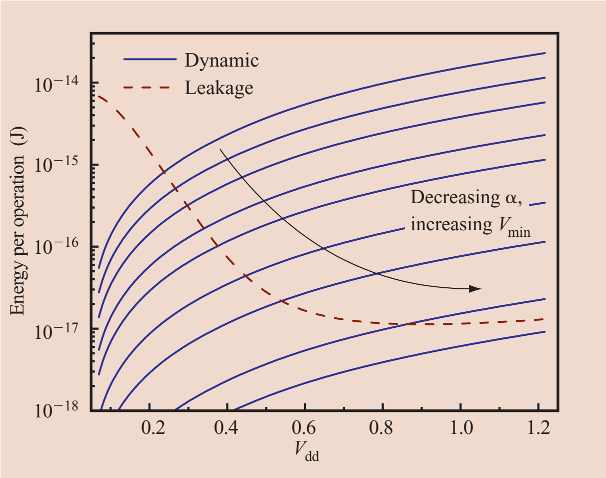

Figure 12

Dynamic and leakage energy are shown for an inverter chain with varying activity. The dynamic-energy curve shifts as the activity factor decreases. V_{min} is entirely determined by the way in which the dynamic and leakage curves interact (130-nm technology).

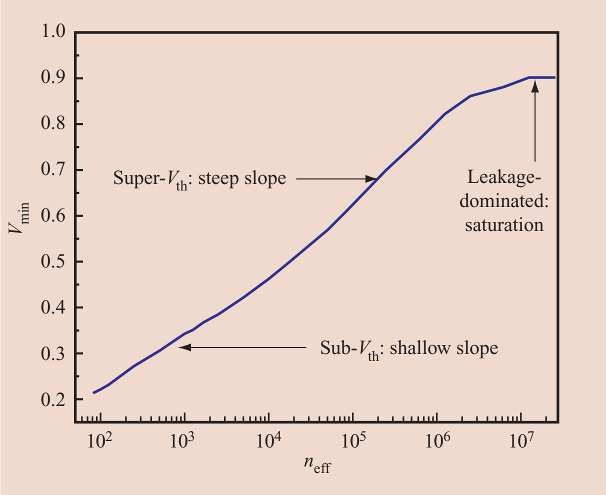

Figure 13

V_{min} is shown for a wide range of n_{eff} values. V_{min} continues to depend logarithmically on n_{eff} in the super-V_{th} region (130-nm technology).

dynamic-energy curves interact. Regardless of whether a system operates in the sub-V_{th} or super-V_{th} regime, all architectural and circuit techniques such as ABB attempt to shift either the leakage-energy or dynamic-energy curve to improve energy efficiency. **Figure 13** shows that V_{min} behaves similarly over a wide range of n_{eff} values. Although Equation (12) was derived for the sub-V_{th} regime, V_{min} continues to exhibit an approximately logarithmic dependence on n_{eff} until V_{min} reaches approximately 600 mV. Above 600 mV, V_{min} is still roughly logarithmically dependent on n_{eff}, but the slope of the line becomes steeper. The location of V_{min} is dependent upon leakage, so the relative insensitivity of leakage to supply voltage in the super-V_{th} region forces larger changes in V_{min} in order to reach energy optimality (and consequently the slope is greater). Also note that saturation of V_{min} occurs at very high n_{eff} values because dynamic energy becomes insignificant compared with leakage energy regardless of V_{dd}.

It is clear that the understanding of α, n_{eff}, and transistor utility (described earlier) developed for sub-V_{th} operation is valuable even when V_{min} shifts into the near-V_{th} and super-V_{th} regions. While the fundamental goal of maximizing transistor utility remains unchanged, circuit design in the near-V_{th} and super-V_{th} regions is clearly different from design in the sub-V_{th} region. Several of the key differences between sub-V_{th} and super-V_{th} operation, including delay characteristics and sensitivity to threshold voltage and device mismatch, have already been discussed extensively. In practical applications, circuit design becomes much easier when the supply voltage rises above the threshold voltage. Circuit speeds increase and variability decreases (and noise margins increase in response) as the supply voltage is raised. The real challenge arises when a circuit is optimized so that the energy-optimal supply voltage lies within the uncertain realm of sub-V_{th} design. The next section shows that one of the most significant challenges is "process-related variability," a phrase that is clarified in the next section.

Variability in the sub-V_{th} regime

Process-related variations, that is, those variations introduced in manufacturing, have become a significant factor that affects circuit performance, even in the super-V_{th} regime. As a result, there has been a movement to develop methodologies and tools for dealing with variation in order to eliminate the pessimism usually associated with "corner-based" design schemes—those that assign "worst-case" parameter values to determine resistance to variability in order to ensure that the circuit functions properly under a range of conditions. The consequences of process variation are far more severe when voltage is scaled into the sub-V_{th} regime because of the exponential dependencies of current, and therefore delay, observed in this regime. In addition to the aforementioned relationship between sub-V_{th} current and threshold voltage, temperature also plays a critical role in determining delay in the sub-V_{th} regime. The strong dependence on threshold voltage, in particular, leads to wild fluctuations in both delay and energy as well as a considerable reduction in noise margins.

Threshold voltage variations can be broadly placed in two categories: *systematic variations* (which include lot-to-lot, wafer-to-wafer, die-to-die, and intra-die spatially correlated variations) and *random variations*. Systematic variations arise from a variety of sources, including gate-length variations, global doping variations, and temperature variations. A number of techniques have been shown to reduce the effects of systematic variation. In [16] and [17], ABB was shown to reduce both frequency and leakage power variations in test circuits. In [18], ABB was used to reduce variation in sub-V_{th} logic. Dynamic voltage scaling (DVS) is another technique that may be used to limit variation. DVS is traditionally discussed in the context of dynamic power management, but it may also be used to improve both frequency and power yields [17] by enabling post-silicon tuning of the supply voltage. It is likely that adaptive systems that incorporate several circuit techniques such as ABB and DVS will be necessary to achieve high yields in sub-V_{th} designs.

