DESIGN CHALLENGES OF TECHNOLOGY SCALING

IS PROCESS TECHNOLOGY MEETING THE GOALS PREDICTED BY SCALING THEORY? AN ANALYSIS OF MICROPROCESSOR PERFORMANCE, TRANSISTOR DENSITY, AND POWER TRENDS THROUGH SUCCESSIVE TECHNOLOGY GENERATIONS HELPS IDENTIFY POTENTIAL LIMITERS OF SCALING, PERFORMANCE, AND INTEGRATION.

Shekhar Borkar
Intel Corporation

Scaling advanced CMOS technology to the next generation improves performance, increases transistor density, and reduces power consumption. Technology scaling typically has three main goals: 1) reduce gate delay by 30%, resulting in an increase in operating frequency of about 43%; 2) double transistor density; and 3) reduce energy per transition by about 65%, saving 50% of power (at a 43% increase in frequency). These are not ad hoc goals; rather, they follow scaling theory. This article looks closely at past trends in technology scaling and how well microprocessor technology and products have met these goals. It also projects the challenges that lie ahead if these trends continue. This analysis uses data from various Intel microprocessors;[1] however, this study is equally applicable to other types of logic designs.

Scaling theory

Scaling a technology reduces gate delay by 30% and the lateral and vertical dimensions by 30%. Therefore, the area and fringing capacitance, and consequently the total capacitance, decrease by 30% to 0.7:

$$\text{Delay} = 0.7,\ \text{frequency} \approx 1.43$$

$$\text{Width} = W = 0.7,\ \text{length} = L = 0.7,\ t_{ox} = 0.7$$
$$(t_{ox}\text{: oxide thickness})$$
$$\text{Area capacitance} = C_A = (0.7 \times 0.7)/0.7 = 0.7$$
$$\text{Fringing capacitance} = C_F \propto L \ \therefore\ C_F = 0.7$$
$$\text{Total capacitance} \Rightarrow C = 0.7$$

Since the dimensions decrease by 30%, the die area decreases by 50%, and capacitance per unit of area increases by 43%:

$$\text{Die area} = X \times Y = 0.7 \times 0.7 = 0.7^2$$

$$\frac{C}{\text{Area}} = \frac{0.7}{0.7 \times 0.7} = \frac{1}{0.7}$$

Frequency-scaling trends (performance)

To evaluate how well past technologies have met the performance goal, we plot product (introduction) frequencies over time, as shown in Figure 1. Assuming that a technology generation spans two to three years, the data show that microprocessor frequency has doubled every generation—not just increased by 43%. Several factors may account for this. Consider the data plotted on the right-hand *y*-axis in Figure 1. The average number of gate delays in a clock period is decreasing because the new microarchitectures use shorter

0272-1732/99/$10.00 © 1999 IEEE

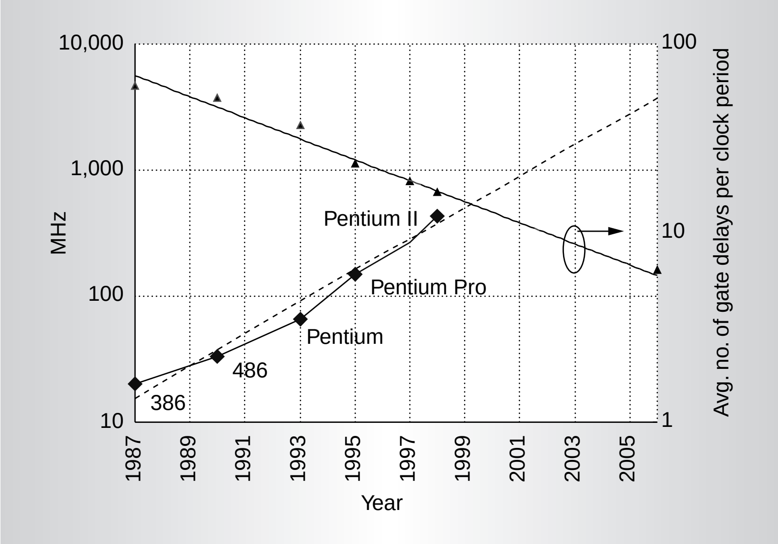

Figure 1. Processor frequency doubles each generation.

