INVITED PAPER

Near-Threshold Computing: Reclaiming Moore's Law Through Energy Efficient Integrated Circuits

Future computer systems promise to achieve an energy reduction of 100 or more times with memory design, device structure, device fabrication techniques, and clocking, all optimized for low-voltage operation.

By Ronald G. Dreslinski, Michael Wieckowski, David Blaauw, *Senior Member IEEE*, Dennis Sylvester, *Senior Member IEEE*, and Trevor Mudge, *Fellow IEEE*

ABSTRACT | Power has become the primary design constraint for chip designers today. While Moore's law continues to provide additional transistors, power budgets have begun to prohibit those devices from actually being used. To reduce energy consumption, voltage scaling techniques have proved a popular technique with subthreshold design representing the endpoint of voltage scaling. Although it is extremely energy efficient, subthreshold design has been relegated to niche markets due to its major performance penalties. This paper defines and explores near-threshold computing (NTC), a design space where the supply voltage is approximately equal to the threshold voltage of the transistors. This region retains much of the energy savings of subthreshold operation with more favorable performance and variability characteristics. This makes it applicable to a broad range of power-constrained computing segments from sensors to high performance servers. This paper explores the barriers to the widespread adoption of NTC and describes current work aimed at overcoming these obstacles.

KEYWORDS | CMOS integrated circuits; computer architecture; energy conservation; parallel processing; VLSI

Manuscript received May 15, 2009; revised September 1, 2009. Current version published January 20, 2010.

The authors are with the Department of Electrical Engineering and Computer Science, University of Michigan-Ann Arbor, MI (e-mail: rdreslin@eecs.umich.edu; wieckows@umich.edu; blaauw@umich.edu; dennis@eecs.umich.edu; tnm@umich.edu).

Digital Object Identifier: 10.1109/JPROC.2009.2034764

I. INTRODUCTION

Over the past four decades, the number of transistors on a chip has increased exponentially in accordance with Moore's law [1]. This has led to progress in diversified computing applications, such as health care, education, security, and communications. A number of societal projections and industrial roadmaps are driven by the expectation that these rates of improvement will continue, but the impediments to growth are more formidable today than ever before. The largest of these barriers is related to energy and power dissipation, and it is not an exaggeration to state that developing energy-efficient solutions is critical to the survival of the semiconductor industry. Extensions of today's solutions can only go so far, and without improvements in energy efficiency, CMOS is in danger of running out of steam.

When we examine history, we readily see a pattern: generations of previous technologies, ranging from vacuum tubes to bipolar-based to NMOS-based technologies, were replaced by their successors when their energy overheads became prohibitive. However, there is no clear successor to CMOS today. The available alternatives are far from being commercially viable, and none has gained sufficient traction, or provided the economic justification for overthrowing the large investments made in the CMOS-based infrastructure. Therefore, there is a strong case supporting the position that solutions to the power conundrum must come from enhanced devices, design styles, and architectures, rather than a reliance on the

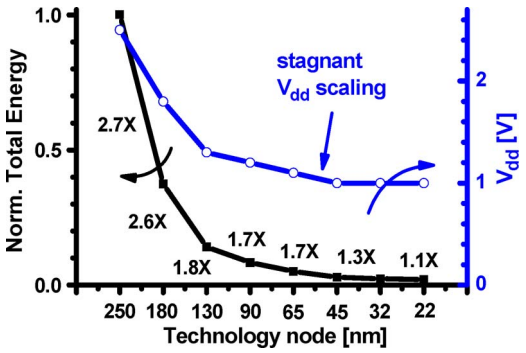

Fig. 1. *Technology scaling trends of supply voltage and energy.*

promise of radically new technologies becoming commercially viable. In our view, *the solution to this energy crisis is the universal application of aggressive low-voltage operation across all computation platforms*. This can be accomplished by targeting so-called "near-threshold operation" and by proposing novel methods to overcome the barriers that have historically relegated ultralow-voltage operation to niche markets.

CMOS-based technologies have continued to march in the direction of miniaturization per Moore's law. New silicon-based technologies such as FinFET devices [2] and 3-D integration [3] provide a path to increasing transistor counts in a given footprint. However, using Moore's law as the metric of progress has become misleading since improvements in packing densities no longer translate into proportionate increases in performance or energy efficiency. Starting around the 65 nm node, device scaling no longer delivers the energy gains that drove the semiconductor growth of the past several decades, as shown in Fig. 1. The supply voltage has remained essentially constant since then and dynamic energy efficiency improvements have stagnated, while leakage currents continue to increase. Heat removal limits at the package level have further restricted more advanced integration. Together, such factors have created a curious design dilemma: *more gates can now fit on a die, but a growing fraction cannot actually be used due to strict power limits*.

At the same time, we are moving to a "more than Moore" world, with a wider diversity of applications than the microprocessor or ASICs of ten years ago. Tomorrow's design paradigm must enable designs catering to applications that span from high-performance processors and portable wireless applications, to sensor nodes and medical implants. Energy considerations are vital over this entire spectrum, including:

- *High-performance platforms*, targeted for use in data centers, create large amounts of heat and require major investments in power and cooling infrastructure, resulting in major environmental and societal impact. In 2006 data centers consumed 1.5% of total U.S. electricity, equal to the entire U.S. transportation manufacturing industry [4], and alarmingly, data center power is projected to double every ~5 years.
- *Personal computing platforms* are becoming increasingly wireless and miniaturized, and are limited by trade-offs between battery lifetimes (days) and computational requirements (e.g., high-definition video). Wireless applications increasingly rely on digital signal processing. While Moore's law enables greater transistor density, only a fraction may be used at a time due to power limitations and application performance is therefore muzzled by power limits, often in the 500 mW–5 W range.
- *Sensor-based platforms* critically depend on ultralow power ($\leq \mu$W in standby) and reduced formfactor (mm^3). They promise to unlock new semiconductor applications, such as implanted monitoring and actuation medical devices, as well as ubiquitous environmental monitoring, e.g., structural sensing within critical infrastructure elements such as bridges.

The aim of the designer in this era is to overcome the challenge of energy efficient computing and unleash performance from the reins of power to reenable Moore's law in the semiconductor industry. Our proposed strategy is to provide 10X or higher energy efficiency improvements at constant performance through widespread application of *near-threshold computing* (NTC), where devices are operated at or near their threshold voltage (V_{th}). By reducing supply voltage from a nominal 1.1 V to 400–500 mV, NTC obtains as much as 10X energy efficiency gains and represents the reestablishment of voltage scaling and its associated energy efficiency gains.

The use of ultralow-voltage operation, and in particular subthreshold operation ($V_{dd} < V_{th}$), was first proposed over three decades ago when the theoretical lower limit of V_{dd} was found to be 36 mV [5]. However, the challenges that arise from operating in this regime have kept subthreshold operation confined to a handful of minor markets, such as wristwatches and hearing aids. To the mainstream designer, ultralow-voltage design has remained little more than a fascinating concept with no practical relevance. However, given the current energy crisis in the semiconductor industry and stagnated voltage scaling we foresee the need for a radical paradigm shift where ultralow-voltage operation is applied across application platforms and forms the basis for renewed energy efficiency.

NTC does not come without some barriers to widespread acceptance. In this paper we focus on three key challenges that have been poorly addressed to date with respect to low-voltage operation, specifically: 1) *10X or greater loss in performance*, 2) *5X increase in performance variation*, and 3) *5 orders of magnitude increase in functional failure rate of memory as well as increased logic failures*.

Overcoming these barriers is a formidable challenge requiring a synergistic approach combining methods from the algorithm and architecture levels to circuit and technology levels.

The rest of this paper is organized as follows. Section II defines the near-threshold operating region and discusses the potential benefits of operating in this region. Section III presents operating results of several processor designs and shows the relative performance/energy tradeoffs in the NTC region. Section IV details the barriers to near-threshold computing while Section V discusses techniques to address them. Section VI provides justification for NTC use in a variety of computing domains. We present future research directions in Section VII and concluding remarks in Section VIII.

II. NEAR-THRESHOLD COMPUTING (NTC)

Energy consumption in modern CMOS circuits largely results from the charging and discharging of internal node capacitances and can be reduced quadratically by lowering supply voltage (V_{dd}). As such, voltage scaling has become one of the more effective methods to reduce power consumption in commercial parts. It is well known that CMOS circuits function at very low voltages and remain functional even when V_{dd} drops below the threshold voltage (V_{th}). In 1972, Meindl *et al.* derived a theoretical lower limit on V_{dd} for functional operation, which has been approached in very simple test circuits [5], [6]. Since this time, there has been interest in subthreshold operation, initially for analog circuits [7]–[9] and more recently for digital processors [10]–[15], demonstrating operation at V_{dd} below 200 mV. However, the lower bound on V_{dd} in commercial applications is typically set to ~70% of the nominal V_{dd} due to concerns about robustness and performance loss [16]–[18].