In addition to systematic variations, random variations (specifically RDF) account for a significant portion of threshold variation [18]. Random variations present a greater threat than systematic variations to delay, power yield, and energy efficiency because global countermeasures such as ABB and DVS cannot be used to effectively address the problem. Researchers have shown that RDF grows in importance as supply voltage is scaled into the sub-V_{th} regime [19]. Furthermore, as voltage is reduced, total variation grows significantly and becomes dominated by RDF. RDF depends strongly on channel area [2] and may therefore be controlled with careful gate sizing. The effects of RDF may also be reduced by increasing the number of gates in a path because random fluctuations "average out" over long paths. Proper selection of gate sizes and architecture (which determines logic depth) is very important in the design of robust sub-V_{th} circuits.

In the subsection on finding the energy minimum, we showed that energy minimization relies on the proper balance of dynamic energy and leakage energy. The strong threshold dependence of leakage makes energy optimization strongly dependent on variability. Researchers have developed statistical models of circuit delay, power, and energy and have shown that variability raises V_{min} by as much as 78 mV and is therefore a threat to energy efficiency [19]. The net voltage and energy shifts are in the positive direction because delay has a lognormal distribution if we assume a normal threshold-voltage distribution. A distribution of delays across many designs is therefore skewed toward longer delays. This effective increase in delay raises the relative importance of leakage, which increases the V_{min} and the total energy of a design. Larger gates and longer logic paths provide the simplest and most powerful solutions to RDF, but these are, in general, contradictory to the goals of energy minimization. Upsizing increases dynamic energy, while larger logic depths lead to lower transistor utility. Designers must therefore carefully strike a compromise between the minimization of variability and the minimization of worst-case energy.

Although energy efficiency is a serious question in the face of variability, robustness is a more pressing concern. Even a small mismatch between p-FET and n-FET threshold voltages (introduced by either systematic or random variations) causes skewed current ratios and a dramatic reduction in noise margins. Table 7 shows how a systematic threshold mismatch can lead to a considerable increase in $V_{\text{dd,limit}}$. We must also consider how noise margins are affected when the supply voltage is greater than $V_{\text{dd,limit}}$. Simulations of an inverter (130-nm-technology node) with $V_{\text{dd}} = 200$ mV show that noise margins are reduced by 19%, 38%, and 79% given p-FET/n-FET threshold mismatches of 25 mV, 50 mV, and 100 mV, respectively (where a mismatch of 2δ means $|V_{\text{th,p-FET}}| = |V_{\text{th,p-FET,nominal}}| - \delta$ and $V_{\text{th,n-FET}} = V_{\text{th,n-FET,nominal}} + \delta$). These reductions are clearly not tolerable, and hence significant effort to control threshold matching is necessary. Static noise margins are of particular importance in SRAM, implying that designers of sub-V_{th} memories must pay special attention to mismatch problems. SRAM design issues are covered in detail in the next section.

Process-related variability is one of the critical barriers that must be overcome for sub-V_{th} logic to have widespread industrial use. A combination of global techniques including ABB and DVS should be employed in combination with careful selection of design parameters, including logic depth and transistor sizing. Well-designed logic, which accounts for variability from the *beginning* of the design cycle, fosters high circuit yields while also minimizing energy.

Sub-V_{th} SRAM design issues

The complications of sub-V_{th} design have been covered extensively for typical logic. We now make special considerations for SRAM. As a result of reduced activity factors, energy consumption due to leakage is especially important in SRAM caches. Depending on the size of the memory, only a small portion may be active at any given time. As shown in Figure 12, such a condition inevitably increases V_{min} because reduced dynamic energy shifts the overall energy minimum. This suggests that for ultralow-power designs, SRAM arrays should be operated at a higher supply voltage than that for logic. **Figure 14(a)** demonstrates this basic concept for a 65-nm process, where the supply voltage that minimizes energy is above 0.4 V because of a low activity factor. With progressively higher levels in the memory hierarchy that are larger or

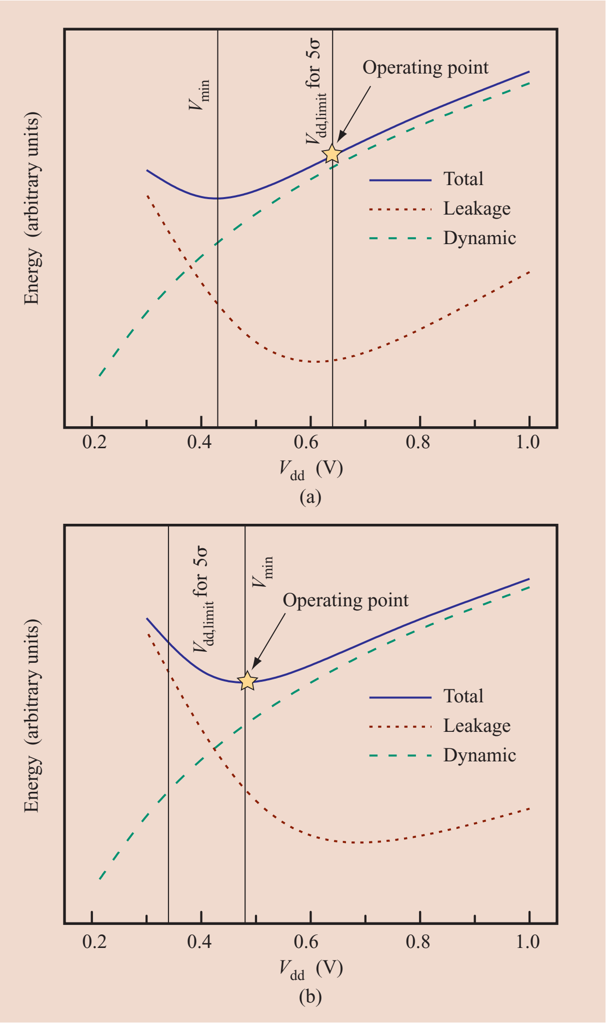

Figure 14

(a) V_{min} for an SRAM is generally higher than that for logic because of a reduced activity factor. However, minimum operating voltage is likely to be limited by variability. For 5σ functionality (such as that needed to yield an ~1-Mb array), $V_{dd,limit}$ may be significantly larger than V_{min}. As a result, optimum energy operation cannot be achieved (65-nm technology). (b) Because of improved robustness to variability, 8T-SRAM can function at lower supply voltages, which allows $V_{dd,limit}$ to be lower than V_{min}. As a result, optimum energy operation can be achieved (65-nm technology).

even lower in the activity factor, V_{min} will increase further.

