Technology and Design Challenges for Low Power and High Performance

(Invited Paper)

Vivek De and Shekhar Borkar
Intel Corporation
MicroComputer Research Labs
Hillsboro, OR 97124

ABSTRACT

We discuss key barriers to continued scaling of supply voltage and technology for microprocessors to achieve low-power and high-performance. In particular, we focus on short-channel effects, device parameter variations, excessive subthreshold and gate oxide leakage, as the main obstacles dictated by fundamental device physics. Functionality of special circuits in the presence of high leakage, SRAM cell stability, bit line delay scaling, and power consumption in clocks & interconnects, will be the primary design challenges in the future. Soft error rate control and power delivery pose additional challenges. All of these problems are further compounded by the rapidly escalating complexity of microprocessor designs. The excessive leakage problem is particularly severe for battery-operated, high-performance microprocessors.

Keywords

Microprocessor, VLSI design, memory, low-power design

1. INTRODUCTION

Advanced logic CMOS technology, when scaled to the next generation, improves (1) transistor and interconnect performance, (2) transistor density, and (3) energy consumed per switching transition. Technology scaling with 30% reduction in minimum feature size per generation has three primary goals: (1) reduce gate delay by 30%, (2) double transistor density, and (3) reduce energy per transition by 30% to 65%, depending on the degree of supply voltage reduction. These technology improvements, coupled with advances in circuits and microarchitecture, are expected to sustain historical trends in clock frequency, die size, functional integration, and power dissipation of high-performance microprocessors.

In this paper we will take a close look at the past trends in technology scaling and examine how well the technology and products have met these goals. We will use data from various known microprocessors [1][2], especially Intel microprocessors, mainly due to the authors' familiarity with microprocessors; however, this study is equally applicable to other types of designs.

We will project these trends into the future and identify the key barriers stemming from device physics, circuit functionality, heat removal, power delivery, battery life, and soft error rates. All of these challenges are further compounded by the rapidly escalating complexity of microprocessor designs.

ISLPED99, San Diego, CA, USA

2. SCALING TRENDS

We examine the trends in transistor characteristics, clock frequency (at product introduction), transistor density, multilevel interconnections, power, and die size of high-performance microprocessors over the last few technology generations to evaluate how well the technology and design goals, as dictated by the 30% scaling theory, have been met.

2.1 Transistor Scaling Trends

Over the last few technology generations, we have pursued the path of constant electric field scaling, where the maximum supply voltage is limited by gate oxide wear-out. Thus, the gate length, effective electrical gate oxide thickness, and supply voltage have scaled by approximately 30% per unit width. The worst-case subthreshold leakage current has remained approximately constant. Thus, reduction of threshold voltage from one technology to the next has been limited by our ability to (1) control short-channel effects through innovative channel and junction engineering, and (2) aggressively scale the control of all lateral and vertical dimensions, especially gate length. In spite of the stringent subthreshold leakage current limitation, we have been able to maintain approximately constant drive current per unit width of the transistor, at reduced supply voltages. All components of device capacitance, except the areal component of the drain-to-body junction capacitance, remain constant per unit width; the areal component of junction capacitance scales by 30% per unit width. As a result, the delay of a gate whose output load is dominated by device capacitances has reduced by 30% per technology generation. If the number of transistors per die remain constant, and no major transistor resizing is performed, the die area reduces by 50%, and the total capacitance on chip reduces by 30%. The transistor capacitance per unit area, however, increases by approximately 43%.

2.2 Clock Frequency

Assuming that a technology generation spans 2-3 years, the data in Figure 1 clearly shows that the microprocessor frequency has been doubling every technology generation—not just by 43%. This could be attributed to several factors. Let us consider clock-period/average-static-gate-delay plotted on the right hand Y axis in Figure 1 (in arbitrary units). The average number of gate delays in a clock period is decreasing, indicating that the new microarchitectures employ shorter pipelines for static gates, and utilize advanced circuit techniques to reduce the critical path delays even further. This could be one of the reasons why the frequency is doubling every technology generation.

One might suspect that this increase in frequency may be at the expense of over-design or over-sizing of transistors. Figure 2

shows total transistor size, and how the transistor size scales across technologies in different microprocessors.

