High-Performance Processors in a Power-Limited World

Sam Naffziger
Intel Corporation, Fort Collins, CO

Abstract

Processor designers and the VLSI industry in general have truly hit the power wall. Many options have been and are being explored to mitigate or circumvent the impact of power limits on performance, but all of these solutions have limited effect and application. The implications of this fundamental limit are far reaching for processor architectures and the shape of computing in coming years. This paper explores the nature of power limitations and some of the implications for the future of processor design.
Keywords: microprocessors, power

Introduction

Power concerns are certainly front and center for most of us in the computing and integrated circuit industry, and the issue has gained attention in the press. However, power consumption as a primary limitation of integrated circuit performance is nothing new, and in fact was identified by Gordon Moore back in 1965 when he charted the course of the IC industry with his eponymous law [1]. Our industry has gone through many transitions starting with the transition from NMOS and bipolar logic to CMOS in the 80's, to the "golden age" of Dennard [2] scaling of the last decade or so. Now that some initial physical limitations on oxide thickness, lithography, and threshold voltage have been hit, this type of scaling is behind us, which is forcing the processor design and architecture community to take power-efficient design seriously. With power consumption as truly a first-order limiter, the bulk of the market will go to the most power-efficient, and therefore highest performance, designs. In this paper I will focus on the impact of power limitations on performance-oriented processor designs which go into desktop computers, servers and workstations.

Process Scaling Background

Many innovations are coming out of the process community that wring more out of standard CMOS than was dreamed possible 10 years ago [3]. There are however some fundamental physical scaling limits that are forcing circuits and architecture to shoulder more of the burden of processor improvements in the future. Primary contributors are a lack of Vt scaling, leakage currents, and variation impacts [4]. The fundamental problem with Vt scaling is that kT/q does not scale, and that leakage currents are set by the transistor's threshold voltage.

The net result is that from the 130nm node forward, Vdd has been scaling slowly (if at all), while leakage as a fraction of total power has been rising (Fig. 1).

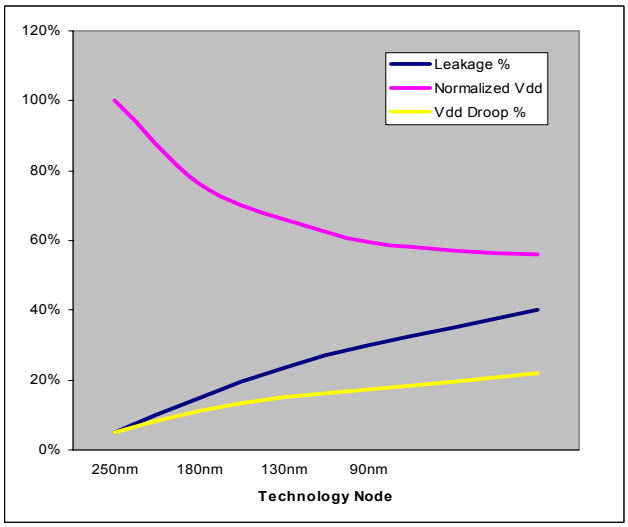

Figure 1: Leakage, Vdd & drop for Hi-Perf Processors

Variability is the other primary contributor to CMOS power challenges. With high frequency designs containing 15 gates per clock cycle or less [7], deviations from nominal device behavior have a huge impact on power efficiency since frequency degrades at the same voltage. One category of variation is die-to-die (D2D), where processing changes between wafers and chips, but transistors on an individual chip are all processed roughly the same way. This type of variation can be compensated in a manufacturing flow change by setting Vdd independently per die. The result can be seen on figure 2 if one slides up and down the channel length axis, a constant power can be achieved by changing the Vdd set point per-part to hold to one of the contour lines.

