Low Power Design Challenges for the Decade

Shekhar Borkar
Microprocessor Research Labs, Intel Corp.,
Hillsboro, OR 97124, USA
e-mail : Shekhar.Y.Borkar@intel.com

Abstract - Technology scaling will become difficult beyond 0.18 micron. For continued growth in performance, transistor density, and reduced energy per computation, circuit design will have to employ a new set of design techniques, with adequate design automation tools support. This paper discusses a few such techniques that reduce active and leakage power, and deliver higher performance. It concludes by pointing out some of the potential paradigm shifts.

I Introduction

Historically, in the 70's the total number of transistors on a chip doubled every 12 months, this trend continued until the 80's where the trend slowed down, and the total number of transistors doubled every 24 months--this is called "Moore's Law". Wafer sizes continue to grow, and die size has increased with wafer size. To make a bigger die economical, the defect density has decreased, as manufacturing matured. Technology scaling also continues to increase transistor density.

To satisfy Moore's Law, die size has been increasing at the rate of 7% per year, or doubles every 10 years. The operating frequency has increased rapidly in the 90's, almost doubling every 24 months. Let us assume that performance demand, and consequently the frequency trend continues. Therefore to meet the performance goal, the supply voltage will scale by only ~15%, rather than the theoretical 30%[1]. Let us further assume that a technology generation lasts approximately two years. Therefore, every two years:
1. Capacitance per node reduces by 30% (scaling)
2. Electrical nodes in a given area increase by 2X
3. Die size grows by 14% (Moore's Law)
4. Supply voltage reduces by 15%
5. And frequency increases by 2X
The above adds up to an increase in active power of 2.7X every two years.

In order to meet the frequency demand of 2X every two years, transistor threshold voltages will have to scale aggressively, in turn resulting in higher subthreshold leakage currents. Thus leakage power will be a substantial component of the total power.

Figure 1 shows the total power consumption of a microprocessor following Moore's Law, applying above arguments. Clearly, the power dissipation is excessive and impractical.

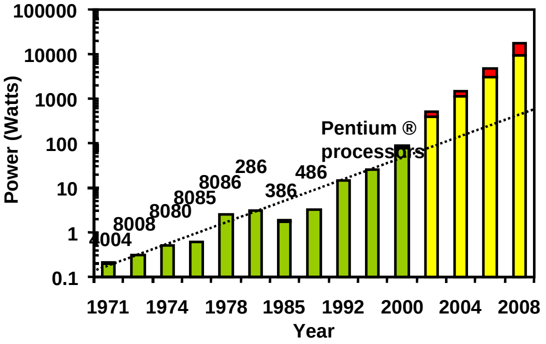

Figure 1: Total power consumption of a microprocessor following Moore's Law.

The scenario painted here is extreme. It assumes that the trends continue, and no innovations occur in microarchitecture and circuits to reduce active and leakage power. We know from our past experience that this is hardly the case in practice—we have dealt with challenges like these before, and we are very optimistic that solutions will be found. This "academic study" is intended to inspire creative thinking and promote research in low power circuits and microarchitecture, by showing the impact if we don't take these challenges seriously.

II. Active Power Reduction

There are two well-known methods for reducing active power: (1) activity reduction, and (2) supply voltage reduction. To reduce activity in synchronous logic, clock gating is employed. Clock signal to a logic block is gated by a control signal, inhibiting clock when the logic block is not in use, thereby reducing the clock signal activity and thus the overall active power consumption. Since clock accounts for substantial activity in the logic, it results in considerable power savings. This technique is not limited to clock signals alone but can be used for other signals having large activity. One has to be careful, however, to ensure that the power consumed in the logic to detect and disable clock gating is considerably lower than the power savings achieved by clock gating. Therefore, in practice clock gating is employed where a logic block is inactive for several clock cycles, limiting effectiveness of this technique.

Since power reduces quadratically with supply voltage, supply voltage reduction can result in substantial power savings. There are two ways to employ supply voltage reduction without compromising performance—static and dynamic.

In dynamic supply voltage scaling, the logic chip is designed to deliver maximum performance at the highest supply voltage. When the performance demand is low, the chip is operated at lower voltage, delivering lower performance but with substantial (quadratic) reduction in power. For example, a mobile microprocessor when running on battery could run at a lower frequency and lower supply voltage, and run at a higher frequency, with higher supply voltage when docked in the desktop docking station [2].

The logic chip could also detect the performance demand and adjust frequency and supply voltage accordingly [3]. A mobile processor when detects peaks in performance demand, adjusts the supply voltage and the frequency to deliver the necessary throughput, thereby saving considerable power and energy.