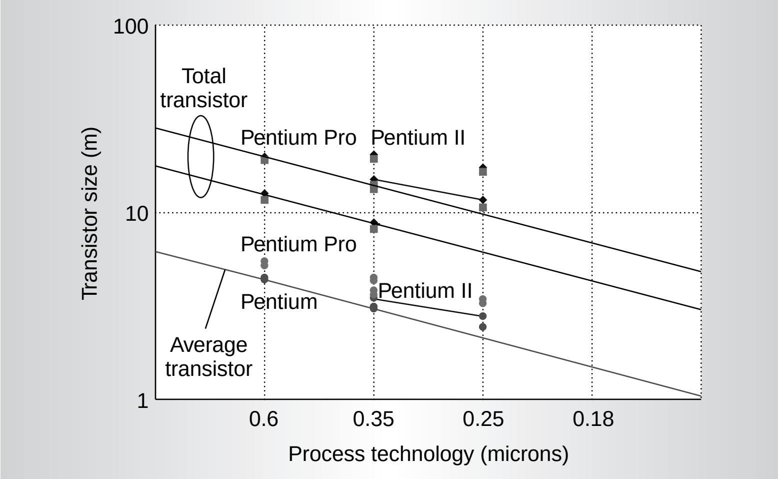

Figure 2. Scaling of transistor size.

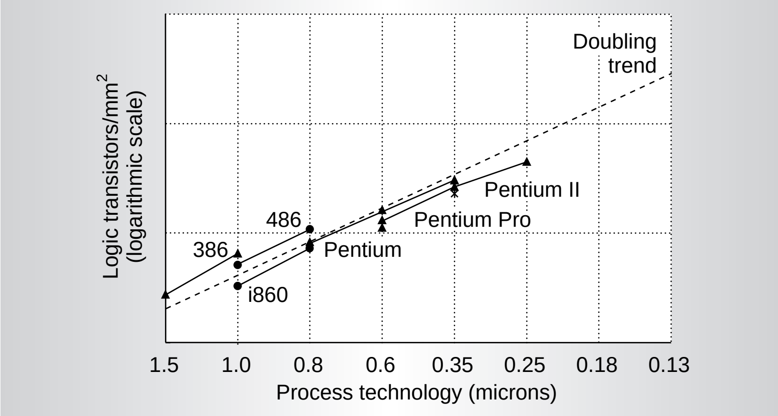

Figure 3. Effect of scaling on logic transistor density.

pipelines for static gates, and advanced circuit techniques reduce critical path delays even further. This could be one reason that frequency is doubling every technology generation.

One might suspect that this frequency increase comes at the expense of overdesign or oversizing of transistors. Figure 2 shows how transistor size scales across technologies in different Intel microprocessors. The dotted lines show transistor size scaling down 30% per generation according to scaling theory. Notice that the total transistor size, which is the sum of all transistor widths, decreases by about 30%. The lower graph shows average transistor sizes (in arbitrary units), which also decrease by about 30%, ruling out any suspicion about oversizing and overdesign.

We can conclude that the twofold frequency improvement each technology generation is primarily due to two factors:

- The reduced number of gates employed in a clock period makes the design more pipelined.
- Advanced circuit techniques reduce the average gate delay beyond 30% per generation.

Transistor density

Transistor density is the number of logic transistors in a unit of area. Transistor density should double every technology generation, since according to scaling theory, area decreases by 50%. Memory density has been scaling as expected, and therefore this study focuses on logic transistor density.

Figure 3 plots the logic transistor density of several microprocessors, showing their compaction across different technologies. The dotted line shows the expected density-doubling trend. Notice that when a processor design is ported to the next process technology, it meets the density goal; however, a new processor microarchitecture implemented in the same technology shows a drop in density. This may be due to the complexity of the new microarchitectures, as well as to the limited resources available to accomplish a more complex design.

