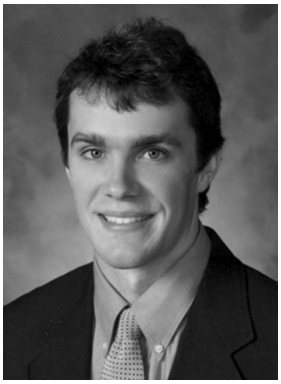

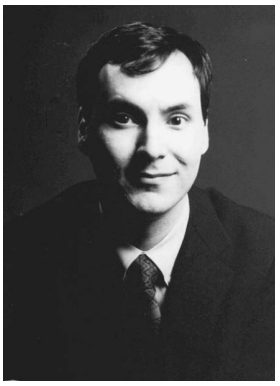

Can Subthreshold and Near-Threshold Circuits Go Mainstream?

BENTON H. CALHOUN
University of Virginia

DAVID BROOKS
Harvard University

Many research teams have demonstrated the ability to operate digital complementary metal-oxide semiconductor (CMOS) chips in the subthreshold or near-threshold region in recent years, but no commercial applications have yet adopted this approach. Subthreshold operation refers to using a supply voltage (V_{DD}) that is less than a single transistor's threshold voltage (V_T). Near-threshold operation refers to using a V_{DD} that is close to (that is, slightly above or below) V_T. Lowering the supply voltage reduces power consumption and also increases energy efficiency by lowering the energy consumed per operation. Subthreshold operation involves using a supply voltage in the 0.2 V to 0.5 V range. This is substantially lower than nominal supply voltages, which fall in the 0.9 V to 1.2 V range for modern process technologies. Lowering the voltage so far can reduce energy consumption by more than 10 times because active energy is proportional to V_{DD}^2.

By the traditional definition (for $V_{GS} < V_T$, $I_D = 0$, where V_{GS} is the MOSFET gate-to-source voltage and I_D is the drain current), transistors in a digital circuit with a subthreshold V_{DD} are always ''off.'' However, the transistor drain current does not immediately fall to zero when V_{DD} drops below V_T. Instead, it decreases exponentially ($I_D \alpha \exp(V_{GS} - V_T)$) in the subthreshold region. Thus, a nonzero gate voltage that is less than V_T will still produce a drain current that is larger than the off current ($I_{OFF} = I_D$ when $V_{GS} = 0$). This positive ratio of I_{ON}/I_{OFF} lets subthreshold digital gates behave statically in a similar fashion to strong inversion. Their transient behavior is much slower because the on-current in the subthreshold region is orders of magnitude less than in strong inversion.

Figure 1 shows circuit delay as a function of voltage. The exponential increase in delay resulting from decaying I_{ON} in subthreshold is obvious. The figure also shows the benefit of lower energy during subthreshold operation. The total energy per operation decreases quadratically (because $E_{ACTIVE} \alpha V_{DD}^2$) until leakage energy becomes dominant (due to the longer delay) and creates a minimum energy point. The voltage that minimizes energy per operation occurs in the subthreshold region for most digital designs, and, as we discuss later, minimizing energy consumption is the primary goal for a broad class of severely energy-constrained applications.

In addition to the lower speed, subthreshold circuits face a few other key challenges. First, the lower I_{ON}/I_{OFF} ratio can lead to functional problems. For example, certain circuit structures with multiple parallel leaking paths (such as SRAM bitlines or wide NOR gates) degrade this ratio to the point that the circuit does not work properly. Second, although variations in transistor V_T affect circuits at nominal V_{DD}, their impact on current is exponential in the subthreshold regime. This means that process variation can severely degrade the already weakened I_{ON}/I_{OFF} ratio. For example, circuits that depend on a ratio of transistor sizes to reliably resolve a fight between two transistors (such as a dynamic logic keeper or an SRAM cell during write) will fail in subthreshold because V_T variation will alter those devices' strengths beyond the ability of size to compensate. Figure 2 demonstrates how V_T variation increases the spread of critical path delay's distribution at lower voltages.

Despite the challenges of subthreshold operation, researchers have successfully

Editors' Note

Recent research has shown the potential benefits of subthreshold or near-threshold operation, which gives up a substantial degree of speed in order to reduce energy per operation. This is an excellent trade-off for many tasks, such as cyberphysical systems. This prolegomenon summarizes the benefits and challenges of subthreshold or near-threshold operation.

