2009 Workshop on New Directions in Computer Architecture held in conjunction with Micro 42nd pp.20-21

OVERCOMING MOORE'S CURSE: TECHNIQUES FOR POWERING LARGE TRANSISTOR COUNTS IN SUB-45NM TECHNOLOGIES

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

Moore's law continues to provide designers with more transistors per chip economically. However, as feature sizes continue to shrink, supply voltages for these devices has stagnated, resulting in higher energy densities. Thus, more transistors will be available on chip, but they can not all be turned on at the same time. Reducing the supply voltage to near-threshold levels reduces energy density and offers new architectural research areas. These include designs where memory cells operate faster than logic, and where 3D integration can occur with less concern for thermal constraints. However, variation in this operating region is exaggerated and techniques like Razor and adaptive body biasing will be needed to reduce design margins. These variation techniques will also help variation of nominal devices at sub-22nm nodes.

I. MOTIVATION

MOORE'S LAW predicts a doubling of the number of transistors that can be placed on a die, within cost limits, roughly every 18 months [1]. This doubling was accompanied by both improvement in frequency and reduction in power dissipation. At first designers leveraged the improvement in frequency to increase performance. Unfortunately the continued reduction in power dissipation with newer technology nodes slowed, leading to a "power wall" where it became infeasible to increase frequency further. In an attempt to sustain performance improvement, architects are now leveraging the additional transistors to create multi-core designs.

Multi-core designs only provide a temporary solution, because the continued reduction in feature size means the industry is again heading in the direction of a power wall. Figure 1 shows the nominal supply voltages of different technology nodes. It is important to note that supply voltage scaling stagnates around the 90nm technology node resulting in dramatic increases in system energy. The equation for energy density is:

$$U \approx \frac{CV_{dd}^2}{A} + \frac{I_{leak}V_{dd}}{Af} \qquad \text{Equation 1}$$

Where C is capacitance, A is gate area, V_{dd} is the supply voltage, I_{leak} is the leakage current, and f is the frequency. As technology scales A scales as $1/s^2$, and C scales as 1/s. Figure 1 shows the resulting impact of technology node scaling on energy density. Of particular note is that energy density quickly grows beyond current limits around the 32nm node.

The net outcome is that designers will be able to place more transistors on a single die, but will be unable to use them all on at a given time—sometimes termed "Dark Silicon". In this sense Moore's law becomes a curse for architects, who will be able to design more logic but unable to power it on. If nothing is done to counter these trends it will inhibit performance gains of future processors.

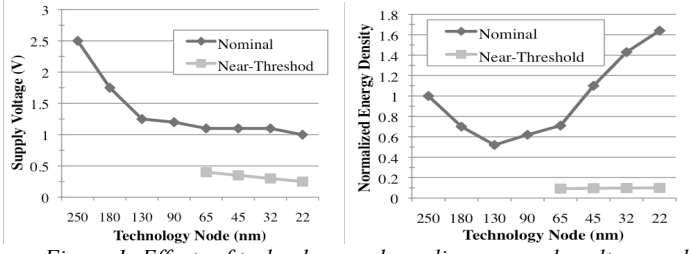

Figure 1: Effects of technology node scaling on supply voltage and energy density. Also ploted are the impacts of NTC operation.

II. NEAR THRESHOLD COMPUTING

IN THE PAST researchers have looked at subthreshold operation as a means to reduce energy consumption [2]. They have shown that an energy minimum occurs in the subthreshold operating region of transistors, where leakage energy increases dominate the gains of dynamic energy reduction. One pitfall of such operation is that as the voltage is reduced, it takes the transistor longer to transition, ultimately leading to longer processing delays. These increased delays make subthreshold computing only viable for low end sensor processors where frequency requirements are often measured in kHz or even Hz. Overall this type of operation leads to a 10,000x power reduction, but with a >500x increase in delay.

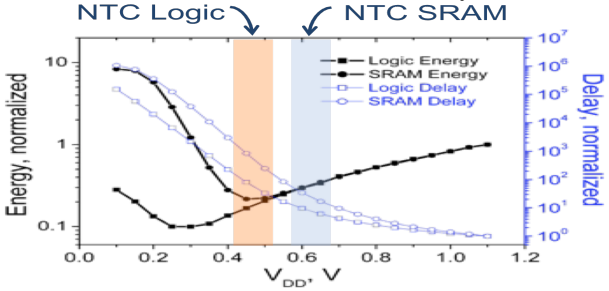

Figure 2: Energy and delay of both logic and SRAM cells.

By increasing the voltage slightly higher than threshold, the transistor is moved into the linear operating region where the energy/performance tradeoffs are more appealing. This operation region is referred to as the near-threshold computing (NTC) region. Researchers have shown that a 100x power reduction can be achieved with only a 10x increase in delay [3]. The energy and delay of different operating regions is presented in Figure 2 for both logic and SRAM based circuits in IBM 90nm technology. Figure 1 includes the energy density of chips using NTC. The difference in SRAM and logic stems from the relatively high leakage component of cache energy, a tradeoff associated with their large size and high density. As leakage increases with respect to switching energy, it becomes more efficient to run faster, and SRAM is shifted higher. In addition, the value of an energy optimal operating voltage for SRAM cache is greatly impacted by reliability issues in the NTC regime, where the need for larger SRAM cells or error correction methods 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.