The dotted lines show transistor size scaling of 30% per generation according to the scaling theory. Notice that the total transistor size, which is the sum of widths of all transistors, scales by about 30%. The lower graph shows average transistor sizes (in arbitrary units), which also scales by about 30%, ruling out any suspicion about over-sizing and over-design.

Therefore we conclude that 2X frequency improvement each technology generation, as opposed to 1.43X, is primarily due to (1) reduced number of gates employed in a clock period, making the design more pipelined, and (2) employing advanced circuit techniques to reduce the average gate delay beyond 30% per generation.

2.3 Transistor Density

We define transistor density as the number of logic transistors packed per unit area. The transistor density is expected to double every technology generation, since the area reduces by 50% (scaling theory). Memory density has been scaling as expected, and therefore we focus our study on the logic transistor density.

Figure 3 plots logic transistor density of some of the microprocessors, and their shrinks and compactions, across different technologies. The dotted line shows the expected 2X density trend. Notice that when a processor design is ported to the next process technology, it meets the density goal; however, a new processor micoarchitecture implemented on the same technology shows a drop in density. We suspect that this is due to the complexity of the new microarchitectures, as well as limited resources available to accomplish a more complex design.

2.4 Interconnect Scaling

In order to meet the technology goals the interconnection system has to scale accordingly. In general, as the width and thickness of the interconnections are reduced, the resistance increases, and as the interconnections get closer, the capacitance increases. The size of a chip should reduce by 30%, which is true for shrinks and compactions of chips onto the next technology; however, to further exploit integration, new designs add more transistors on the chip. As a result, the average die size of a chip tends to increase over time. To account for increased parasitic (R & C), and increased integration and complexity, more interconnect layers are added. The thinner, tighter interconnect layers get used for local interconnections, and the new thicker and sparser layers get used for global interconnections and power distribution.

Figure 4 shows the interconnection scaling trends. From left to right, and top to bottom, the graphs show interconnection stack, widths, spacing, and pitch on a relative scale. Notice that the interconnections seem to be scaling normally.

There is always a question about interconnection distribution—does advancement in the microarchitecture make interconnect system more complex? If so, then this could explain why new microarchitectures drop density as shown in Figure 3.

Figure 5 shows interconnection distribution extracted from several microprocessor chips, employing different microarchitectures. The Y-axis (log scale) has the number of interconnections plotted against the length of the interconnections on the X-axis. It shows that the 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 the interconnection distribution seems to follow the trend.

2.5 Power

Maximum power consumption of a chip depends on the technology as well as implementation. According to the scaling theory, a design when ported to the next generation technology would operate at 43% higher frequency. Since the total capacitance and the maximum supply voltage reduce by 30%, the total energy consumed in a clock cycle must reduce by 65%, assuming that the average number of switching transitions per cycle has remained unchanged. Then, the power should reduce by 50%. If the electric field is below the maximum sustainable for gate oxide reliability, then the supply voltage may not reduce by 30%. In the extreme case, both the maximum supply voltage and threshold voltage will remain constant. The energy consumed per clock cycle then reduces by 30%, and the power remains constant. Of course, there is then no room for adding more transistors to the chip without increasing the power budget from previous generation.

Maximum thermal power dissipation of several microprocessors is plotted against technologies in Figure 6. The technologies employed constant voltage scaling until 0.8µ, and started constant electric field scaling thereafter. That is why the power has increased dramatically until 0.8µ, and the growth has slowed down afterwards. Notice that microprocessors ported to next generation technologies with constant voltage scaling do not show decrease in power; the power remains constant. On the other hand, microprocessors ported to constant electric field scaled technologies do reduce power. This is consistent with the scaling theory discussed before.

The power dissipation of a chip not only depends on the technology, but also on the implementation. That is, the power depends on the size of the chip, circuit style, microarchitecture, frequency of operation, and so on. Hence to better understand trends in power it is necessary to normalize. We introduce the notion of active capacitance, which is a fictitious equivalent "switched" capacitance responsible for power dissipation. This active capacitance is measured as "Power/($V_{DD}^2 \times$ Frequency)".