Published by the IEEE Computer Society

developed techniques to build robust low-voltage circuits for logic, memory, and processors. We believe that subthreshold and near-threshold circuits will make their way into commercial products. They will first appear in ASICs for ultralow power and low-energy applications as a means of meeting extremely limited energy budgets. Next, they will show up as integrated components or special modes in high-performance chips, including mainstream processors. Finally, near-threshold operation will emerge as a competitive choice to address the stringent power crisis for highly parallel peta- or exascale processing.

Subthreshold circuits for ultralow energy and ultralow power

Most applications of subthreshold circuits so far have focused on operating scenarios that are severely energy or power constrained. For example, wearable body sensors or wireless sensor nodes must be small (tiny battery with little energy stored) and long lasting (low energy), but they do not require high operating frequencies. These sorts of energy-constrained applications are perfect for subthreshold circuits, which can provide digital signal processing up to several tens of megahertz at extremely low energy per operation.

For example, we have demonstrated a 0.13-μm CMOS mixed-signal system on chip (SoC) that implements an electrocardiogram (ECG). The SoC includes an analog front end (instrumentation amp [IA] and analog-to-digital converter [ADC]) and a subthreshold microprocessor.[1] The processor consumes only 1.5 pJ at 280 mV (700 nW), and we use it for signal processing and for controlling the analog circuits. This chip is ideal for monitoring patients for cardiac arrhythmias, whose presence can be identified by abnormalities in the heart rate, making full ECG transmission only necessary when actual arrhythmic events occur. By extracting the heart rate interval on chip, we can reduce the wireless data rate by more than 500 times, allowing for reduced use of a power-hungry radio.[1] The heart rate algorithm can continue to operate accurately even when the analog components operate at lower voltages to save power. The SoC consumes only 2.6 μW while providing either computed heart rate or raw ECG data.

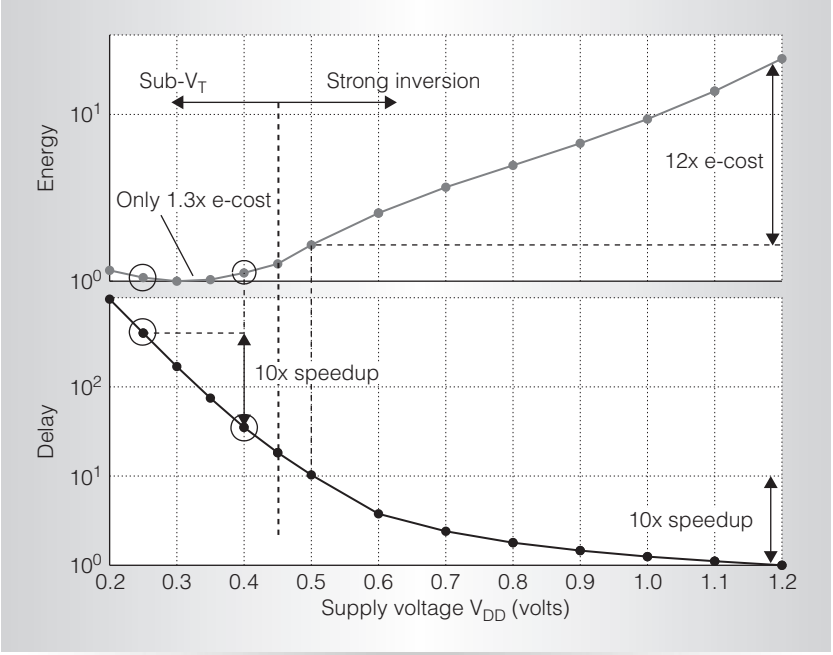

Figure 1. Normalized energy per operation (top) and normalized delay (bottom) of a digital circuit as a function of V_{DD}. Subthreshold circuits minimize energy consumption at a cost of slower speed.

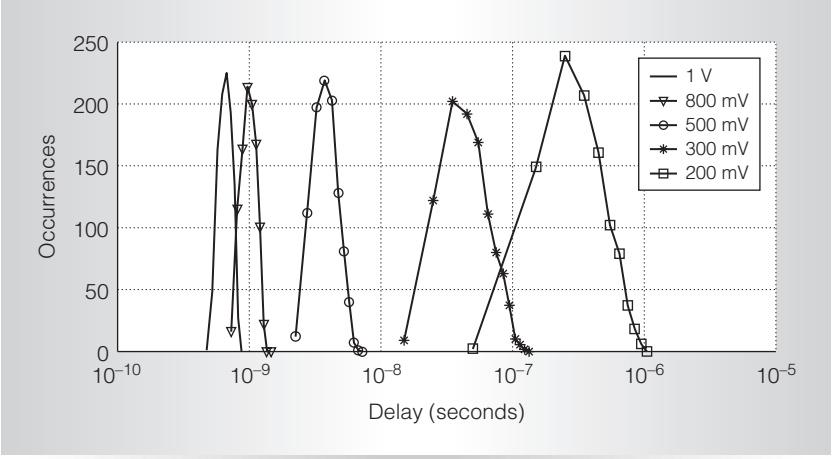

Figure 2. Distribution of a critical path's delay at different V_{DD} values showing the increase of variability at low voltage.

