A look into the future of nanoelectronics

Gilbert Declerck

IMEC
Kapeldreef 75
3001 Leuven, Belgium

Abstract

On the occasion of the 25th anniversary of the VLSI Symposium it is appropriate to reflect on the past and peer into the future. It is clear that continuing scaling in the coming decade will no longer be an evolution but rather a revolution involving materials science based engineering at the device level and computer science at the systems level – not "living apart together" but intensely interlinked in the design world. The challenges are daunting: materials and device breakthroughs, innovations in circuit and system architecture, new design tools and skills are urgently needed if we are to reconcile nanoscale realities with the promises of nomadic connectivity and "embedded-everywhere" systems. Close interactions among a multitude of disciplines are mandatory. But while "getting it all together" is not for the faint of heart, life has never been more exciting for the scientist with a sharp eye and an open mind.

The era of happy scaling and its mechanisms

Ever since Moore's original statement in 1965 [1] the spectacular growth of the IC sector has been associated with "Moore's law". In his own words this law was associated with "unit cost falling as the number of components per circuit rises". Forty years later, unit cost is still falling with the number of components, and it appears that the "law" will remain in place for some time to come (fig. 1).

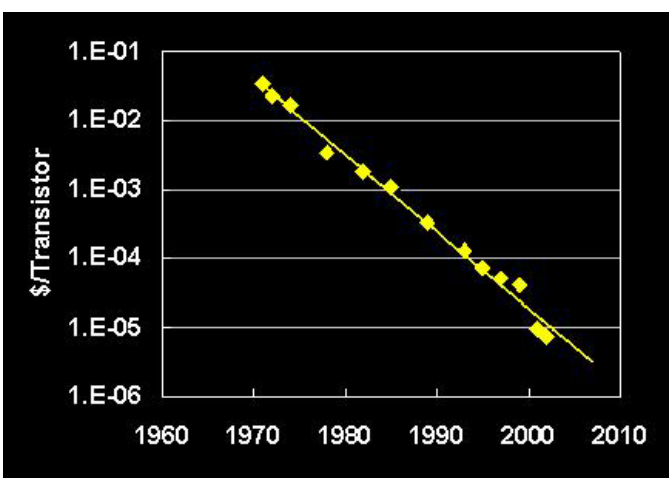


Figure 1: Decrease of average transistor price over the period 1968 – 2002 (Source: Intel/ Dataquest).

The underlying mechanism can be understood using a simple model which we call "Moore's clock". Its two main features are found in any well-behaved watch i.e. a spring and a pendulum. The spring provides the driving force that keeps the wheelwork running. In Moore's clock, this drive is provided by the set of MOSFET scaling rules first put forward by Dennard et al. [2] some 6 years after Moore's initial paper, and which has shown almost the same remarkable endurance over time as the "law" itself. With size reduction now spanning over two orders of magnitude, the persistence of scaling algorithms as applied to CMOS is a truly unique occurrence in the history of technology.

It should be realized that scaling rules only apply to spatial dimensions and do not define any timescale for the miniaturization process. Therefore, Moore's clock also needs a pendulum, the periodic motion of which will determine the timescale of miniaturization. In contrast to the spring, the pendulum is not solely based on technology but also on business development. As such, it is closely linked with the microeconomic base cycle of the IC industry.

Until 2000, this cyclical evolution was driven by the growing computing power of the PC and the capability of the related software making full use of this enormous power and ever increasing functionalities (more data storage, performing games, internet access, …). These times of "happy scaling" were characterized by:

1. downscaling of component size resulting in a decreasing cost/function;
2. maintaining the structure of the basic transistor building block;
3. increasing performance measured in clock frequency, while maintaining the basic architecture and instruction set;
4. fully maintaining the "0" and "1" abstraction for digital operations, which includes the dominance of the transistor delay/energy over the local interconnect contribution. As a result, the design tools and design methods did not need to change; hence, technology and design could be decoupled;
5. yield mainly determined by process quality (i.e. not by design or variability);
6. total energy per function (at the application level) decreasing.

