Energy Optimality and Variability in Subthreshold Design

Scott Hanson Bo Zhai David Blaauw Dennis Sylvester
University of Michigan
{hansons,bzhai,blaauw,dmcs}@umich.edu

Andres Bryant Xinlin Wang
IBM T.J. Watson Research Center
{bryanta,xinlinw}@us.ibm.com

ABSTRACT

Recent progress in the development of subthreshold circuit design techniques has created the opportunity for dramatic energy reductions in many applications. However, energy efficiency comes at the price of timing and energy variability due to process variations. We explore energy optimality in the subthreshold regime, discuss variability in this region, and highlight the energy and variability characteristics of a real subthreshold design.

Categories and Subject Descriptors:
B.7.1 Integrated Circuits - Types and Design Styles

General Terms: Design

Keywords: Subthreshold circuits, variability, ultra-low energy

1. INTRODUCTION

The low voltage design community has grown explosively during the past few years. New research has built a theoretical foundation for subthreshold ($V_{dd} < V_{th}$) design strategies and has begun to address the more critical problems of memory design and process variability. Subthreshold designs, which can improve energy efficiency by several orders of magnitude, will play an important role in the continuing drive toward mobile, pervasive computing. Applications like environmental monitoring, biomedical sensing, and supply chain management will be bolstered by complex subthreshold logic. High performance servers may also receive significant benefit from the subthreshold design community through low power, massively parallel systems of subthreshold processors. In this paper, we address three key points. In the first section, we show how subthreshold operation can be used to achieve energy optimality. We then highlight one of the primary problems facing subthreshold circuit designers: variability. In the final section, we relate the topics of the first two sections to measurements of a real subthreshold design.

2. MINIMIZING ENERGY

Though power has received more attention than energy in optimizing high performance and embedded processors, energy is a more suitable metric for mobile applications. Both theory [1] and hardware measurements [2][3] have shown that the energy consumed by a processor per operation is typically minimized in the subthreshold regime, where the supply voltage (V_{dd}) is smaller than the device threshold voltage (V_{th}). As V_{dd} is reduced toward V_{th}, the transistor drive current evolves from drift-dominated strong-inversion current to diffusion-dominated weak-inversion current. Though the magnitude of the weak-inversion current is small, it is large enough to charge and discharge nodes in the manner required by digital logic.

It is well known that digital logic may function correctly at extremely low voltages, but it is worthwhile to consider whether there are energy benefits to low voltage operation. The energy consumed by an inverter chain is described by Equation 1 and plotted in Figure 1 for a 130nm technology. Energy is composed of two components: dynamic energy and leakage energy. Short-circuit energy is commonly ignored but may be accounted for through appropriate adjustment of the activity factor.

$$E = E_{dyn} + E_{leak} = \frac{1}{2}\cdot C_s \cdot V_{dd}^2 \cdot \alpha + I_{leak}\cdot V_{dd}\cdot t_p \quad \textbf{(EQ 1)}$$

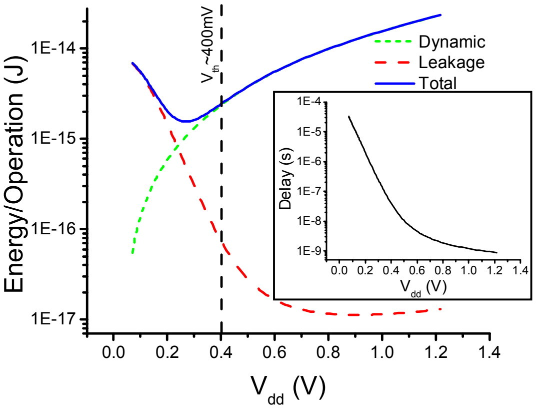

Figure 1: Delay and energy consumption for a chain of 50 inverters with activity factor of 0.2. For this circuit, energy reaches a minimum at V_{dd}=266 mV (130nm technology)

Figure 1 also shows how the delay of an inverter chain increases dramatically as V_{dd} scales into the subthreshold regime. For many applications, the performance penalty paid for subthreshold operation is tolerable. However, the implications of increased delay extend beyond performance concerns. As Figure 1 shows, the leakage energy of the inverter chain under test increases significantly as voltage reduces. Though leakage *power* ($I_{leak}V_{dd}$) reduces with supply voltage, delay increases exponentially and forces leakage *energy* ($I_{leak}V_{dd}t_p$) to increase. Consequently, total energy reaches a minimum value at a voltage called V_{min}. The expression for V_{min} was first solved numerically in [1] as:

$$V_{min} = \left[1.587\cdot\ln\left(\eta\cdot\frac{n}{\alpha}\right) - 2.355\right]\cdot m\cdot v_T \quad \textbf{(EQ 2)}$$