In an SRAM cell, a proper ratio of device strengths must be maintained among the pass-gate, pull-down, and pull-up transistors to ensure functionality under write, read, and standby conditions. When the supply voltage is reduced, acceptable ratios must be maintained: The drive strength of the pass-gate must be greater than that of the pull-up transistor to allow writing of the cell, and the pull-down must be stronger than the pass-gate to avoid accidental flipping of the cell during a read event. In the standby mode, functionality constraints are the same for logic: The two inverter transfer characteristics must be maintained. In general, the read and write requirements provide more severe constraints on cell functionality at low supply voltages. With proper setting of device threshold voltages and widths, however, cell functionality can be maintained even at extremely low voltages. Measurements of a 6T-SRAM cell shown in Figure 5(b) confirm that bi-stable operation is possible with V_{dd} as low as 70 mV. In such a case, $V_{dd,limit}$ can be lower than V_{min}, and optimum energy can be achieved.

The rapid increase in random process-related variability in recent technology generations, especially variability due to random dopant fluctuations, can have a severe impact on $V_{dd,limit}$ for complete SRAM arrays. Such variability is particularly important in SRAM because of the widespread use of minimum-size devices and aggressive technology ground rules. While a single cell with perfectly matched devices can function at very low voltages, from a statistical standpoint, cells with significant threshold-voltage mismatch will exist in a large array. Such threshold-voltage variation can effectively degrade noise margins such that cell functionality is compromised [20, 21]. When variability is considered, $V_{dd,limit}$ rises dramatically for a complete SRAM array. As an example, Figure 14(a) shows that the variability characteristic of a 5σ cell (as needed to yield an ~1-Mb cache under random variation) increases $V_{dd,limit}$ to well beyond V_{min} for a 65-nm technology. In addition, this increase is expected to become more serious as technology scales [22]. As such, optimal energy consumption cannot be achieved because an SRAM array cannot function at the optimal voltage for energy consumption under the presence of variability.

Error-correction codes and redundancy are often used to address variability in today's SRAM arrays, but these techniques will most likely be insufficient to completely address the widespread problems expected at low voltages. New circuit techniques must also be used to counter variability. By modifying the traditional 6T-SRAM cell circuit, a more variation-tolerant design can be attained. With the 8T cell, as depicted in **Figure 15(b)**, significantly improved read noise margins can be realized [23]. This improvement occurs because the read-disturb condition [as depicted in **Figure 15(a)**, in which the stored "0" node of the cell is pulled above ground] is eliminated by the introduction of a separate read port in the SRAM cell. With discrete read and write ports in the cell, the device ratio constraints of the traditional 6T-SRAM cell for read and write functionality are removed, which allows for simultaneous improvement of both read and write noise margins. As a result, variability tolerance is greatly enhanced, and $V_{dd,limit}$ can be reduced to less than

$V_{\min}$. As shown in **Figure 14(b)**, employment of an 8T-SRAM design can allow for operation at the supply voltage for optimal energy, thus making it a desirable design option for ultralow-power SRAM caches.

4. Architectural choices at ultralow-voltage operation

Power-aware microarchitectures are able to have a substantial impact on power consumption, and can be more valuable than transistor-level remedies in reducing total power. This section covers several techniques that may be used to shift $V_{\min}$ and tune energy efficiency at the architectural level. The role of multi-core processing in recovering some of the speed penalty paid for sub-V_{th} operation is also discussed in this section. In order to understand the value of energy-efficient architectural techniques, we first discuss several metrics that are commonly used to describe energy and power efficiency.

Metrics

Common architecture metrics capture the influence of energy minimization design for chip logic. Energy- and power-driven design affects area and performance as well as power; capturing these influences in selected benchmark conventions becomes important. Below, we describe selected common benchmarks that provide insight into energy use.

- *MIPS per watt:* The number of instructions completed by a processor is often described by the millions of instructions per second ("MIPS") completed at peak load. This benchmark, divided by the power consumption (in watts), describes the power cost of an architecture throughput technique. More effective power- and energy-aware architectures exhibit higher values of MIPS per watt. Supplemental logic circuitry is typically added in high-performance microprocessors to architecturally improve throughput. These performance accelerators add more circuits to execute the same logic and reduce power efficiency. Thus, the total energy cost of an instruction rises as these innovations are added.
- *Energy–delay product (EDP):* For a given benchmark logic path, the product of the path delay and its total ac and dc energy consumption provides a measure of the effectiveness of the architecture. Large relative values of EDP indicate increased delays and/or high energy consumption, neither of which is tolerable in a power-constrained machine. EDP is often computed for common benchmarks such as a ring oscillator comprising the inverter driving a fan-out of four additional inverters ("INVFO4"). The utility value of EDP is realized when evaluating the design "return" as a result of accepting reduced performance. Active and static power per INVOF4 is often quoted in the literature [24].
- *Transactions per CPU cycle (TPCC):* Energy-reduction techniques often have a negative impact on microprocessor throughput; for this reason, the designer must consider changes to the number of transactions per CPU cycle ("TPCC") that can be retired. TPCC is a measure of microprocessor logic efficacy and is purely an architecture performance metric. Nonetheless, TPCC is a superb bellwether that the system designer can use to assess overall system impacts to power management.