The near-perfect synchronization of power x delay product improvement, cost/bit reduction and functionality increase

resulted in a steadfast grip of the technological spring on the commercial pendulum, thereby insuring the smooth running of Moore's clock through several generations of scaling [3]. However, we now face the issue of Moore's clock falling apart, as the downsizing of the components no longer guarantees the combined bonuses of higher performance and lower cost. The happy scaling days are over!

Changing paradigms challenge Moore

Power consumption is now a major constraint for our industry. Another challenge is the set of emerging applications that introduce new performance metrics. And as we are entering the "late-CMOS" age, we see a diversification of process options while component behavior becomes less predictable, which necessitates a close relationship between system designer and process engineer. We will discuss each of these aspects. A schematic representation of some coming challenges as a function of time is shown in fig. 2.

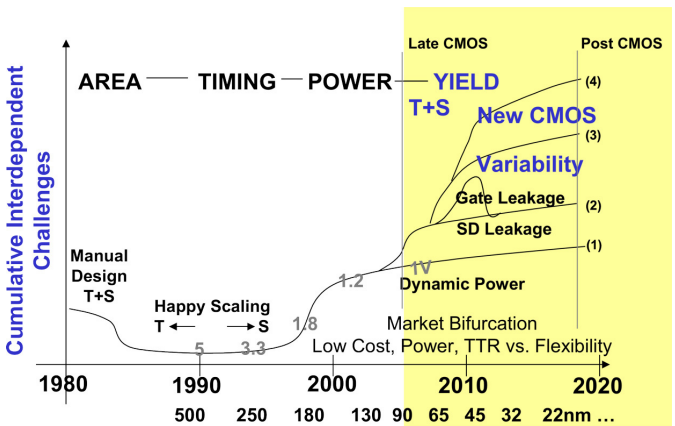


Figure 2: Cumulative interdependent challenges as a function of time (and technology generation).

A. "Power is the only limiter"

From the 0.35µm generation onwards supply voltage V_{DD} has been scaled, primarily for intrinsic transistor reliability reasons, from 5V all the way down to 1.2V in recent production technologies. In order to maintain the average 25% performance increase from generation to generation, threshold voltage V_t had to be dropped in order to maintain sufficient overdrive (V_{DD}-V_t). However, this resulted in a steady increase in S/D subthreshold leakage which has led to a continuous increase in I_{off} for the successive generations. There are two practical limits to the increase in I_{off}:

1. an application maximum power limit, e.g. in PDAs, cell phones, portables;
2. maximum package power limit, e.g. in servers and desk tops.

As an example, for large SRAM arrays scaling below 1V is challenging. For desk tops the power limit is of the order of 100W [4].

A plethora of techniques for leakage control have been proposed [5], but they all affect library design, chip density, and process technology, and require tight system/technology interaction. From recent conferences, it becomes obvious that reducing gate delay is no longer the sacrosanct performance metric: the alternative is parallelism but this leads to a drastic change in architecture (multi-core processors). This is a true paradigm shift. It impacts the software and slows down the positive feedback loop of Moore's pendulum, since the cost is no longer a matter of area alone. Perhaps more importantly, it shifts the focus of productivity from processing tools to people.

B. To the post-PC world in the late CMOS age

The transition from a technology push to a market pull in the post-PC world is dominated by (1) a growing consumerization (cfr. home and car appliances); (2) a shift from computation to communication (which moreover is nomadic and wireless) and user interfaces for applications such as: interactive audio-visual infotainment, broadcast on the move, recognition, augmented reality; and (3) the introduction of smart objects observing and controlling our surroundings and our body functions, and forming ad-hoc communication networks. In the coming decade we will witness the introduction of increased inter-device communication for new applications allowing more mobility, safety, living comfort, services and health monitoring. The economic challenge is that average selling price has decreased from 200$ to 5-10$ per computational device, which has reduced profit margins and put a tremendous pressure on unit cost.

A bifurcation of the market has occurred [6]. We have witnessed a shift from general purpose (GP) processors designed for raw performance at 100W, to devices with required performance for two-orders-of-magnitude lower power for a given task set. This results in a required 100x increase in Power Efficiency (PE) for Ambient Intelligence (AmI) [7] or pervasive computation devices, which in turn requires a rethinking of domain-specific computation architectures (see fig. 3). Computational power for the consumer applications can reach 1 Tops in the future, but packaging and cooling costs limit power for such consumer products to less than 5-10W, resulting in a PE of 100 Gops/W. On the other hand autonomous wireless transducers require a computational power of perhaps 10 Mops, for a power input of 100 µW/cm^2, resulting in a similar PE of 100Gops/W.