Most notably absent from Equation 2 is V_{th}. Assuming that the circuit is operating safely in the subthreshold regime, V_{min} does not depend on V_{th}. Equation 2 also highlights the fact that V_{min} is closely tied to the dynamic to leakage ratio through the n and α terms. Circuits with longer paths (larger n) tend to consume more leakage and tend to have larger values of V_{min}. Similarly, circuits with lower switching activity tend to be more leakage

ISLPED'06, October 4–6, 2006, Tegernsee, Germany.
Copyright 2006 ACM 1-59593-462-6/06/0010...$5.00.

dominated and have larger values of V_{min}. These sensitivities are important when thinking about V_{dd} selection for a larger system. Caches, for example, will achieve optimality at much higher V_{dd} than general logic due to differences in switching activity.

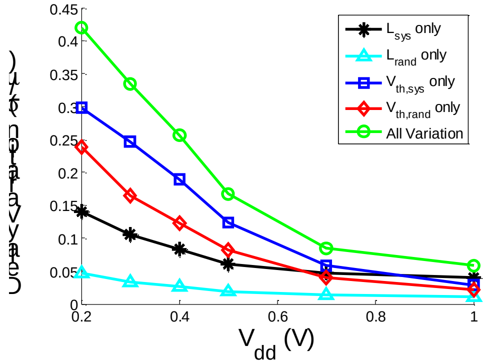

Figure 2: Delay variability (σ/μ) as a function of supply voltage (65 nm technology)

3. ADDRESSING VARIABILITY

The last section showed that energy/operation may be, in theory, minimized by operating in the subthreshold regime. In practice, achieving energy optimality is not as simple as reducing V_{dd} to V_{min}. Process-induced variability leads to problems with both functionality and energy efficiency.

We can take a simplistic but accurate view of variability by assuming that there are four types of variation: systematic V_{th} variation, random V_{th} variation, systematic gate length variation, and random gate length variation. There is also some component of threshold and gate length variation that varies from region to region on the chip, but this can be safely grouped with global systematic variation for our simple discussion.

Figure 2 shows how the delay variation of a chain of 10 inverters changes as V_{dd} scales in a 65 nm technology. Since subthreshold current is exponentially dependent on V_{th}, variation in V_{th} becomes more problematic at subthreshold voltages. Conversely, subthreshold current is inversely proportional to gate length and has a relatively weak exponential dependence through DIBL-induced V_{th} variations. V_{th} variations are therefore the most important concern for subthreshold designers.

3.1 Timing Variability

The increased sensitivity to threshold voltage fluctuations in the subthreshold regime leads to dramatic variations in gate delay, which may result in both late-mode and early-mode timing failures. Late-mode failures occur when a circuit path delay exceeds the clock period and may be fixed by increasing the clock period. Monte Carlo simulations show that the clock period for a 10-inverter chain in a 65 nm technology must increase by 10% at V_{dd}=1V and an astonishing 230% at V_{dd}=300mV to eliminate late-mode errors introduced by variability. The performance and energy implications of addressing late-mode failures are clearly undesirable. Early-mode failures can occur when excessive clock skew allows data to be latched one clock cycle early at a receiving latch. Monte Carlo simulations suggest that clock skew (in terms of FO4 delay) can increase by more than 10X as voltage is scaled from 1V to 300mV. Early-mode failures must be fixed by adding delay elements to short paths or by designing variation-tolerant clock distribution networks.

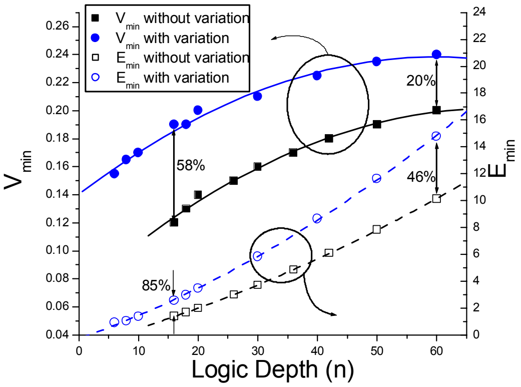

Figure 3: Variation in worst-case (μ+3σ) V_{min} and E_{min} for an inverter chain of length n gates. The relative increases in V_{min} and E_{min} are less severe at large n (130 nm technology).