Interconnection-scaling trends

To meet technology goals, the interconnection system must scale accordingly. In general, as interconnection width and thickness

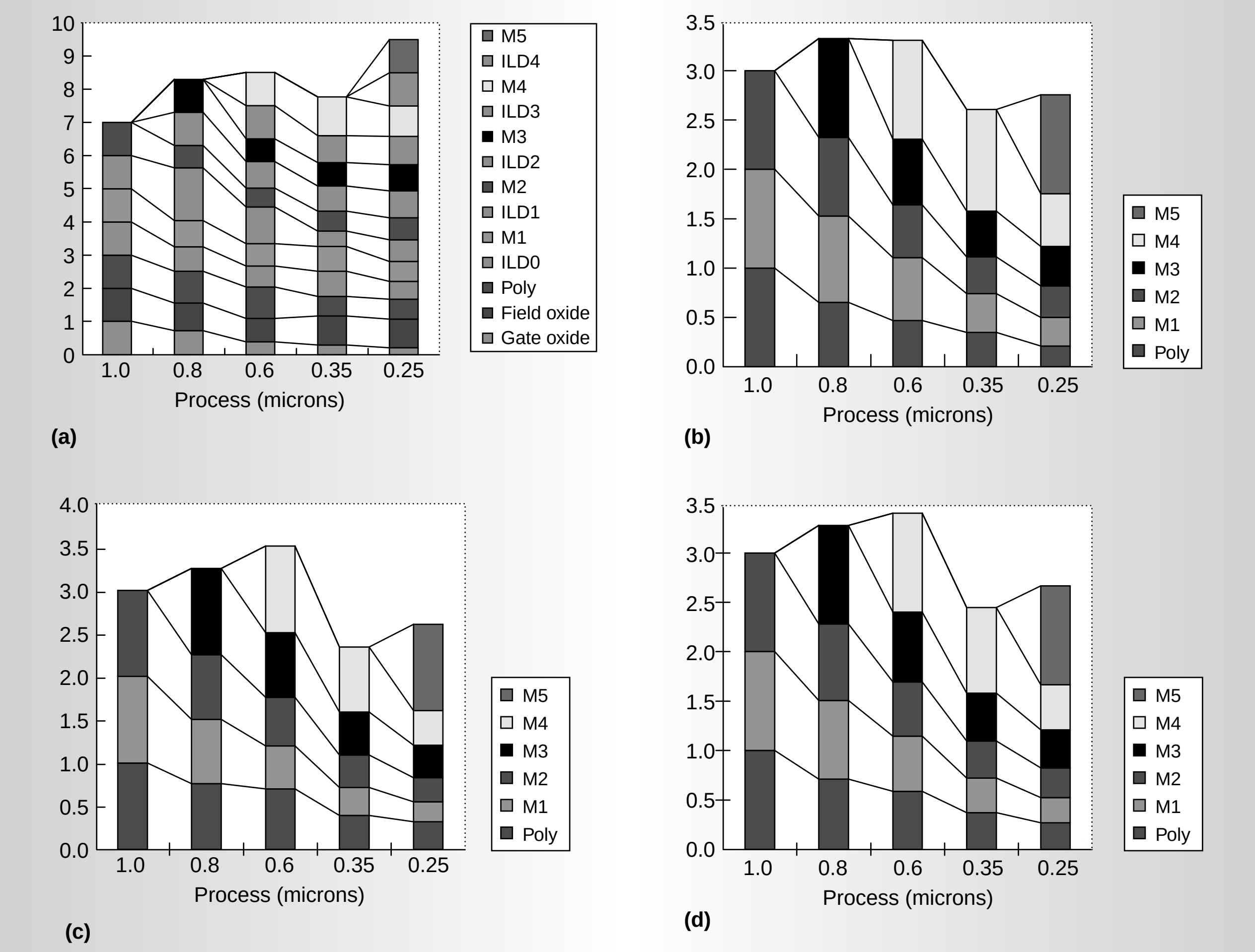

Figure 4. Interconnection scaling: interconnection stacking (a); relative minimum width (b); relative minimum spacing (c); relative minimum pitch (d).