According to the AmI vision, we are entering the "embedded-everywhere" world which surrounds us with intelligent microsystems interacting with each other and with people, through wireless sensors and actuators. As a result, systems are becoming increasingly heterogeneous. Moreover, we have moved from GP programmability to embedded software (i.e. from hardwired ASIC to embedded programmable platforms).

The latter devices increasingly rely on technologies emerging around CMOS, also called "More than Moore", such as 3D interconnection and packaging, MEMS, and polymer devices. Nanoscale biosensors will connect electronics to biotechnology and create new opportunities for healthcare. Introduction of novel technologies besides Si-technology results in another parameter that slows down Moore's pendulum. At

the same time, the post-PC world witnesses the unfolding of a new technological dimension allowing the optimization of overall system functionality.

All of these factors emphasize the simultaneous occurrence of architectural and technological paradigm shifts. As a result, technology and design should be optimized together, similar to what happened with the PC and the ITRS roadmap until recently. But now the roadmap needs to be made specific for the major segments that result from the aforementioned market bifurcation. Both IP design and platform architecture are affected, at a time when NRE costs are exceeding $50M (for the 90nm node). This increases the need for finding economies-of-scale and hence for alliance formation. Here again we are facing increasing "gravitational forces" that threaten to slow down the ticking of Moore's clock.

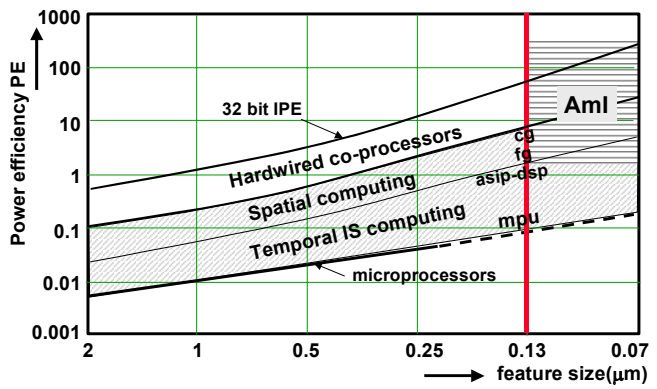


Figure 3: Power Efficiency as a function of feature size for various architectures (source: T. Claasen [8]).

C. Facing nanoscale realities in the "late-CMOS" era

"Scaling as usual" will need to cope with the law of diminishing returns unless we introduce new device architectures using high-k gate dielectrics, metal gates, strained silicon, and multiple-gate devices.

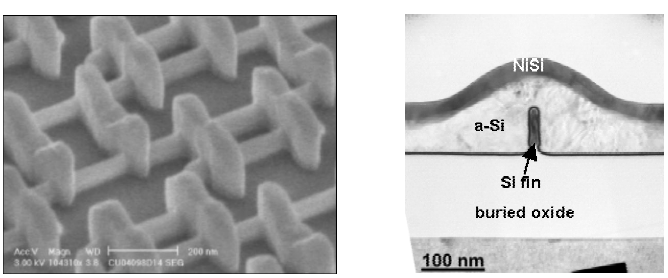
For some applications in AmI it may be feasible to postpone or even completely avoid the introduction of new gate stack materials and maintain gate delay by using mobility enhancement techniques, such as strained silicon, and architectural innovation used with a slightly thicker gate oxide. However, ultimately high-k gate dielectrics combined with fully-silicided or metal gates will be highly desired to solve the gate-leakage problem in most high-performance and low standby power applications [9]. On the other hand, it is not until 45nm that yielding high-k solutions are expected to appear in production.

Probably the greatest challenge caused by nano-level scaling is the increase of the intra-die variability of threshold voltage, drive and leakage current as they become dependent on the statistical distribution of parameters such as physical gate length and dopant concentration (see curve 3 in fig. 2). Especially for on-chip memories this already causes problems in the 65-nm node.