We can further normalize the active capacitance to the chip area, called the active capacitance density—capacitance per unit area responsible for power dissipation of the chip. Figure 7 plots active capacitance density of several microprocessors in different technologies.

From scaling theory we expect active capacitance density to increase by 43% per technology generation. Figure 7 shows that the increase is of the order of 30-35% and not 43%, which may be attributed to lower density. That is, logic transistor density improvement of 2X between technologies is not achieved in practice, resulting in lower active capacitance density.

2.6 Die Size

We have not only taken advantage of increased transistor density, but have also increased the size of the chip (die) to further the level of integration.

Figure 8 plots die size of lead microprocessors in mils (1 mil = 1/1,000 inch) over time, and shows that the die size tends to grow about 25% per technology generation. Loosely speaking, this satisfies Moore's Law.

3. PROJECTIONS

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

Figure 9 plots computed power of future microprocessor chips if the trends continue. The power dissipation will increase from 100 watts now, to about 2,000 watts in the future if the supply voltage is scaled, otherwise it would be of the order of 10,000 watts! So far the analysis considers only the active power and neglects the leakage power since the leakage has not been that significant in the past; however, it will be in the future.

Figure 10 shows the supply current projections. The supply current will grow from 100 amps now, to about 3,000 amps if the supply voltage is scaled, otherwise it will be even higher.

To bring the power dissipation within reasonable range the die size will have to be restricted. If the die size is restricted to about 15mm (small die) then the power will stay around 100 watts, and the supply current will grow to about 300 amps. A larger die, about 22mm, will consume about 200 watts of power with supply current of about 500 amps. These are reasonable targets and can be realized in practice.

4. ENERGY-DELAY TRADEOFFS

The power can be reduced by scaling down any or all of (1) supply voltage, (2) frequency, and (3) die size. All of these reduce the performance of the chip. Thus, the analysis indicates that one has to tradeoff performance to reduce the power. Therefore one could argue whether the primary technology goal (delay reduction by 30%) makes sense? Why not set a goal that comprehends both delay as well as power? A good metric would be to use energy delay product, set goals to achieve lower ExD product, and make right technology decisions as discussed in [3].

5. BARRIERS TO VOLTAGE SCALING

Clearly, constant electric field scaling (supply voltage scaling) gives the lowest energy delay product (ignoring leakage energy), and hence it is preferred. However, this requires scaling threshold voltage (V_t) as well, which increases the subthreshold leakage current. This impacts both the power consumption, and circuit robustness. Plus, reducing the amount of charge at storage nodes by a factor of two per generation exponentially increases the vulnerability to single-event upsets or soft errors.

5.1 Leakage Power

Now we will attempt to estimate the subthreshold leakage power of the future chips, starting with the 0.25μ technology described in [4], and projecting subthreshold leakage currents for 0.18μ, 0.13μ, and 0.1μ technologies. Assume that 0.25μ technology has V_t of 450 mV, and I_{off} is around 1na/μ at 30°C. Also assume that subthreshold slopes are 80 and 100 mV/decade at 30°C and 100°C respectively, V_t scales by 15% per generation, and I_{off} increases by 5X each technology generation. Since I_{off} increases exponentially with temperature, it is important to consider leakage currents and leakage power as a function of temperature. Figure 11 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 15mm die (small die), and compare the leakage power with the active power. The total transistor width on the die increases by ~50% each technology generation, hence the total leakage current increases by ~7.5X. This results in leakage power of the chip increasing by ~5X each technology generation. Since the active power remains constant (scaling theory) for constant die size, the leakage power will become a significant portion of the total power.

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

5.2 Impact on Circuits

Supply voltage scaling increases subthreshold leakage currents, increases leakage power, and also poses numerous circuit design challenges for special circuits.

Domino circuits, shown in Figure 12, are widely used for high performance. A domino gate typically provides 30% delay reduction over a static gate; however, it consumes 50% more power than a static gate. Domino circuit also takes less space since the logic is implemented using N transistors and the most of the complementary P stack is absent. As the threshold voltage reduces, the noise margin will reduce. To compensate for this, the size of the keeper P transistor needs to be increased, which increases the contention current, consequently reducing performance of the gate. Overall, the effectiveness of Domino over static logic will continue to decrease.