Other recent examples apply similar principles and rely on subthreshold circuits. These include the integration of an analog front end, ADC, and digital seizure classifier for a complete electroencephalogram channel at 120 μW[2]; and an 8.75 mm^2 multichip module that includes a lithium battery, solar cell, and intraocular

pressure sensing chip with 7.7 μW of active power.[3]

Power-constrained systems differ subtly from energy-constrained applications, but they also benefit from subthreshold operation. For example, RFIDs or energy-harvesting systems, where power is limited but ample time is available, might require circuit operation at voltages below even the minimum energy point to reduce power consumption. Because the power source remains present, the fact that the operation takes more energy is less important than keeping the circuit power below the available power constraint.

We believe that subthreshold and near-threshold operations are critical for enabling new energy- or power-constrained systems. However, engineers might have to adjust their design philosophies to incorporate these operating modes. The most important realization to help with this adjustment is that electronic goods consumers want exciting functionality, novel products, and better performance. To consumers, better performance does not just mean frequency, which has been the dominant circuit-related metric (at least from a marketing viewpoint) for many years. In the context of portable computing (cell phones and so on) and emerging applications such as ubiquitous computing and healthcare, performance means longer lifetimes, less conspicuous form factors, or more functionality for the same size or lifetime. For example, a user might want his or her cell phone to provide personal health information, but the device can only do so if sensors on the user's shirt cuff, shoe insert, or chest bandage are sending it relevant data. In addition, the sensors should be so inconspicuous (small and long lasting) that the user is not even aware of them. As long as the system does what the user wants, the user will not care whether the sensor's processor achieves some target clock frequency. If designers prioritize correct system behavior over component-specific circuit metrics (such as frequency, data rate, or operating V_{DD}), they will see that subthreshold operation is a valuable tool for healthcare sensors and other such systems.

For subthreshold circuits to help realize this vision, they will first be designed as tightly integrated pieces of a complete system, such as a sensor node. In this way, the designer can ensure that the chip provides the desired functionality without being tied to a specific V_{DD} or frequency, and different chips can operate at different points in the energy-delay space based on their variation, workload, current operating mode, and so on. Making subthreshold components available (for example, as standalone chips, IP blocks, layers in a 3D integrated circuit, or system in package) will require a similar shift in component specifications. Traditional data sheets carefully specify operating parameters such as V_{DD} and f_{CLK}. However, a subthreshold chip's speed will vary dramatically at a single voltage.

Figures 1 and 2 reveal one option for solving this problem. For V_{DD} well above V_T, delay and energy are both roughly linear with V_{DD}, such that a tenfold change in delay costs roughly 10 times more energy. In the subthreshold region, however, raising V_{DD} slightly increases speed exponentially with only a minor energy penalty. Further, Figure 2 shows that raising V_{DD} rapidly reduces the delay variability. The supply voltage is therefore a powerful knob for adjusting speed and variability in subthreshold circuits. Thus, rather than defining a subthreshold part by specifying a certain V_{DD} and operating frequency, we should define it by its primary function and let each chip set its own V_{DD} and f_{CLK} to meet its functionality obligations.

Subthreshold circuits for high performance?

Although subthreshold circuits limit operating frequency, they can still play a role in high-performance circuits. When high-speed circuits enter a standby state to reduce idle leakage power, some circuits must remain active to monitor the system, decide when to wake up, and oversee power management. One existing example is a standby leakage-reduction system that uses a feedback loop and canary SRAM cells to set the standby V_{DD} to a value that minimizes leakage while protecting data in the main SRAM.[4] Additionally, many high-performance systems spend a large fraction of time in which their processing load is much lower than the maximum. Subthreshold operation during these times can reduce the processor's energy consumption, which can lower on-die temperature and abate energy costs. One practical implementation combines multi-V_{DD} with a small set of block-level power switches to implement a flexible form of dynamic voltage scaling that easily supports a subthreshold mode.[5,6]

Finally, the growing push toward large-scale chip multiprocessors combined with hard power constraints even for high-performance systems has opened a new research avenue for subthreshold and near-threshold computing. The design style's energy-efficiency advantages make subthreshold and near-threshold computing a natural fit for highly parallel architectures such as many-core or wide-SIMD processors.[7] Such architectures are amenable to high-throughput applications with abundant parallelism; however, given power-scaling trends for conventional design styles, subthreshold and near-threshold designs might be the only practical way to simultaneously power up all the cores on future many-core chips.