Given such wide voltage scaling potential, it is important to determine the V_{dd} at which the energy per operation (or instruction) is optimal. In the superthreshold regime ($V_{dd} > V_{\text{th}}$), energy is highly sensitive to V_{dd} due to the quadratic scaling of switching energy with V_{dd}. Hence voltage scaling down to the near-threshold regime ($V_{dd} \sim V_{\text{th}}$) yields an energy reduction on the order of 10X at the expense of approximately 10X performance degradation, as seen in Fig. 2 [19]. However, the dependence of energy on V_{dd} becomes more complex as voltage is scaled below V_{th}. In subthreshold ($V_{dd} < V_{\text{th}}$), circuit delay increases exponentially with V_{dd}, causing leakage energy (the product of leakage current, V_{dd}, and delay) to increase in a near-exponential fashion. This rise in leakage energy eventually dominates any reduction in switching energy, creating an energy minimum seen in Fig. 2.

The identification of an energy minimum led to interest in processors that operate at this energy optimal supply voltage [13], [15], [20] (referred to as V_{min} and typically 250 mV–350 mV). However, the energy minimum is relatively shallow. Energy typically reduces by only ~2X when V_{dd} is scaled from the near-threshold regime (400–500 mV) to the subthreshold regime, though delay rises by 50–100X over the same region. While acceptable in ultralow energy sensor-based systems, this delay penalty is not tolerable for a broader set of applications. Hence, although introduced roughly 30 years ago, ultralow-voltage design remains confined to a small set of markets with little or no impact on mainstream semiconductor products.

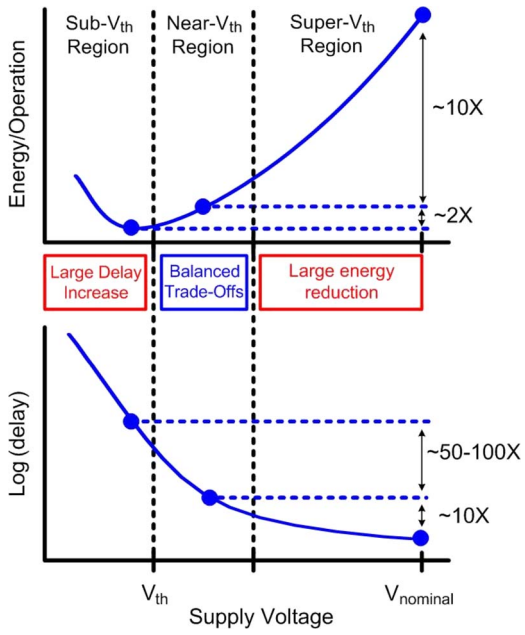

Fig. 2. ***Energy and delay in different supply voltage operating regions.***

III. NTC ANALYSIS

Recent work at many leading institutions has produced working processors that operate at subthreshold voltages. For instance, the Subliminal [20] and Phoenix processors [21] designed by Hanson *et al.* provide the opportunity to experimentally quantify the NTC region and how it compares to the subthreshold region. Figs. 3 and 4 present the energy breakdown of the two different designs as well as the clock frequency achieved across a range of voltages. As discussed in Section II, there is a V_{min} operating point that occurs in the subthreshold region where energy usage is optimized, but clock frequencies are limited to sub-1 MHz values (not pictured for Phoenix as testing was not conducted in subthreshold). On the other hand, only a modest increase in energy is seen operating at the NTC region (around 0.5 V), while frequency characteristics at that point are significantly improved. For example, at nominal voltages, the Subliminal processor runs at 20.5 MHz and 33.1 pJ/inst, while at NTC voltages, a 6.6X reduction

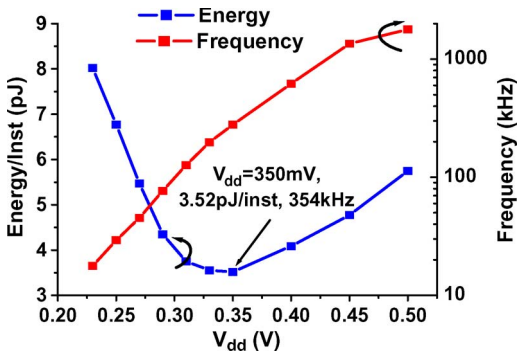

Fig. 3. *Subliminal processor frequency and energy breakdowns at various supply voltages.*

in energy and an 11.4X reduction in frequency are observed. For the Phoenix processor a nominal 9.13 MHz and 29.6 pJ/inst translate to a 9.8X reduction in energy and a 9.1X reduction in frequency. These trade-offs are much more attractive that those seen in the subthreshold design space and open up a wide variety of new applications for NTC systems.

IV. NTC BARRIERS

Although NTC provides excellent energy-frequency trade-offs, it brings its own set of complications. NTC faces three key barriers that must be overcome for widespread use; performance loss, performance variation, and functional failure. In the following subsections we discuss why each of these issues arises and why they pose problems to the widespread adoption of NTC. Section V then addresses the recent work related to each of these barriers.

A. Performance Loss

The performance loss observed in NTC, while not as severe as that in subthreshold operation, poses one of the most formidable challenges for NTC viability. In an industrial 45 nm technology the fanout-of-four inverter delay (FO4, a commonly used metric for the intrinsic speed of a semiconductor process technology) at an NTC supply of 400 mV is 10X slower than at the nominal 1.1 V. There have been several recent advances in architectural and circuit techniques that can regain some of this loss in performance. These techniques, described in detail in Section V-A, center around aggressive parallelism with a novel NTC oriented memory/computation hierarchy. The increased communication needs in these architectures is supported by the application of 3-D chip integration, as made feasible by the low power density of NTC circuits. In addition, new technology optimizations that opportunistically leverage the significantly improved silicon wearout characteristics (e.g., oxide breakdown) observed in low-voltage NTC can be used to regain a substantial portion of the lost performance.

B. Increased Performance Variation

In the near-threshold regime, the dependencies of MOSFET drive current on V_{th}, V_{dd}, and temperature approach exponential. As a result, NTC designs display a dramatic increase in performance uncertainty. Fig. 5 shows that performance variation due to global process variation alone increases by approximately 5X from ~30% (1.3X) [22] at nominal operating voltage to as much as 400%, (5X) at 400 mV. Operating at this voltage also heightens sensitivity to temperature and supply ripple, each of which can add another factor of 2X to the performance variation resulting in a total performance uncertainty of 20X. Compared to a total performance uncertainty of ~1.5X at nominal voltage, the increased performance uncertainty of NTC circuits looms as a daunting challenge that has caused most designers to pass over low-voltage design entirely. Simply adding margin so that