In static supply voltage reduction method, multiple supply voltages are used. Figure 2 shows such a scheme employing two supply voltages. Higher supply voltage is used for performance critical logic, which runs at higher speeds, and consumes higher power.

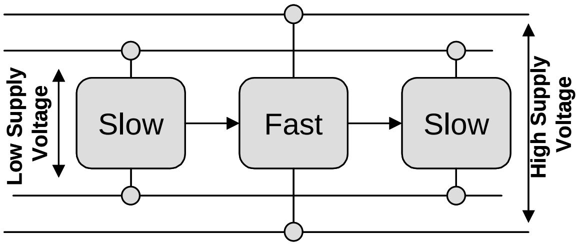

Figure 2: Logic with multiple supply voltages

There are several issues with multiple supply voltages that need careful evaluation. When a logic signal that emerges from a slow block is connected to the fast block, the signal levels are closer to the threshold voltages of the transistors, and could consume excessive leakage power and reduce noise margin. This scheme also requires additional power supply grid, and associated support such as decoupling capacitors, to ensure error free operation.

III. Leakage Power Reduction

Supply voltage will continue to reduce each technology generation to lower power dissipation. But to improve transistor and circuit performance by at least 30% per technology generation, transistor threshold voltage (Vt) must reduce at the same rate, so that a sufficiently large gate overdrive (Vcc/Vt) is maintained. However, reduction in Vt causes transistor subthreshold leakage current (I_{off}) to increase exponentially. Large leakage can (1) severely degrade noise immunity of dynamic logic circuits, (2) compromise stability of 6T SRAM cells, and (3) increase leakage power consumption of the chip to an unacceptably large value. In addition, degradation of short channel effects, such as Vt roll-off & Drain Induced Barrier Lowering (DIBL), in conventional bulk MOSFET's with low Vt can pose serious obstacles. Figure 3 estimates subthreshold leakage current, I_{off}, of transistors in future technologies. Note that I_{off} is a strong function of junction temperature.

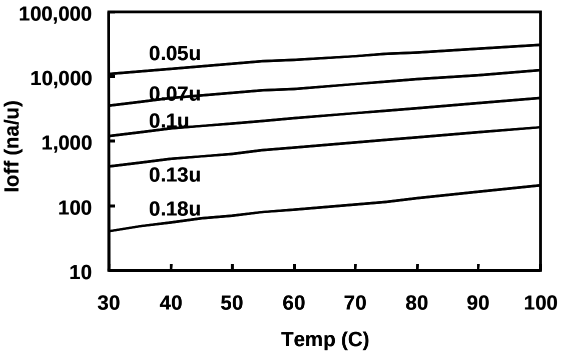

Figure 3: Estimated subthreshold leakage currents of deep submicron transistors

Using these subthreshold leakage numbers, Figure 4 estimates total leakage power of a large chip at high temperature. Notice that almost half the power of the chip can be from subthreshold leakage. That is why leakage power reduction techniques will be necessary in the future designs.

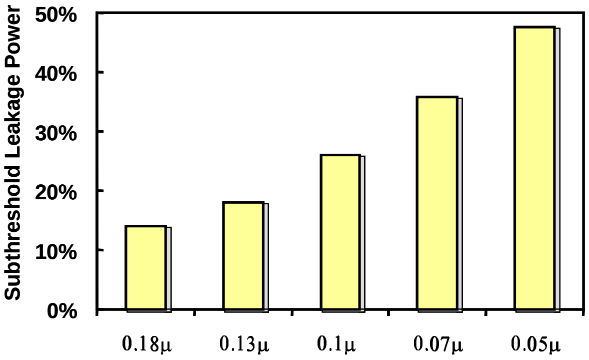

Figure 4: Half the power of a chip could be in subthreshold leakage.

Dual Vt design technique will be widely used to reduce the subthreshold leakage power. In this technique, the process technology provides two flavors of transistors: high threshold voltage (High Vt), and low threshold voltage (Low Vt). The High Vt transistors yield slower logic, but lower leakage, whereas Low Vt transistors yield faster logic, but higher (~10X) leakage.

Figure 5 shows the dual Vt design methodology for a typical logic block using path delay distribution. If we use high Vt transistors everywhere, then it yields higher delay, or lower frequency. On the other hand, if we use low Vt transistors exclusively, then it yields higher frequency, but ~10X higher leakage power. A selective insertion of low Vt transistors yields higher frequency with lower leakage power.