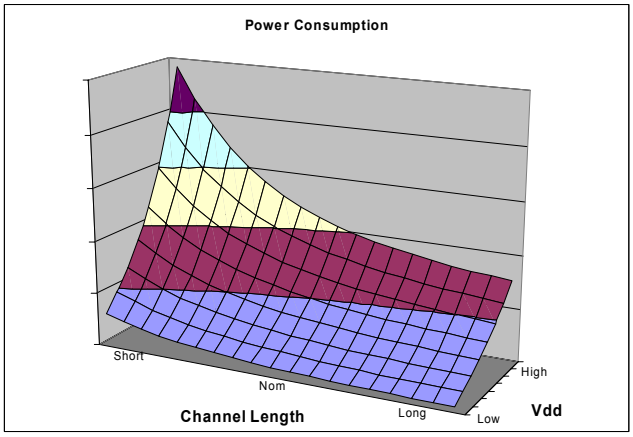

Figure 2: Processor Power vs. Vdd vs. Channel Length

This technique is widely applied in processors shipping today. So D2D variation can be compensated for to some extent, but within-die (WID) variation is a more difficult problem. A couple manifestations are the degradations in maximum frequency due to variation, and significant increases in minimum operating voltage due leakage and Vt

fluctuations causing storage node upsets. This last issue obviously impacts our ability to compensate for D2D variations through voltage changes. The former issue can be illustrated by looking at the effect of WID Le variation on die leakage as shown in figure 3 [23].

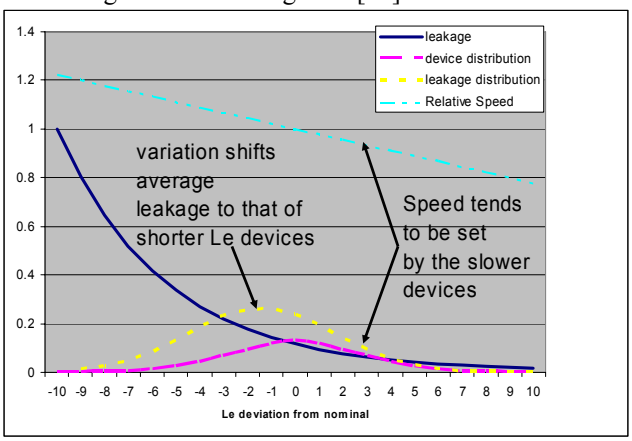

Figure 3: WID Var. impact on Leakage power and Speed

With no variation, the leakage characteristics and the performance of the processor are both set by devices with the same characteristics. But WID variations result in sizeable fractions of the transistors having shorter than nominal Le's, and the leakage of the entire chip ends up averaging to that of a shorter Le transistor, as can be seen by the shift left of the leakage distribution curve from the device distribution curve. Not only do short Le's leak exponentially more, but long Le's are slower as well. In a processor where perhaps hundreds of timing paths may limit frequency, the statistics are such that a sizeable number of gates in just one of those paths will be of the longer Le sort. Thus, we end up with a processor where the *leakage* is set by a device with shorter Le than nominal, and the *speed* of the processor is limited by longer Le devices. For 90nm processors, the result is that leakage can be several times higher than would be predicted by an analysis unaware of variation.

AC voltage droop is another form of variation that is hurting power efficiency. This is a result of a combination of higher current densities, increased on-chip metal resistance and higher di/dt. If one plots voltage droop seen by the same processor core scaled through each technology generation, the result is as in Fig. 1 with almost a tripling of voltage droop. Key assumptions behind the plot are that droop is half inductive and half resistive and that each generation provides .7X area, 1/.7X frequency and one additional metal layer. The effect is somewhat mitigated by leakage current. Measured results corroborate this calculation [11]. The net result of voltage droop is a lower frequency at a given average voltage. Power is determined by average voltage, speed by the minimum, so a droop of 16% results in a power efficiency degradation of 29% from an ideal supply due to the square law dependence of power on voltage.

Future Process Improvements Less Dramatic

There are promising improvements on the horizon from the process community such as hi-K gates, 3D stacking and finFETs. Hi-K and FD-SOI promise some relief from the Vt scaling issue mentioned above, but unfortunately, none of these provide the kind of exponential improvement in efficiency that was achieved by Dennard scaling. 3D stacking can reduce interconnect capacitance for global wires, but these are typically only about 20% of switched capacitance which limits the benefit[24]. Hi-K gates enable some continued channel length scaling without causing an oxide leakage explosion, and finFETs promise improved short channel effects. Each of these improvements provides relief for perhaps one generation and will only enable a continuation of the trend of the past few years where for instance, strain engineering came to the rescue [3]. So process scaling has slowed down, which means that we must look to improved circuit and architectural approaches to manage power in addition to taking advantage of process improvements.