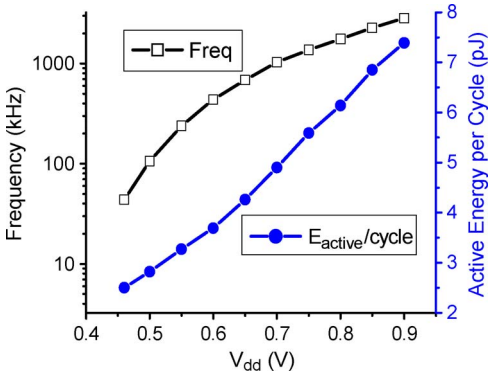

Fig. 4. *Phoenix frequency and energy breakdowns at various supply voltages.*

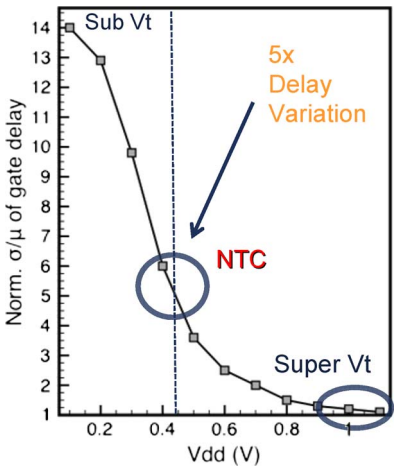

Fig. 5. *Impact of voltage scaling on gate delay variation.*

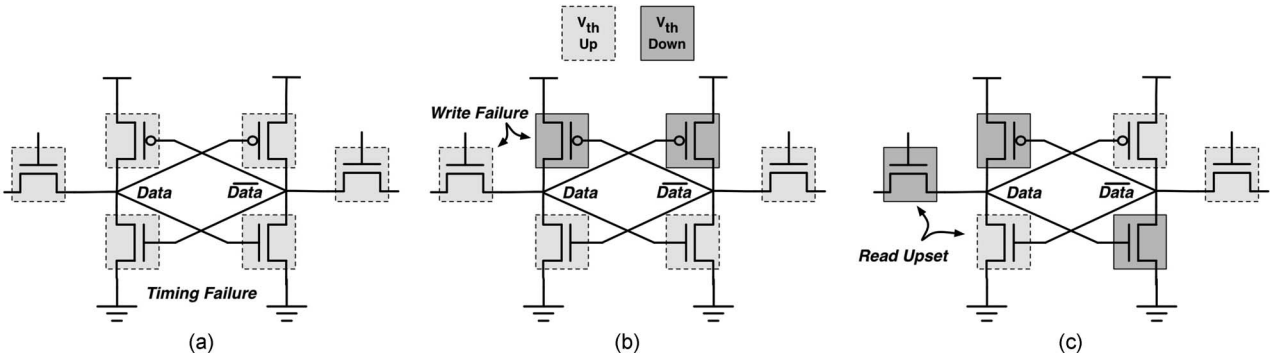

Fig. 6. *Effects of global and local variation on a standard 6 T SRAM cell. (a) Global V_{th} reduction resulting in timing failure. (b) Global V_{th} P-N skew resulting in write failure. (c) Local V_{th} mismatch resulting in read upset.*

all chips will meet the needed performance specification in the worst case is effective in nominal voltage design. In NTC design this approach results in some chips running at 1/10th their potential performance, which is wasteful both in performance and in energy due to leakage currents. Section VII presents a new architectural approach to dynamically adapting the performance of a design to the intrinsic and environmental conditions of process, voltage, and temperature that is capable of tracking over the wide performance range observed in NTC operation. This method is complemented by circuit-level techniques for diminishing the variation of NTC circuits and for efficient adaptation of performance.

C. Increased Functional Failure

The increased sensitivity of NTC circuits to variations in process, temperature and voltage not only impacts performance but also circuit functionality. In particular, the mismatch in device strength due to local process variations from such phenomena as random dopant fluctuations (RDF) and line edge roughness (LER) can compromise state holding elements based on positive feedback loops. Mismatch in the loop's elements will cause it to develop a natural inclination for one state over the other, a characteristic that can lead to hard functional failure or soft timing failure. This issue has been most pronounced in SRAM where high yield requirements and the use of aggressively sized devices result in prohibitive sensitivity to local variation.

Several variation scenarios for a standard 6 T SRAM cell are shown in Fig. 6. In (a), global process variation has resulted in both P and N devices being weakened by a V_{th} increase resulting in a potential timing failure during both reads and writes. In (b), a similar global effect has introduced skew between the P and N device strengths. This is particularly detrimental when the P is skewed stronger relative to the N resulting in a potential inability to write data into the cell. In (c), random local mismatch is considered and the worst case is shown for a read upset condition. The cell is effectively skewed to favor one state over another, and the weak pull-down on the left side cannot properly combat the strong access device at its drain. As such, the *Data* node is likely to flip to the "1" state during normal read operations. While these examples are shown in isolation, a fabricated circuit will certainly experience all of them simultaneously to varying degrees across a die and with different sensitivities to changes in supply voltage and temperature. The resulting likelihood of failure is potentially very high, especially as supply voltage is reduced and feature sizes are shrunk.

For instance, a typical 65 nm SRAM cell has a failure probability of $\sim 10^{-7}$ at nominal voltage, as shown in Fig. 7. This low failure rate allows failing cells to be corrected for using parity checks or even swapped using redundant columns after fabrication. However, at an NTC voltage of 500 mV, this failure rate increases by $\sim$5 orders of magnitude to approximately 4%. In this case, nearly every row and column will have at least one failing cell, and possibly multiple failures, rendering simple redundancy methods completely ineffective. Section V-C therefore presents novel approaches to robustness ranging from the

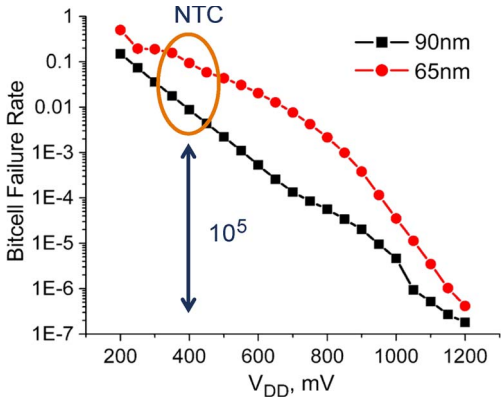

Fig. 7. *Impact of voltage scaling on SRAM failure rates.*

architectural to circuit levels that address both memory failures and functional failure of flip-flops (FFs) and latches.

V. ADDRESSING NTC BARRIERS

A. Addressing Performance Loss

To enable widespread NTC penetration into the processor application space, the ~10X performance loss must be overcome while maintaining energy efficiency. This section explores architectural and device-level methods that form a complementary approach to address this challenge.

1) Cluster-Based Architecture: To regain the performance lost in NTC without increasing supply voltage, Zhai *et al.* [23], [24] propose the use of NTC-based parallelism. In applications where there is an abundance of thread-level parallelism the intention is to use 10 s to 100 s of NTC processor cores that will regain 10–50X of the performance, while remaining energy efficient. While traditional superthreshold many-core solutions have been extensively studied, the NTC domain presents unique challenges and opportunities in these architectures. Of particular impact are the reliability of NTC memory cells and differing energy optimal voltage points for logic and memory, as discussed below.

Zhai's work showed that SRAMs, commonly used for caches, have a higher energy optimal operating voltage (V_{min}) than processors, by approximately 100 mV [23]. This stems from the relatively high leakage component of cache energy, a trade-off associated with their large size and high density. As leakage increases with respect to switching energy, it becomes more efficient to run faster, hence V_{min} is shifted higher. In addition, the value of an energy optimal operating voltage for SRAM cache is greatly effected by reliability issues in the NTC regime, where the need for larger SRAM cells or error correction methods (see Section V-C) further increases leakage. The cumulative result of these characteristics is that SRAM cache can generally run with optimal energy efficiency at a higher speed than it's surrounding logic. Hence, there is the unique opportunity in the NTC regime to exploit this effect and design architectures where multiple processors share the same first level cache.

More specifically this observation suggests an architecture with n clusters and k cores, where each cluster shares a first level cache that runs k times faster than the cores (Fig. 8). Different voltage regions are presented in different colors and use level converters at the interfaces. This architecture results in several interesting trade-offs. First, applications that share data and communicate through memory, such as certain classes of scientific computing, can avoid coherence messages to other cores in the same cluster. This reduces energy from memory coherence.

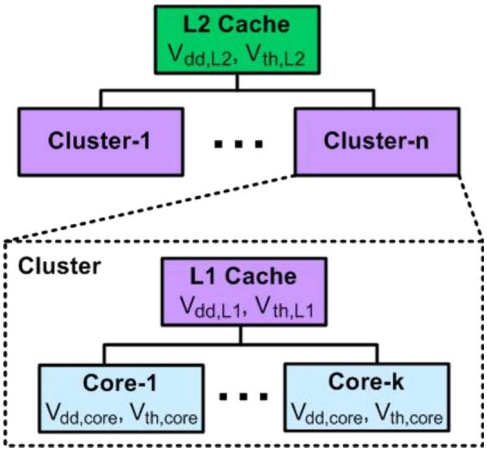

Fig. 8. *Cluster-based architecture.*

However, the cores in a cluster compete for cache space and incur more conflict misses, which may in turn increase energy use. This situation can be common in high performance applications where threads work on independent data. However, these workloads often execute the same instruction sequences, allowing opportunity for savings with a clustered instruction cache. Initial work on this architecture [21] shows that with a few processors (6–12), a 5–6X performance improvement can be achieved.

