Near Threshold Computing: Overcoming Performance Degradation from Aggressive Voltage Scaling

Ronald G. Dreslinski, Michael Wieckowski, David Blaauw, Dennis Sylvester, and Trevor Mudge
Department of Electrical Engineering and Computer Science
University of Michigan - Ann Arbor, MI

Abstract

Power has become the primary design constraint for chip designers today. While Moore's law continues to provide additional transistors, power budgets are beginning to prohibit those devices from actually being turned on. To reduce energy consumption, voltage scaling techniques have proved a popular technique with subthreshold design representing the endpoint of voltage scaling. However, while extremely energy efficient, subthreshold design has been relegated to niche markets due to its major performance penalties. In this paper we explore Near Threshold Computing (NTC), a design space where the supply voltage is approximately set 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. In this paper we briefly discuss several barriers to the wide spread adoption of near threshold computing and focus, in detail, on techniques to overcome the performance barrier of aggressive voltage scaling.

1. Intro:

Over the past four decades, the number of transistors on a chip has increased exponentially in accordance with Moore's law [1]. This has lead 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 to NMOS-based technologies, were replaced by their successors when their energy overheads were 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 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 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 ultra-low voltage operation to niche markets.

CMOS-based technologies have continued to march in the direction of miniaturization as per Moore's law. New silicon-based technologies such as FinFET devices [2] and 3D 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 65nm node, device scaling no longer delivers the energy gains that drove the semiconductor growth of the past several decades, as shown in Figure 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 larger portion 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 application spanning 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, personal computing platforms, and sensor based platforms.

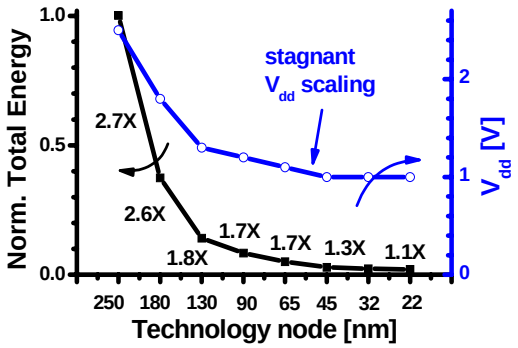

Figure 1: Technology scaling trends of supply voltage and energy.

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 recapture Moore’s law in the semiconductor industry. The 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 nominal 1.1V to 400-500mV, NTC obtains as much as 10X energy efficiency gains and represents the re-establishment of voltage scaling and its associated energy efficiency gains.

The use of ultra-low 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 36mV[4]. However, the challenges that arise from operating in this regime have kept subthreshold operation confined to a handful of niche markets, such as wristwatches and hearing aids. To the mainstream designer, ultra-low 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 ultra-low voltage operation is applied ubiquitously across application platforms and forms the basis for renewed energy efficiency.

However, NTC does not come without some barriers to widespread acceptance. Three of the key challenges that have been poorly addressed to date are: 1) *10X or greater loss in performance*; 2) *5X increase in performance variation*, and 3) *5 orders of magnitude increase in functional failure 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. In this paper we will focus on the first barrier, performance degradation.

2. 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 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 [4,5]. Since this time, there has been interest in subthreshold operation, initially for analog circuits [6,7,8] and more recently for digital processors [9,10,11,12,13,14], demonstrating operation at V_{dd} below 200mV. However, the lower bound on V_{dd} in commercial applications is usually reduced to no lower than ~70% of the nominal V_{dd} due to concerns about performance loss and robustness [15,16].

Given such wide voltage scaling potential, it is important to determine the V_{dd} at which the energy per instruction is optimal. In the superthreshold regime ($V_{dd} > V_{th}$), 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 ($V_{dd} \sim V_{th}$) yields an energy reduction on the order of 10X at the expense of approximately 10X performance degradation, as seen in Figure 2. However, the dependence of energy on V_{dd} becomes more complex as voltage is scaled below V_{th}. In subthreshold ($V_{dd} < V_{th}$), 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 Figure 2.