decrease, resistance increases, and as interconnections become denser, capacitance increases. Although chip size should decrease by 30% in each successive technology, new designs add more transistors to further exploit integration. As a result, the average die size of a chip tends to increase over time. To account for increased parasitics (resistance and capacitance), integration, and complexity, manufacturers add more interconnection layers. The thinner, tighter layers are used for local interconnections, and the new thicker and sparser layers are used for global interconnections and power distribution. Figure 4 shows interconnection-scaling trends on a relative scale. Notice that interconnections seem to be scaling normally.

Does advancement in a microarchitecture make the interconnection system more complex? If so, this could explain why new microarchitectures decrease in density. Figure 5 shows interconnection distribution extracted from

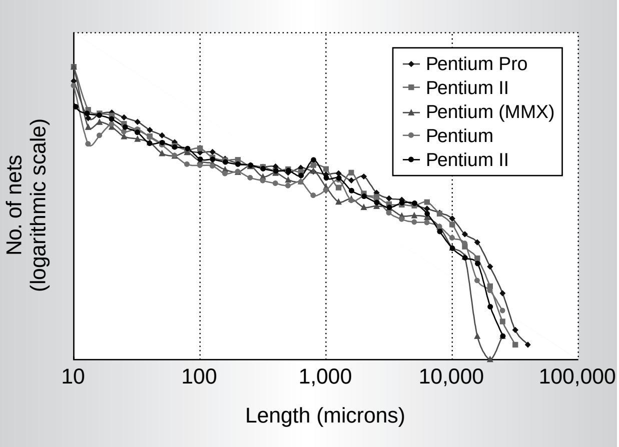

Figure 5. Interconnection distribution.

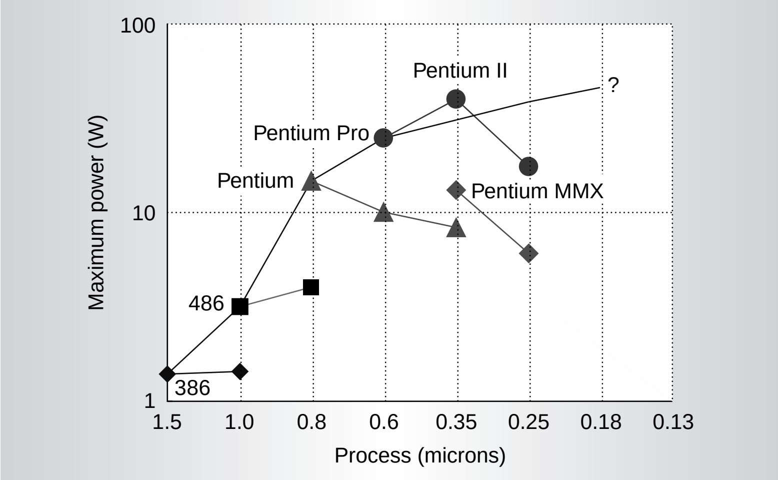

Figure 6. Maximum thermal power dissipation.

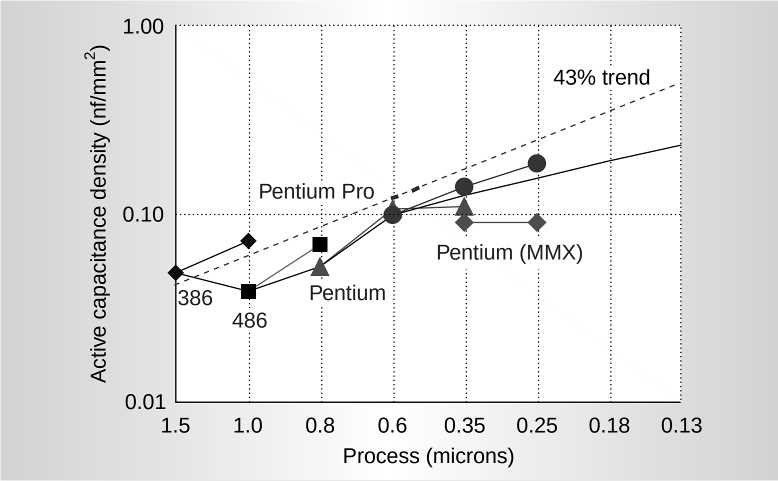

Figure 7. Active capacitance density.