Such designs are not without a wide range of research challenges. These challenges include traditional issues such as programming models, scalable architecture design, and memory bandwidth. However, subthreshold and near-threshold design introduces additional challenges for high-performance systems, particularly with respect to the impact of design variations. As we mentioned earlier, reduced supply voltage exacerbates process, voltage, and temperature variations. Although low-performance systems might be relatively insensitive to these variations, the

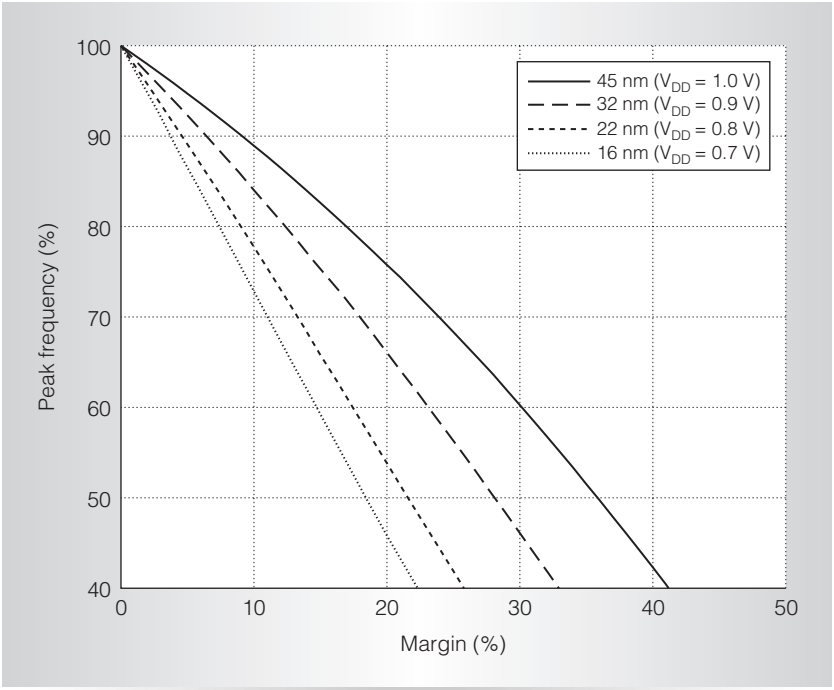

Figure 3. Reduction in peak achievable frequency as a function of the voltage margin imposed to protect against functionality problems that might result from V_{DD} droop.

impact on high-performance systems could be substantial. For example, Intel's 80-core TeraFlops processor reported an increase in the standard deviation (sigma) of critical path delay of 45 percent when moving from 1.2 V to 0.8 V.[8]

For high-performance subthreshold and near-threshold design, voltage noise might be one of the most difficult challenges to overcome. For a fixed power budget, current delivery requirements increase linearly with reduced supply voltage. Increased current draw results in higher voltage swings due to nonzero impedance of the power distribution network. The traditional approach of dealing with voltage noise has been to introduce voltage margins, effectively operating the processor at a higher V_{DD} to accommodate droops, sacrificing energy efficiency.

Figure 3 shows a relatively simple example of how larger voltage swings can significantly impact performance even for nominal design styles. The figure shows peak F_{MAX} while sweeping voltage margins across four predictive technology model (PTM)[9] technology nodes. These simulations, based on an 11-stage ring oscillator consisting of fanout-of-4 inverters, show that at today's 32-nm node, a 20 percent voltage margin translates to a 33 percent frequency degradation, and at future technology nodes the situation gets much worse.

The large amount of voltage noise present in high-performance processors will likely make margin-based design approaches impractical. However, alternative solutions might ameliorate the noise problem. Recent research suggests that voltage noise within a single processor core is highly predictable and heavily correlated with certain microarchitectural events and code paths.[10] Given that the root cause of noise events is high-level activity, current smoothing through voltage-noise-aware code scheduling could reduce the problem.[11] For highly parallel systems, voltage noise will likely be heavily correlated with core-to-core and core-to-memory activity patterns. Thus, high-level solutions at the architecture and software layers have significant potential to tackle this seemingly low-level design problem.