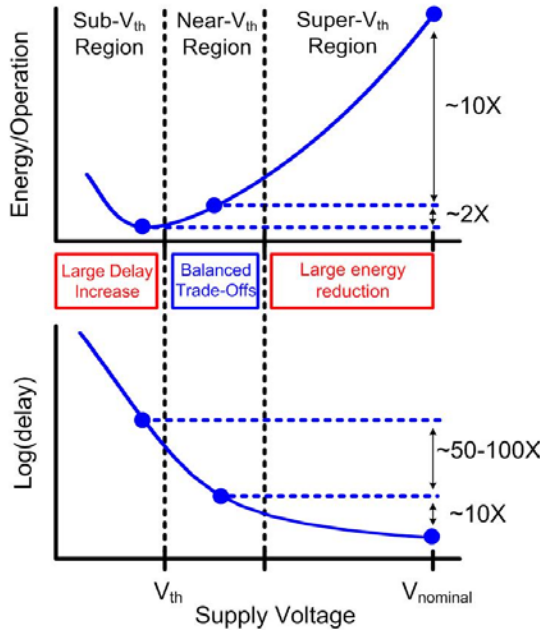

Figure 2: Energy and delay in different supply voltage operating regions.

The identification of an energy minimum has led to interest in processors that operate at this energy optimal supply voltage [12,14,17] (referred to as V_{min} and typically 250mV-350mV). However, the energy minimum is relatively shallow. Energy typically reduces by only ~2X when V_{dd} is scaled from the near-threshold regime (400-500mV) to the subthreshold regime, though delay rises by 50-100X over the same region. While acceptable in ultra-low energy sensor-based systems, this delay penalty is not tolerable in a broad set of applications. Hence, although introduced roughly 30 years ago, ultra-low voltage design remains confined to a small set of markets with little or no impact on mainstream semiconductor products.

3. NTC Analysis:

Recent work at many leading institutions has produced working processors that operate at subthreshold voltages. For instance, the Subliminal processor [17] designed by Hanson et al. provides the opportunity to clearly quantify the NTC region and how it compares to the subthreshold region. Figure 3 presents the energy breakdown of the design as well as the operating frequency achieved across a range of voltages. As was discussed in Section 2, there is a V_{min} operating point that occurs in the subthreshold operating region but is tied to operating points of less than 1MHz. On the other hand, only a modest increase in energy is seen operating at the NTC region (around .5V), while frequency characteristics at that point are significantly better. At nominal operating points Subliminal operates at 20.5 MHz and 33.1 pJ/inst, showing approximately a 6.6x reduction in energy and an 11.4x reduction in frequency at the NTC operating point.

4. NTC Barriers:

Although NTC provides for excellent energy-frequency tradeoffs, it doesn't come without its own set of complications. NTC faces three key barriers that must be overcome for widespread use, performance loss, performance variation, and even functional failure. In the following subsections we will discuss why each of these exist and why they pose problems to the wide spread adoption of NTC.

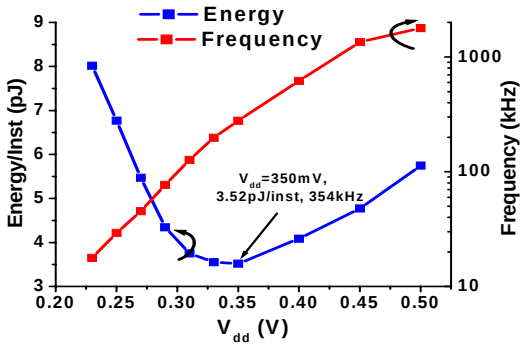

Figure 3: Subliminal processor frequency and energy breakdowns at various supply voltages.

4.1. 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 45nm technology the fanout-of-four inverter (FO4) delay at 400mV is 10X slower than at the nominal 1.1V. There have been several recent advances of architectural and circuit techniques that can be used to improve performance in the NTC regime. These techniques, described in detail in Section 5.1, 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 3D chip integration, as made feasible by the low power density of NTC circuits. In addition a new

technology optimization that opportunistically leverages 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.