several microprocessor chips employing different microarchitectures. On the *y*-axis (log scale), the number of interconnections is plotted against the length of the interconnections on the *x*-axis. The graph shows that interconnection distribution does not change significantly with advances in microarchitecture. Hence, complexity can be ruled out as the reason for the drop in density, and interconnection distribution seems to follow the trend.

Power

A chip's maximum power consumption depends on its technology as well as its implementation. According to scaling theory, a design ported to the next-generation technology should operate at 43% higher frequency. If the supply voltage remains constant (constant voltage scaling), the power should remain the same. On the other hand, if the supply voltage scales down by 30% (constant electric field scaling), the power should decrease by 50%:

$$V_{DD} = 1 \text{ (constant voltage scaling)}$$
$$\text{Power} = C \times V^2 \times f = 0.7 \times 1 \times (1/0.7) = 1$$
$$V_{DD} = 0.7 \text{ (constant electric field scaling)}$$
$$\text{Power} = C \times V^2 \times f = 0.7 \times 0.7^2 \times (1/0.7) = 0.5$$

Figure 6 plots the maximum thermal power dissipation of several microprocessors in successive technologies. The technologies used constant voltage scaling until reaching 0.8 microns and used constant electric field scaling thereafter. Therefore, power increased dramatically up to 0.8 microns, when the increase slowed. Notice that microprocessors ported to next-generation technologies with constant voltage scaling do not show a decrease in power; the power remains constant. On the other hand, microprocessors ported to technologies using constant electric field scaling decrease in power. This is consistent with scaling theory.

The power dissipation of a chip depends not only on its technology, but also on its implementation—that is, on size, circuit style, microarchitecture, operation frequency, and so on. Hence, to better understand power trends, it is necessary to normalize by introducing the notion of active capacitance, a fictitious equivalent capacitance responsible for power dissipation. We can further normalize the active capacitance to the chip area, in a metric called the active capacitance density—capacitance per unit of area responsible for power dissipation.

$$\begin{matrix}\text{Active}\\ \text{capacitance}\end{matrix} = \frac{\text{Power}}{V_{DD}{}^2 \times \text{frequency}}$$

$$\begin{matrix}\text{Active}\\ \text{capacitance}\\ \text{density}\end{matrix} = \frac{\text{Active capacitance}}{\text{Area}}$$

Figure 7 plots the active capacitance density of several microprocessors in different technologies. From scaling theory, we expect active capacitance density to increase by 43% per technology generation. The figure shows that the increase is on the order of 30% to 35%, not 43%, due to lower density. That is, in practice the microprocessors do not achieve a twofold improvement of logic transistor density between technologies, resulting in a lower active capacitance density.

Die size trends

Manufacturers have not only taken advantage of increased transistor density, but have also increased chip (die) size to further the level of integration. Microprocessor die size tends to grow about 25% per technology generation. Loosely speaking, this satisfies Moore's law.

Projections

So far, we have seen trends in several aspects of microprocessor technologies and characteristics. Let's assume that these trends continue—that is, frequency doubles, supply voltage scales down 30%, active capacitance grows 30% to 35%, and die size grows 25%. Now we can speculate on power dissipation and supply currents.

Figure 8 plots the computed power dissipation of future microprocessor chips if the trends continue. If supply voltage scales down, power dissipation will increase from 100 W in 1999 to about 2,000 W in 2010; otherwise, it could reach approximately 10,000 W! So far, this analysis considers only active power and neglects leakage power. Leakage has not been highly significant in the past, but it will be significant in the future.