This effect is not restricted to Domino logic alone, but most of the special circuits, such as sense amplifiers and PLA's will be affected. 6T SRAM cells are also vulnerable to "read stability" problems. As the transistor Vt is reduced, the maximum current sinking capability of the pulldown NMOS device, which is in the linear region of operation during read, increases at a weaker rate than the maximum saturation drain current through the access NMOS transistor. Furthermore, the low-Vt devices are prone to larger parameter mismatch, which further degrades stability of the cell. This also poses serious challenges to scaling the offset voltage and delay of the sense amplifiers.

Increase in subthreshold leakage current also impacts bit line delay in large on-chip caches, which use differential low voltage swing sensing. The effective drain current available for differential bitline swing development is the saturation drain current of a single "on" access transistor minus the total subthreshold leakage current of the other 100-200 "off" access transistors hanging from the bitline pair. As Vt reduces, the combined subthreshold leakage current through the 100-200 access transistors becomes comparable to the drive current of a single access transistor; thus the effective drain current available for the bitline discharge reduces drastically. Furthermore, since the offset voltage of the sense amp does not reduce by 30% per generation, the voltage swing on the bit line cannot be scaled at the same rate as the supply voltage. A combination of the above factors causes the bit line delay to scale at a significantly slower

rate than the rest of the circuits. This divergence of logic and cache performance on chip can significantly impact overall processor performance in the future.

5.3 Single Event Upsets – Soft Errors

Soft errors are caused by alpha particles in the material and cosmic rays from space. Since capacitance will reduce, voltages will reduce, a smaller charge ($Q=C\times V$) will be needed to flip a bit in the memory. Therefore the soft error rate will increase. An attempt to reduce the soft error rate by increasing the capacitance on the node will result in performance reduction, which is not desirable.

Data stored in memory is typically protected by parity or ECC (error correcting codes), but latches and flip-flops storing state in random logic are not protected. Increased soft error rate on logic latches would have detrimental effect and needs more investigation.

6. BARRIERS TO TRANSISTOR SCALING

Device physics poses several challenges to future scaling of the bulk MOSFET structure. One key challenge is controlling short-channel effects which are manifested as Vt-roll off and Drain-Induced Barrier Lowering (DIBL). Furthermore, a large contributor to within-die parameter variation is critical dimension (CD) variation. Short-channel effects also increase the sensitivity of device characteristics to CD control.

In order to minimize degradation of device behavior at short channel lengths, the lateral-to-vertical aspect ratio of the device (defined as $L_{eff}/\{(t_{ox}.\varepsilon_{si}/\varepsilon_{ox})^{1/3}(d)^{1/3}(d_j)^{1/3}\}$ [5], L_{eff}: effective channel length; t_{ox}: gate oxide thickness; d: channel depletion depth; d_j: effective junction depth; ε_{si}, ε_{ox}: permittivities of silicon and oxide) must be made as large as possible; then the gate terminal of the transistor has more control on the channel than the source or drain. Ideally, the aspect ratio must never be smaller than what is allowed by a combination of (1) the worst-case subthreshold leakage current constraint, and (2) the CD control capability of a technology generation. If a constant aspect ratio can be maintained from one generation of the technology to next, then short-channel effects remain unchanged; as a result, all the benefits of improved CD control or relaxing worst-case subthreshold leakage constraint can be directly translated to improved drive current. On the other hand, if the aspect ratio degrades, the threshold voltage at nominal channel length would be larger and the available drive current would reduce.