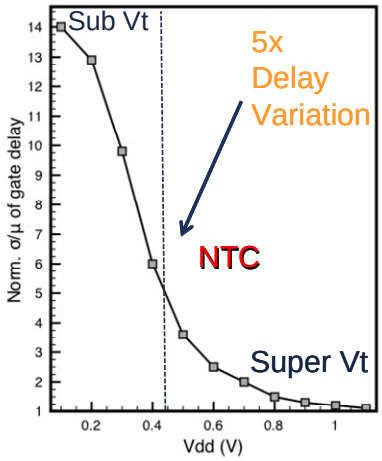

Figure 4: Impact of voltage scaling on gate delay variation.

4.2. 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. Figure 4 shows that performance variation due to process variation alone increases by approximately 5X from ~30% (1.3X) [18] at nominal operating voltage to as much as 150%, (2.5X) at 400mV. When combined with approximately 2X performance variation due to supply voltage ripple and 2X variation due to temperature, a total performance uncertainty of 10X emerges. Given 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 all chips will meet the needed performance specification in the worst-case is effective in nominal voltage design. However, in NTC design this approach results in some chips running at 1/10th their potential performance, which is wasteful both in performance and energy due to leakage currents. Several techniques exist to help mitigate these problems including Adaptive Body Biasing [19], soft edge clocking [20].

4.3. Increased functional failure. The increased sensitivity to process, temperature and voltage variations not only impacts circuit performance but also circuit functionality. In particular the mismatch in device strength due to process variations such as random dopant fluctuations (RDF) can compromise state elements as the feedback loop develops a natural inclination for one state over the other. This issue has been most pronounced in SRAM where high yield requirements and the use of minimum sized devices limit variability tolerance. For instance, a typical 45nm SRAM cell has a failure probability of ~10^{-7}at nominal voltage, see Figure 5. This low failure rate allows failing cells to be readily swapped using spare columns after fabrication. However, at an NTC voltage of 500mV, this failure rate increases by ~5 orders of magnitude to approximately 4% ($4x10^{-2}$). 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. There are many alternative SRAM approaches to help address this variability [21,22,23], new failure rate estimation techniques [24], and alternative cache designs [25]. All these techniques are used to help overcome these failures.

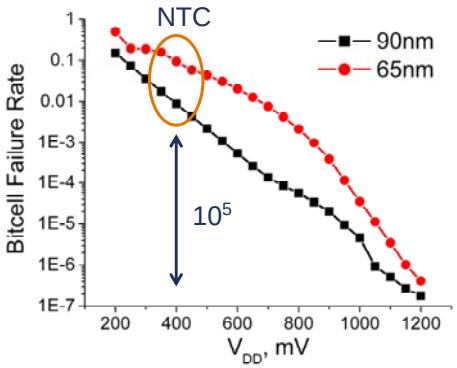

Figure 5: Impact of voltage scaling on SRAM failure rates.

5. Addressing Performance Loss:

One of the most formidable challenges to widespread NTC penetration is overcoming the ~10x performance loss associated with NTC operation while maintaining energy efficiency. Below, we explore architectural and device level methods that form a complementary approach to address this challenge.

5.1 Cluster Based Architecture

In order to regain the performance lost in NTC without increasing supply voltage, Zhai et. al [26,27] propose the use of NTC based parallelism. In applications where there is an abundance of thread-

level parallelism the intention is to use 10s to 100s of NTC processor cores that will regain 10-50X the performance, while remaining energy efficient. While traditional superthreshold many-core solutions have been 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

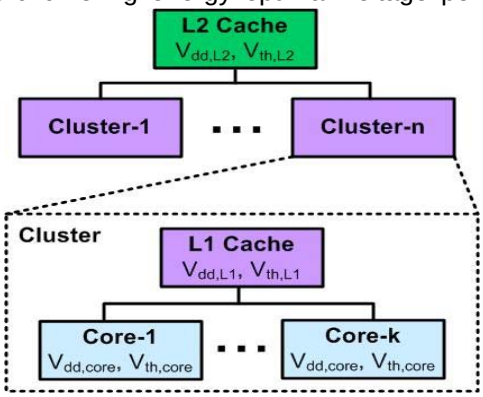

Figure 4:Cluster-based architecture.

