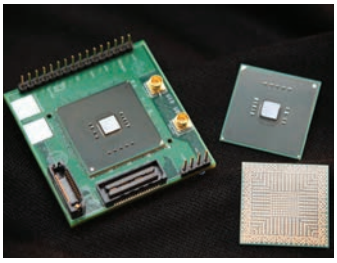
# Radu Teodorescu

## Designing Low-Power Microprocessors in the Era of Unpredictable Transistors

For many decades we have taken for granted the fact that computer performance has been doubling every one or two years. This extraordinary growth created an industry that has impacted almost every aspect of our lives–from the way we work to the way we play, and how we communicate or provide healthcare. This revolution was enabled in no small part by one of the computer's core technologies: the microprocessor. Over the last fifty years, microprocessors have benefited tremendously from technology innovations that have delivered more and faster transistors with every new generation. Unfortunately, that technology has reached an impasse in recent years, as transistors have approached low-nanometer dimensions. Transistors are so small in the latest technology that about 6 million of them would fit in the period at the end of this sentence. These transistors are less predictable, less reliable and their energy efficiency is increasing very slowly. Building chips with these minute transistors is likely the most significant manufacturing challenge humans have ever undertaken.



*Intel NTC chip. (source: Intel Corp.)*

These technological challenges are happening at a time when the need for energy efficient computing is greater than ever. The recent explosive growth of ultra-portable computing devices like smartphones and tablets is reshaping the consumer computing landscape. Expectations of battery life for these devices are now measured in days rather than hours. At the same time, these ultra-portable devices are rapidly becoming gaming consoles and productivity platforms with high performance demands. Meeting these performance requirements, while keeping power consumption under control, requires dramatic improvements in the energy efficiency of computation. Power consumption is now the main roadblock facing one of our fastest-growing industries.

The work conducted at Ohio State in the Computer Architecture Research Lab, led by Assistant Professor Radu Teodorescu is taking on the challenge of designing faster and more energy efficient microprocessors under the most adverse technological challenges our industry has ever faced. Overcoming these challenges will enable a new class of energy-conscious microprocessors that deliver the performance of supercomputers in mobile form factors, and enable environmentally-responsible growth in computing.

The principal approach used in this undertaking is a new computing paradigm generally referred to as "Near-Threshold Computing" (NTC). This technique relies on lowering the supply voltage (Vdd) of a chip to a level only slightly higher than the threshold voltage (Vth)–the level at which transistors begin conducting current. Vdd is the most powerful lever for improving energy efficiency because it impacts both dynamic and static power super-linearly. Even though NTC significantly reduces chip speed, it allows for many more computation units (cores) to be powered on simultaneously for the same power cost. Multi-threaded workloads that can take advantage of the increased parallelism can run much more efficiently at NTC. Experimental data shows these applications can attain 8x to 10x higher energy efficiency at NTC compared to conventional super-threshold computing (STC). A recent prototype of a low-voltage chip from Intel Corp. is showing very promising results.

Unfortunately, Near-Threshold Computing faces multiple challenges before it can become a mainstream technology. This is because NTC is less reliable than conventional technology, requiring additional protection against failures. NTC also amplifies the effects of process (post-manufacturing) and runtime variability.

Process variability is caused by the relative imprecision in the manufacturing process of chips with very large numbers of minute transistors. This imprecision makes transistor dimensions and properties somewhat uncertain and with a high degree of variability–not unlike cells in the human body. This variability affects crucial transistor parameters such as threshold voltage and leads to heterogeneity in transistor speed and power consumption. A microprocessor, which is generally homogeneous by design, will instead behave more like a heterogeneous system with cores that have different speeds and power consumption profiles. This heterogeneity is very large at NTC, with design-identical cores having 2-3x variation in top speed (*Figure 1*). This very large and unpredictable heterogeneity can be very detrimental to the performance

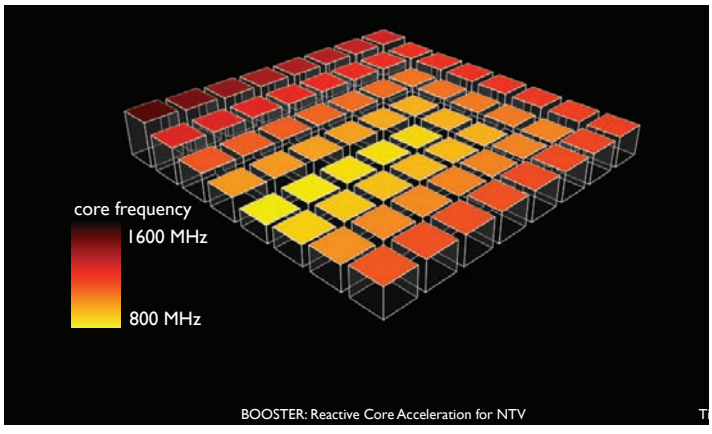



*Figure 1: Variation in core frequency in a simulated 64-core NTC chip.*

and energy efficiency of synchronized parallel applications such as those used in scientific computing.

