Design and CAD Challenges in 45nm CMOS and beyond

David J. Frank   Ruchir Puri
IBM T.J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598
djf@us.ibm.com, ruchir@us.ibm.com

Dorel Toma
US Technology Development Center
Tokyo Electron US Holdings
Austin, TX 78741
dorel.toma@us.tel.com

ABSTRACT

With semiconductor industry's aggressive march towards 45nm CMOS technology and introduction of new materials and device structures in sight for 32nm and 22nm nodes, it is crucial for the IC design and CAD community to understand the challenges posed by these potential technology changes. This tutorial will focus on these challenges starting from front end of line (devices) to the back end of line (interconnects) and finally the impact on CAD. We will discuss the impact of various device technology options/improvements, such as high-κ, metal gate, low temperature operation, increased mobility and reduced variability, on the overall chip performance in the context of power-constrained technology optimization. This will show that power constraints limit, but do not eliminate, the performance improvements available from new technology. The integration issues related to low-κ materials for interconnects in 45nm and beyond will be examined in the context of advanced IC design. Ultra low-κ materials, evolution of etch and chemical mechanical polishing (CMP), and techniques to limit damage during processing and their impact on design performance will be discussed in detail. These advanced device and interconnect structures and materials including 3D technology have tremendous impact on the direction of the CAD industry. We will discuss the design methodology and CAD implications of these imminent technology changes.

1. Introduction

As CMOS technology progresses to the 45nm generation and beyond, a variety of significant changes are being studied and developed for the materials, processes and structures. These options include high-κ gate dielectrics, metal gates, increased mobility, low-κ wiring dielectrics, liquid cooling, sub-ambient temperature operation, and multiple layers of circuitry ("3D"). In addition, unwanted effects are becoming prominent and require design attention, including increased random variability and quantum mechanical tunneling currents. One way to address the impact of these various device technology options/improvements on overall microprocessor performance is to use power-constrained technology optimization. The results tend to show that power constraints limit, but do not eliminate, the performance improvements available from technology enhancements.

ICCAD'06, November 5-9, 2006, San Jose, CA

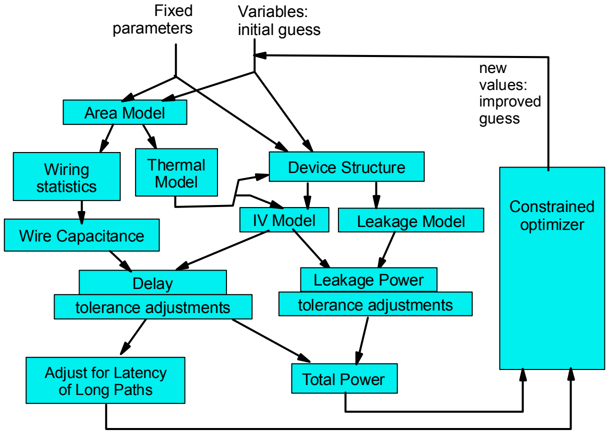

Figure 1. Schematic block diagram of the chip-oriented technology optimization tool.

The tool used for some of the analysis here consists of a collection of simplified submodels, as indicated schematically in Fig.1, which are described in more detail in [4,5].

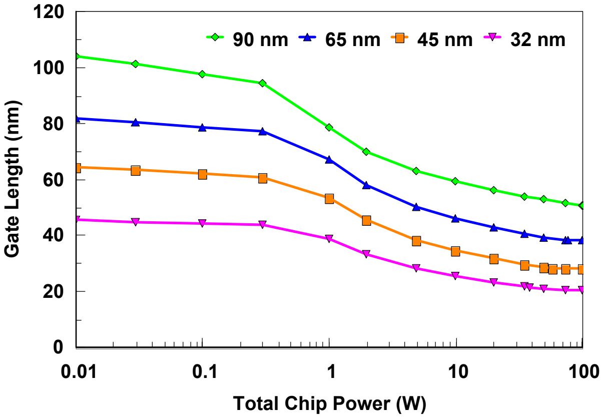

Figure 2. Optimum gate length versus total chip power constraint for several technology generations.

These models capture many of the phenomena associated with the proposed technology changes. The device model captures 2D effects associated with very short channel FETs, allowing the optimizer to choose the optimum gate lengths. The thermal model accounts for typical on-chip power distributions and hot spots and can accommodate a wide range of heatsink technologies. Random on-chip tolerances and noise are accounted for in the delay and power calculations. Long paths are given repeaters that are included in the overall optimization by taking into account these paths' impact on average latency.

