International Journal of High Speed Electronics and Systems
Vol. 16, No. 1 (2006) 3–8
© World Scientific Publishing Company
World Scientific
www.worldscientific.com

ARE WE AT THE END OF CMOS SCALING?

GHAVAM G. SHAHIDI

IBMT.J. Watson Research Center
Yorktown Heights, NY 10598, USA

Received 6 February 2005
Accepted 2 June 2005

CMOS scaling enabled by advances in lithography has been behind the information revolution. Over the last 15 years, there has been a new CMOS technology node approximately every two years. The key feature of every node has been 2X density shrink and ~35% performance gain per technology node. At 90 nm node a number usual knobs that have enabled the scaling have approached their limits. Furthermore chip power (both active and stand-by) has been increasing rapidly, approaching air cool limit. Chip stand-by power, which was negligible a few years ago, is now about the same order of magnitude as the active power in high end microprocessors.

In this talk it will be argued that because of power density limitation of 90 nm, 65 nm, and beyond nodes, performance and ability to shrink are more than ever linked, and in fact if the performance gain would significantly slow down (for the designs that operate at the existing cooling limit). It is more than ever critical to come up with technology features that will enhance the performance, at a given device leakage.

1. Introduction

Over the last two decades, Silicon CMOS scaling had been the foundation of the information revolution. CMOS scaling refers to the advances in the semiconductor technology, which are in turn responsible for enhancing the system performance. There are two key aspects to the scaling: density and performance. Density scaling or Moore's Law, enabled by advances in lithography, is based on the observation in 1975 that the number of components in a chip has been doubling every year [1]. The trend has been holding to this date. The second aspect of scaling is the performance improvement, based on Dennard's scaling theory [2]: As the MOS transistor becomes smaller, it becomes faster and consumes less power. Over the last few CMOS generations, 2X density shrink and ~35% performance gain per technology node were obtained. At 90 nm node a number usual knobs that have enabled the scaling have approached their limits. Chip stand-by power is now in 10's of Watts and about the same order of magnitude as the active power. Total chip power and power density are approaching the air-cool limit. In this paper, it is argued that the "device performance scaling" will be slowed down at 65 nm and beyond nodes, because a number of enabling knobs have approached their limits. If there significant slowdown in performance gain, then scaling will become limited to density improvement (for the designs that are not power density limited). Unless a number of technology features increase the performance without increasing the leakage come to fruition, power density may approach unmanageable value, which in turn may limit density scaling also.

2. Performance Scaling

Over the last 15 years, CMOS device performance has been improving by about 34% per generation (Fig. 1). This performance improvement is based on Dennard's theory [2] that states as the MOS transistor dimensions and the supply voltage are scaled (channel length, junction, depth, gate oxide) by α, its speed improves by α, while the power density remains constant. Table 1 lists a number of the key features covering nodes 250 nm to 90 nm. Channel length and the oxide thickness scaling are as expected. Indeed the channel length and dielectric thickness scaling sped up in 180 and 130 nm nodes (compared to the earlier predictions of ITRS).

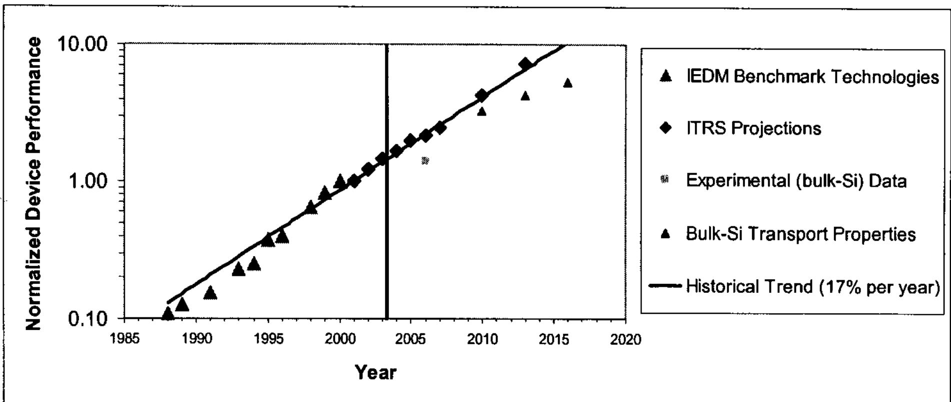

Figure 1- Device performance improvement as function of time: 17% per year, or about 34% per node, assuming a node per 2 years [courtesy of D. A. Antoniadis].