Deep nanometer scaling causes increased complexity with respect to lithography techniques, requiring highly regular cell and interconnect architectures to reduce mask/design cost and litho-friendly layouts to improve printability. In addition, line edge roughness causes variances in line width of the order of 5nm [10] due to the granularity of resists and photon beams. This edge effect is another contributor to the collection of variability issues, and needs innovative solutions. One way could be to use nanotechnology to self-assemble structures at the atomic scale rather than using top-down techniques, but this is far from exploitation today [11].

Taken together, these effects have a significant impact on circuit operation. Threshold disturbances can be modeled by the V_t-variability $\sigma_{\Delta Vt}$. Pelgrom's law [12] states that $\sigma_{\Delta Vt}$ = A/sqrt(W.L) which shows its deterioration with scaling. Voltage headroom V_{DD}-V_t (and thus I_{on} and t_d) becomes very unpredictable even for neighboring identically designed transistors and gate delay becomes a stochastic variable. This jeopardizes timing-closure techniques and requires statistical timing-analysis methods instead.

In SRAMs, increased transistor mismatch prevents V_{DD} scaling below 0.8V during read/write operation for yield and noise-margin reasons, especially for bulk CMOS [13]. The use of FDSOI and MUGFETs is expected to improve this situation from 65nm downwards [14,15], but this can only be verified through real designs. In view of the increasing difficulties resulting from scaling, there is a need to investigate other scalable non-volatile memory components with SRAM-like properties but with better cost/performance trade-offs. In the coming years, memory technology will go through changes that will strongly affect circuit architecture, IP, and design methods.



(a) (b)

Figure 4: (a) Tilted SEM of a 45nm node SRAM cell (0.314µm²) made with finFET; (b) Cross-sectional TEM of a finished device consisting of metal gate on a 10nm narrow fin [16].

Interconnects are also cause of concern as scaling results in smaller gate delays but slower signal transmission across global interconnects. And strong capacitive interline coupling leads to poor signal integrity. Novel low-k dielectric materials can only partially address this problem and the resistivity per unit length of the Cu wires will increase due to additional scattering effects. New compromises will have to be found between speed, energy, noise and density of wires. This will impact the way to design and lay out on-chip communication. Global bus structures do not scale well to higher complexity, and global synchronism will have to be abandoned in favor of Globally-Asynchronous Locally-Synchronous (GALS) architectures. But also the local interconnect issues are pro-

blematic as the highest activities lie there for well-optimized application designs, so also the energy bottlenecks reside in these shorter lines. They dominate the transistor contributions which has a major impact on the circuit and communication architecture design (no technology/design decoupling any longer).

Larger computational power must increasingly come from more transistors rather than faster ones. Architecture and technology must be tuned to find an optimum trade-off between clock frequency, degree of parallelism, total power in active mode, and leakage power in idle mode.

Meanwhile yield concern from functional losses is compounded by parametric (or circuit-limited) losses. In this mode circuits fail due to technology-induced variations that impact the spread of intra-die device parameters. Relative process variations increase with scaling in the nanometric regime, with a rapid degradation of circuit-limited yield as a corollary. Most importantly, platform architects will have to come up with new methods to design reliable electronic systems with uncertain components, and worst-case design must be avoided. The answer to this challenge is a yield-aware modeling approach, replacing the worst-case algorithms by a probabilistic design methodology. One way to do this is by providing a run-time controller that minimizes the impact of the variability of the individual system components. To implement this strategy detailed variability models are needed, linking the description of scaling related parameter variations with yield issues. This methodology impacts all stages of design, and continuing "more Moore" will critically depend on the availability of platform architects skilled in these new design methods.

In the nano-era, the list of technological options is growing rapidly and the number of alternative pathways to be explored appears staggering. But Moore's pendulum does not leave much time to make the right choice, create cost effective yielding processes, develop IP libraries and learn how it impacts design of giga-scale architectures.