Power Efficient Operating Points

Before looking at specific approaches, it is instructive to observe a model of processor power vs. performance trade-offs. One way to look at process design and operating voltage tradeoffs is illustrated by a 3D contour where gate channel length and Vdd are independently varied, as shown in Fig. 2 using parameters typical of a 90nm processor. Leakage increases exponentially with shorter gate channels, but transistor speed increases less than linearly. Similarly, leakage and switching power have a super-linear dependence on Vdd. The absolute fastest devices at short channel and highest Vdd pay an extreme power price for that performance. Up until the last two years, this was the operating point chosen for high-performance processors. Perhaps more interesting is to observe the performance/watt (or energy/operation) across this optimization space, as shown in Fig. 4.

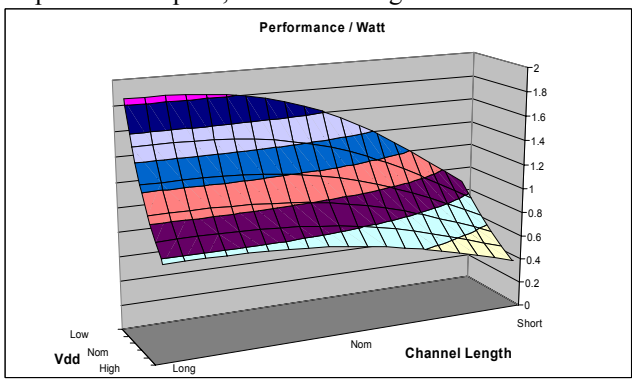

Figure 4: Energy Efficiency vs. Vdd vs. Channel Length

This surfaces the interesting conclusion that the most power efficient processors are the ones with processing and operating conditions that produce the lowest maximum frequency. The ratio in frequency between the shortest channel length, highest Vdd device to the longest channel length, lowest Vdd device is 4:1. So power efficiency is at odds with peak performance.

One final observation is that when comparing processor power and performance, we need to take pains to normalize out not only processing, but operating point as well. As was seen in Fig. 4, the choice of Vdd and channel length can produce wildly different power efficiencies for the same design. Arguably the best way to compare the

operating point-independent, intrinsic power efficiency of a design is by looking at performance^3/Watt [5]. This effectively accounts for the cubic power tradeoff (CV^2F) associated with voltage scaling. Compare Fig. 5 to Fig. 4 to observe the relative flatness of this metric across operating point – still imperfect, but for reasonable processing and operating conditions this is a good metric.

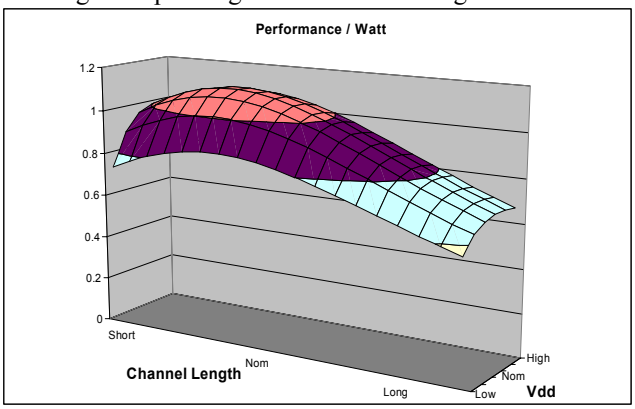

Figure 5: Energy Delay^2 vs. Vdd vs. Channel Length

Dynamic scaling of Vdd is a partial solution to the optimized operating point problem [6,7], but as can be seen by the contour in Fig. 4, the channel lengths chosen have a big impact. Body bias has been discussed as a solution to the leakage vs. speed trade-off [8], but this capability has not made it into a high performance commercial product and the diminished impact of the silicon body on device behavior in today's transistors (or the inaccessibility for SOI) make that solution impractical.