To ensure that the aspect ratio does not degrade severely, the gate oxide thickness, the junction depth and the depletion depth must all scale by 30% per generation. Leakage through the gate oxide by direct band-to-band tunneling limits physical oxide thickness scaling. For example, if the oxide leakage current density is $1A/cm^2$ at 1V, and 20% of a large die area (22mm) is occupied by on-chip oxide decoupling capacitors, then the leakage power due to oxide alone is ~1W. For a battery-operated processor, this amount of leakage power consumption during standby mode would be unacceptable! Poly depletion and quantum effects add ~1nm to the physical gate oxide thickness; thus the smallest electrical oxide thickness achievable is limited to ~1nm even with very high permittivity gate dielectric material. Of course, metal gates can extend this limit to 0.5nm by eliminating poly depletion effects. However, the main difficulty is related to finding silicon-compatible metals with suitable work functions that will yield low-Vt devices of both n- & p-type without overly complicating the fabrication process. Barriers to scaling of the source/drain junction extensions are posed by increase in the parasitic resistance and solid-solubility limited maximum junction doping [6]. Reducing junction depth below 30nm degrades drive current, even though short-channel effect is improved. Finally, the requirement of 15%-30% Vt scaling per technology generation (to maintain a sufficiently large Vdd/Vt ratio) poses limits to scaling of the channel depletion depth by 30%. For uniform channel doping, the depletion depth will be constant for 30% Vt scaling, and will reduce by 18% for 15% Vt scaling. Using an ideal retrograde doping profile reduces the depletion depth by 50% for same Vt, improves channel mobility, but degrades body effect. Halo doping can mask the Vt roll-off behavior, but does not improve DIBL significantly. Furthermore, the lateral fall-off distance of the halo doping profile must scale by 30% per generation in order for it to remain effective.

7. POWER DENSITY

Power density is defined as the power dissipated by the chip per unit area, in $Watts/cm^2$. Figure 13 plots power density of several microprocessor chips in different technologies. The arrow shows power density of a hot plate, about 10 $Watts/cm^2$. Chips in 0.6μ technology generation have surpassed the power density of a hot plate, and clearly the trend is increasing. It is essential to control the die temperature and keep it low for better performance and lower leakage. Controlling power density will be even more crucial for leakage control in the future deep sub-micron technologies.

8. IMPACT ON PRODUCTIVITY

Figure 14 shows complexity growth in terms of logic transistors in a chip. The complexity is increasing at the rate of 58% per year. The design productivity, on the other hand, is improving only at the rate of 21% per year. The circuit design challenges described earlier will further impact productivity, necessitating advanced circuit design capabilities and tools.

9. CONCLUSION

This paper has evaluated past trends in technology. It shows that trends in performance, density, and power have followed the scaling theory. If these trends continue, then power delivery and dissipation will be the biggest limiters. To overcome these limiters, die size growth will have to be constrained, and supply voltage scaling will have to continue. The threshold voltage will have to scale to meet the performance demand, resulting in higher subthreshold leakage current, limiting functionality of special circuits, increasing leakage power, soft error susceptibility, short channel effects, and device parameter variations. These are some of the major challenges that circuit designers will face in the future technologies.

10. ACKNOWLEDGMENTS

We would like to thank our colleagues K. Soumyanath, Ali Keshavarzi, Siva Narendra, Kevin Zhang, Ian Young, and several others for providing insight into topics discussed in this paper. Bill Holt, Ricardo Suarez, Fred Pollack, and Richard Wirt provided continued support and encouragement.

11. REFERENCES

[1] Web sites: Intel Corp*, Digital Equipment Corp (Compaq), IBM.

[2] Microprocessor Reports (MDR).

[3] Gonzalez et al, IEEE Journal of Solid State Circuits, Aug 97.

[4] M. Bohr et al, IEDM 96.

[5] C. Wann et. al., IEEE Trans. Electron Dev, Oct 96.

[6] S. Thompson et. al., 1998 Symp. VLSI Tech, Jun 98

* Pentium® is a registered trademark of Intel Corporation. Third-party trademarks are property of their respective owners.