Post-CMOS: on to new frontiers

Several scenarios have been put forward to describe what could become the end game of Moore's law. They broadly fall into two categories: the ultimate scalable MOS and the post-CMOS extension(s). Ultimate scaling challenges to be faced by CMOS include: (i) limiting off-state power leakage and short-channel effects; (ii) increasing saturation current while reducing the power supply; (iii) controlling the variability across the chip and from chip to chip. Starting from these prerequisites, the general trend is towards more compact 3D transistor structures, in which at least two critical device dimensions are scaled to the nanometer size. In this sense the finFET scheme of today represents the first step towards the "*quantum wire"* paradigm, with possible physical implementation involving carbon nanotubes or semiconductor nanowires. These are considered as possible gateways to the final shrink that will end the scaling game around 5 nm physical gate length.

In contrast to the prognoses for ultimate CMOS, which for now seem to focus on just a few alternatives, the post-CMOS game remains much more open. The challenges faced when attempting to build appliances with capabilities that exceed those of circuits based on the CMOS transistor, are (i) designing architectures compatible with the novel nanodevices plus their interconnect options, and (ii) developing technologies to fabricate and assemble such devices inexpensively. The bottom-up approach that builds nanometer-scale structures from the atom and molecule level upwards has recently received attention [11]. It allows in principle very precise positioning of atomic structures. It is fair to assume that the bottom-up approach will play a role in the future, but in order to become competitive several issues must be resolved, such as developing strategies for time-efficient self-assembly, as well as creating self-organization schemes up to the level of complex patterns. Self-assembly is already a hot research topic for molecular nanoelectronics and for the fabrication of quantum dots. The challenge is that many of the proposed nanodevices operate at low current levels and / or low ambient temperature and hence are not compatible with present circuit architectures.

If the new technologies are successfully implemented, the combination of top-down and bottom-up manufacturing will create the first wide-scale industrial application area for nanotechnology in ICT. At this stage, nano-CMOS may still qualify as the common platform on which both approaches will be implemented. However, once the nano-scale domain has been reached, there is no room for further downsizing. Hence, it can be stated that nanotechnology will be the ultimate fulfillment of Moore's law. From that point on, the evolution of nano-CMOS will shift from scaling to a systematic exploitation of its huge potential as enabler for the implementation of AmI. Unleashing the full power of CMOS as the central AmI platform will become even more challenging in an era of *"nanoelectronics with giga-complexity"*, yet its socio-economic rewards should be comparable to the benefits currently achieved by Moore's law. A striking example can be found in the field of biological applications: the combination of bioengineering (bottom-up) with microelectronics (top-down) offers some fascinating perspectives for new electronic devices based on self-assembled structures. For about a decade, ULSI technology has been put to use in fabricating (bio)MEMS, whereas microelectrodes and FETs serve as transducers in a wide variety of (bio)sensors, resulting in commercially available products with even more exciting applications around the corner. However, this requires multidisciplinary research teams of engineers, physicists, chemists, biologists and medical doctors, expanding CMOS technology beyond its original boundaries in order to interact with the biological world. Direct interfaces between neurons and electronic devices that allow two-way communication between biological entities and the extra-corporal world of computers would create a synergy between biology and electronics that goes beyond mere bio-sensing.

Major challenges lie in interfacing inorganic electron-conducting semiconductors with an organic liquid phase in which signals are relayed by chemicals. An implantable transducer that is selective and sensitive enough, that allows interacting on the scale of individual neurons, and is suited

for chronic interaction, proves to be the largest remaining obstacle. Faith is put into nanotechnology and surface chemistry to bridge the gap between the seeming chaos of living tissue and the planar geometry of microelectronics.

Experimental applications are rapidly advancing the field, including the development of algorithms to drive cortical neural prostheses in animal test subjects, and visual prostheses that partially restore sight of persons suffering from retinal degeneration. However, it will probably take at least another decade before brain-controlled prostheses with real-time sensory-motor feedback will become common in treating neurological or sensory pathology. In the long term also on-line diagnosis and even decision-taking will be feasible through nanoelectronics.

Conclusions

We find ourselves at this very moment amidst two paradigm shifts occurring simultaneously: the shift of application drivers from the PC to the AmI node and the CMOS process diversification.

Managing giga-complexity is faced with the moving target of widely diversified CMOS process technologies. CMOS has become a complex materials system: choices are crucial and very different for each application domain. The corresponding IP blocks adapted to the AmI node type must be developed. Hence, a big burden rests on IP development and reuse.