Figure 9 shows supply current projections. Supply current will grow from 100 A to about 3,000 A if supply voltage scales down; otherwise, it will become even higher.

To bring power dissipation within reasonable range, we will have to restrict die size. With die size restricted to about 15 mm (a small die), power will stay around 100 W, and supply current will grow to about 300 A. A larger die, about 22 mm, will consume about 200 W of power with a supply current of about 500 A. These are reasonable targets that we can realize in practice.

Energy-delay trade-offs

Reducing supply voltage, frequency, and/or die size will reduce power. All of these reductions reduce chip performance. That is, one has to trade off performance to reduce power. Therefore, we must ask whether the primary technology goal (30% delay reduction) makes sense. Why not set a goal that includes delay as well as power? A good metric for this purpose would be the energy-delay product. We can set goals to achieve a lower energy-delay product and make technology decisions as discussed in Gonzalez et al.[2] The choices are as follows:

$V_{DD} = 1$ (constant voltage scaling)
Energy $= C \times V^2 = 0.7 \times 1 = 0.7$, delay $= 0.7$
Energy $\times$ delay $= 0.7^2 = 0.5$

$V_{DD} = 0.7$ (constant electric field scaling)
Energy $= C \times V^2 = 0.7 \times 0.7^2 = 0.7^3$, delay $= 0.7$
Energy $\times$ delay $= 0.7^4 = 0.25$

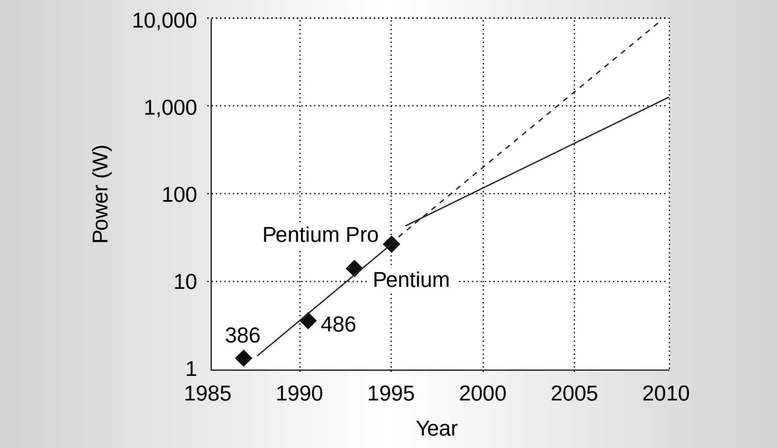

Figure 8. Power dissipation projections.

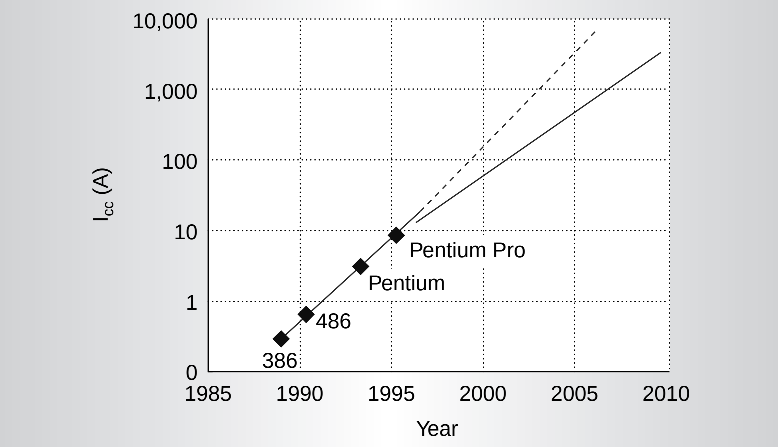

Figure 9. Supply current projections.