Node	Features	Density	Performance	Enhancements	Reference
250	Tox: 28 A Lpoly: 120 nm Leakage: ~2 nA	0.5X	34%	Cu	[3]
180	Tox: 21 A Lpoly: 85 nm Leakage: ~20 nA	0.5X	34%	Thinner SOI Cu	[4], [5]
130	Tox: 15 A Lpoly: 55 nm Leakage: ~100 nA	0.5X	34%	Thinner SOI	[6], [7]
90	Tox: 11-12 A Lpoly: 45 nm Leakage: ~300 nA	~0.5X	~34%	Strain Low K BEOL	[8],[9]

Table 1: Key features of CMOS technology nodes, 250 nm to 90 nm

Performance gain and the density scaling also follow the historical trends. In addition to the scaling the device, a number of companies have introduced technology elements (copper, Silicon on Insulator, strained Silicon, low-k back-end of the line dielectric) to enhance the chip performance beyond that enabled by the simple device scaling. At 90 nm, the gate oxide thickness is about ~1.1-1.2 nm, and has leakage of 100's of A/cm^2 and probably can not be much reduced beyond its value in 90 nm. Scaling of channel length, without degradation in its performance, is increasingly difficult due to finite as implanted dopant gradient. High end microprocessors have been the driver of high performance CMOS technology. Their power and power density is approaching 100 W/cm^2, which is the air cooling limit (Fig. 2).

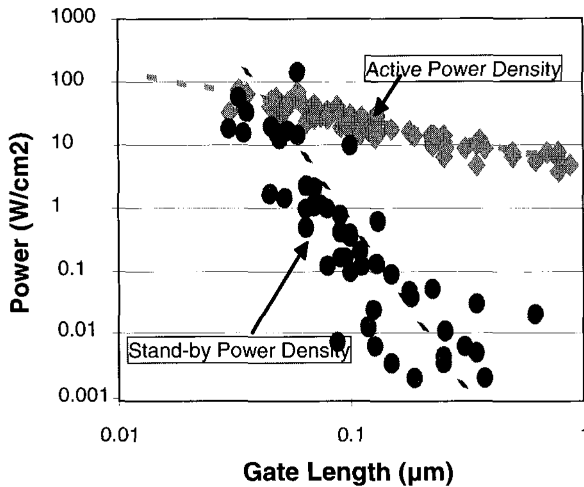

Figure 2- Stand-by and active power density of various chips.

Indeed, both power and power density have been increasing steadily over the last few years (Fig. 2). The increase in active power is the result of increased switching activity and parallelism in every clock cycle. Increase in device off current has been another knob used in increasing the chip performance. Operating the chip at lower threshold (i.e. higher off current), has become a way of running the chip at higher performance (and higher stand-by power). This knob which is responsible for the unsettling trend is the increase in stand-by power, is the result of the non-scalability of threshold voltage.: Until the 250 nm node, the device off current has been steady at about 1-2 nA/μm. Since 180 nm, the off current has been increasing, where in 90 nm, at minimum channel length, the off current is a few 100 nA/μm.

The rapid rise in the off current, when coupled with the variability of channel length (i.e. Across Chip Line-width variation), causes a shaper increase in the stand-by power of chip: Figure 3a, is a typical off current in a generic 130 nm technology (nominal channel length of 55 nm). Assuming a Gaussian distribution of gate length with a sigma of 1, 3 and 5 nm, Fig. 3b shows the normalized (to nominal stand-by power at nominal L) stand-

by power as the chip is run at shorter channel length. It is noted that the stand-by power rises very rapidly. In other words, if the stand-by power of a chip is in 10's of Watts, as many chips are today, the chip can not be run at much shorter L due to the rapid increase in the stand-by power, and the off current at nominal or minimum L can not be much more increased in order to get more performance, without tighter poly distribution.

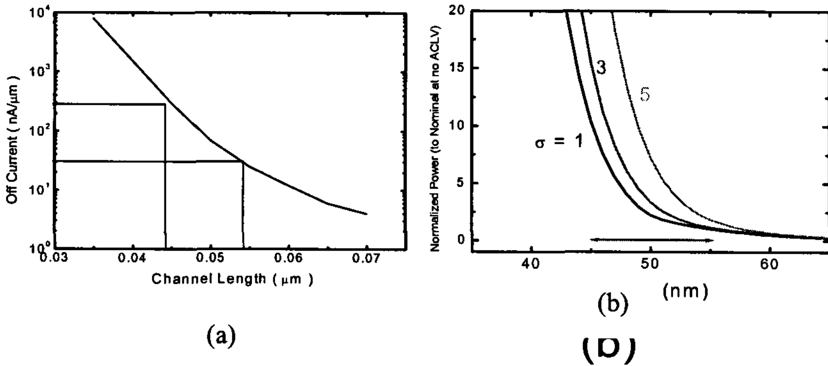

Figure 3- Device off current for a typical 130 nm technology with nominal L of 55 nm, and the corresponding normalized chip stand-by power, assuming a Gaussian distribution for the gate length with sigma of 1, 3 and 5 nm.

Considering the limitations posed by the power, and saturation of many knobs used in scaling, it is becoming increasing difficult to increase the chip performance (i.e. frequency) by simple scaling. An example of this limitation has been the challenge of increasing Intel's Pentium IV frequency to 4 GHz in 90 nm (Fig 4). Indeed in going from 180 to 130, Pentium IV had a significant increase in frequency. Migrating to 90 nm, the gain gas been much smaller. This is caused by power and the slowdown in CMOS scaling.