Circuit Design Improvements

A number of circuit related approaches to power reduction have been considered and applied to processor design in recent years. These include dynamic voltage and frequency scaling, clock gating, sleep transistors, multiple Vt (or Le [20]) device insertion, and the tried and true power optimization of device sizes. These approaches all have a place and will be used extensively by competitive processors in the future. They have the effect of driving power consumption down to the minimum needed for a given computation – only fire clocks when needed, only power up circuits that are in use, only use fast devices where they buy performance, and reduce voltage for non-critical applications or operating modes.

Variation adds a new dimension to the challenges of circuit design. Approaches to managing variation effects have been arriving more recently and mitigation can be achieved through a number of methods. Latching methodologies with soft clock edges help hide some of the problem [10], but have the drawback of worsening hold time exposure – which is also subject to variation concerns. A promising development is the use of very local clock delay vernier circuits [11] that can be empirically tuned for a particular design and potentially per part. It is not unreasonable to envision a self-tuning processor where D2D and WID variations are tuned out at manufacturing test. Such innovations are likely required to contain the impact of variation on power efficiency.

The growing problem of voltage variation can be mitigated with adaptive frequency approaches [14], with active decoupling caps [21] and with logic changes that limit the instantaneous change in power consumption. More complex and expensive process steps can also be added such as MIM cap [15], but these need to be traded off with other competing cost adders on the process designer's plate. So refining the above approaches and inventing new ones will likely occupy the best and brightest circuit designers in coming years.

Architectural Directions for Power Efficiency

For processor architecture approaches to improving power efficiency, there are three main directions. The first is to normalize around the "architectural sweet spot" of relatively short pipelines and narrow issue design, the second is the exploitation of multi-core designs, and finally there is the integration of more system components. Recently, architectures are consolidating around an optimal architectural point where the best processor architectures provide both good peak performance and power efficiency.

This data is supported by recent studies which have shown that increasing pipelines beyond a certain length in today's high-performance microprocessors provides only a small performance return, and a negative return if power is taken into account [12]. Long pipelines and short cycle times not only increase clock and latch switching power overhead, but result in a design that is much more susceptible to process variations in both clock and logic paths. Fewer gates per cycle mean it is more likely that one of them is very slow and the spread between the speed limiting Le and the leakage Le grows (Fig. 3). In addition, the higher power of a processor reduces its power efficiency in and of itself, compounding the issue. For instance, high power means that di/dt is higher as the workload of the processor changes dynamically. The resulting voltage variation requires a higher average voltage for the same device speed as discussed previously. Higher temperature resulting from higher power also slows devices down and increases the leakage power. The net is that a low power processor core reaps compounded benefits; resiliency to variation, improved supply integrity, and lower operating temperature – all of which contribute to further improving power efficiency.

The next obvious architectural approach to power efficiency is to exploit the greater density from process scaling to add more cores and run them at a more energy-efficient point. The benefit is easily seen through the cubic relationship between power and frequency/voltage in Fig. 2. To fit two cores in the same power envelope as a single higher frequency/higher power core, the frequency only needs to be dropped to $.5^{(1/3)}$ or by 21%. If we have perfect parallelism in our code, we get 2*.79 or a nearly 1.6X performance benefit simply through greater density. There are however a couple flies in this ointment (Fig. 6), the first of which is that most applications don't parallelize easily, and those that do generally have a non-trivial portion of "serial" code that limits the benefits from multiple cores according to Amdahl's law [16]. The next issue with multi-core is that the demand on memory and IO bandwidth into the chip scales up roughly with the performance as per Amdahl's balanced system law[22]. This means that the

power consumed by I/O will grow as multi-core is exploited. If we start with a typical 10% fraction of die power in I/O for a single core, and assume a constant BW/Hz/core demand with efficiency of modern high-speed I/Os of 10-20mW/Gb/s [17,18], then the power available for all those extra cores diminishes.