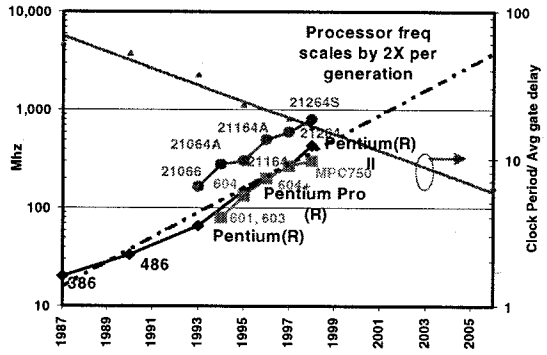

Figure 1: Microprocessor Frequency

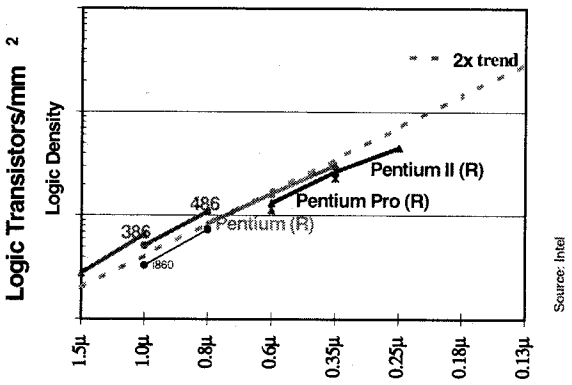

Figure 3: Logic transistor density

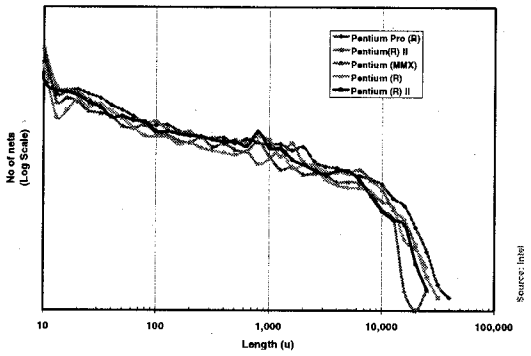

Figure 5: Interconnection distribution

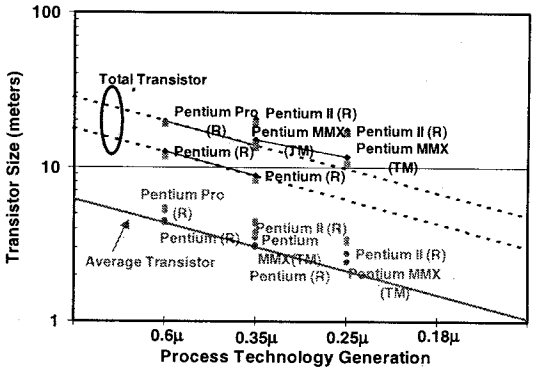

Figure 2: Transistor size scaling

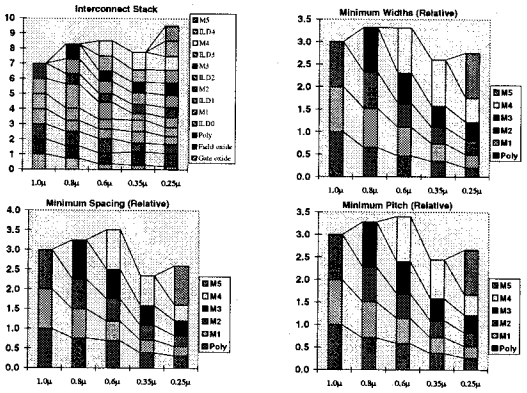

Figure 4: Interconnection scaling

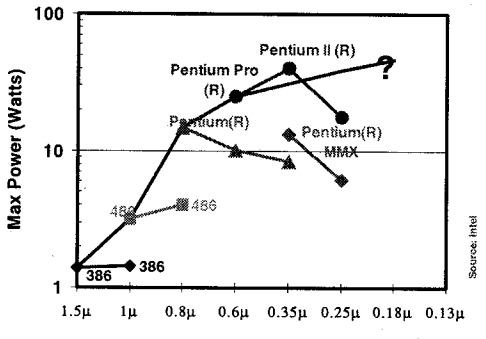

Figure 6: Power

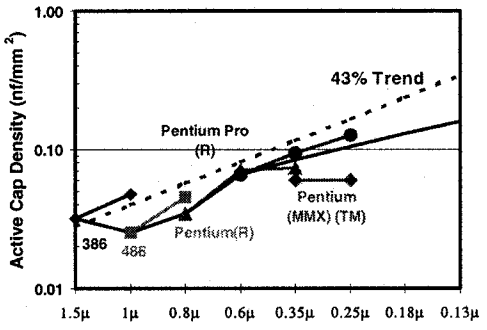