3.2 Energy Variability

For most subthreshold designs, energy will be the most important metric. It is therefore very important to understand how V_{min}, identified in Section 2, is affected by variability. While dynamic energy remains relatively constant with variability, worst-case delay and worst-case leakage energy increase dramatically [4]. Assuming a fixed frequency and supply voltage across all chips, process variation leads to a lower operating frequency and consequently, a dramatic increase in leakage. Figure 3 shows the variation in the worst-case V_{min} and energy for an inverter chain of length n in a 130 nm technology.

3.3 Mitigating Variability

Though variability is one of the primary problems in the subthreshold regime, careful design will help alleviate its effects. We begin by considering device-level optimizations. The FinFET promises to reduce threshold variation (δV_{th}) induced by random dopant fluctuations (RDF) as well as δV_{th} induced by short channel effects (SCEs). With its double gate and thin body, the FinFET can achieve excellent SCE control down to very short channels in the absence of channel doping [5]. Figures 4(a-c) show FIELDAY simulation results comparing the SCEs of a planar poly-gate PDSOI FET, a poly-gate FinFET with body doping, and a mid-gap work-function metal gate FinFET without body doping. Even without body doping, the FinFET V_{th}, drain barrier lowering, and subthreshold slope dependence on gate length are much weaker than those observed for planar PDSOI FETs. Thus, the FinFET enables the elimination of RDF-induced δV_{th} while still reducing SCE-induced δV_{th}. However, the elimination of body doping will require the development of metal gates of varying work-functions to set thresholds to their optimum values. In Figure 4(a), the mid-gap metal gate FinFET has a V_{th} that may be too high for some applications. A quarter-gap metal gate would reduce V_{th} to a more reasonable level. An alternative approach would be to use a split-gate FinFET, where one gate is active and the other gate is used to adjust V_{th}. With this "back-gate" approach, FETs with multiple thresholds can coexist in the same circuit, and thresholds can be adjusted dynamically to optimize power-performance tradeoffs [5].

In addition to novel device design, there are several well-studied circuit techniques that can reduce variability significantly. Systematic variations can be addressed globally with techniques like dynamic voltage scaling (DVS) and adaptive body biasing (ABB), though both techniques incur area and complexity overheads. Random variations are somewhat more difficult to address. Figure 3 shows that relative energy (E_{min}) and V_{min} variations can be reduced dramatically by increasing path lengths in a circuit. Recent research has also shown that timing and energy fluctuations due to random variation can be reduced by using larger gate sizes [4][6]. Both longer paths and larger gates attempt to "average out" variability over larger gate area. When using these techniques, designers must be careful to weigh the variability benefits against the accompanying energy penalties.

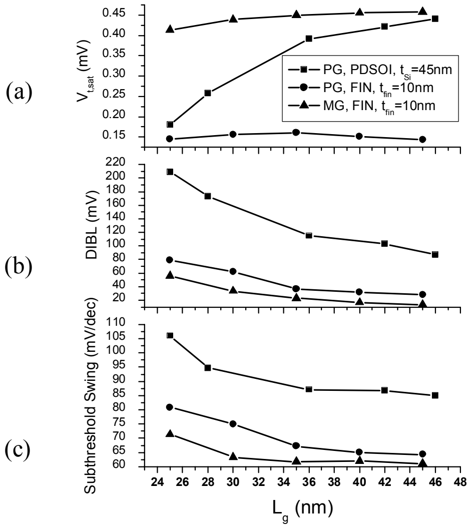

Figure 4: (a) $V_{th,sat}$ (b) DIBL and (c) subthreshold swing characteristics for a poly-gate (PG) partially-depleted SOI (PDSOI) device, a poly-gate FINFET (FIN) with body doping, and a metal-gate (MG) FINFET without body doping

Successful subthreshold operation will undoubtedly require the development of error-tolerant architectures. Error correction codes (ECC) and memory redundancy have been studied extensively, but it will be important to create whole architectures that can dynamically detect and correct errors. Without such techniques, designers will need to incorporate large margins in design parameters to avoid timing errors and will need to resort to expensive statistical techniques like Monte Carlo simulation.