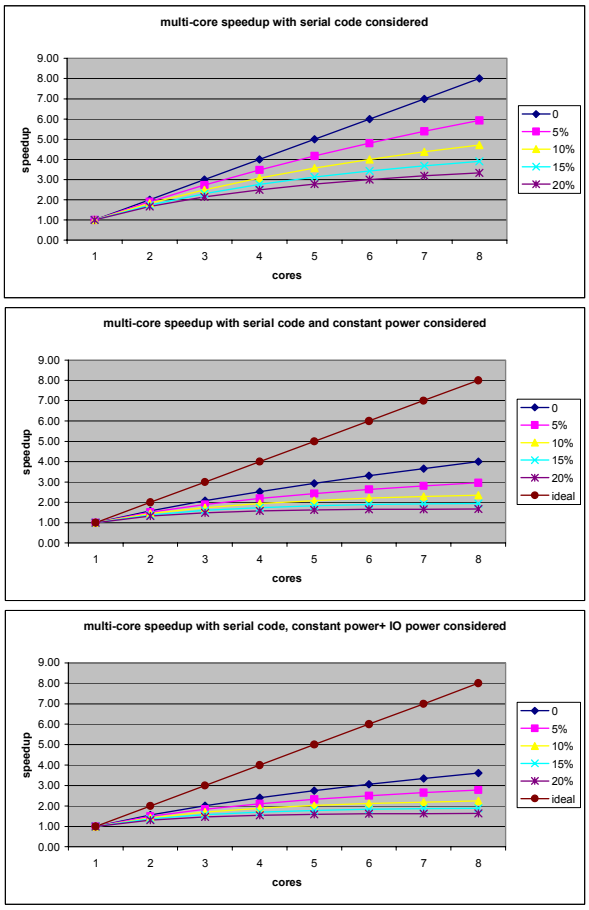

Figure 6: Performance Benefits of Multi-Core Designs With Power Limitations Considered (iso-Process)

Serial code limits speedup significantly. If we consider 10% to be a typical value, then power limitations take a further bite and I/O power another one. It is clear that while multi core will improve computing, it is not a panacea. However, applications and operating systems that make efficient use of increased thread counts will achieve a performance advantage over the traditional single-threaded approach, which will motivate software development along those lines.

The final architectural direction involves the integration of more system components onto the processor die. This has continued for some time starting with floating point units in the early 90s, to memory controllers in the last few years. The embedded space has seen the integration of a large number of system I/O and acceleration functions with a large return in system level power efficiency[19] While this approach does not necessarily drive the CPU power itself down, it does free up more power at the platform level – which in many applications is what the consumer cares about. An additional benefit of integration is the reduction by roughly an order of magnitude the power cost of the I/O circuitry, resulting in greater efficiencies. As the integration of additional cores and system functionality on a single die continues, the computational capacity of a single box and even socket will satiate the appetite of most applications. It is likely that this kind of integration will drive the scale-out architectures to aggressively displace the more complex scale-up designs which have increased I/O, area, and power overhead with diminishing returns in performance.

Conclusion

Power efficiency will occupy the most creative minds in the circuit and processor design community for the foreseeable future. The challenges of designing with near-limit CMOS transistors present new opportunities for creative circuit designers to work around leakage issues and adapt to variations in order to reap the benefits of further process development. Process improvements will continue to provide density increases, with less energy per operation reduction and frequency improvement than in the past. This fact will drive changes in architecture including a consolidation of processor micro-architecture, a growth in the number of cores per die, and a further integration of system components. There are practical limits to all of these trends due to inherent limitations in parallel code and integration opportunities, but it will be at least 10 years before these tricks are fully played out. The final frontier is in the system design and software space where sophisticated power management approaches will need to be developed to extract the greatest performance per watt. Processors with the most efficient adaptive and reconfigurable capabilities will provide the best performance as they conform to changing workloads, environments and applications.

References

[1] Gordon Moore, “Cramming more components onto integrated circuits”, Electronics, Volume 38, Number 8, April 19, 1965

[2] R.H. Dennard, F.H. Gaensslen, V.L. Rideout, E. Bassous, and A.R. LeBlanc, “Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions,” IEEE Journal of Solid-State Circuits, Oct. 1974.