2. Front End of Line (Device) Challenges

The primary device characteristic that changes with each generation of CMOS is the gate length. When power constraints dominate, however, one can gain performance by optimizing the gate length, as shown in Fig. 2, which shows the results of optimizing 7 variables, including the gate length, at each power level. The results clearly show that one should design longer gate lengths for lower power chips, because it reduces short-channel effects. The ideal oxynitride thickness is also 20-30% higher for the lower power designs, and the supply voltage is lower.

Perhaps the most widely discussed new technology feature is the use of strain to increase the mobility of the channel in FETs and thus to increase the drive current and the speed. The strain can originate from dielectric layers on top of the FETs and/or from strained materials within the device. It is expected that mobility increases will continue in future generations of technology, possibly even through the use of alternate semiconductors such as Ge and/or III-Vs. The power constrained optimization makes it clear, however, that the chip level performance only modestly increases even when the mobility increases a great deal, as shown in Fig. 3.

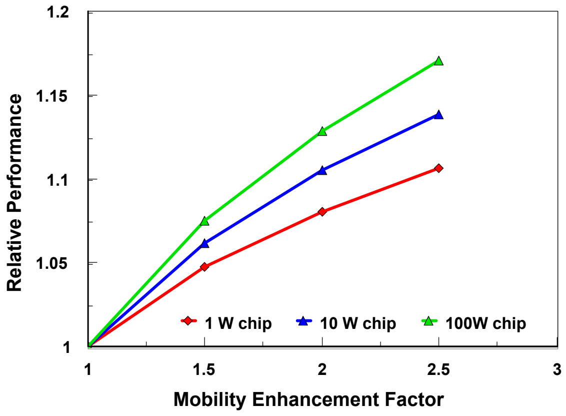

Figure 3: Dependence of optimized chip performance on mobility enhancement, for the 45nm node [5]

The second most anticipated technology change is probably the introduction of high-κ gate insulators (most likely hafnium silicates), in conjunction with metal gates [7]. The goal is to enable greater charge control by the gate while simultaneously reducing the gate tunneling current. The use of metal gates eliminates poly-Si depletion and controls charge trapping problems at the interface. The use of metal gates also offers the opportunity to adjust the threshold voltage by means of the workfunction. The optimizations results shown in Fig. 4 make it clear, however, that the 30-40% performance advantage of high-κ disappears very quickly if the workfunction isn't close enough to that of poly-Si.

Optimizations can also provide information about how power in future systems should be allocated between static and dynamic dissipation, as shown in Fig. 5. The calculations show that gate leakage should always be a small fraction, but sub-threshold dissipation may reach 50% for high power systems.

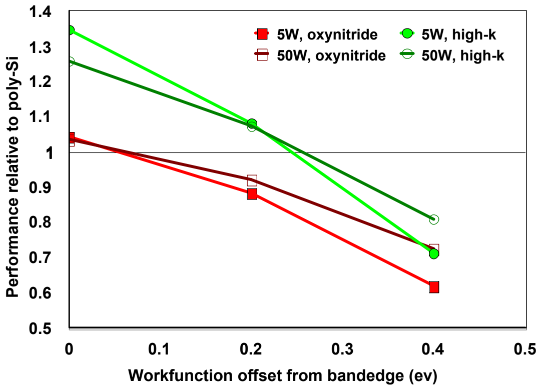

Figure 4. Relative chip performance versus workfunction offset from the bandedge, for oxynitride and high-κ gate dielectrics at the 45 nm node, and for two different chip power levels [5].

Perhaps the most important unwanted feature of upcoming technology generations is random variability. Random statistical effects due to the quantization of matter and energy include the number and placement of dopant atoms, dose variations in photolithography, line edge roughness and poly-crystalline grain boundaries. These effects are often characterized in terms of the threshold voltage variation that they cause [1]. Fig. 6 shows the significant performance loss caused by these local random variations, which occur because the chip design must be degraded to guarantee functionality in the presence of these variations.

Another variability concern relates to the contacts between the FET and the first layer of metal. As scaling continues, these contacts may significantly disturb the stress of the dielectric films over the FETs, causing layout-dependent variations in strain, in mobility, and in drive current of individual FETs [2,8]. It may become necessary for design tools to take into account the detailed contact placement geometry in calculating the FET drive characteristics.