4. A 2.6 PJ SUBTHRESHOLD PROCESSOR

A number of subthreshold circuits have been successfully demonstrated recently [2][3][7]. In this section, we focus on the design described in [2] to put the topics from Sections 2 and 3 into the context of a real design. The chip under test is a subthreshold sensor network processor with an 8-bit CISC architecture and a 2-kbit memory. The processor, fabricated in a 130nm technology, occupies an area of 85,022 μm^2.

Figure 5(a) shows the energy consumption of the processor for a typical instruction stream. The total energy consumption clearly reaches the minimum predicted by Equation 2. The energy minimum of 2.6 pJ per instruction occurs at V_{dd}=400 mV, which is significantly higher than the minimum energy voltage observed for the inverter chain in Figure 1. Recall that V_{min} tends to be higher for circuits with lower switching activity. The memory, which has a very low switching activity compared to typical logic, accounts for 65% of the total transistor area and is largely responsible for the higher V_{min}.

Figure 5(b) shows measured energy distributions for 26 chips under different frequency and voltage selection schemes. As expected, the worst-case energy is minimized by allowing each chip to operate at its own V_{min} and frequency. V_{min} varies by 80 mV across the measured chips with worst-case energy variation ($3\sigma/\mu$) of 16.99%. These fluctuations in V_{min} and energy are in agreement with the trends discussed in the last section but are not nearly as large as one might expect when looking at the variability characteristics of a single gate (where $3\sigma/\mu$ for delay can be greater than 200% in the subthreshold regime).

Allowing each chip to operate at the optimal supply voltage and frequency requires adaptive voltage and frequency tuning and significant design overhead. This design overhead can be minimized to some extent by fixing the supply voltage at a mean voltage but allowing frequency to vary. As shown in Figure 5(b), the energy penalty for choosing this scheme is small. The energy minimum is shallow, so the choice of supply voltage can deviate from V_{min} without a significant penalty. However, fixing both V_{dd} and frequency (also shown in Figure 5(b)) imposes a significant energy penalty and is only desirable when the overheads of adaptive voltage and frequency tuning are large.

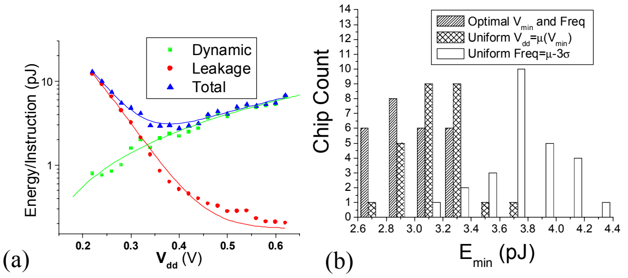

Figure 5: (a) Energy consumption of an 8-bit subthreshold processor (130 nm technology) (b) Energy distributions for 26 measured subthreshold processors using different supply voltage and frequency selection schemes (130 nm technology)

5. CONCLUSIONS

In this paper, we used both simulated and measured hardware to demonstrate that energy efficiency can be maximized by operating in the subthreshold regime. One of the primary problems that must be addressed by subthreshold designers is the management of variability. Threshold variability, in particular, becomes problematic at low voltage. Existing device, circuit and architectural techniques offer some reprieve to this problem, but further innovation will be necessary to ensure that subthreshold design gets widespread acceptance.

6. REFERENCES

[1] B. Zhai, D. Blaauw, D. Sylvester, K. Flautner, "Theoretical and Practical Limits of Dynamic Voltage Scaling," *DAC*, 2004, pp. 868-873.

[2] B. Zhai, et al., "A 2.60pJ/Inst Subthreshold Sensor Processor for Optimal Energy Efficiency," *VLSI Circuits Symposium,* 2006.

[3] B. Calhoun, A. Wang, A. Chandrakasan, "Device sizing for minimum energy operation in subthreshold circuits," *CICC*, 2004, pp. 95-98.

[4] B. Zhai, S. Hanson, D. Blaauw, D. Sylvester, "Analysis and Mitigation of Variability in Subthreshold Design," *ISLPED*, 2005, pp. 20-25.

[5] W. Haensch, et al., " Silicon CMOS devices beyond scaling," *IBM J. Res. & Dev.* ***50***,No. 4, pp. 339-362.

[6] J. Kwong, A. Chandrakasan, "Variation-Driven Device Sizing for Minimum Energy Sub-threshold Circuits," *ISLPED*, 2006.

[7] C. Kim, H. Soeleman, K. Roy, "Ultra-Low-Power DLMS Adaptive Filter for Hearing Aid Applications," *IEEE Trans on VLSI Systems* ***11*****,** No. 6, pp. 1058-1067.