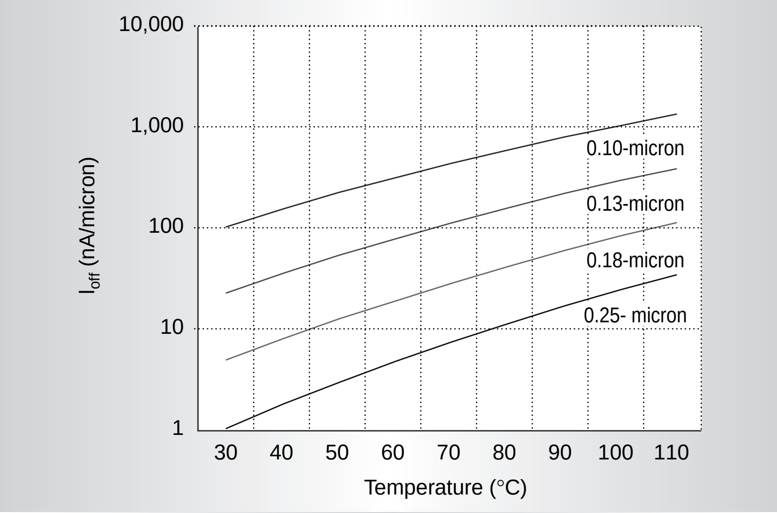

Figure 10. Projected I_{off}.

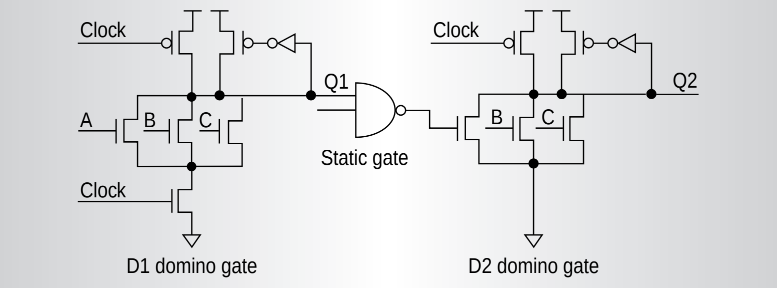

Figure 11. Domino circuit.

Clearly, constant electric field scaling (supply voltage scaling) gives the lower energy-delay product (ignoring leakage energy) and hence is preferable. However, it requires scaling threshold voltage (V_T) as well, which increases the subthreshold leakage current, thus increasing the chip's leakage power.

Subthreshold leakage

Now we attempt to estimate the subthreshold leakage power of future chips, starting with the 0.25-micron technology described in Bohr et al.,[3] and projecting subthreshold leakage currents for 0.18-, 0.13-, and 0.1-micron technologies. Assume that 0.25-micron technology has a V_T of 450 mV, and I_{off} is around 1 nA per micron at 30°C. Also assume that subthreshold slopes are 80 and 100 mV per decade at 30°C and 100°C respectively. Assume that V_T decreases by 15% per generation, and I_{off} increases by 5 times each generation. Since I_{off} increases exponentially with temperature, it is important to consider leakage currents and leakage power as a function of temperature. Figure 10 shows projected I_{off} (as a function of temperature) for the four different technologies.

Next we use these projected I_{off} values to estimate the active leakage power of a 15-mm die and compare the active leakage power with the active power. The total transistor width on the die increases around 50% each technology generation; hence, the total leakage current increases about 7.5 times. This results in the chip's leakage power increasing about 5 times each generation. Since active power remains constant (according to scaling theory), leakage power will become a significant portion of total power.

Notice that it is possible to substantially reduce leakage power, and hence overall power, by reducing the die temperature. Therefore, better cooling techniques will be more critical in advanced deep-submicron technologies to control both active leakage power and total power.

Impact of scaling on circuits

Supply voltage scaling increases subthreshold leakage currents, increases leakage power, and poses numerous challenges in the design of special circuits.

Domino circuits (Figure 11), for example, are widely used to achieve high performance. A domino gate typically reduces delay 30% compared with a static gate, but it consumes 50% more power. A domino circuit also takes less space because the logic is implemented with N transistors, and most of the complementary P stack is absent. As the threshold voltage decreases, the noise margin decreases. To compensate, the size of the keeper P transistor must increase, in turn increasing the contention current and consequently reducing the gate's performance. Overall, the domino's advantage over static logic will continue to decrease. This effect is not restricted to domino logic alone; supply voltage scaling will affect most special circuits, such as sense amplifiers and programmable logic arrays.