Figure 7: Active capacitance density

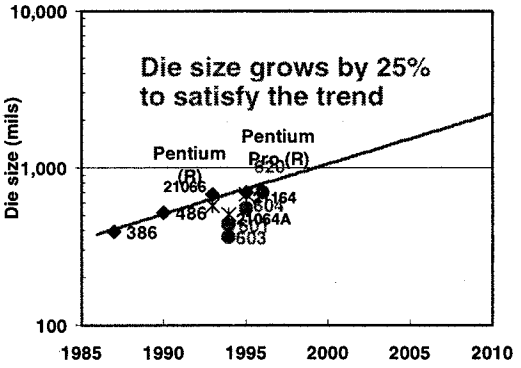

Figure 8: Die size growth

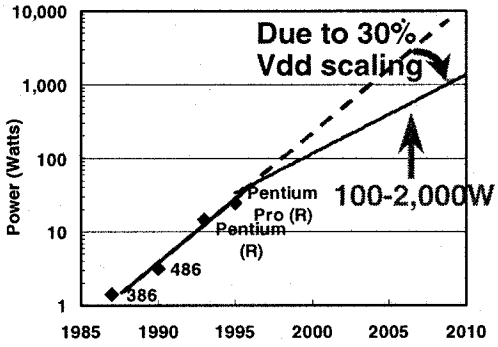

Figure 9: Power projections

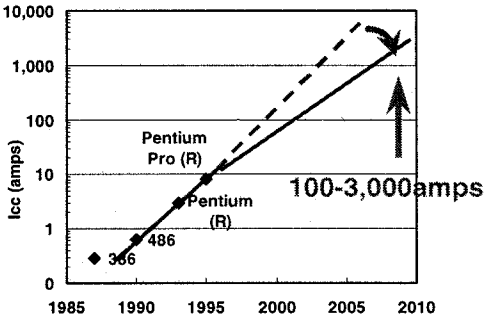

Figure 10: Supply current

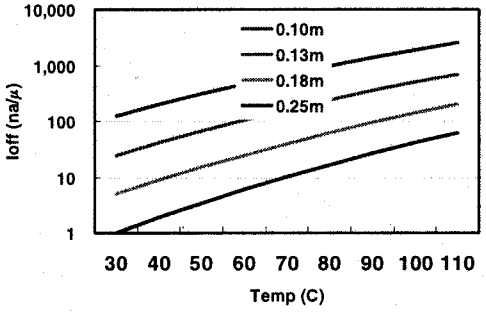

Figure 11: Projected off currents

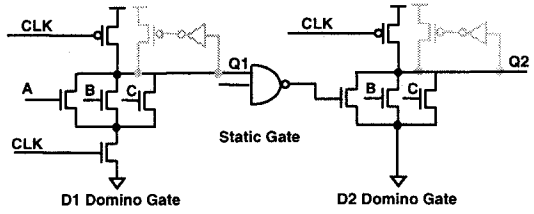

Figure 12: Domino Logic

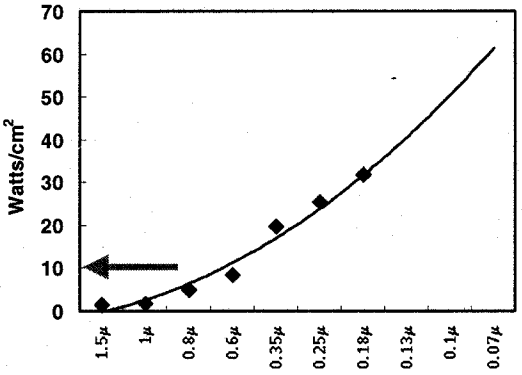

Figure 13: Power density

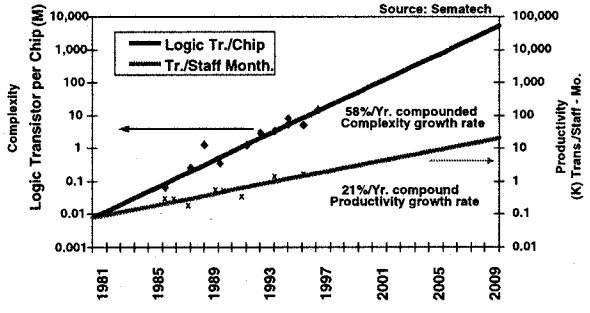

Figure 14: Design productivity