[3] Bai, P. et al, “A 65nm logic technology featuring 35nm gate lengths, enhanced channel strain, 8 Cu interconnect layers, low-k ILD and 0.57 /spl mu/m/sup 2/ SRAM cell”, IEDM 2004

[4] Horowitz et al, “Scaling and the Future of CMOS”, IEDM 2005

[5] Viji Srinivasan, David Brooks, Michael Gschwind, Pradip Bose, Victor Zyuban, Philip N Strenski, and Philip G Emma, "Optimizing Pipelines for Power and Performance," 35th International Symposium on Microarchitecture (MICRO-35), November 2002

[6] Lawrence T Clark, et al, "An embedded 32b microprocessor core for low-power and high performance applications" IEEE JSSC vol 36, pp1599-1608, Nov 2001

[7] S. Das, S. Pant, D. Roberts, S. Lee, D. Blaauw, T. Austin, T. Mudge, and K Flautner, “A Self-Tuning DVS Processor Using Delay-Error Detection and Correction,” IEEE Symposium on VLSI Circuits, June 2005

[8] J.W. Tschanz, S.G. Narendra, Y. Ye, B. A. Bloechel, S. Borkar, and V. De, “Dynamic Sleep Transistor and Body Bias for Active Leakage Power Control of Microprocessors”, IEEE Journal of Solid-State Circuits, Nov. 2003.

[9] B. Curran et al, “4GHz+ Low-Latency Fixed-Point and Binary Floating-Point Execution Units for the POWER6 Processor”, IEEE International Solid State Circuits Conference, Feb 2006
[10] H. Partovi, R. Burd, U. Salim, F. Weber, L. DiGregorio, and D. Draper. Flow-through latch and edge-triggered flip-flop hybrid elements. IEEE International Solid-State Circuits Conference, pages 138--139, 1996
[11] Patrick Mahoney, Eric Fetzer, Bruce Doyle, Sam Naffziger, “Clock Distribution on a Dual-Core, Multi-Threaded Itanium®-Family Processor”, IEEE International Solid State Circuits Conference, Feb 2005
[12] H. P. Hofstee, “Future Microprocessors and Off-Chip SOP Interconnect,” IEEE Transactions on Advanced Packaging, Vol. 27, No. 2, May 2004, pp.301-303.
[13] E. Alon, V. Stojanović, and M. Horowitz, “Circuits and Techniques for High-Resolution Measurement of On-Chip Power Supply Noise,” IEEE Symposium on VLSI Circuits, June 2004.
[14] Fischer et al, “A 90nm Variable Frequency Clock System for a Power Managed Itanium-Family Processor”, IEEE International Solid State Circuits Conference, Feb 2005
[15] Sanchez et al, “Increasing Microprocessor Speed by Massive Application of On-Die High-K MIM Decoupling Capacitors”, IEEE International Solid State Circuits Conference, Feb 2006
[16] Amdahl, G.M. Validity of the single-processor approach to achieving large scale computing capabilities. In AFIPS Conference Proceedings vol. 30 (Atlantic City, N.J., Apr. 18-20). AFIPS Press, Reston, Va., 1967, pp. 483-485
[17] Kromer et al, “100-mW 4 10 Gb/s Transceiver in 80-nm CMOS for High-Density Optical Interconnects”, IEEE Journal of Solid State Circuits, Vol 40 No 12, Dec 2005
[18] S. Naffziger, B. Stackhouse, T. Grutkowski, “The Implementation of a 2-core Multi-Threaded Itanium®-Family Processor”, ISSCC 2006
[19] PA semi, http://www.pasemi.com/, 3/3/2006
[20] S. Rusu, S. Tam, H. Muljono, D. Ayers, J. Chang, “A Dual-Core Multi-Threaded Xeon Processor with 16MB L3 Cache”, ISSCC 2006
[21] M. Ang, R. Salem, A. Taylor, “An On-chip Voltage Regulator using Switched Decoupling Capacitors”, ISSCC 2000
[22] Gray, J.; Shenoy, P.; “Rules of Thumb in Data Engineering”, 2000. Proceedings. 16th International Conference on Data Engineering 29 Feb. March 2000
[23] Sirisantana, N.; Wei, L.; Roy, K, “High-performance low-power CMOS circuits using multiple channel length and multiple oxide thickness”.; Proceedings of the International Conference on Computer Design, Sept. 2000
[24] Nir Magen, Avinoam Kolodny, Uri Weiser, Nachum Shamir, “Interconnect power dissipation in a processor” SLIP2004 conference proceedings