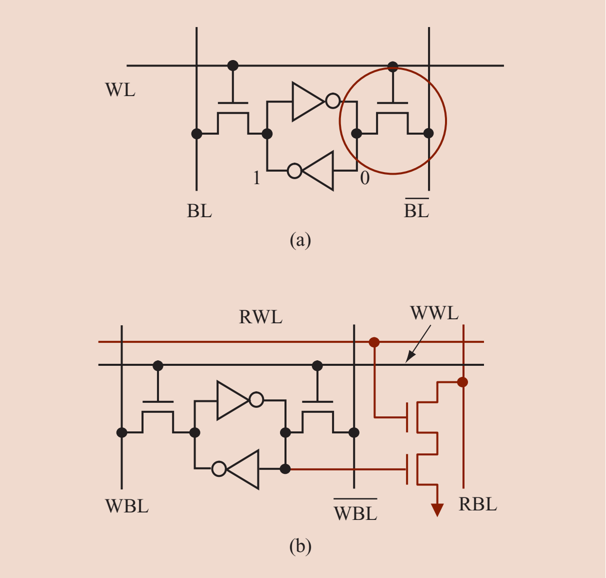

Figure 15

(a) Traditional 6T-SRAM cell, depicting read-disturb condition; (b) 8T-SRAM cell with added read port to eliminate any read disturbs. The 8T cell thus has a much improved static noise margin in the read condition, and is more tolerant of variability.

Each of these metrics offers valuable information to the designer. However, no single metric is best for all applications. Instead, designers must be careful to choose the metric that best describes the particular power, area, and delay requirements of an application. Though different applications may be driven by different metrics, the applications will use very similar energy-reduction techniques. The following two subsections describe some of these architectural techniques.

Semitransparent and structural alterations for power savings

Semitransparent uniprocessor power-management techniques reduce power through consideration of under-utilized resources. Note that we use the term *semitransparent* to refer to a technique that is managed at the hardware level and that does not require support from the operating system. Adaptive body biasing, or ABB, modulates the voltage of the substrate in order to dynamically change the static-power-vs.-active-performance tradeoff asserted by threshold voltage. This technique was mentioned in Section 3 as a tool that may be used to limit systematic threshold variability in a sub-V_{th} system. In ABB, machine state requirements are anticipated via instruction-lookahead techniques, and substrate voltages are adjusted to optimize power without dramatically affecting performance [16]. The use of ABB in the super-V_{th} regime has recently fallen out of favor. In [25], it was shown that the application of a reverse body bias (RBB) becomes less effective with technology scaling. Short-channel effects (SCEs) worsen with shrinking transistor dimensions, so threshold control via ABB becomes less effective. In the sub-V_{th} regime, this problem is much less severe because SCEs, in particular DIBL, are much reduced. It is important to note that ABB is in accordance with the energy model derived in Section 3. By reducing the standby-mode leakage, ABB lowers the penalty paid for idle time, raises transistor utility, and lowers V_{min}.

Alternatively, logic may be partitioned for placement in specific *voltage islands* with independent supplies that are compiler-controlled [26]. In Section 3, it was observed that different circuit blocks tend toward different energy-optimal supply voltages. Voltage islands are attractive because they enable a designer to target the energy-optimal conditions on a block-by-block basis rather than on a system level. Memory, in particular, is expected to have a much higher V_{min} than conventional logic. With memory and logic on different voltage islands, memory would be able to operate at higher voltages to address both functionality and leakage problems, and logic would be allowed to scale voltage more aggressively to yield lower dynamic energy.

Just as different blocks have different V_{min} values, different runtime conditions (such activity factor and temperature) may require the scaling of supply voltages to maintain energy optimality. Fine-grained dynamic voltage scaling (DVS) allows tuning of a design for runtime conditions and therefore holds promise for use in dedicated energy-optimal design. DVS may also be used to achieve combined performance and energy targets [27]. In [28], DVS was used in combination with ABB to effectively minimize power under a performance constraint, and the system was found to be functional at voltage levels as low as 175 mV. Such an architecture has the potential to target energy optimality given both variable runtime conditions and process variations. A DVS system that operates in both sub-V_{th} and super-V_{th} regions requires careful design because, as we saw in Section 2, n-FET/p-FET mismatch is a function of supply voltage.

Voltage gating, or multiple-threshold CMOS (MTCMOS), is yet another power-management technique in which large on-board header and footer MOSFETs provide power to specific domains of the microprocessor chip. When the resource in these domains is no longer needed, their supply access is cut by these devices [29]. Typically, these domain-based power-management architectures require latches surrounding the domain. These latches are powered independently so that they can retain machine state after the supply voltage to that region is reduced. Voltage gating is very similar to ABB because both techniques seek to lower V_{min} by minimizing leakage during idle periods. In super-V_{th} MTCMOS, power is gated by high-threshold transistors during idle periods. The high-threshold transistor limits the performance of a circuit, and, as a result, sizing of the header and footer transistors can be very challenging. The performance impact of using high-threshold sleep transistors is even greater in sub-V_{th} operation. A 100-mV increase in threshold voltage results in a delay increase of more than 10x for a sub-V_{th} inverter (compared to only ~1.3x for a super-V_{th} inverter). A 10x increase in delay is unacceptable for energy-optimal operation, so sub-V_{th} MTCMOS will probably use a single threshold voltage for all transistors. Sub-V_{th} MTCMOS could leverage low-threshold sleep transistors that are supplied with a reverse body bias (RBB) during idle periods [30].

Each of the techniques discussed in the preceding paragraphs accounts for the fact that resources are not fully utilized at all times. In other words, transistor utility is a function of many runtime conditions, and may be lower at certain times. Flexible designs that are able to accommodate changing transistor utility will be very important for robust energy-efficient design. Discrete levels of compute activity have been considered in microprocessors for many years to save power. IBM integrated "Nap," "Doze," and "Sleep" modes into the PowerPC 750* Microprocessor [31]. Each mode supports a subset of the full resource available. Architecturally, sleep modes are more invasive to the design than the semitransparent techniques, and require support from the operating system. More recently, machine architectures have attempted to improve throughput by speculatively executing selected operations based on past history. Power is conserved by disabling this feature as needed.