2) Device Optimization: At the lowest level of abstraction, performance of NTC systems can be greatly improved through straightforward modifications and optimizations of the transistor structure and its fabrication process. This follows directly from the fact that commercially available CMOS processes are universally tailored to sustaining the superthreshold trends forecasted by Moore's law. In most cases, this results in a transistor that is clearly suboptimal for low-voltage operation. Recently, optimizing for low voltage has generated substantial interest in the academic community because of the potential performance gains that could be obtained by developing a process flow tailored for subthreshold operation. In large part, these gains would be comparable for NTC operation since the devices in question still operate without a strongly inverted channel. For example, Paul *et al.* [25] demonstrate a 44% improvement in subthreshold delay through simple modifications of the channel doping profile of a standard superthreshold device. Essentially, the nominal device is doped with an emphasis on reducing short channel effects at standard supply voltage such as DIBL, punchthrough, and V_{th} roll-off. These effects are much less significant when the supply is lowered below about 70% of the nominal. This allows device designers to instead focus on a doping profile that minimizes junction capacitance and subthreshold swing without negatively impacting the device off current.

Entirely new device structures based on fully depleted silicon-on-insulator (FDSOI) technologies are also being considered as candidates for enabling subthreshold applications [26]. The naturally higher subthreshold slope in FDSOI along with its reduced parasitic capacitances make it an attractive option for enhancing performance with little power penalty. Further modifications to the established bulk process methodology, such as using an undoped body with a metal gate structure and removal of the source-drain extensions, serves to improve speed while maintaining standard threshold voltage targets. When these devices are combined using thin-metal interconnect for low-capacitance, the energy-delay product in the subthreshold can be comparable to low power designs operating in the super-threshold. This level of performance makes tailored FDSOI devices highly desirable for NTC design and offers a viable solution for mainstream applications as the process matures.

With similar goals in mind, Hanson *et al.* [27] showed that the slow scaling of gate oxide relative to the channel length yields a 60% reduction in I_{on}/I_{off} between the 90 nm and 32 nm nodes. This on to off current ratio is a critical measure of stability and noise immunity, and such a reduction results in static noise margin (SNM) degradation of more than 10% between the 90 nm and 32 nm nodes in a CMOS inverter. As a solution, they have proposed a modified scaling strategy that uses increased channel lengths and reduced doping to improve subthreshold swing. They developed new delay and energy metrics that effectively capture the important effects of device scaling, and used those to drive device optimization. Based on technology computer-aided design (TCAD) simulations they found that noise margins improved by 19% and energy improved by 23% in 32 nm subthreshold circuits when applying their modified device scaling strategy. Their proposed strategy also led to tighter control of subthreshold swing and off-current, reducing delay by 18% per generation. This reduction in delay could be used in addition to the parallelism discussed in Section V-A1 to regain the performance loss of NTC, returning it to the levels of traditional core performance.

B. Addressing Performance Variation

As noted in Section IV-B, the combined impact of intrinsic process variations and extrinsic variations, such as fluctuations in temperature and supply voltage, results in a spread in the statistical distribution of NTC circuit performance of ~10X compared to designs at nominal supplies. Traditional methods to cope with this issue, which are largely centered on adding design margin, are inadequate and hugely wasteful when voltage is scaled, resulting in a substantial portion of the energy efficiency gain from NTC operation being lost. Hence, in this section architectural and circuit solutions to provide variation tolerance and adaptivity are discussed.

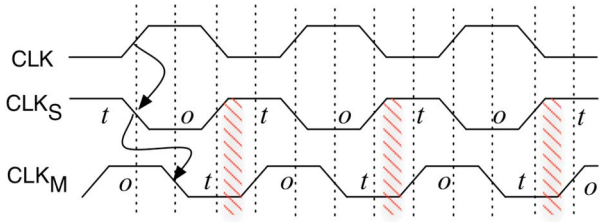

Fig. 9. ***Delaying the master clock creates a window of transparency.***

1) Soft Edge Clocking: The device variation inherent to semiconductor manufacturing continues to increase from such causes as dopant fluctuation and other random sources, limiting the performance and yield of ASIC designs. Traditionally, variation tolerant, two-phase, latch-based designs have been used as a solution to this issue. Alternatively, hard-edge data flip-flops (DFF) with intentional or "useful" skew can be used. Both of these techniques incur a significant penalty in design complexity and clocking overhead.

One potential solution to address timing variation while minimizing overhead is a type of soft-edge flip-flop (SFF) that maintains synchronization at a clock edge, but has a small transparency window, or "softness." In one particular approach to soft-edge clocking, tunable inverters are used in a master–slave flip-flop to delay the incoming master clock edge with respect to the slave edge as shown in Fig. 9.

As a result of this delay, a small window of transparency is generated in the edge-triggered register that accommodates paths in the preceding logic that were too slow for the nominal cycle time—in essence allowing time borrowing within an edge-triggered register environment. Hence, soft edge clocking results in a trade-off between short and long paths and is effective at mitigating random, uncorrelated variations in delay, which are significant in NTC. In theoretical explorations at a nominal super-threshold supply voltage, it was shown that soft-edge clocking reduced the mean (standard deviation) clock period in benchmark circuits by up to 22% (25%). Joshi *et al.* [28] furthered this work by developing a library based on these soft flip-flops and providing a statistical algorithm for their assignment. In the work by Wieckowski *et al.* [29], this technique was employed in silicon to show that small amounts of softness in a FIR filter achieved improvements in performance of 11.7% over a standard DFF design and improvement of 9.2% compared to a DFF with useful skew. These increases in performance, shown in Fig. 10, demonstrate a greater tolerance to intradie variation that becomes even more important in the NTC operating region.

2) Body Biasing: At superthreshold supply voltages, body biasing (BB) is a well known technique for adapting

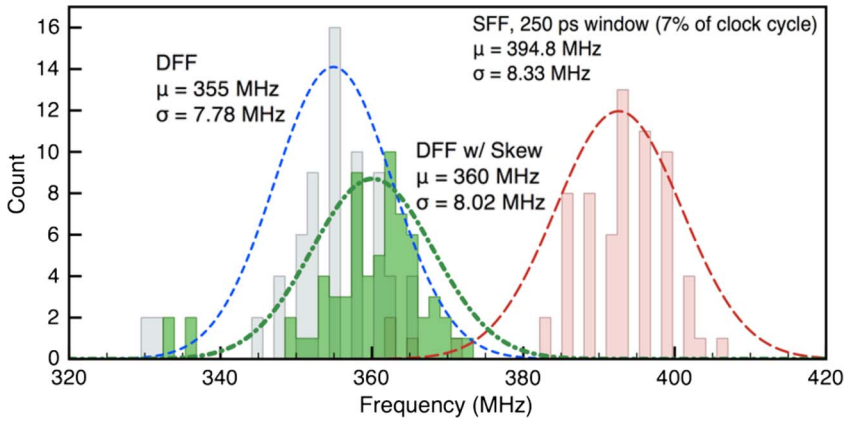

Fig. 10. *FIR filter with soft edge clocking compared to standard flip-flops (SFF); presented with and without useful skew.*

performance and leakage to global variation of process, voltage, and temperature. It's use is becoming more widespread and was recently demonstrated in silicon within a communication processor application [30]. While effective in the superthreshold domain, the influence of body-biasing becomes particularly effectual in the NTC domain where device sensitivity to the threshold voltage increases exponentially. Body-biasing is therefore a strong lever for modulating the frequency and performance in NTC, and is ideally suited as a technique for addressing the increased detriments of process variation in NTC. Further, because P and N regions can be adapted separately using body biasing, and because the relative drive strength of P and N transistors can change dramatically from superthreshold to NTC, body biasing has the added advantage of allowing the P to N ratio of a design to be optimally adjusted.

Hanson *et al.* [6] show that the extreme sensitivity to process variation in NTC design tends to raise $V_{\min}$ and reduce energy efficiency. They explore the use of adaptive body-bias (ABB) techniques to compensate for this variation both locally and globally. Indeed, their later work on a subthreshold processor [20] implements these techniques in silicon and demonstrates their effectiveness. They further showed that the body bias voltages that tune the P to N ratio for optimal noise margin also minimizes energy. Hence, one tuning can be used to both increase robustness of the design as well as to reduce its energy consumption. They found that skewing P and N body biases in increments of 5 mV to match strengths enabled them to improve the minimum functional voltage by 24%. For global performance they improved the variability for several target voltages, as seen in Fig. 11. This directly demonstrates the effectiveness of ABB in dealing with variation, especially in low-voltage designs, and is a technique that can be directly leveraged in NTC systems to cope with these same issues.

C. Addressing Functional Failure

The variations discussed in Section IV-C not only impact design performance but also design functionality. In NTC the dramatically increased sensitivity to process, temperature and voltage variations lead to a precipitous rise in functional failure (the likelihood that a data bit will be flipped), particularly due to drive strength mismatch. In this section, architectural and circuit-level techniques for addressing SRAM robustness in NTC operation are discussed.