Two additional parameters threaten to slow down "Moore's clock" for future applications: we have to come up (i) with architectures that fit the embedded software, and (ii) with autonomous microsystems of unprecedented power efficiency and equipped for connectivity with the physical environment.

Managing systems of 10 to 100 BT-complexity requires grand alliances – both in industry and in the research environment – that share the multi-disciplinary R&D costs and are capable of educating and organizing hosts of scientists and engineers to design and process the AmI products of the future.

But the outlook is fascinating: more than ever will computing and communicating devices, their networks and everything they connect help humans reach their dreams. Never has the future been so challenging yet so exciting.

Acknowledgements

The author gratefully acknowledges the support from R. De Keersmaecker, H. De Man and M. Van Rossum and useful discussions with and/or contributions from E. Beyne, S. Biesemans, G. Borghs, F. Catthoor, K. De Keersmaecker, M. Heyns, R. Lauwereins, K. Maex, R. Mertens, K. Ronse, M. Stucchi and L. Van den hove.

References

[1[G. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, pp. 114-117, 1965; see also G. Moore, "Progress in digital integrated electronics," *IEEE IEDM Tech. Digest*, pp. 11-13, 1975

[2] R. Dennard et al. "Design of ion-implanted MOSFETs with very small dimensions," *IEEE J. Solid-St. Circuits*, vol. 9, 256, 1974.

[3] D.L. Critchlow, "MOSFET scaling – The driver of VLSI technology", *Proc. IEEE*, vol. 87, pp. 659-667, 1999.

[4] P. Gelsinger, "Microprocessors for the new millennium: challenges, opportunities and new frontiers", *ISSCC Digest of Techn. Papers*, pp. 22-25, 2001.

[5] T. Sakurai, "Perspectives on power-aware electronics", *ISSCC Digest of Techn. Papers*, pp. 26-29, 2003.

[6] J. Rabaey, "Design at the end of the silicon roadmap", Keynote Presentation, ASPDAC, Shanghai, January 2005.

[7] E. Aarts, R. Harwig and M. Schuurmans, "Ambient Intelligence" in *The Invisible Future*, P. Denning, Ed., McGraw Hill, New York, pp. 235-250, 2001.

[8] T. Claasen, "High speed: not the only way to exploit the intrinsic computational power of silicon", *ISSCC Digest of Techn. Papers*, pp. 22-25, 1999.

[9] P.M. Zeitzoff, J.A. Hutchby and H.R. Huff, "MOSFET and front-end process integration: scaling trends, challenges, and potential solutions through the end of the roadmap," *Int. J. of High-Speed Electronics and Systems*, vol. 12, pp. 267-293, 2002.

[10] P. Leunissen, M. Ercken and G. Patsis, "Determining the influence of statistical fluctuations on resist line edge roughness", presented at MNE 2004 (Rotterdam, The Netherlands), to be published in *Microelectronic Engineering*, 2005.

[11] Babak Amir Parviz, Declan Ryan and George M. Whitesides, "Using self-assembly for the fabrication of nano-scale electronic and photonic devices", *IEEE Trans. On Adv. Packaging*, vol. 26, pp. 233 – 241, 2003.

[12] M. Pelgrom et al., "Matching properties of MOS transistors", *IEEE J. Solid-St. Circuits*, vol. 24, pp. 1433-1440, 1989.

[13] K. Itoh et al., "Review and future prospects of low-voltage embedded RAMs", *Digest of CICC*, Oct 2004.

[14] M. Yamaoka et al., "Low power SRAM menu for SoC application using ying-yang feedback memory cell", *Digest Symp. VLSI Circuits*, pp. 288-291, June 2004.

[15] L. Chang et al., "Moore's law lives on", *IEEE Circuits and Dev. Mag.*, vol. 19, pp. 35-42, 2003.

[16] (a) A. Nackaerts et al., "A 0.314 μm^2 6T-SRAM cell built with tall triple-gate devices for 45nm node applications using 0.75 NA 193nm lithography", IEEE IEDM Tech. Digest, pp. 269-272, 2004; (b) L. Witters et al., this symposium.