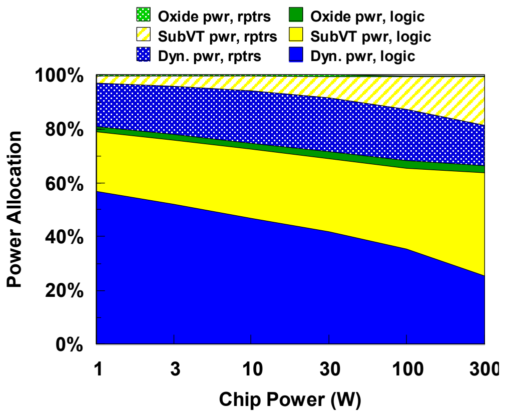

Figure 5. Cumulative power allocation fractions for the logic in processor cores, for 45nm node [5].

3. Back End of Line (Wiring) Challenges

In sub 0.25um technology, interconnect technology migrated from Al/SiO_2 interconnect integration schemes to Cu/low-κ to keep up with front end of the line scaling. However, simple geometric scaling of the Cu/low-κ is not keeping pace with the transistor speed. This requires introduction of new ultra low-κ (ULK) materials with $\kappa = 2.2$ or below. The new dielectric materials being investigated have a high degree of porosity. This presents process integration issues due to weak mechanical and chemical characteristics of the film. The issues are compounded by post-deposition processing such as cure, etch, ash, cleaning and others, which induce a degradation of film properties and added complexity in integration. New dielectrics present lower electrical performances (e.g. current leakage and time-dependent dielectric breakdown (TDDB)) and lower thermal characteristics. All these new / weak characteristics are reflected in a narrowing process integration window with a large impact on design process.

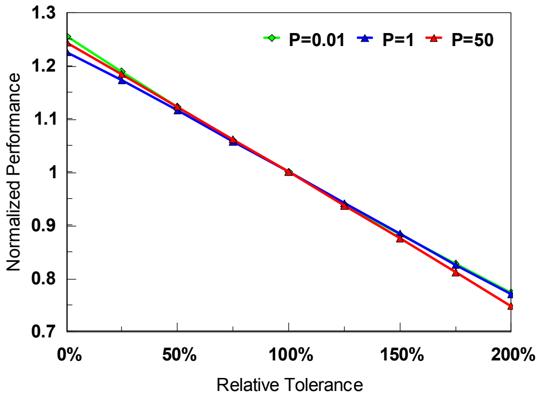

Figure 6. Relative tradeoff between on-chip variability and chip performance at 3 power levels, 65nm node.

For 32nm technology node and beyond, the conventional processes are unlikely to be sufficient. Innovative solutions are needed to eliminate dielectric degradation and minimize integration concerns. In addition, different technology paths have varying influence on the integration of new low-κ and ultra low-κ materials. The issues related to post deposition processes are important because it is crucial to identify areas with minimum negative effect on the ULK, including methods for process degradation recovery. Obviously, the work on decreasing the permittivity of the wiring insulators is very important from a chip performance point of view, since by lowering the capacitance it both decreases the switching energy and increases the speed. Design tools will need to be able to take advantage of the dielectric constants of the different wiring levels.

4. Challenges for CAD

Limits to technology [9][10] and power dissipation in 65nm and beyond is resulting in a fundamental shift in the architecture of VLSI circuits from a high frequency single core SoCs to relatively scaled back frequencies multiple core, multiple thread SoCs. Therefore, exploiting parallelism and special purpose hardware (accelerators) has become a new mantra in architecture of VLSI systems [11]. This presents a significant opportunity for CAD tools for early design space exploration that can concurrently address: Power, system performance, physical, and thermal issues concurrently. Specific issues to be addressed are: number of core and heterogeneous nature of the architecture; what power management scheme should be implemented; inter-processor communication scheme, e.g., based on a standardized bus etc.; memory hierarchy; physical integration, power delivery and clocking schemes; and control and programming models.

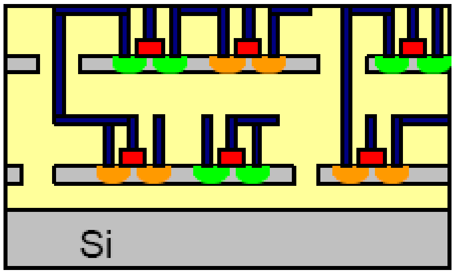

Figure 7. Schematic drawing of a 3D IC with active circuits on two different layers [6].