1) Alternative SRAM Cells: As mentioned previously, SRAM cells require special attention when considering cache optimization for the NTC design space. Even though it is clear that SRAM will generally exhibit a higher $V_{\min}$ than logic, it will still operate at supply level significant lower than the nominal case. This in turn reduces cell stability and heightens sensitivity to V_{th} variation, which is generally high in SRAM devices to begin with due to the particularly aggressive device sizing necessary for high density. This problem is fundamental to the standard 6 T

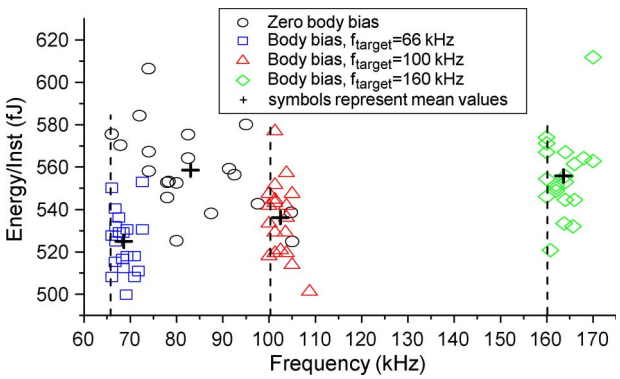

Fig. 11. *Body biasing techniques for three target frequencies.*

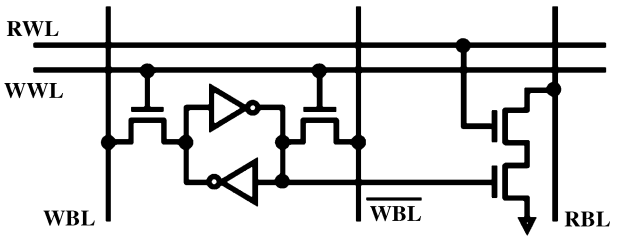

Fig. 12. *Alternative 8 T SRAM cell, decoupling the read and write [32].*

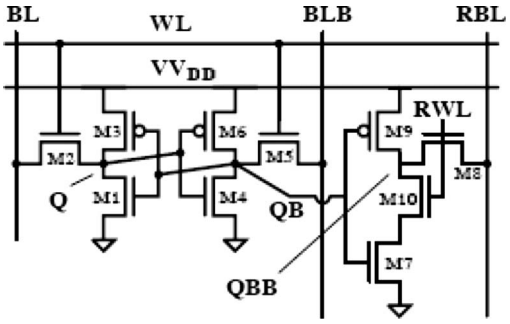

Fig. 13. *Alternative 10 T SRAM cell [33].*

design, which is based on a carefully balancing act of relative device sizing to optimize read/write contention. The only solution to keeping SRAM viable for NTC applications is to trade-off area for improved low-voltage performance. The question then becomes how best to do this—resize and optimize the 6 T devices, or abandon the 6 T structure completely?

One example in which the basic 6 T structure was maintained can be seen in the work by Zhai *et al.* [31]. The cell itself is optimized for single-ended read stability, and a supply modulation technique is used on a per column basis to improve writeability. Thus, the read and write operations are effectively decoupled by relying on extra complexity in the periphery of the core array. The result is a cell that is functional below 200 mV and that achieves relatively high energy efficiency.

There have also been a number of alternative SRAM cells proposed that are particularly well suited for ultralow-voltage operation. For example, Chang *et al.* [32] developed an 8 T design, in Fig. 12, with the premise of decoupling the read and write operations of the 6 T cell by adding an isolated read-out buffer, as shown in Fig. 6. This effectively allows the designer to optimize the write operation sizing independently of the output buffer and without relying on supply modulation or wordline boosting. This greatly enhances cell stability, but incurs area overhead in the core array to accommodate the extra devices and irregular layout.

Similarly, Calhoun and Chandrakasan [33] developed a 10-transistor (10 T) SRAM cell also based on decoupling read and write sizing and operation. The 10 T cell is pictured in Fig. 13 and offers even better low-voltage operation due to the stacking of devices in the read port, though it suffers a commensurate area penalty. Such alternative SRAM cell designs successfully cope with the difficulty of maintaining proper operation at high yield constraints in the subthreshold operating region, and offer promising characteristics for realizing reliable cache in NTC-based systems.

2) SRAM Robustness Analysis Techniques: The importance of robustness for NTC systems means that credible analyses techniques must be available. This is particularly true for the case of large level-2 and level-3 (L2 and L3) caches where low bitcell failure rates are required to achieve high yield. Inaccurate estimates of robustness in such cases would lead to wasted die space, in the case of oversized cells, or large portions of unusable memory, when they are undersized. Chen *et al.* [34] have developed a technique to determine proper cell sizing to maintain the same SRAM cell robustness at NTC voltages as traditional cells have at nominal. The technique they developed uses importance sampling as a means to determine the cell device sizes needed for a given robustness to variation. Using importance sampling for yield estimation, Chen compared a 6 T, single-ended 6 T with power rail drooping and an 8 T bitcell at an iso-robustness and iso-delay condition. This condition requires that both cells be designed to tolerate the same level of process variation before functional failure while operating with the same nominal delay. The results for a 20 cycle latency in terms of SRAM bitcell area and energy consumption are presented in Figs. 14 and 15. At higher V_{dd}, the differential 6 T bitcell has the smallest area. The 8 T bitcell becomes smaller below a V_{dd} of 450 mV and a twenty-cycle latency. As V_{dd} approaches V_{th}, all bitcells must be sized greatly to maintain robustness unless delay is relaxed, making large arrays impractical. The differential 6 T bitcell has the

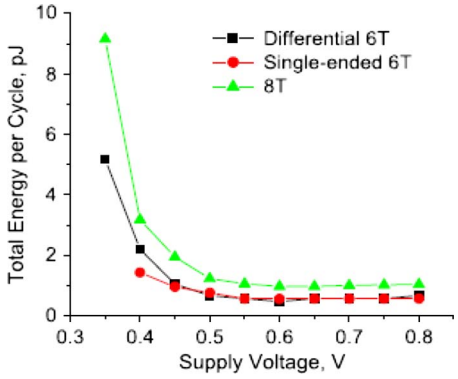

Fig. 14. *Energy of SRAM topologies for 20-cycle L2 cache across voltages.*

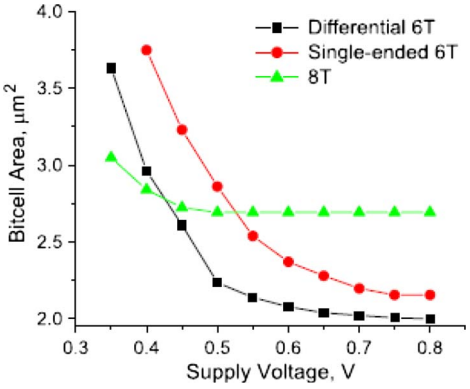

Fig. 15. *Size of SRAM topologies for 20-cycle L2 cache across voltages with iso-robustness.*

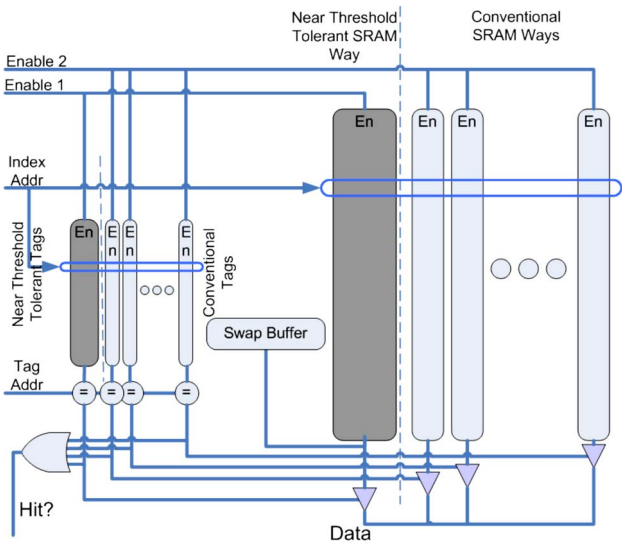

Fig. 16. *Alternative L1 cache design with one cache way NTC enabled.*

lowest dynamic energy consumption at most supply voltages. The single-ended 6 T bitcell has the lowest leakage per cycle. $V_{\min}$ increases with cache size and bank size, and decreases with associativity, activity factor, and cache line length. For common cache configurations, $V_{\min}$ may be near or even above V_{th} and is significantly higher than reported in previously literature. By comparing SRAM bitcells at an iso-robustness and iso-delay condition, the best SRAM architecture and sizing for a design can be quickly and accurately chosen. This work will be valuable in assessing the viability of new SRAM designs, particularly in the NTC domain.