III. NTC Architectures

Differences in optimal operating voltage for logic and SRAM allow for new architectural decisions to help regain the performance lost in NTC using 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 of the performance, while remaining energy efficient. Due to the differences in SRAM and logic 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. 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. Initial work on this architecture [3] shows that with a few processors (6-12), a 5-6X performance improvement can be achieved.

In addition using NTC, 3D die stacking becomes more feasible. The idea of integrating dies through stacked technology shows promising new designs where DRAM and logic are tightly coupled through wide and fast interfaces. Typically thermal constraints are a major concern in these stacks, but with the decreased energy density of NTC operation more dies can be stacked within the same thermal constraints.

IV. Variation Concerns

Process variation that occurs in the fabrication of semiconductors is becoming worse as minimum feature size is reduced. These variations impact many different device parameters, but of particular concern is the threshold voltage (V_t). V_t variation has become a major concern for designers, because significant changes will drastically increase or decrease transistor speed. With a wider variance of gate speeds, the number of critical paths in the system is larger and more diverse. In order to guarantee designs will meet operating requirements, manufacturers need to place safe margins on timing that will grow with variation. To get a scope of the problem, Li et al. [4] measured the total V_t standard deviation in a 35nm process for various gate lengths (L_g). The equation for which is given by:

$$\sigma^2_{Vt,total} = \sigma^2_{Vt,RD} + \sigma^2_{Vt,Lg} + \sigma^2_{Vt,LER} \qquad \textit{Equation 2}$$

where *Vt,RD* is the random-dopant-induced fluctuation (RDF), *Vt,Lg* and *Vt,LER* are fluctuations caused by the gate length deviation and line edge roughness, respectively. The results of their simulation based results from ITRS roadmap projections, shown in Figure 3, indicate that variation is significantly worse at smaller feature sizes. With this in mind, architects and circuit designers will need to develop techniques to overcome variation or suffer from highly margined designs.

In addition the use of NTC operation increases the sensitivity to variations in threshold voltage, and supply noise. This means that if NTC techniques are to become mainstream, architectural and circuit level techniques will be necessary to overcome the impact of variation.

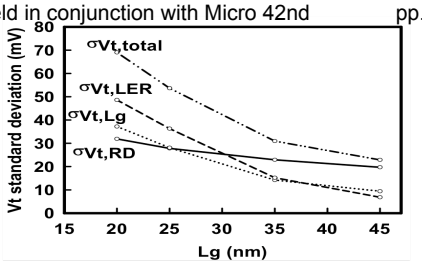

Figure 3: Threshold variation at various gate lengths in 35nm.[4]

V. Variation Reduction Techniques

Systems that can detect and dynamically adapt to variation provide a promising technique to reduce the design margins caused by V_t. variation. Systems like Razor [5] and soft edge clocking [6] are techniques designed to push the limits of frequency even in the presence of variation. These types of systems provide techniques for detecting hold time violations an indication that a critical path provided the data to late, Figure 4. Once the violation is detected the system initiates a recovery sequence and adapts its frequency. The system can then be pushed past the point of traditional failure and recover. This allows for smaller design margins, as chips can be operated at the point of timing failure without error.

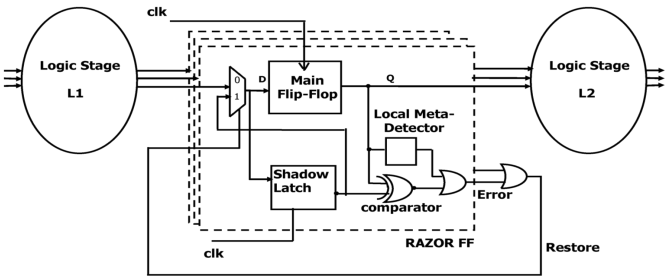

Figure 4: Diagram of Razor latch for timing violation detection.[5]

At the same time new circuit level techniques can be employed to compensate for variation. Body biasing (BB), a technique where the body of the transistor has a voltage applied to adjust the gate P/N ratio, is a well known technique for adapting performance and leakage to global variation of process, voltage, and temperature. Hanson et al. [7] explore the use of adaptive body-bias (ABB) techniques, where sub regions of the chip are designated as critical paths and adjusted separately. This is done to compensate for both local and global variation.

References

[1] G.E. Moore, "Cramming More Components onto Integrated Circuits," *Electronics*, vol. 38, number 8,, Apr. 19, 1965.

[2] A. Wang,B. Calhoun, H. Benton, A.P. Chandrakasan,"Sub-threshold Design for Ultra Low-Power Systems". Book. 2006,

[3] B. Zhai, R.G. Dreslinski, D. Blaauw, T.N. Mudge, D. Sylvester,"Energy Efficient Near-Threshold Chip Multi-Processing". ISLPED 2007: 32-37

[4] Y. Li et al., "Process-variation- and random-dopants-induced threshold voltage fluctuations in nanoscale CMOS and SOI devices," *Microelectronic Engineering*, Vol. 84, Issues 9-10, 2007, Pp. 2117-2120

[5] S. Das et al., "A Self-Tuning DVS Processor Using Delay-Error Detection and Correction." IEEE Journal of Solid-State Circuits, Vol. 41, No. 4, pp. 792-804, April 2006.

[6] M. Wieckowski et al., "Timing Yield Enhancement Through Soft Edge Flip-Flop Based Design," *IEEE Custom Integrated Circuits Conference (CICC), September 2008*

[7] S. Hanson *et al*., "Performance and variability optimization strategies in a sub-200mV, 3.5pJ/inst, 11nW subthreshold processor," *Symposium on VLSI Circuits*, pp. 152-153, 2007.