Multiprocessing

This paper has so far focused on energy-optimal operation; each of the previous sections targeted primarily dedicated low-voltage operation. Because many applications place both power and performance requirements on designs, we now discuss how the design strategies from the previous sections may be used within a framework that optimizes speed and power.

The advent of single-chip multi-core processors ("CMP") offers a specific opportunity for the application of the aggressive power-reduction techniques advocated in this work. Industry engineers know that transactions that are easily given to instruction-level parallelism (ILP) are more quickly retired in CMPs. For ultralow power, however, CMP also holds opportunity. In ILP-friendly transactions, where power reduction rather than throughput is required, the parallelism of the CMP can be exploited to trade speed for power.

This response to voltage can also be leveraged for transactions that are given to parallel solutions. A multiplicity of cores, running at a lower clock rate and at reduced operating supply voltage, has the opportunity to decrease power consumption because of its quadratic voltage and linear frequency dependence. Multiple identical cores operated at low frequencies can replace a single high-power uniprocessor operated at high frequency and can produce a comparable output data rate.

Alternatively, the uniprocessor may also be replaced by a *heterogeneous* multi-core architecture. In this scenario, multiple cores on a single chip all execute the same set of instructions, but with each core emphasizing a different power–performance point [32]. The cores are single-threaded; each runs at its own characteristic frequency. They each have dedicated L1 memory arrays and typically share L2 array space. Under the control of the compiler, software determines which core is most energy-optimal in order to retire a given instruction, and issues an instruction to that unit. Architectural features that improve the transaction retirement rate on big processors are removed in the lower-performance, lower-power cores. These performance enhancers, which include multiple fixed-point and floating-point units, speculative execution engines, instruction lookahead facilities, and out-of-order instruction queuing, all indeed improve performance, but with fairly punitive power costs. Criteria for selecting the most appropriate core include application urgency, energy availability, and thermal limitations. Other cores may be power-gated off. Overall savings of up to 63% in energy–delay product have been quoted in the literature for multi-core architectures. Implicitly, the lowest-power core could be designed to operate at substantially reduced supply voltage.

The notion of operating different cores at different voltages in a CMP is especially powerful when integrating the memory subsystem. SRAM circuits, in both silicon-on-insulator (SOI) and bulk fabrication processes, have known cell stability issues when the feature size and accompanying technology drop below sizes associated with the 130-nm-lithography node. These stability issues are associated with noise, defects, history effect, electrical parameter variability, device tracking, supply-voltage variation, and temperature. For this reason, memory arrays are already operated at elevated supply voltages that are above the nominal logic voltage, and do not easily tolerate dramatically reduced V_{dd} [33]. A multiplicity of low-power, low-voltage cores that share a common L1.5 or L2 cache puts substantial bandwidth and latency demands on that array. Therefore, by operating the shared L1.5 or L2 array resource at elevated voltage, and the CMP cores at reduced supplies, the stability concerns of the memory and the bandwidth required by the core subsystem are simultaneously satisfied. The shared resource cannot collectively be operated at reduced voltage: by sharing the only component that requires elevated voltage, we extend the energy savings realized by the processor.

In a conventional operating space, specific compute-intensive operations are supplanted with hardware accelerators. Commonly used in streaming data applications such as graphics rendering and imaging, hardware accelerators realize specific instructions in specialized hardware, and these instructions are performed repeatedly. The general processor is capable of executing these instructions, but inefficiently and with greater overhead. The presence of the accelerator frees the processor to handle other requests in parallel. We anticipate that hardware-instruction accelerators will become even more valuable on processors operated at low voltage, when execution inefficiency becomes especially expensive with respect to latency.

Conclusion

Aggressive voltage scaling into the sub-V_{th} region holds great promise for applications with strict energy budgets. For many circuits, energy consumption reaches an absolute minimum in the sub-V_{th} regime that is of the order of 20x improvement over super-V_{th} operation. In this paper, we examined the evolution of system design as voltages enter the sub-V_{th} regime. We began by considering device-level changes, and then we discussed the implications of behavioral changes on circuit and architectural design. Sub-V_{th} devices are characterized by an increased sensitivity to changes in threshold voltage, supply voltage, and temperature. Device sensitivity to threshold voltage is particularly important, so sub-V_{th} devices should use longer gate lengths in conjunction with low-workfunction metal gates. Dual-gated and back-gated FETs are also attractive for sub-V_{th} operation. The heightened sensitivity of device current to the threshold voltage, in combination with a low $I_{\mathrm{on}}/I_{\mathrm{off}}$ ratio, leads to serious circuit-level robustness concerns in the face of process-related variability. Threshold variability, in particular, poses a great threat to both functionality and energy efficiency. Variability is particularly threatening to the design of robust SRAM arrays, but alternative structures such as the 8T-SRAM cell offer increased robustness. A simple and powerful model for energy efficiency underlies all of these trends and was discussed extensively in this paper. In this model, the minimization of energy relies on the proper balance between leakage and dynamic energies. Furthermore, circuit- and architectural-design optimizations should drive $V_{\min}$, the energy-optimal supply voltage, toward $V_{\mathrm{dd,limit}}$, the lowest supply voltage that guarantees functionality. Several existing architectural techniques (ABB, MTCMOS) may be effectively applied in the sub-V_{th} regime for energy efficiency. Architectural techniques that exploit parallelism have the potential to recover much of the performance penalty paid as a result of voltage scaling. Sub-V_{th} logic will likely play a key role in many future energy-efficient designs, but designers must first dedicate all of their efforts to developing variability-resistant designs. The future success of sub-V_{th} design will depend on the combined effort of device, circuit, and architecture designers to develop robust, variability-aware, low-leakage technologies.

*Trademark, service mark, or registered trademark of International Business Machines Corporation.