3) Reconfigurable Cache Designs: For designs that do not require on-chip L2 or L3 caches, such as mobile embedded applications or sensor processors, implementing an energy efficient L1 is important. On the architectural design front, recent work by Dreslinski *et al.* [35] addresses cache robustness for small L1 caches. The work is focused on single core systems with moderate amounts of cache requirements. In these situations, converting the entire cache to larger cells to maintain robustness would limit the total cache space by effectively cutting it in half. To maintain the excellent energy efficiency of the NTC SRAM, but with minimal impact on die space a cache where only a subset of the cache ways are implemented in larger NTC tolerant cells is proposed. This modified cache structure, shown in Fig. 16, can dynamically reconfigure access semantics to act like a traditional cache if needed for performance and act like a filter cache to balance energy in low power mode. When performance is not critical, power can be reduced by accessing the low-voltage cache way first, with the other ways of the cache only accessed on a miss. This technique is similar to that of filter caches, and while providing power savings it does increase access time for hits in the higher-NTC cache way voltages. When performance is critical, the access methodology is changed to access all ways of the cache in parallel to provide a fast single cycle access to all data. The work resulted in a system where in low power mode (10 MHz) energy savings of greater than 70% were seen for typical embedded workloads with less than a 5% increase in runtime while operating in high performance mode (400 MHz).

VI. NTC COMPUTING SEGMENTS

A. NTC Integration in Ultra Energy-Efficient Servers

The exponential growth of the web has yielded a dramatic increase in the demand for server style computers with the installed base of servers expected to exceed 40 million by 2010 [36]. Server growth is accompanied by an equally rapid growth in the energy demand to power them. For example, it is estimated that the five largest internet sites consume at least 5 MW each [37].

The tier 1 of a data center that serves web pages provides a perfect opportunity for NTC. The requests in these servers represent the bulk of requests to these data centers [38], consuming 75% of the overall energy. The workload is a stream of independent requests to render web pages that can be naturally executed in parallel. HTML is fetched from memory, subjected to relatively simple operations, and returned to memory without requiring extensive shared data. To achieve this, *10–100 s of NTC cores on a single die* can be used to obtain very high throughput with unprecedented energy efficiency.

B. NTC Integration in Personal Computing

The personal computing platform continues to evolve rapidly. WiFi is a standard on laptops but other mobile wireless communications are also starting to be supported. Future devices must be able to move seamlessly

among communication alternatives. Such systems will combine a high level of processing power along with signal processing capabilities integrated into a much smaller form factor than today's laptops. Battery life is expected on the order of days, while functionality requirements are extreme and may include high-definition video, voice recognition, along with a range of wireless standards. The features of PC platforms that distinguish them from the two other systems are the dual needs to cope with variable workloads and energy efficient wireless communication.

In the PC platform space, cores may run at widely varying performance/energy points. The voltage and frequency of the cores and their supporting peripherals can be dynamically altered in real time to meet the constraints of performance and power consumption. This dynamic voltage and frequency scaling technique (DVFS) can be leveraged to enable adaptive NTC circuits in the personal computing space [39]. The scaling method may be driven by operating system commands and/or distributed sensors. Exploiting phase variations in workloads [40], efficient phase detection techniques need to be established for multicore multithreaded processors to enable power management schemes in achieving savings without significantly compromising performance.

C. NTC Integration in Sensor Networks

With advances in circuit and sensor design, pervasive sensor-based systems, from single to thousands of nodes, are quickly becoming a possibility. A single sensor node typically consists of a data processing and storage unit, off-chip communication, sensing elements, and a power source. They are often wirelessly networked and have potential applications in a wide range of industrial domains, from building automation to homeland security to biomedical implants. The versatility of a sensor is directly linked to its form factor—for a sensor to be truly useful in many new application areas, a form factor on the order of 1 mm^3 is desirable while maintaining a lifetime of months or years.

To meet the above requirements, the key limiting constraint is energy. Both sensors and electronics are readily shrunk to $< 1\ mm^3$ in modern technologies. However, current processors and communication systems require batteries that are many orders of magnitude larger than the electronics themselves (e.g., 50 mm^3 processor die in a laptop vs. 167 cm^3 4-cell lithium-ion battery). Hence, whether a sensor node is powered through batteries, harvesting, or both, power consumption will limit overall system size. To integrate a sensor node in $< 1\ mm^3$, energy levels must be reduced by 4–7 orders of magnitude. Processing speed is not a major constraint in most sensor applications [41], easing the integration of NTC. Initial investigations showed simple sensor architectures coupled with NTC can obtain an active energy reduction of $\geq$ 100X [15].

VII. FUTURE DIRECTIONS

In addition to the techniques discussed above, significant momentum has developed in the area of adaptivity: processors and mixed-signal circuits that dynamically adjust to meet the constraints imposed by process variation, changing environments, and aging. Often this has been achieved using so-called "canary" circuits. These circuits employ specialized structures that predict the delay failure of a pipeline using a critical path replica, ring oscillator, or canary flip-flop [42]–[44]. The fundamental idea is to design the replica circuit such that it will fail *before* the critical path elements in the pipeline, thus providing an indicator that retuning to the current operating condition is required. While these implementations are relatively noninvasive, the replica circuits themselves can suffer from mistracking under temperature and voltage variations and are unable to assess the impact of local process/voltage/temperature (PVT) variations on the actual critical paths, particularly in NTC where variation between paths is greatly amplified.

A second category of adaptive designs has been based upon directly monitoring the variation-constrained logic using *in situ* circuitry [45], [46]. The Razor approach is one recent example [47]–[49]. A novel flip-flop structure is used to detect and correct for timing errors dynamically. This allows reduction of timing margins via dynamic voltage scaling (DVS) to meet an acceptable error correction rate. While effective, the Razor technique suffers from three difficulties. First, the flip-flop structure introduces two-sided timing constraints due to the large Razor flip-flop hold times. This adds significant complexity to the design cycle of Razor systems and incurs a power overhead due to the buffer insertion required for its mitigation. Second, due to the large process variations in NTC, a larger speculation window is needed in NTC. However, this increases the hold time constraint and overhead, making Razor less suitable for NTC operation. Third, the hardware required for correcting an error is complex and highly specialized for a given application. The system must be able to roll back the pipeline to a state before the errors were detected. This reduces the portability of the Razor approach and hinders the development of a Razor framework for general-purpose applications.

It is clear that current approaches are either highly invasive, such as the error detection and correction methods or, as in the simple canary-type predictor circuits, still require substantial margins at design time and do not fully exploit the potential gains provided by true run-time adaptation of frequency and voltage. We propose to resolve this by using *in situ* delay monitoring combined with worst case vector recognition and control. A basic vision of the proposed system is presented in Fig. 17. On the left, a simple pipeline constrained by delay variation is shown. The basic idea is to directly sample the transition edges or glitches of each stage of logic using ultrawide transition

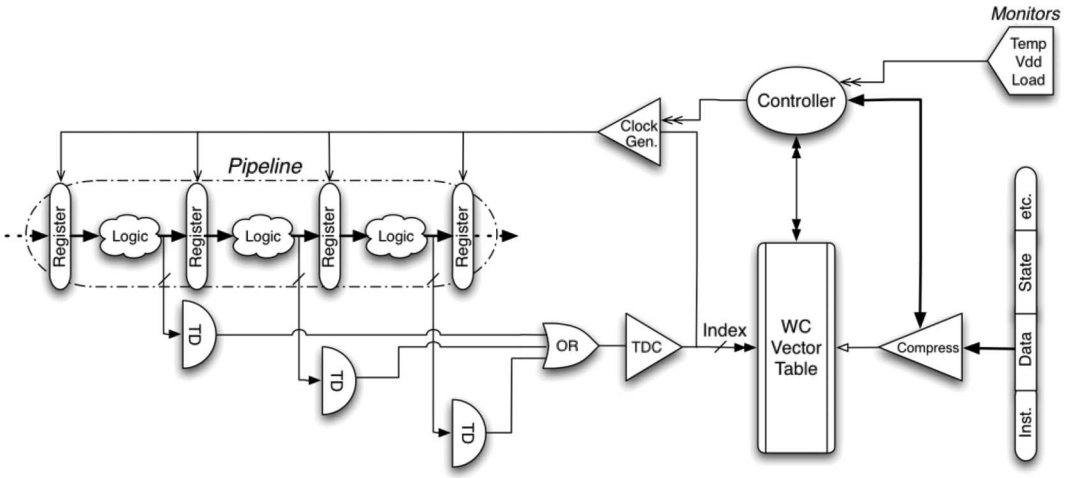

Fig. 17. ***In situ delay monitoring DVFS system.***

detectors (TDs). Each detector provides a measure of the distance in time between the most recent logic transition and the clock edge. The output of these detectors is combined and converted into a digital representation using a time-to-digital converter (TDC) for use by the adaptive control system.