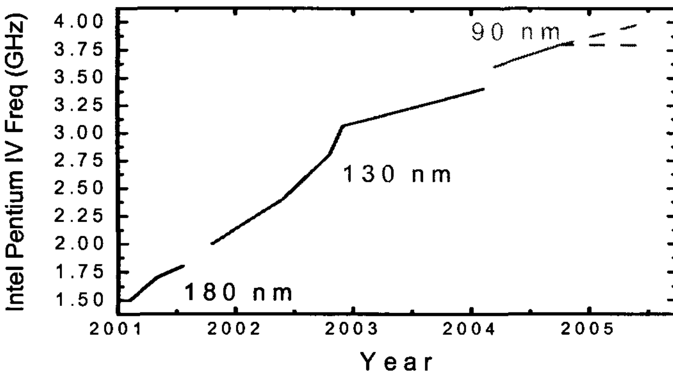

Figure 4- Frequency of Intel's Pentium 4, at 180, 130, and 90 nm.

3. Scaling Beyond 90 nm

A number of companies have announced the elements of their 65 nm node technology. The key feature of 65 nm technologies that have been announced so far is that they have the same oxide thickness as the 90 nm technology, and they enhance a number of knobs that were employed in 90 nm (i.e. more strain.). At the device level, device performance enhancement of 15-30% as compared to 90 nm has been reported. It remains to be seen how much of the device enhancement can be translated to the enhanced chip performance. Because of higher density, unless the technology has noticeable device enhancement and the variability is under control, it is very difficult to obtain chip level performance gain, especially for high power chips in 90 nm technology.

Node	Features	Density	Performance	Enhancements	Reference
90	Tox: 11-12 A Tinv: ~19-20A Lpoly: 45 nm Leakage: ~300 nA	~0.5X	~34%	Strain Low K BEOL	[8],[9]
65	Tox: Same tInv: Same Lpoly: 35-45 nm Leakage: Same	>0.5X	15-30%	More strain	[10],[11]
45	Tinv: 14-15 A (?) Lpoly: ~35 nm Leakage: Same	>0.5X	?		
32	Tinv: 12 A (?) Lpoly: ~30 nm Leakage: Same	>0.5X	?		

Table 2: Key features of CMOS technology nodes, beyond 90 nm

At 45 nm, a number of companies have expressed their intention to use high K dielectric (and thus operate at lower inversion thickness). As of this writing, the task of introducing high K for the 45 nm node appears to be daunting (no company as of yet have introduced gate stack, which uses high K and meets the requirements for inversion thickness, mobility, and the work function). Going beyond high K, a number of new substrate material and device structures are in work (i.e. hybrid orientation technology, ultra-thin SOI, FINFET's, biaxial strain, Ge substrate, etc.). It remains to be seen if any of these newer materials and structures can be brought to the level in time that they will meet the requirements for a post 45 nm node.

4. Summary

Looking at scaling as means to enhance the system performance through density and/or device performance enhancement, we are approaching a milestone: The knobs that have

enabled the "classical scaling" are at their limits. Furthermore, power limitations are another barrier to take advantage of "area-scaling" aspect of CMOS scaling (especially to high end microprocessors). A number of novel material and structures are in works, which may result in extending the device enhancement and thus scaling. It remains to see if they will be ready in time for the immediate upcoming nodes. Many companies are considering design and system approaches to enhance the system performance (such as lower power multi-cores).

5. References

1- G. Moore, "Progress in digital integrated electronics", IEDM Tech. Dig., p. 11, 1975.

2- R. H. Dennard, *et al*, "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions," IEEE J. Solid-State Circuits SC-9, p. 256, 1974.

3- D. Edelstein, et al, "Full copper wiring in a sub-0.25μm CMOS ULSI technology", IEDM Tech. Dig., pp. 773, 1997.

4- S. Yang, et al, "A high performance 180 nm generation logic technology", IEDM Tech. Dig., pp. 197, 1998.

5- D. J. Schepis, et al, "A 0.25μm CMOS SOI technology and its application to 4 Mb SRAM", IEDM Tech. Dig., p. 587, 1997.

6- E. Leobandung, et al, "Scalability of SOI technology into 0.13μm 1.2V CMOS generation", IEDM Tech. Dig., p. 403, 1998.

7- S. Thompson, et al, "An enhanced 130 nm generation logic technology featuring 60 nm transistors optimized for high performance and low power at 0.7 - 1.4 V", IEDM Tech. Dig., p. 257, 2001.

8- H. S. Yang, et al, "Dual stress liner for high performance sub-45nm gate length SOI CMOS manufacturing", Tech. Dig., p. 1075, 2004.

9- T. Ghani, et al, "A 90nm high volume manufacturing logic technology featuring novel 45nm gate length strained silicon CMOS transistors", IEDM Tech. Dig., p. 978, 2003.

10- E. Leobandung, et al, "High Performance 65 nm SOI Technology with Dual Stress Liner and low capacitance SRAM cell", VLSI Tech Dig, 2005.

11- P. Bai, et al, "A 65nm logic technology featuring 35nm gate lengths, enhanced channel strain, 8 Cu interconnect layers, low-k ILD and 0.57 μm2 SRAM Cell", IEDM Tech. Dig., p. 657, 2004.