Soft errors

Soft errors (single-event upsets) are caused by alpha particles in the chip material and by cosmic rays from space. Since capacitance and voltages will decrease in future technologies, a smaller charge (Q = C × V) will be needed to flip a bit in memory. Therefore, the soft error rate will increase. Attempting to reduce the soft error rate by increasing capacitance on the node will result in reduced performance.

Typically, we protect data in memory with parity or error-correcting codes, but there is no mechanism to protect latches and flip-flops that store state in random logic. The increased soft error rate will have a detrimental effect on logic latches, a problem that needs more investigation.

Power density

Power density is the power dissipated by the chip per unit of area, in W/cm^2. Figure 12 plots the power density of microprocessor chips in different technology generations. Chips in 0.6-micron technology have surpassed the power density of a kitchen hot plate's heating coil, and clearly the trend is increasing. Controlling die temperature and keeping it low are essential for better performance and lower leakage. Controlling power density will be even more crucial for leakage control in deep-submicron technologies.

Trends in performance, density, and power follow scaling theory. If these trends continue, power delivery and dissipation will be the primary limiters of performance and integration. To overcome these limiters, we must constrain die size growth and continue to scale supply voltage. We will have to scale threshold voltage to meet performance demands, thus increasing subthreshold leakage current, limiting functionality of special circuits, increasing leakage power, and increasing soft error susceptibility. These are among the major challenges that circuit designers will face in future technologies. MICRO

Acknowledgments

I thank my colleagues Vivek De, K. Soumyanath, Ali Keshavarzi, Ian Young, and several others for providing insight into topics discussed in this article. I also thank Bill Holt, Ricardo Suarez, Fred Pollack, and Richard Wirt for their continued support and encouragement.

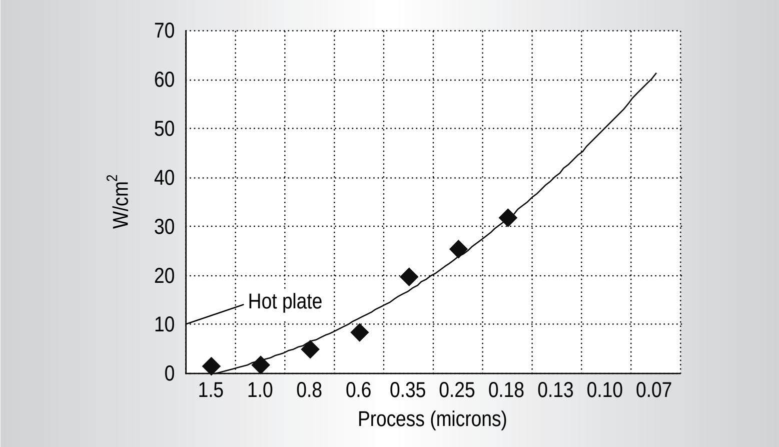

Figure 12. Chip power density.

References

1. Intel Corp., http://www.intel.com.
2. R. Gonzalez, B. Gordon, and M. Horowitz, "Supply and Threshold Voltage Scaling for Low-Power CMOS," *IEEE J. Solid-State Circuits*, Vol. 32, No. 8, Aug. 1997, pp. 1210-1216.
3. M. Bohr et al., *Proc. Int'l Electron Devices Meeting*, IEEE Electron Device Society, 1996, pp. 843-850.

Shekhar Borkar is the director of Intel's Circuit Research Laboratory, where he leads research on low-power circuits and high-speed signaling. He is also an adjunct faculty member at the Oregon Graduate Institute, teaching digital CMOS VLSI design. Borkar received a master's degree in physics from the University of Bombay and an MSEE from the University of Notre Dame.

Send questions and comments to Shekhar Borkar, 5200 NE Elam Young Parkway, EY2-07, Hillsboro, OR, 97124; shekhar@hf.intel.com.