At the heart of the adaptive control system is a worst case vector table to keep track of the pipeline vectors that result in worst case delays in the critical logic paths. This table is initially populated after fabrication by executing an extensive postsilicon qualification test over different voltage and temperature conditions and detecting and recording those vectors that result in the critical delays. This process is performed once and the results are stored in a table of worst case vectors for each possible voltage/temperature condition. This provides the system an optimized starting point that compensates for global process variation. During normal execution of the processor, temperature and voltage will change over time, forming environmental epochs of operation. Monitoring and control of the circuit delay and test vectors will be completely transparent to the operation of the processor. On-chip sensors will be used to detect and signal the start of each new epoch, at which time the controller will exercise the corresponding worst case vectors in the pipeline and the optimal clock period will in turn be generated. Alternative to frequency tuning, the voltage can instead by tuned while keeping the frequency constant. Such a system will be able to achieve near-optimal energy efficiency over a wide range of operating conditions.

VIII. CONCLUSION

As Moore's law continues to provide designers with more transistors on a chip, power budgets are beginning to limit the applicability of these additional transistors in conventional CMOS design. In this paper we looked back at the feasibility of voltage scaling to reduce energy consumption. Although subthreshold operation is well known to provide substantial energy savings it has been relegated to a handful of applications due to the corresponding system performance degradation. We then turned to the concept of near-threshold computing (NTC), where the supply voltage is at or near the switching voltage of the transistors. This regime enables energy savings on the order of 10X, with only a 10X degradation in performance, providing a much better energy/performance trade-off than subthreshold operation. The rest of the paper focused on the three major barriers to widespread adoption of NTC and current research to overcome them. The three barriers addressed were: 1) performance loss; 2) increased variation; and 3) increased functional failure. With traditional device scaling no longer providing energy efficiency improvements, our primary conclusion is that the solution to this energy crisis is the universal application of aggressive low-voltage operation, namely NTC, across all computation platforms. ■

REFERENCES

[1] G. Moore, "No exponential is forever: But 'forever' can be delayed!" in *Proc. IEEE Int. Solid-State Circuits Conf.*, 2003, Keynote address.

[2] X. Huang, W.-C. Lee, C. Kuo, D. Hisamoto, L. Chang, J. Kedzierski, E. Anderson, H. Takeuchi, Y.-K. Choi, K. Asano, V. Subramanian, T.-J. King, J. Bokor, and C. Hu, "Sub 50-nm p-channel FinFET," *IEEE Trans. Electron Devices*, pp. 880–886, May 2001.

[3] A. W. Topol, J. D. C. La Tulipe, L. Shi, D. J. Frank, K. Bernstein, S. E. Steen, A. Kumar, G. U. Singco, A. M. Young, K. W. Guarini, and M. Ieong, "Three-dimensional integrated circuits," *IBM J. Res. Develop.*, vol. 50, no. 4/5, pp. 491–506, Jul./Sep. 2006.

[4] "Report to congress on server and data center energy efficiency," U.S. Environmental Protection Agency. [Online]. Available: http://www.energystar.gov/ia/partners/prod_development/downloads/EPA_Datacenter_Report_Congress_Final1.pdf

[5] R. Swanson and J. Meindl, "Ion-implanted complementary MOS transistors in low-voltage circuits," *IEEE J. Solid-State Circuits*, vol. 7, no. 2, pp. 146–153, 1972.

[6] S. Hanson, B. Zhai, K. Bernstein, D. Blaauw, A. Bryant, L. Chang, K. Das, W. Haensch, E. Nowak, and D. Sylvester, "Ultra low-voltage, minimum energy CMOS," *IBM J. Res. Develop.*, pp. 469–490, Jul./Sep. 2006.

[7] E. Vittoz and J. Fellrath, "CMOS analog integrated circuits based on weak inversion operations," *IEEE J. Solid-State Circuits*, vol. 12, no. 3, pp. 224–231, 1977.

[8] R. Lyon and C. Mead, "An analog electronic cochlea," *Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 7, pp. 1119–1134, 1988.

[9] C. Mead, *Analog VLSI and Neural Systems*. Boston, MA: Addison-Wesley, 1989.

[10] H. Soeleman and K. Roy, "Ultra-low power digital subthreshold logic circuits," in *Proc. ACM/IEEE Int. Symp. Low Power Electronics Design*, 1999, pp. 94–96.

[11] B. Paul, H. Soeleman, and K. Roy, "An 8 × 8 sub-threshold digital CMOS carry save array multiplier," in *Proc. IEEE Eur. Solid-State Circuits Conf.*, 2001.

[12] C. Kim, H. Soeleman, and K. Roy, "Ultra-low-power DLMS adaptive filter for hearing aid applications," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 11, no. 6, pp. 1058–1067, 2003.

[13] A. Wang and A. Chandrakasan, "A 180 mV FFT processor using subthreshold circuit techniques," in *Proc. IEEE Int. Solid-State Circuits Conf.*, 2004, pp. 292–529.

[14] B. Calhoun and A. Chandrakasan, "A 256 kb sub-threshold SRAM in 65 nm CMOS," in *Proc. IEEE Int. Solid-State Circuits Conf.*, 2006, pp. 628–629.

[15] B. Zhai, L. Nazhandali, J. Olson, A. Reeves, M. Minuth, R. Helfand, S. Pant, D. Blaauw, and T. Austin, "A 2.60 pJ/Inst subthreshold sensor processor for optimal energy efficiency," in *IEEE Symp. VLSI Circuits*, 2006, pp. 154–155.

[16] *Transmeta Crusoe*. [Online]. Available: http://www.transmeta.com/

[17] *Intel XScale*. [Online]. Available: http://www.intel.com/design/intelxscale/

[18] *IBM PowerPC*. [Online]. Available: http://www.chips.ibm.com/products/powerpc/

[19] B. Zhai, D. Blaauw, D. Sylvester, and K. Flautner, "Theoretical and practical limits of dynamic voltage scaling," in *Proc. Design Automation Conf.*, Jan. 1, 2004, pp. 868–873.

[20] S. Hanson, B. Zhai, M. Seok, B. Cline, K. Zhou, M. Singhal, M. Minuth, J. Olson, L. Nazhandali, T. Austin, D. Sylvester, and D. Blaauw, "Performance and variability optimization strategies in a sub-200 mV, 3.5 pJ/inst, 11 nW subthreshold processor," in *Symp. VLSI Circuits*, 2007, pp. 152–153.

[21] M. Seok, S. Hanson, Y. Lin, Z. Foo, D. Kim, Y. Lee, N. Liu, D. Sylvester, and D. Blaauw, "The phoenix processor: A 30 pW platform for sensor applications," in *IEEE Symp. VLSI Circuits*, 2008, pp. 188–189.

[22] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De, "Parameter variations and impact on circuits and microarchitecture," in *Proc. ACM/IEEE Design Automation Conf.*, 2003, pp. 338–343.

[23] B. Zhai, R. Dreslinski, T. Mudge, D. Blaauw, and D. Sylvester, "Energy efficient near-threshold chip multi-processing," in *Proc. ACM/IEEE Int. Symp. Low-Power Electronics Design*, 2007, pp. 32–37.

[24] R. Dreslinski, B. Zhai, T. Mudge, D. Blaauw, and D. Sylvester, "An energy efficient parallel architecture using near threshold operation," in *Parallel Architectures and Compilation Techniques (PACT)*, Sep. 2007.

[25] B. Paul, A. Raychowdhury, and K. Roy, "Device optimization for ultra-low power digital sub-threshold operation," in *Proc. Int. Symp. Low Power Electronics and Design*, 2004, pp. 96–101.

[26] N. Checka, J. Kedzierski, and C. Keast, "A subthreshold-optimized FDSOI technology for ultra low power applications," in *Proc. GOMAC*, 2008.

[27] S. Hanson, M. Seok, D. Sylvester, and D. Blaauw, "Nanometer device scaling in subthreshold circuits," in *Proc. Design Automation Conf.*, 2007, pp. 700–705.

[28] M. Wieckowski, Y. Park, C. Tokunaga, D. Kim, Z. Food, D. Sylvester, and D. Blaauw, "Timing yield enhancement through soft edge flip-flop based design," in *Proc. IEEE Custom Integrated Circuts Conf. (CICC)*, Sep. 2008.