Professor Teodorescu's group at OSU has developed mechanisms for addressing variation-induced heterogeneity in near-threshold chips. One such solution relies on a microprocessor design that provides two power supply lines to each core set at two very low but different voltages. Each core in the CMP can be dynamically assigned to either of the two power rails using a gating circuit. This allows each core to periodically switch between two different maximum frequencies, on a predetermined schedule. The schedule is different for each core and is chosen so that core frequencies average to the same value over a finite interval. This means that cores that are inherently slow are scheduled to spend more time on the high voltage rail while those that are fast will spend more time on the low voltage rail. The result is a CMP that achieves performance homogeneity from an underlying heterogeneous fabric.

Low-voltage operation slows down transistors and makes them more likely to behave unpredictably which can lead to computation or memory retention errors. Large memory blocks such as those used in large on-chip caches are especially vulnerable. They are optimized for density and therefore built using the smallest transistors, which are the most sensitive to low-voltage operation.

Professor Teodorescu's group has developed a new error correction technique designed for near-threshold caches. The basic idea is to apply a simple error correction code repeatedly, in multiple permutations, to a data block to achieve an exponential increase in error correction capability. The proposed solution trades off correction strength for decoding time. While this iterative decoding process takes longer than traditional schemes, it is a pay-as-you-go approach (depending on the severity of the errors and the correctness requirements). This means applying more decoding iterations to lines that suffer larger numbers of errors and fewer decoding iterations to the others, while keeping the storage overhead low. This technique allows near-threshold chips to operate reliably in the presence of high error rates.

Another significant challenge of NTV operation is the increased sensitivity to voltage fluctuations. These fluctuations are caused by abrupt changes in power demand triggered by processor activity variation with workload. If the voltage deviates too much from its nominal value, it can lead to so-called "voltage emergencies," which can cause timing and memory retention errors. *Figure 2* illustrates voltage variability across a 4-core chip at some time instance T. The heat map represents percentage drop in supply voltage relative to nominal values. Voltage variability correlates with chip activity. Sections of the chip that are inactive show no voltage drop (purple areas on the map), while areas of intense activity exhibit significant voltage drops (red and yellow areas).

Professor Teodorescu's group is developing

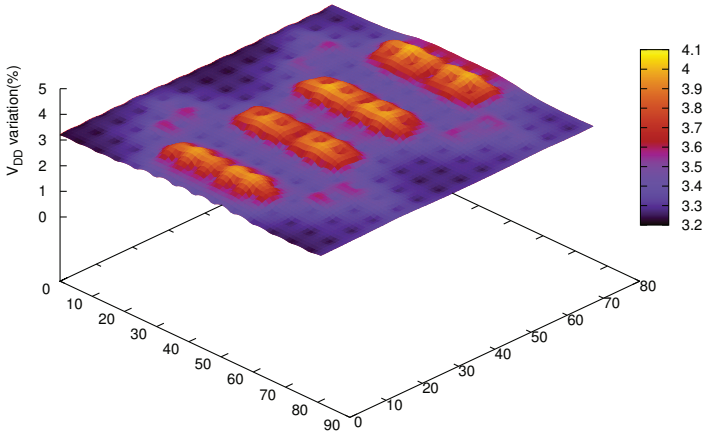



*Figure 2: Voltage variability in a 4-core chip multiprocessor.*

hardware/software co-design solutions that enlist software support to address these hardware vulnerabilities. For example, by redesigning synchronization libraries to include knowledge about voltage variability effects, the synchronization primitives can smooth-out power spikes and eliminate voltage emergencies.

Near-Threshold Computing has the potential to sustain the performance and energy efficiency growth of computing systems for another decade or more. Overcoming its significant challenges however requires a new level of innovation that spans multiple layers of the computing stack including circuits, microarchitecture, runtime systems and software. Professor Teodorescu's team brings together experts in each of these areas to work on making NTC a reality.

The applied nature of this research benefits tremendously from close collaboration with industry partners interested in NTC such as Intel and Mentor Graphics. As part of a GOALI (Grant Opportunities for Academic Liaison with Industry) project funded by the National Science Foundation Professor Teodorescu's team is working with Mentor Graphics researchers to design and enhance the capabilities of modeling and CAD tools to accurately support NTC development.

Professor Teodorescu's research is contributing to and is supported in part by the Defense Advanced Research Projects Agency (DARPA) through the Power Efficiency Revolution for Embedded Computing Technologies (PERFECT) program, an ambitious five and a half year project that seeks to build an embedded computing system for military applications with a computation efficiency of 75 gigaflops/watt. By comparison, a typical system today achieves a computation efficiency of about 1 gigaflop/watt, almost two orders of magnitude lower. The driving technology behind this embedded system is Near-Threshold Computing. Professor Teodorescu's team together with teams from University of Illinois and University of Wisconsin are responsible for designing circuits, architecture and runtime components that mitigate and tolerate parameter variations at Near Threshold.

> "Near-Threshold Computing has the potential to sustain the performance and energy efficiency growth of computing systems for another decade or more."
> -Radu Teodorescu