References

1. B. Paul, A. Raychowdhury, and K. Roy, "Device Optimization for Ultra-Low Power Digital Sub-Threshold Operation," *Proceedings of the International Symposium on Low Power Electronics and Design (ISLPED)*, Newport, RI, August 2004, pp. 96–101.
2. K. Bernstein, D. J. Frank, A. E. Gattiker, W. Haensch, B. L. Ji, S. R. Nassif, E. J. Nowak, D. J. Pearson, and N. J. Rohrer, "High-Performance CMOS Variability in the 65-nm Regime and Beyond," *IBM J. Res. & Dev.* **50**, No. 4/5, 433–449 (2006, this issue).
3. A. Bryant, J. Brown, P. Cottrell, M. Ketchen, J. Ellis-Monaghan, and E. J. Nowak, "Low Power CMOS at Vdd=4KT/q," *Proceedings of the Device Research Conference*, Notre Dame, IN, June 2001, pp. 22–23.
4. W. Haensch, E. J. Nowak, R. H. Dennard, P. M. Solomon, A. Bryant, O. H. Dokumaci, A. Kumar, X. Wang, J. B. Johnson, and M. V. Fischetti, "Silicon CMOS Devices Beyond Scaling," *IBM J. Res. & Dev.* **50**, No. 4/5, 339–361 (2006, this issue).
5. J. Kim and K. Roy, "Double Gate-MOSFET Subthreshold Circuit for Ultralow Power Applications," *IEEE Trans. Electron Devices* **51**, No. 9, 1468–1474 (September 2004).
6. J. Meindl and J. Davis, "The Fundamental Limit on Binary Switching Energy for Terascale Integration (TSI)," *IEEE J. Solid-State Circuits* **35**, No. 10, 1515–1516 (October 2000).
7. R. Swanson and J. Meindl, "Ion-Implanted Complementary MOS Transistors in Low-Voltage Circuits," *IEEE J. Solid-State Circuits* **7**, No. 2, 146–153 (April 1972).
8. J. B. Burr, "Cryogenic Ultra Low Power CMOS," *Proceedings of the IEEE Symposium on Low Power Electronics,* October 1995, pp. 82–83.
9. B. Zhai, D. Blaauw, D. Sylvester, and K. Flautner, "The Limit of Dynamic Voltage Scaling and Insomniac DVS," *IEEE Trans. VLSI Syst.* **13**, No. 11, 1239–1252 (2005).
10. B. Calhoun and A. Chandrakasan, "Characterizing and Modeling Minimum Energy Operation for Subthreshold Circuits," *Proceedings of the International Symposium on Low Power Electronics and Design (ISLPED)*, Newport, RI, August 2004, pp. 90–95.
11. A. Wang, A. Chandrakasan, and S. Kosonocky, "Optimal Supply and Threshold Scaling for Subthreshold CMOS Circuits," *Proceedings of the IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, Pittsburgh, PA, April 2002, pp. 5–9.
12. L. Nazhandali, B. Zhai, J. Olson, A. Reeves, M. Minuth, R. Hefland, S. Pant, T. Austin, and D. Blaauw, "Energy Optimization of Subthreshold-Voltage Sensor Network Processors," *Proceedings of the International Symposium on Computer Architecture*, Madison, WI, June 2005, pp. 197–207.
13. S. Mutoh, T. Douseki, Y. Matsuya, T. Aoki, S. Shigematsu, and J. Yamada, "1-V Power Supply High-Speed Digital Circuit Technology with Multithreshold-Voltage CMOS," *IEEE J. Solid-State Circuits* **30**, No. 8, 847–854 (August 1995).
14. J. Halter and F. Najm, "A Gate-Level Leakage Power Reduction Method for Ultra-Low-Power CMOS Circuits," *Proceedings of the Custom Integrated Circuits Conference (CICC)*, Santa Clara, CA, May 1997, pp. 475–478.
15. T. Kuroda, T. Fujita, S. Mita, T. Nagamatsu, S. Yoshioka, K. Suzuki, F. Sano, M. Norishima, M. Murota, M. Kako, M. Kinugawa, M. Kakumu, and T. Sakurai, "A 0.9-V, 150-MHz, 10-mW, 4 mm^2, 2-D Discrete Cosine Transform Core Processor with Variable Threshold-Voltage (VT) Scheme," *IEEE J. Solid-State Circuits* **31**, No. 11, 1770–1779 (November 1996).
16. J. Tschanz, J. Kao, S. Narendra, R. Nair, D. Antoniadis, A. Chandrakasan, and V. De, "Adaptive Body Bias for Reducing Impacts of Die-to-Die and Within-Die Parameter Variations on Microprocessor Frequency and Leakage," *IEEE J. Solid-State Circuits* **37**, No. 11, 1396–1402 (November 2002).
17. T. Chen and S. Naffziger, "Comparison of Adaptive Body Bias (ABB) and Adaptive Supply Voltage (ASV) for Improving Delay and Leakage Under the Presence of Process Variation," *IEEE Trans. Very Large Scale Integration (VLSI) Syst.* **11**, No. 5, 888–899 (October 2003).
18. T. Mizuno, J. Okamura, and A. Toriumi, "Experimental Study of Threshold Voltage Fluctuation Due to Statistical Variation of Channel Dopant Number in MOSFET's," *IEEE Trans. Electron Devices* **41**, No. 11, 2216–2221 (November 1994).
19. B. Zhai, S. Hanson, D. Blaauw, and D. Sylvester, "Analysis and Mitigation of Variability in Subthreshold Design," *Proceedings of the International Symposium on Low Power Electronics and Design (ISLPED)*, San Diego, CA, August 2005, pp. 20–25.
20. E. Seevinck, F. J. List, and J. Lohstroh, "Static-Noise Margin Analysis of MOS SRAM Cells," *IEEE J. Solid-State Circuits* **22**, No. 5, 748–754 (October 1987).
21. A. J. Bhavnagarwala, X. Tang, and J. D. Meindl, "The Impact of Intrinsic Device Fluctuations on CMOS SRAM Cell Stability," *IEEE J. Solid-State Circuits* **36**, No. 4, 658–665 (April 2001).
22. H. Qin, Y. Cao, D. Markovic, A. Vladimirescu, and J. Rabaey, "Standby Supply Voltage Minimization for Deep Sub-Micron SRAM," *Microelectron. J.* **36**, No. 9, 789–800 (September 2005).
23. L. Chang, D. M. Fried, J. Hergenrother, J. W. Sleight, R. H. Dennard, R. K. Montoye, L. Sekaric, S. J. McNab, A. W. Topol, C. D. Adams, K. W. Guarini, and W. Haensch, "Stable SRAM Cell Design for the 32 nm Node and Beyond," *Proceedings of the Symposium on VLSI Technology,* June 2005, pp. 128–129.
24. R. Gonzalez and M. Horowitz, "Energy Dissipation in General Purpose Microprocessors," *IEEE J. Solid State Circuits* **31**, No. 9, 1277–1284 (September 1996).