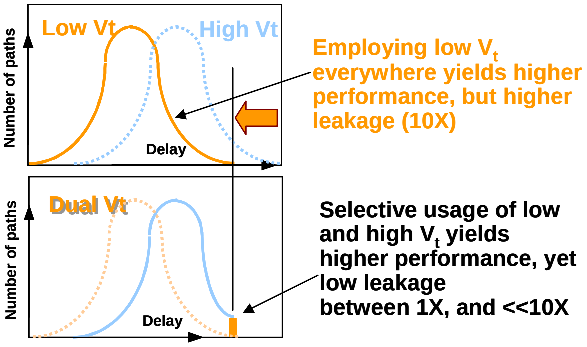
Low Vt
High Vt
Number of paths
Delay
Employing low V_t everywhere yields higher performance, but higher leakage (10X)
Dual Vt
Number of paths
Delay
Selective usage of low and high V_t yields higher performance, yet low leakage between 1X, and <<10X

Figure 5: Dual Vt design methodology

For standby leakage reduction, stack effect can be exploited [4]. This technique uses the fact that an "off" transistor stack has an order of magnitude lower subthreshold leakage than the individual transistors. Figure 6 shows subthreshold leakage of an NMOS stack when both transistors are turned off. The second graph shows that as the transistor leakage is increased (X axis), the stack leakage increases at a lower rate than the transistor leakage, making this technique effective even with high leakage currents.

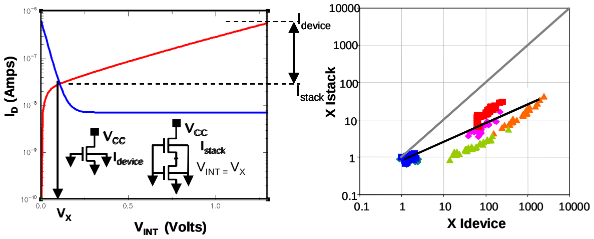

Figure 6: Stack effect to reduce subthreshold leakage

To exploit stack effect in the standby mode, the logic block needs to be placed in a state where all stacked transistors are turned off. This can be done manually, however, design tool innovations are needed to automate this process. Studies show that 1.5 to 2.5X reduction in leakage can be achieved using this technique [4].

Supply gating or "sleep transistor" is another technique to reduce active and standby subthreshold leakage power. This is similar to clock gating, where power supply is "gated" using a high threshold transistor, to cut off power to the logic block as shown in Figure 7. This technique could reduce the leakage power by 1,000X; however, there are several issues. First, the high Vt transistor in series with the supply causes performance degradation. Second, the virtual supply rails could couple to noise, reducing noise immunity. Third, the virtual supply rails demand a local power grid design. Fourth, the virtual grid needs careful design to ensure that logic state is not lost when the virtual supplies are collapsed.

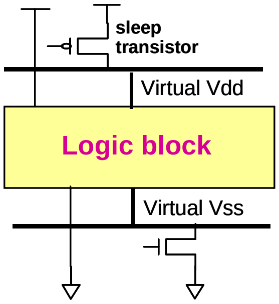

Figure 7: Supply gating or Sleep Transistor technique

All these leakage control techniques will be employed in the near future, and design automation support will be needed. A detailed discussion on sources of leakage power and reduction techniques may be found in [5].

IV. Low Power Microarchitecture Techniques

Low power microarchitecture will play a key role in exploiting low power circuit techniques. To further discuss effectiveness of "general purpose" logic, consider Pollack's rule [6]. He observed that 2X growth in "general purpose" logic provides only 1.4X increase in the performance—a square law. Therefore, "general purpose" logic is not power efficient in delivering performance. Future applications will lend themselves to inherent parallelism, and could be easily served by special purpose hardware, tailored for the applications, and thus power efficient. Microarchitecture will play a key role in identifying special purpose solutions for such applications.

Figure 8 compares estimated active power density of logic and static memory in a given technology. Memory power density tends to be an order of magnitude lower than that of logic. This is because only a part of the memory is accessed at any given time. Also, memory transistors can withstand relatively higher threshold voltages, reducing the leakage power compared to logic. To make up for the loss of transistor performance, memory operations can be pipelined, with modest increase in latency.