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 100mV [26]. This results from the lower activity in caches, which amplifies leakage effects. As discussed, SRAM designs also face reliability issues in the NTC regime, leading to a need for larger SRAM cells or error correction methods (see Section 5.3), further increasing leakage and the energy optimal operating voltage. Due to this higher V_{min}, SRAMs remain energy efficient at higher supply voltages, and thus at higher speeds, compared to logic. Hence, there is the unique opportunity in the NTC regime to run caches faster than processors for energy efficiency, which naturally leads to architectures where multiple processors share the same first level cache.

It follows to suggest an architecture with *n* clusters and k cores, where each cluster shares a first level cache that runs *k* times faster than the cores (Figure 6). Different voltage regions are presented in different colors and use level converters at the interfaces. This architecture results in several interesting tradeoffs. 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. 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 research of this architecture shows that with a few processors (6-12), a gain of 5-6X performance improvement can be achieved.

5.2 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 super-threshold trends forecasted by Moore's law. In most cases, this results in a transistor that is decidedly sub-optimal for low voltage operation. Recently, this has generated substantial interest in the academic community because of the potential performance gains that could be gleaned by developing a process flow tailored for sub-threshold 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, Raychowdhury, and Roy [28] demonstrate that a 44% improvement in sub-threshold delay can be realized through simple modifications of the channel doping profile of a standard super-threshold device. Essentially, the nominal device is doped with an emphasis on reducing short channel effects at standard supply voltage such as DIBL, punch-through, and V_{th} roll-off. These effects are much less significant when the supply is lowered below about 70% of the nominal. This allows one to instead focus their efforts on a doping profile that minimizes junction capacitance and sub-threshold swing without negatively impacting the device off current.

Similarly, Hanson et al. [29] showed that the slow scaling of gate oxide relative to the channel length has lead to a 60% reduction in I_{on}/I_{off} between the 90nm and 32nm nodes. This on to off ratio is a critical measure of stability and noise immunity, and such a reduction results in SNM degradation of more than 10% between the 90nm and 32nm 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 sub-threshold swing. They developed new delay and energy metrics that effectively capture the important effects of device scaling, and used those to drive device optimization. They found that noise margins improved by 19% and

energy improved by 23% in 32nm subthreshold circuits when applying their modified device scaling strategy. Their proposed strategy also used tight control of sub-threshold swing and off-current to reduce delay by 18% per generation. This reduction in delay could be used in addition to the parallelism discussed in Section 5.1.1 to regain the performance loss of NTC, returning it to the levels of traditional core performance.

8. 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 has shown in the past to provide vast amounts of energy savings it has been relegated to a handful of applications due to the performance degradation of the system. We then turn to the Near Threshold Computing (NTC) region of operation, where the supply voltage is at or near the switching voltage of the transistors. In this region we show that energy savings on the order of 10x can be achieved, with only a 10x degradation in performance. This degradation is much less than the 500x of subthreshold operation, providing a region with excellent tradeoffs in terms of energy savings and performance.

Bibliography

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

[2] X. Huang *et al*, "Sub 50-nm p-channel FinFET" *IEEE Transactions on Electron Devices*, pp. 880-886, May 2001.

[3] A.W. Topol *et al*, "Three-dimensional integrated circuits," *IBM Journal of Research and Development*, v. 50, no. 4/5, pp. 491-506, July/September 2006.

[4] R. Swanson, J. Meindl, "Ion-implanted complementary MOS transistors in low-voltage circuits," JSSC, Vol. 7, No. 2, pp. 146-153, 1972.

[5] 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 Journal of Research and Development*, pp. 469-490, July/September 2006.

[6] E. Vittoz, J. Fellrath, "CMOS analog integrated circuits based on weak inversion operations," *IEEE Journal of Solid-State Circuits*, Vol. 12, No. 3, pp. 224-231,1977.

[7] R. Lyon, C. Mead, "An analog electronic cochlea," *Transactions on Acoustics, Speech, and Signal Processing*," Vol. 36, No. 7, pp. 1119-1134, 1988.

[8] C. Mead, *Analog VLSI and neural systems*, Addison-Wesley, Boston, 1989.

[9] H. Soeleman, K. Roy, "Ultra-low power digital subthreshold logic circuits," *ACM/IEEE International Symposium on Low Power Electronics Design*, pp. 94-96, 1999.