25. A. Keshavarzi, S. Ma, S. Narendra, B. Bloechel, K. Mistry, T. Ghani, S. Borkar, and V. De, “Effectiveness of Reverse Body Bias for Leakage Control in Scaled Dual Vt CMOS ICs,” *Proceedings of the International Symposium on Low Power Electronics and Design (ISLPED)*, Huntington Beach, CA, August 2001, pp. 207–212.
26. D. E. Lackey, P. S. Zuchowski, T. R. Bednar, D. W. Stout, S. W. Gould, and J. M. Cohn, “Managing Power and Performance for System-on-Chip Designs Using Voltage Islands,” *Proceedings of the International Conference on Computer Aided Design*, San Jose, CA, November 2002, pp. 195–202.
27. T. Burd and R. Broderson, “Design Issues for Dynamic Voltage Scaling,” *Proceedings of the International Symposium on Low Power Electronics and Design (ISLPED)*, Rapallo/Portacino Coast, July 2000, pp. 9–14.
28. M. Miyazaki, J. Kao, and A. Chandrakasan, “A 175mV Multiply–Accumulate Unit Using an Adaptive Supply Voltage and Body Bias (ASB) Architecture,” *Proceedings of the International Solid-State Circuits Conference (ISSCC)*, February 2002, pp. 58–59 and p. 444.
29. S. Shigematsu, S. Mutoh, Y. Matsuya, Y. Tanabe, and J. Yamada, “A 1-V High-Speed MTCMOS Circuit Scheme for Power-Down Application Circuits,” *IEEE J. Solid-State Circuits* **32**, No. 6, 861–869 (June 1997).
30. S. Kosonocky, M. Immediato, P. Cottrell, T. Hook, R. Mann, and J. Brown, “Enhanced Multi-Threshold (MTCMOS) Circuits Using Variable Well Bias,” *Proceedings of the International Symposium on Low Power Electronics and Design (ISLPED)*, Huntington Beach, CA, August 2001, pp. 165–169.
31. S. Gary, C. Dietz, J. Eno, G. Gerosa, Sung Park, and H. Sanchez, “The PowerPC 603 Microprocessor: A Low-Power Design for Portable Applications,” *Proceedings of Compcon Spring ’94*, February/March 1994, pp. 307–315.
32. R. Kumar, K. I. Farkas, N. P. Jouppi, P. Ranganathan, and D. M. Tullsen, “Single-ISA Heterogeneous Multi-Core Architectures: The Potential for Processor Power Reduction,” *Proceedings of the IEEE/ACM International Symposium on Microarchitecture (MICRO)*, San Diego, CA, December 2003, pp. 81–92.
33. T. B. Hook, M. Breitwisch, J. Brown, P. Cottrell, D. Hoyniak, Chung Lam, and R. Mann, “Noise Margin and Leakage in Ultra-Low Leakage SRAM Cell Design,” *IEEE Trans. Electron Devices* **49**, No. 8, 1499–1501 (August 2002).

Received October 3, 2005; accepted for publication January 9, 2006; Internet publication August 9, 2006

Scott Hanson *Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan 48109 (hansons@eecs.umich.edu).* Mr. Hanson received his B.S. degree in electrical engineering *summa cum laude* from the University of Michigan at Ann Arbor in 2004. He is currently pursuing his Ph.D. degree at the University of Michigan and is the recipient of a fellowship from the Semiconductor Research Corporation (SRC). Mr. Hanson’s research is focused on the development of energy-optimal circuit design strategies, with a particular emphasis on subthreshold CMOS design techniques.

Bo Zhai *Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan 48109 (bzhai@eecs.umich.edu).* Mr. Zhai received his B.S. degree in microelectronics from Peking University, China, in 2002 and his M.S. degree in electrical engineering from the University of Michigan at Ann Arbor in 2004. He is currently a Ph.D. candidate in electrical engineering at the University of Michigan. Mr. Zhai is a research assistant in the Advanced Computer Architecture Laboratory at the University of Michigan, working with Prof. David Blaauw. His research focuses on low-power VLSI design.

Kerry Bernstein *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (kbernste@us.ibm.com).* Mr. Bernstein is a Senior Technical Staff Member at the IBM Thomas J. Watson Research Center. He is currently responsible for future product technology definition, performance, and application. He received his B.S. degree in electrical engineering from Washington University in St. Louis, joining IBM in 1978. He holds 50 U.S. patents and is a coauthor of three college textbooks and multiple papers on high-speed and low-power CMOS. Mr. Bernstein is currently interested in the area of high-performance, low-power advanced circuit technologies. He is a Senior Member of the IEEE, and is a staff instructor at RUNN/Marine Biological Laboratories, Woods Hole, Massachusetts.