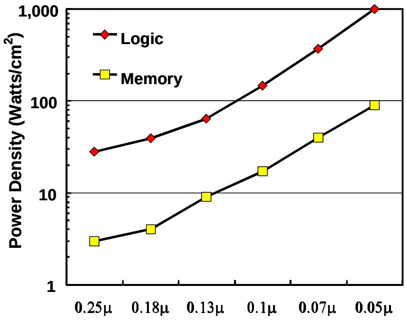

Figure 8: Comparing logic and memory power density

Therefore future microarchitectures could exploit lower power density of memory to stay on the performance trend, and yet lower active and leakage power.

V. Platform Level Low Power Techniques

Power delivery at the package and platform level will face the biggest challenge of all. Figure 9 plots supply current for the future microprocessors following Moore's Law. Supply voltages will decrease, yet supply currents will increase. Hence resistive voltage drop (IR drop) in the power delivery grid will have to be reduced by lowering the effective resistance. Since the processor frequency will increase, supply current will increase, resulting in higher inductive (Ldi/dt) noise. To solve the inductive noise problem, we will have to continually reduce the effective inductance in the power delivery network, and employ larger decoupling capacitors on the die.

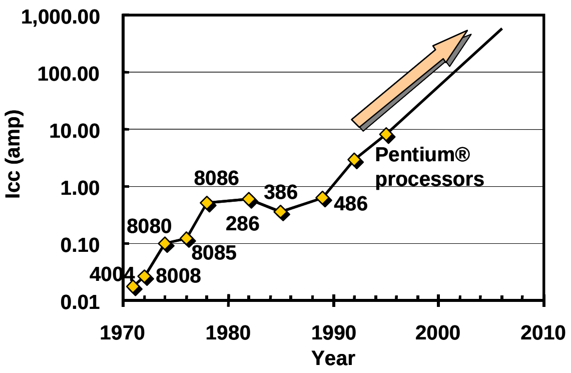

Figure 9: Power supply currents will grow tremendously

Power dissipation tradeoffs are even more challenging. With close to 50% of the power being dissipated in subthreshold leakage, which is exponentially dependent on temperature. The chip could dissipate less power if its die temperature is reduced, as shown in Figure 10.

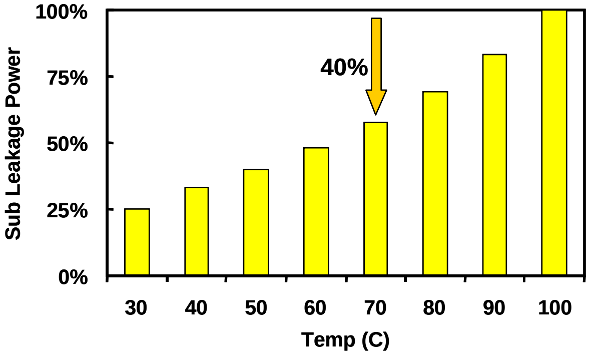

Figure 10: Subthreshold leakage power reduces with lower temperature.

There is an interesting tradeoff between power dissipated into an "active cooling" device, savings in the chip power dissipation, and the performance gain due to lower junction temperature. Therefore, active cooling of chips, to bring junctions at room temperature, or even below, will become attractive.

VI. Summary

We have studied active and leakage power dissipation trend. If the trend continues then we will be posed with numerous design challenges in power dissipation, power delivery, and power density. We also visited several circuit, microarchitecture, and platform level techniques to reduce active and leakage power for the future VLSI chips. We are confident that industry and academia will find solutions to these challenges. If we don't make significant breakthroughs in these areas, "general purpose" logic growth will have to reduce, restricting die size growth. This will result in different design challenges than we thought.

Acknowledgements

I would like to thank Vivek De, Yibin Ye, Siva Narendra, Dinesh Somasekhar, and Ali Keshavarzi for helpful discussions on active and leakage power control, Fred Pollack for bringing insight into low power microarchitectures, and Justin Rattner for continued support and encouragement.

References

[1] S. Borkar, "Design Challenges of Technology Scaling," IEEE Micro, July/Aug, 1999.

[2] Intel SpeedStep™ technology, *http://www.intel.com*.

[3] D. Ditzel, "Transmeta's Crusoe," Proceedings of COOL Chips III conference, Tokyo, Japan, April 24-25, 2000.

[4] Y. Ye, S. Borkar, V. De, "A New Technique for Standby Leakage Reduction in High-Performance Circuits," *1998 Symposium on VLSI Circuits*, June 1998.

[5] A. Chandrakasan, W. Bowhill, F. Fox, "Design of High Performance Microprocessor Circuits," *IEEE Press*, Chapter 3.

[6] F. Pollack, "New Microarchitecture Challenges in the Coming Generations of CMOS Process Technologies", Micro32, 1999.