[10] B. Paul, H. Soeleman, K. Roy, "An 8x8sub-threshold digital CMOS carry save array multiplier," *IEEE European Solid-State Circuits Conference*, 2001.

[11] C. Kim, H. Soeleman, K. Roy, "Ultra-low-power DLMS adaptive filter for hearing aid applications," *IEEE Transaction on VLSI Systems*, Vol. 11, No. 6, pp. 1058-1067, 2003.

[12] A. Wang and A. Chandrakasan, "A 180mV FFT processor using subthreshold circuit techniques," *IEEE International Solid-State Circuits Conference*, pp. 292-529, 2004.

[13] B. Calhoun and A. Chandrakasan, "A 256kb Sub-threshold SRAM in 65nm CMOS," *IEEE International Solid-State Circuits Conference*, pp. 628-629, 2006.

[14] B. Zhai, L. Nazhandali, J. Olson, A. Reeves, Michael Minuth, R. Helfand, S. Pant, D. Blaauw, T. Austin, "A 2.60pJ/Inst subthreshold sensor processor for optimal energy efficiency," *IEEE Symposium on VLSI Circuits*, pp. 154-155, 2006.

[15] Transmeta Crusoe. http://www.transmeta.com/.

[16] Intel XScale. http://www.intel.com/design/intelxscale/.

[17] S. Hanson *et al*., "Performance and variability optimization strategies in a sub-200mV, 3.5pJ/inst, 11nW subthreshold processor," *Symposium on VLSI Circuits*, 2007.

[18] S. Borkar *et al*., "Parameter Variations and Impact on Circuits and Microarchitecture," *ACM/IEEE Design Automation Conference*, pp. 338-343, 2003.

[19] S. Hanson, B. Zhai, K. Bernstein, D. Blaauw, A. Bryant, L. Chang, K. Das, W. Haensch, E. Nowak, D. Sylvester, "Ultra-Low Voltage Minimum Energy CMOS," *IBM Journal of Research and Development, Vol. 50, No. 4/5, July/September 2006, pg. 469-490.*

[20] M. Wieckowski, Y. Park, C. Tokunaga, D. Kim, Z. Food, D. Sylvester, D. Blaauw, "Timing Yield Enhancement Through Soft Edge Flip-Flop Based Design," *IEEE Custom Integrated Circuts Conference (CICC), September 2008.*

[21] L. Chang et al., "A 5.3 GHz 8T-SRAM with Operation Down to 0.41 V in 65nm CMOS," *IEEE Symposium on VLSI Circuits*, pp. 252-253, 2007.

[22] B. Calhoun and A. Chandrakasan, "A 256kb Sub-threshold SRAM in 65nm CMOS," *International Solid-State Circuits Conference*, pp. 2592-2601,2006.

[23] B. Zhai, D. Blaauw, D. Sylvester, S. Hanson, "A sub-200mV 6T SRAM in 130nm CMOS," *IEEE International Solid-State Circuits Conference (ISSCC), February 2007*

[24] G.K. Chen et al., "Yield-driven near-threshold SRAM design," *International conference on Computer aided design*, 2007, pp. 660-666.

[25] R. Dreslinski, G. Chen, T. Mudge, D. Blaauw, D. Sylvester, K. Flautner. "Reconfigurable Energy Efficient Near Threshold Cache Architectures," Proceedings of the 41st annual MICRO, 2008**.**

[26] B. Zhai, R. Dreslinski, T. Mudge, D. Blaauw, and D. Sylvester, "Energy efficient near-threshold chip multi-processing," *ACM/IEEE International Symposium on Low-Power Electronics Design*, pp. 32-37, 2007.

[27] R. Dreslinski, B., T. Mudge, D. Blaauw, D. Sylvester, "An Energy Efficient Parallel Architecture Using Near Threshold Operation," *Parallel Architectures and Compilation Techniques (PACT), September 2007*

[28] B. Paul, A. Raychowdhury, K. Roy, "Device optimization for ultra-low power digital sub-threshold operation, "*International Symposium on Low Power Electronics and Design*, pp. 96-101, 2004.

[29] S. Hanson, M. Seok, D. Sylvester, D. Blaauw, "Nanometer device scaling in subthreshold circuits, "*Design Automation Conference*, pp. 700-705, 2007.