David Blaauw *Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan 48109 (blaauw@eecs.umich.edu).* Dr. Blaauw received his B.S. degree in physics and computer science from Duke University in 1986, his M.S. degree in computer science from the University of Illinois, Urbana, in 1988, and his Ph.D. degree in computer science from the University of Illinois, Urbana, in 1991. He worked for the IBM Corporation as a Development Staff Member until August 1993. From 1993 until August 2001, he worked for Motorola, Inc. in Austin, TX, where he was the manager of the High Performance Design Technology group. Since August 2001, he has been on the faculty at the University of Michigan as an Associate Professor. Dr. Blaauw’s work has focused on VLSI design and CAD, with particular emphasis on circuit design and optimization for high-performance and low-power designs. He was the Technical Program Chair and General Chair for the International Symposium on Low Power Electronics and Design in 1999 and 2000, respectively, and he was the Technical Program Co-Chair and member of the Executive Committee of the ACM/IEEE Design Automation Conference in 2000 and 2001.

Andres Bryant *IBM Systems and Technology Group, 1000 River Street, Essex Junction, Vermont 05452 (bryanta@us.ibm.com).* Dr. Bryant received his B.S.E.E. degree from the University of Maine in 1982 and his Ph.D. degree in electrical engineering from Stanford University in 1986, joining IBM in Burlington, Vermont, that same year. His work areas have ranged from surface-acoustic-wave gas sensors and scanning tunneling microscopy to CMOS transistor design. He has also worked on DRAM transistor design and high-performance-logic transistor design. Dr. Bryant's current interests include energy-efficient, ultralow-voltage transistor and circuit design.

Leland Chang *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (lelandc@us.ibm.com).* Dr. Chang received his B.S., M.S., and Ph.D. degrees in electrical engineering and computer science in 1999, 2001, and 2003, respectively, from the University of California at Berkeley. He joined the IBM Thomas J. Watson Research Center in 2003 as a Research Staff Member in the area of silicon technology. He is currently bridging the gap between device fabrication and circuit design to confront issues concerning continued technology scaling and the introduction of new device structures. Dr. Chang's technical research has encompassed topics ranging from advanced silicon device technology and nonvolatile memory devices to RF MEMS. With colleagues at U. C. Berkeley and Advanced Micro Devices, he helped demonstrate the FinFET double-gate structure down to record gate lengths. With IBM researchers, he has worked to demonstrate SRAM cells at record small sizes and has proposed alternative cell designs to improve SRAM stability. He has authored or coauthored more than 35 technical papers.

Koushik K. Das *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (kkdas@us.ibm.com).* In 1998, Dr. Das received his B. Tech. (honors) degree in electronics and electrical communications engineering from the Indian Institute of Technology (IIT), Kharagpur, India, and his M.S. and Ph.D. degrees in electrical engineering from the University of Michigan at Ann Arbor in 2000 and 2003, respectively. At IIT in 1998, he won the President of India's Gold Medal for being the most outstanding undergraduate student among all branches of engineering and sciences. Dr. Das has been a Research Staff Member at the IBM Thomas J. Watson Research Center since 2003. His research interests include high-speed low-power VLSI circuit design, with an emphasis on SOI technology. Dr. Das has authored numerous papers and holds several U.S. patents. He is currently serving on technical program committees of the IEEE International System-on-Chip Conference, the ACM Great Lakes Symposium on VLSI, the IEEE International Conference on Microelectronic Systems Education, and the IBM Watson PAC2 conference. He has also served as session chair/co-chair in multiple ACM/IEEE conferences.

Wilfried Haensch *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (whaensch@us.ibm.com).* In 1981, Dr. Haensch received his Ph.D. degree from the Technical University of Berlin, Germany, in the field of theoretical solid-state physics. In 1984 he joined Siemens Corporate Research in Munich to investigate high-field transport in MOSFET devices, and in 1988 he joined the DRAM development team at the Siemens Research Laboratory to investigate new cell concepts. In 1990, he joined the DRAM alliance between IBM and Siemens to develop quarter-micron 64M DRAM. In this capacity, Dr. Haensch was involved with device characterization of shallow-trench bounded devices and cell-design concerns. In 1996, he moved to a manufacturing facility to build various generations of DRAM. His primary mission was to transfer technologies from development into manufacturing and to guarantee a successful yield ramp of the product. In 2001, he joined the IBM Thomas J. Watson Research Center to lead a group concerned with novel devices and applications. He is currently responsible for post-45-nm-node device design and its implication for circuit functionality.

Edward J. Nowak *IBM Systems and Technology Group, 1000 River Street, Essex Junction, Vermont 05452 (ejnowak@us.ibm.com).* Dr. Nowak received his B.S. degree in physics in 1973 from M.I.T., and M.S. and Ph.D. degrees, also in physics, from the University of Maryland in 1975 and 1978, respectively. In 1981, following postdoctoral research at New York University, he joined IBM in Essex Junction, Vermont, to work on DRAM development. Since 1985, Dr. Nowak has worked in high-performance CMOS device design. His current interests include energy-driven device design and FinFET device architectures.

Dennis M. Sylvester *Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan 48109 (dennis@eecs.umich.edu).* Dr. Sylvester received his B.S. degree in electrical engineering *summa cum laude* from the University of Michigan in 1995. He received his M.S. and Ph.D. degrees in electrical engineering from the University of California at Berkeley in 1997 and 1999, respectively. After completing research with the Advanced Technology Group of Synopsys and Hewlett-Packard Laboratories, Dr. Sylvester joined the University of Michigan, where he is currently Associate Professor of Electrical Engineering and Computer Science. He has published numerous articles along with one book and several book chapters in his field of research, which includes low-power circuit design and design automation techniques, design-for-manufacturability, and on-chip interconnect modeling. Dr. Sylvester received an NSF CAREER award, the 2000 Beatrice Winner Award at ISSCC, a 2004 IBM Faculty Award, and several best paper awards and nominations. He has served on the technical program committees of numerous design automation and circuit design conferences and was general chair of two ACM/IEEE workshops. He is currently an Associate Editor for *IEEE Transactions on VLSI Systems* and *IEEE Transactions on Computer-Aided Design.* He also helped define the circuit and physical design roadmap as a member of the International Technology Roadmap for Semiconductors (ITRS) U.S. Design Technology Working Group from 2001 to 2003.