[29] V. Joshi, D. Blaauw, and D. Sylvester, "Soft-edge flip-flops for improved timing yield: Design and optimization," in *Proc. Int. Conf. Comput.-Aided Design*, 2007, pp. 667–673.

[30] G. Gammie, A. Wang, M. Chau, S. Gururajarao, R. Pitts, F. Jumel, S. Engel, P. Royannez, R. Lagerquist, H. Mair, J. Vaccani, G. Baldwin, K. Heragu, R. Mandal, M. Clinton, D. Arden, and K. Uming, "A 45 nm 3.5 G baseband-and-multimedia application processor using adaptive body-bias and ultra-low-power techniques," in *Proc. Int. Solid-State Circuits Conf.*, 2008.

[31] B. Zhai, D. Blaauw, D. Sylvester, and S. Hanson, "A sub-200 mV 6 T SRAM in 130 nm CMOS," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, Feb. 2007.

[32] L. Chang, Y. Nakamura, R. K. Montoye, J. Sawada, A. K. Martin, K. Kinoshita, F. H. Gebara, K. B. Agarwal, D. J. Acharyya, W. Haensch, K. Hosokawa, and D. Jamsek, "A 5.3 GHz 8 T-SRAM with operation down to 0.41 V in 65 nm CMOS," in *IEEE Symp. VLSI Circuits*, 2007, pp. 252–253.

[33] B. Calhoun and A. Chandrakasan, "A 256 kb Sub-threshold SRAM in 65 nm CMOS," in *Int. Solid-State Circuits Conf.*, 2006, pp. 259–260.

[34] G. K. Chen, D. Blaauw, T. Mudge, D. Sylvester, and N. S. Kim, "Yield-driven near-threshold SRAM design," in *Int. Conf. Comput.-Aided Design*, 2007, pp. 660–666.

[35] R. Dreslinski, G. Chen, T. Mudge, D. Blaauw, D. Sylvester, and K. Flautner, "Reconfigurable energy efficient near threshold cache architectures," in *Proc. 41st Annu. MICRO*, 2008.

[36] *IDC's Worldwide Installed Base Forecast, 2007–2010*. Framingham, MA: IDC, Mar. 2007. IDC.

[37] R. Katz, "Research directions in internet-scale computing," in *Proc. 3rd Int. Week Management of Networks and Services*, 2007, Keynote presentation.

[38] [Online]. Available: http://www.spec.org/web2005

[39] T. Pering, T. Burd, and R. Brodersen, "The simulation and evaluation of dynamic voltage scaling algorithms," in *Proc. ACM/IEEE Int. Symp. Low Power Electronics Design*, 1998, pp. 76–81.

[40] L. Bircher and L. John, "Power phases in a commercial server workload," in *Proc. Int. Symp. Low Power Electronics and Design*, 2006.

[41] L. Nazhandali, M. Minuth, B. Zhai, J. Olson, T. Austin, and D. Blaauw, "A second-generation sensor network processor with application-driven memory optimizations and out-of-order execution," in *ACM/IEEE Int. Conf. Compilers, Archit., Synthesis Embedded Syst.*, 2005.

[42] M. Elgebaly and M. Sachdev, "Efficient adaptive voltage scaling system through on-chip critical path emulation," in *Proc. Int. Symp. Low Power Electronics and Design*, 2004, pp. 375–380.

[43] A. Raychowdhury, S. Ghosh, and K. Roy, "A novel on-chip delay measurement hardware for efficient speed-binning," in *Proc. Int. On-Line Testing Symp.*, 2005, pp. 287–292.

[44] B. H. Calhoun and A. P. Chandrakasan, "Standby power reduction using dynamic voltage scaling and canary flip-flop structures," *IEEE J. Solid-State Circuits*, vol. 39, pp. 1504–1511, 2004.

[45] T. Kehl, "Hardware self-tuning and circuit performance monitoring," in *Proc. IEEE Int. Conf. Computer Design*, 1993, pp. 188–192.

[46] T. Sato and Y. Kunitake, "A simple flip-flop circuit for typical-case designs for DFM," in *Proc. Int. Symp. Quality Electronic Design*, 2007, pp. 539–544.

[47] T. Austin, V. Bertacco, D. Blaauw, and T. Mudge, "Opportunities and challenges for better than worst-case design," in *Proc. Asia South Pacific Design Automation Conf.*, 2005, pp. 2–7.

[48] T. Austin, D. Blaauw, T. Mudge, and K. Flautner, "Making typical silicon matter with Razor," *IEEE Comput.*, vol. 37, pp. 57–65, 2004.

[49] S. Das, D. Roberts, S. Lee, S. Pant, D. Blaauw, T. Austin, K. Flautner, and T. Mudge, "A self-tuning DVS processor using delay-error detection and correction," *IEEE J. Solid-State Circuits*, vol. 41, no. 4, pp. 792–804, Apr. 2006.

ABOUT THE AUTHORS

Ronald G. Dreslinski received the B.S.E. degree in electrical engineering, the B.S.E. degree in computer engineering, and the M.S.E. degree in computer science from the University of Michigan, Ann Arbor. He is currently working toward the Ph.D. degree at the University of Michigan.

Mr. Dreslinski is a member of the ACM. His research focuses on architectures that enable emerging low-power circuit techniques.

Michael Wieckowski received the Ph.D. degree in electrical and computer engineering from the University of Rochester, NY, in 2007.

He is currently a Postdoctoral Research Fellow at the University of Michigan, Ann Arbor. His work is focused on low-power mixed-signal design to enable energy constrained computing platforms. His recent research interests include variation tolerant low-voltage memory, inductorless power management systems, and dynamically tuned low-voltage pipelines.

David Blaauw (Senior Member, IEEE) received the B.S. degree in physics and computer science from Duke University, Durham, NC, in 1986, and the Ph.D. degree in computer science from the University of Illinois, Urbana, in 1991.

Until August 2001, he worked for Motorola, Inc., Austin, TX, were 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, Ann Arbor, where he is a Professor. His work has focused on VLSI design with particular emphasis on ultralow power and high-performance design.

Prof. Blaauw was the Technical Program Chair and General Chair for the International Symposium on Low Power Electronic and Design. He was also the Technical Program Co-Chair of the ACM/IEEE Design Automation Conference and a member of the ISSCC Technical Program Committee.

Dennis Sylvester (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the University of California, Berkeley, where his dissertation was recognized with the David J. Sakrison Memorial Prize as the most outstanding research in the UC-Berkeley EECS department.

He is an Associate Professor of Electrical Engineering and Computer Science at the University of Michigan, Ann Arbor. He previously held research staff positions in the Advanced Technology Group of Synopsys, Mountain View, CA, Hewlett-Packard Laboratories in Palo Alto, CA, and a visiting professorship in Electrical and Computer Engineering at the National University of Singapore. He has published over 250 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 interconnect modeling. He also serves as a consultant and technical advisory board member for electronic design automation and semiconductor firms in these areas.

Dr. Sylvester received an NSF CAREER award, the Beatrice Winner Award at ISSCC, an IBM Faculty Award, an SRC Inventor Recognition Award, and numerous best paper awards and nominations. He is the recipient of the ACM SIGDA Outstanding New Faculty Award and the University of Michigan Henry Russel Award for distinguished scholarship. He has served on the technical program committee of major design automation and circuit design conferences, the executive committee of the ACM/IEEE Design Automation Conference, and the steering committee of the ACM/IEEE International Symposium on Physical Design. He is currently an Associate Editor for IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS and previously served as Associate Editor for IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS. He is a member of ACM and Eta Kappa Nu.

Trevor Mudge (Fellow, IEEE) received the B.Sc. degree from the University of Reading, England, in 1969, and the M.S. and Ph.D. degrees in Computer Science from the University of Illinois, Urbana, in 1973 and 1977, respectively.

Since 1977, he has been on the faculty of the University of Michigan, Ann Arbor. He recently was named the first Bredt Family Professor of Electrical Engineering and Computer Science after concluding a ten year term as the Director of the Advanced Computer Architecture Laboratory—a group of eight faculty and about 70 graduate students. He is author of numerous papers on computer architecture, programming languages, VLSI design, and computer vision. He has also chaired about 40 theses in these areas. His research interests include computer architecture, computer-aided design, and compilers. In addition to his position as a faculty member, he runs Idiot Savants, a chip design consultancy.

Prof. Mudge is a Fellow of the IEEE, a member of the ACM, the IET, and the British Computer Society.