Highly parallel near-threshold high-performance multiprocessors will also require on-die tuning similar to what we described for the ultralow energy applications. In addition to those techniques, we need more research in methods to allow task assignment and task migration with an awareness of process variation, temperature, and so on.

Although challenges remain, subthreshold and near-threshold computing provide energy efficiency that is increasingly critical for modern integrated circuits, ranging from the most energy-constrained applications to massively multicore supercomputers. These operating regimes will inevitably become part of future products, but they will require designers at the circuit, architecture, system, and even software levels to modify traditional design practices to exploit their benefits.

References

1. S. Jocke et al., ''A 2.6-μW Subthreshold Mixed-signal ECG SoC,'' *Proc. Int'l Symp. Low Power Electronics and Design* (ISLPED 09), ACM Press, 2009, pp. 117-118.
2. N. Verma et al., ''A Micro-Power EEG Acquisition SoC With Integrated Feature Extraction Processor for a Chronic Seizure Detection System,'' *IEEE J. Solid-State Circuits,* vol. 45, no. 4, 2010, pp. 804-816.
3. G. Chen et al., ''Millimeter-Scale Nearly Perpetual Sensor System with Stacked Battery and Solar Cells,'' *Proc. IEEE Int'l Solid-State Circuits Conf.* (ISSCC 10), IEEE Press, 2010, pp. 288-289.
4. J. Wang and B.H. Calhoun, ''Techniques to Extend Canary-based Standby VDD Scaling for SRAMs to 45nm and Beyond,'' *IEEE J. Solid-State Circuits,* vol. 43, no. 11, 2008, pp. 2514-2523.
5. B.H. Calhoun and A. Chandrakasan, ''Ultra-Dynamic Voltage Scaling

(UDVS) Using Subthreshold Operation and Local Voltage Dithering,'' *IEEE J. Solid-State Circuits,* vol. 41, no. 1, 2006, pp. 238-245.
6. B.H. Calhoun et al., ''Flexible Circuits and Architectures for Ultra Low Power,'' *Proc. IEEE,* vol. 98, no. 2, 2010, pp. 267-282.
7. B. Zhai et al., ''Energy Efficient Near-Threshold Chip Multi-processing,'' *Proc. Int'l Symp. Low Power Electronics and Design* (ISLPED 07), ACM Press, 2007, pp. 32-37.
8. S. Dighe et al., ''Within-Die Variation-Aware Dynamic Voltage-Frequency Scaling, Core Mapping and Thread Hopping for an 80-Core Processor,'' *Proc. IEEE Int'l Solid-State Circuits Conf.* (ISSCC 10), IEEE Press, 2010, pp. 174-175.
9. W. Zhao and Y. Cao, "New Generation of Predictive Technology Modeling for Sub-45nm Early Design Exploration," *IEEE Trans. Electron Device,* vol. 53, no. 11, 2006, pp. 2816-2823.
10. V.J. Reddi et al., ''Voltage Emergency Prediction: A Signature-Based Approach To Reducing Voltage Emergencies,'' *Proc. Int'l Symp. High-Performance Computer Architecture* (HPCA 09), IEEE CS Press, 2009, pp. 18-27.
11. V.J. Reddi et al., ''Software-Assisted Hardware Reliability: Abstracting Circuit-level Challenges to the Software Stack,'' *Proc. 46th Design Automation Conf.* (DAC), ACM Press, 2009, pp. 788-793.

Benton H. Calhoun is an assistant professor in the Charles L. Brown Department of Electrical and Computer Engineering at the University of Virginia. His research interests include low-power digital circuit design, subthreshold digital circuits, SRAM design for end-of-the-roadmap silicon, variation-tolerant circuit design methodologies, and low-energy electronics for medical applications. Calhoun has a PhD in electrical engineering from the Massachusetts Institute of Technology. He is a member of IEEE.

David Brooks is a Gordon McKay Professor of Computer Science in the School of Engineering and Applied Sciences at Harvard University. His research interests include power-efficient computer system design, variation-tolerant computer architectures, and embedded system design. Brooks has a PhD in electrical engineering from Princeton University. He is a member of IEEE and ACM.

Direct questions or comments about this article to Benton Calhoun, 351 McCormick Road, PO Box 400743, Charlottesville, VA 22904-4743; bcalhoun@virginia.edu.