Power and variability has created a perfect storm at 65nm node which are being handled by various efforts in the CAD research [12][13]. To address variability, these efforts so far have focused on statistical analysis [14] and techniques to address systematic variation through measure, model, mitigate approach (DFM) [15]. However, in order to address variability directly, new system and circuit techniques must be devised that can sense variability and dynamically critical circuit parameters to adapt the behavior of the circuit. The long term goal must be to move away from the concept of building margins (whether worst case or statistical) and focus on adaptive techniques for addressing variability and building robust circuits. CAD techniques must be developed that can analyze various circuit characteristics such as power, performance, temperature etc (that change dynamically) and synthesize the control and adaptive management circuits for the "adaptive methods" to be widely utilized beyond certain custom microprocessor designs [16]. In the long term, these adaptive techniques must also be extended to resilient designs where a relatively large percentage of transistors are either out of specification or simply don't work. Advances in CAD research are needed to enable the design of such systems.

New and emerging technologies such as double gate devices and FinFETs are presenting new challenges and opportunities to optimize the circuits [17]. In addition, another technology option under investigation that will require significant design changes is 3D integration – the building of integrated circuits with more than one layer of active circuitry, as suggested schematically in Fig. 7 [6]. There are at least 3 types of benefits to 3D integration: (1) it may allow different types of technology to be placed on the same chip (e.g., III-V RF circuitry on top of Si logic), (2) it may increase performance by decreasing the interconnect length as well as latency of critical interconnects/buses since different parts of a circuit can be brought closer together if they are on different levels, and (3) it may permit more compact ICs for applications for which space is at a premium.

Traditionally, CAD research in 3D design has focused on Physical design challenges associated with physical placement and thermal issues [18]. Given the constraints of 3D vias and technology, further CAD research is needed to ensure that techniques in this area are targeting the right level of partitioning. As shown in Figure 8, to extract maximum benefit out of 3D implementations, the sweet spot of partitioning is at a relatively coarse level, i.e., at the level of processing unit or beyond. This pushes the problem of 3D design into the domain of system level analysis and demands early analysis exploration tools targeted for 3D designs with concurrent analysis in physical, performance, and thermal domains.

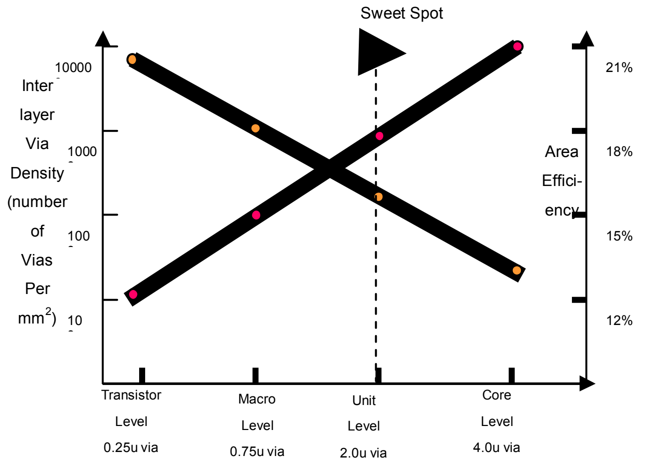

Figure 8. Tradeoff of 3D Partitioning Level wrt Area efficiency and interlayer via density.

Given the continuation of leakage trend in 45nm and beyond, power will continue to be one of the most challenging issues in 65nm technology and beyond. In a sense, there are two approaches to dealing with the increasing power dissipation of future chips: One can work to implement low power circuit techniques, and one can improve the cooling. Significant progress has taken place in both of these areas. Different low power techniques from Architecture level power management to Logic level and down to circuit level methods are being aggressively pursued. To achieve the goal of lower power circuits, adaptive power management methods and circuits and will be crucial in 45nm and beyond.

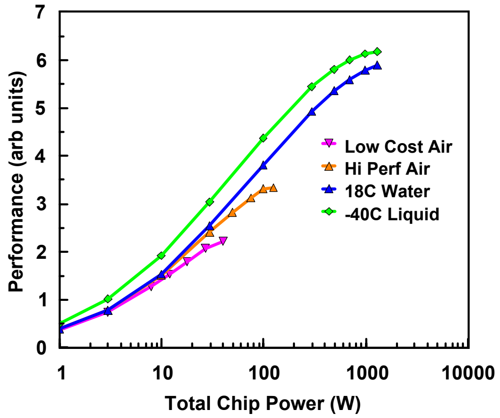

Figure 9. Dependence of total optimized chip performance on cooling technology, for the 45nm node, 4-core processor [5]

This will require significant change in the design and analysis tool infrastructure. There has been significant recent work on improved cooling technologies using liquid coolants and microchannel heatsink structures [3]. Emulating some of these cooling options in the optimizer gives the curves shown in Fig. 9, which suggest that improved cooling can indeed increase performance, but since the performance only increases as roughly the log of the power, this appears to be a very costly way to increase performance.

5. REFERENCES

[1] Asenov, A., Brown, A., