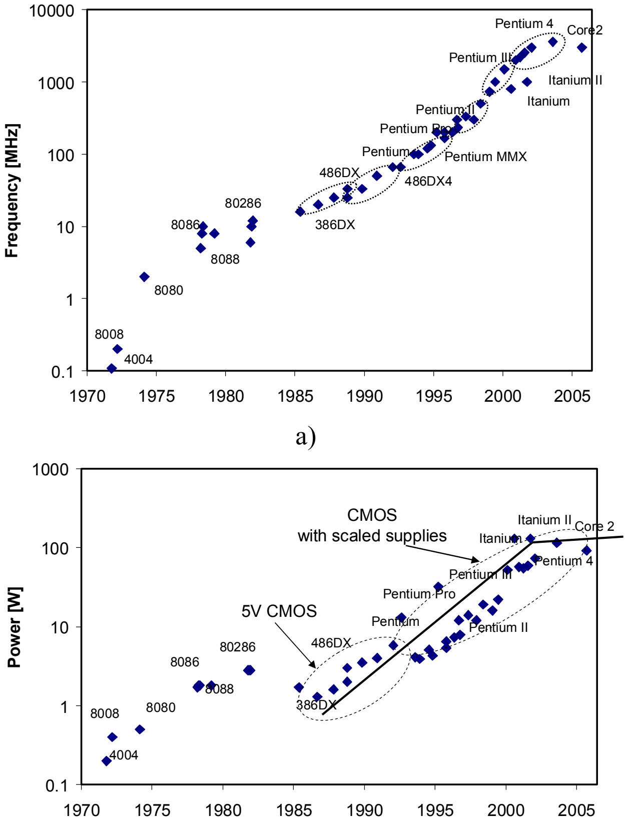
# Power-Limited Design

Borivoje Nikolić
Electrical Engineering and Computer Sciences
University of California,
Berkeley, CA 94720-1770, USA

*Abstract*— **Technology scaling has entered a new era, where the chip performance is constrained by its power dissipation. Although the power limits vary with the application domain, they dictate the choice of technology, and architecture, and dictate the use of implementation techniques that trade off performance for power savings. This paper examines the technology options in the power-limited scaling regime, and reviews sensitivity-based analysis that can be used for the optimal selection of power-performance tradeoffs, to achieve the best performance under the power constraints. These tradeoffs are examined on the techniques for power minimization at the technology, circuit, logic, and architecture levels.**

## I. Introduction

Technology scaling reduces the minimum physical dimensions of transistors by a factor of $S = 0.7$ in each generation, and interconnect scaling follows a similar trend. In turn, the area needed to implement digital functions and memory have been reducing roughly by a half with the introduction of each new technology node. Additionally, scaled devices have had increased switching speeds, with simultaneously lowered switching energy. Ideal scaling scenario proposed by Dennard, *et al* [1], requires that all the voltages scale with the same factor of 0.7, in order to maintain constant fields. One of the consequences of this scaling model is that the switching energy per transistor has been scaling with a factor of $S^3$. resulting in constant power for a chip with the same area. The original paper also points out that one of the limitations of this scaling regime is in the fact that $kT/q$ doesn't scale, resulting in non-scaling of device subthreshold characteristics. Ideal scaling does not account for gate tunneling currents, which are significant with very thin gate oxides. Practical scaling has not always followed this ideal principle. Initially, supply voltages were maintained at high levels, particularly at 5V for an extended period of time, to maintain compatibility of chip-to-chip interfaces. Supply voltage scaling has started with approximately the 0.5µm technology node, and, until very recently, has roughly followed the scaling of linear dimensions. However, designers and manufacturers have often used somewhat higher supply voltages above the ideal values of $V_{DD}$ = feature size * 10V/µm, to boost the performance within the reliability constraints. In addition, chip dimensions have traditionally been increasing, rather than staying constant. In the microprocessor design in particular, architectural changes have resulted in faster scaling of operating frequencies, beyond the gains achieved by technology scaling alone [2][3]. Figure 1.a, as an example, illustrates the frequency trends in lead Intel's microprocessors over time. All this factors have resulted in a rapid increase in power dissipation that continued until many of the designs reached the limits. The power dissipation in high-performance applications is limited by the practicality and the cost of cooling; in case of microprocessors with



a)

Figure 1. a) Frequency and b) power trends in Intel's microprocessors.

The NSF grant No. 0403427, provided the computing infrastructure. The research has been sponsored in part by the Center for Circuit & System Solutions (C2S2) Focus Center, one of five research centers funded under the Focus Center Research Program, a Semiconductor Research Corporation program, and an NSF CAREER grant ECCS No. 0238572.

1-4244-1378-8/07/$25.00 ©2007 IEEE.

forced-air cooling systems this limit is in the 100-150W range. Chips for portable applications often do not allow for the use of fans and are limited to about 2W of power with plastic packaging. As a result, most of the designs today and all of the designs of the future are power limited. Figure 1.b. illustrates the increase in power dissipation in Intel's processors, and the data form the other manufacturers follows a very similar trend. The power dissipation in the lead microprocessors introduced in the past 10 years has been increasing with a factor of 2.5 per generation, until saturating at about 100W levels. Mobile applications are often limited by the battery life, which dictates constraints on both active and leakage power during the standby and sleep modes.

These trends in the technology scaling and design have made the power dissipation a primary design constraint for both high-performance and mobile applications. In contrast to the past, fitting within the power budget today is as important for the designers as is achieving the maximum performance. Chip designs have become power limited, and instead of targeting the absolute maximum performance, the designers need to maximize the performance for the given power budget. There are many degrees of freedom in the design for trading off performance and power, and they can be performed at the technology selection, circuit and logic design and the architecture. Many of the decisions in the system design are dependent on each other and can involve optimization of both discrete and continuous variables.

## II. Power-Limited Scaling

### A. Constant-field scaling

The constant field scaling regime keeps the active power density constant, by scaling the active power per device with a factor of $S^2$. The leakage power for the chip with a constant area scales with a factor of

$$S_L = \frac{1}{S^2} 10^{\frac{V_{Th}(1-S)}{S_{subth}}} . \qquad (1)$$

The relative increase in leakage current is dependent on the actual threshold voltage. Threshold reduction by the factor of $S = 0.7$ increases the chip leakage power density by several orders of magnitude with high values of threshold voltage (>0.5V). However, this traditionally did not affect the overall power consumption, as the subthreshold leakage was a very small component of the total power, even smaller than reverse bias junction leakage currents. Continued exponential increase in leakage currents has brought it to a level where it significantly contributes to the overall power budget, with low thresholds. In sub-100nm processes this increase in leakage is less than an order of magnitude with each technology generation, since $SV_{Th} < S_{subth}$.

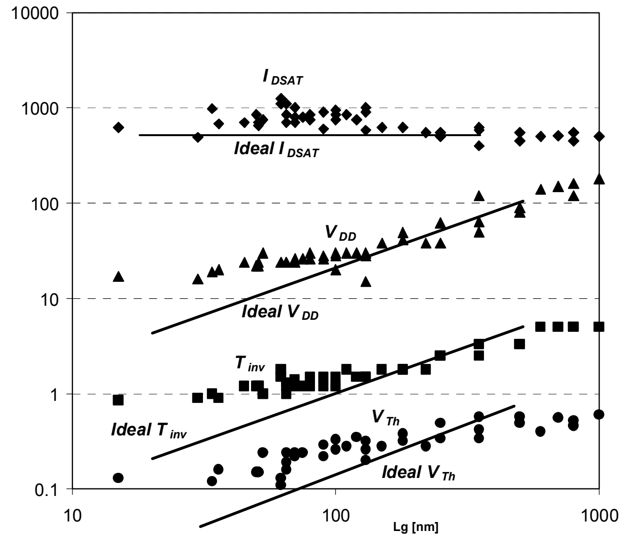


Figure 2. Historic trends of scaling the saturation current, oxide thickness, supply voltage and threshold voltage.

### B. Power-Limited Scaling

With constant subthreshold slopes, threshold voltage scaling results in the exponential increase in transistor drain leakage currents. While the leakage currents were negligible in the past, they have been on a steady increase, and present a significant portion of the overall power budget today. With scaling of both the supply and the threshold voltage, the minimum power is achieved when a balance is struck between the active and leakage power components. This optimum is at the point where leakage contributes to about 30-40% of the total power during active operation of the circuit [4], [5]. Many of the high performance designs have reached this point around 130nm or 90nm technology nodes. As a result, continued scaling in the 90nm, 65nm, 45nm nodes and beyond, departs form the constant-field model and enters the power-limited scaling regime. Still, the continued scaling of technology outlined by ITRS introduces new devices with lower thresholds [6]. The power-limited scaling regime is characterized by the use of multiple devices in the design optimized for different performance/power targets, together with slowed down supply and threshold voltage scaling, and dramatic changes in chip architectures.

### C. Recent scaling trends

Although the technology scaling from the 0.5μm down to the 0.13μm technology has involved both the reduction in device dimensions and voltages, it has not closely been following the ideal constant-field scaling rules. Practical scaling data is plotted against the ideal requirements in Figure 2. Both the supply voltage and the transistor thresholds have been scaling with feature sizes, but have been generally falling behind the ideal values. Particularly, the slowdown in the threshold scaling has been recently additionally slowed down, resulting in the reduced $V_{DD}/V_{Th}$ ratios. On the other hand, shortened channel lengths and

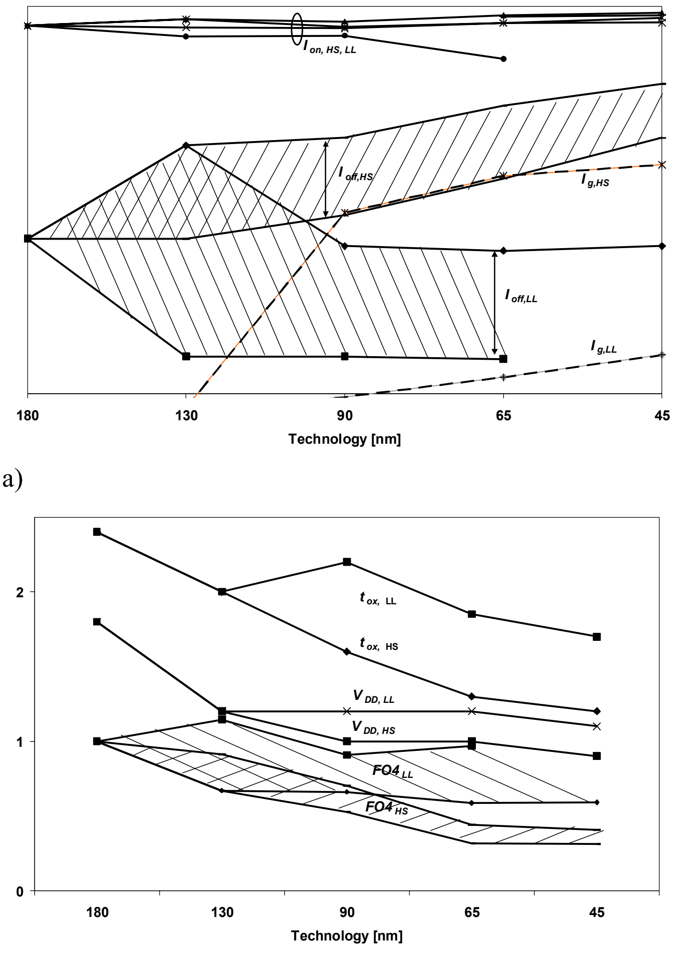


Figure 3. a) Trends in on-currents, $I_{on}$, drain-to-source leakage currents, $I_{off}$, and gate tunneling currents, $I_g$ for foundry deep submicron processes. b) Trends in oxide thicknesses, $t_{ox}$, supply voltage, $V_{DD}$ and fanout-of-4 inverter delays, *FO4*. HS represents a high-speed (or general purpose) process option, and LL represents a low-leakage (or low-power) process option.

mobility enhancement techniques have been increasing the transistor saturation currents. Similarly the gate capacitances have been reducing, instead of staying constant. Both of these trends have been contributing to improvements in transistor switching speeds, despite the slowdown in threshold scaling.

Present technology scaling is characterized by the availability of multiple devices, as outlined by the ITRS. Figure3.a. and Figure 3.b. illustrate trends in on-currents, $I_{on}$, drain-to-source leakage currents, $I_{off}$, gate tunneling currents, $I_g$, oxide thicknesses, $t_{ox}$, supply voltage, $V_{DD}$ and corresponding fanout-of-4 inverter delays, FO4, for one foundry. In today's technologies, generally a choice of one of the two oxide thicknesses is available for chip implementation; the thinner oxide is used for high-speed (HS) applications, and the thicker oxide is used in the applications that require lower leakages (LL). This second option is often denoted as 'low-power' because those applications usually have tight standby power requirements. Within each of the two process options, generally two, out of 2 or 3 offered threshold voltages are available for implementation. The gate oxide thickness for the high-speed process option generally follows the historic trend from Figure 2, for oxide scaling by about 20-25% per technology generation, down to the 45nm node. Oxide thickness scaling has been slowed down with increased tunneling currents; further scaling of the effective oxide thickness will be enabled by commercialization of the high-k dielectrics. Scaling of the thicker oxide follows the same trend, lagging by about 1.5 technology generations. In the high-speed process options, the threshold voltages continue to scale, resulting in a continued off-current increase in lead, standard-threshold devices by a factor of 2-2.5 per technology generation. In contrast, in low-leakage process option, the threshold voltages are held approximately constant, which is dictated by the battery life requirements in mobile devices. In turn, 65nm and 45nm processes offer a wide variety of devices, whose, on-currents can vary by a factor of 4, off-currents can vary by three orders of magnitude, and FO4 delays could vary by a factor of 3. These process options open up a possibility for power-performance optimization at circuit and architecture levels using a number of different design variables. By simply mapping a design into a different technology option, large tradeoffs in power-performance can be made. For example up to 3× in delay can be traded off for three orders of magnitude of leakage savings.

## III. Optimal design

Maximizing the performance under energy constraints can be formulated as a constrained optimization problem and has been studied recently. The system can be optimized to maximize the performance under energy constraints, or to minimize the energy under performance constraints. In our recent work we use sensitivities to formalize the tradeoff between energy and performance. Sensitivity is defined as the absolute gradient of energy to delay with respect to a change in some design variable.

There are usually several tuning variables that can be exploited to trade off energy for performance at various levels of design hierarchy. The tradeoff achieved by tuning some design variable $x$ is given by the energy/delay sensitivity to variable $x$:

$$S_x(X) = \left.\frac{\partial E/\partial x}{\partial D/\partial x}\right|_{x=X}. \qquad (2)$$

This quantity represents the amount of energy that can be traded for delay by tuning variable *x,* around the design point *X*. The energy-efficient design is achieved when the relative sensitivities to all the tuning variables are balanced [7][8]. This result is more general than optimization for a single design point, such as a minimum energy-delay (ED) product.

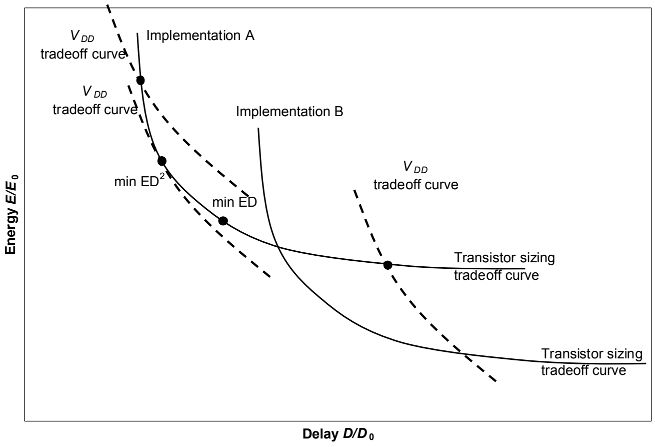


Figure 4. Illustration of the energy-delay tradeoff s.

Figure 4 illustrates some of the tradeoffs at the circuit level. In general, the energy-delay tradeoff curve obtained by continuously changing one design variable is convex. An example such a curve is shown in Figure 4 for a complex digital function (such as 64-bit addition). A continuous curve can be achieved by optimally changing transistor widths to maximize the performance under varying energy constraints. A different implementation of such a curve (e.g. using different logic design for an adder, or different circuit style) would result in a different curve (implementation B). Each point on these curves corresponds to a different design, with different transistor sizes. The slope of the curves changes in each point, and the curves may or may not intersect. The curve that is closer to the coordinate origin is more energy efficient. Dashed lines in the figure illustrate the energy-delay tradeoffs with changing $V_{DD}$, for a fixed sizing. The optimal design is achieved when sensitivity to sizing is equal to the sensitivity to $V_{DD}$, as is the case for the design that corresponds to $ED^2$ minimum.

Balancing sensitivities requires optimization of supplies and thresholds. In the optimum for a design, leakage contributes to about 30-40% of the total energy [4][5].

Tradeoff variables differ with the abstraction level. Design variables accessible to the circuit designer include the transistor sizing, and choice of supply and threshold voltages. Logic designers and architects can affect the logic design, logic depth, pipeline depth, parallelism etc to achieve these tradeoffs. These variables can be continuous, such as supply voltage, or discrete, such as the threshold selection or a choice of the architecture. A classical example of interaction of a higher level, architectural, design variable, such as the degree of parallelism or pipelining, with a technology design variable, such as a supply voltage, has been studied in [9]. By using parallelism or deeper pipelining, a datapath can have the same throughput at a lower supply voltage, reducing the active power dissipation. With optimized transistor thresholds and supplies, parallelism and pipelining reduces the total power.

## IV. Conclusion

Power limitations are today as important in the design as is the performance. Design in the power-limited scaling regime requires continuous changes in the architectures, circuit implementation and technology choices to maximize the performance under the power constraints. Many techniques for lowering power consumption are well known, but their implementation often incurs a performance penalty. An optimum implementation is achieved when the energy/delay sensitivity of the design is equal, for all the design and technology variables. Implementation of low power techniques increase the design and verification complexity, and often requires special technology features, which increases the design cost. Ultimately, technology scaling will end when the increase in the design cost stops being manageable.

## Acknowledgment

The author acknowledges the contributions of the students, faculty and sponsors of the Berkeley Wireless Research Center. Most of the sensitivity-based optimization has been done by Dejan Markovic, Vladimir Stojanovic, and Radu Zlatanovici, in collaboration with Mark Horowitz and Bob Brodersen.

## References

[1] R.H. Dennard, F.H. Gaensslen, V.L. Rideout, E. Bassous, A.R. LeBlanc, "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE Journal of Solid-State Circuits*, vol. SC-9, no. 5, pp. 256-268, October 1974.

[2] S. Borkar, "Design challenges of technology scaling," *IEEE Micro*, vol.19, no.4, pp. 23-29, July-Aug. 1999.

[3] P.P. Gelsinger, "Microprocessors for the new millennium: Challenges, opportunities and the new frontiers," *International Solid-State Circuits Conf, Digest of Technical Papers, ISSCC 2001*, San Francisco, February 5-7, 2001, pp. 22-25

[4] R. Gonzalez, B. Gordon, M.A. Horowitz, "Supply and Threshold Voltage Scaling for Low Power CMOS," *IEEE J. Solid-State Circuits*, vol. 32, no. 8, pp. 1210-1216, Aug. 1997.

[5] K. Nose and T. Sakurai, "Optimization of $V_{DD}$ and $V_{TH}$ for Low-Power and High-Speed Applications," in *Proc. Asia South Pacific Design Automation Conf.*, Jan. 2000, pp. 469-474.

[6] *International Technology Roadmap for Semiconductors*, 2006 data, available at http://public.itrs.net

[7] V. Zyuban, et al, "Integrated analysis of power and performance for pipelined microprocessors," *IEEE Transactions on Computers*, vol. 53, no. 8, pp. 1004 – 1016, Aug. 2004.

[8] D. Marković, V. Stojanović, B. Nikolić, M.A. Horowitz, R.W. Brodersen, "Methods for True Energy-Performance Optimization," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 8, pp. 1282-1293, August 2004.

[9] A.P. Chandrakasan, R.W. Brodersen, "Minimizing power consumption in digital CMOS circuits," *Proceedings of the IEEE*, vol. 83 no. 4, pp. 498 –523, April 1995.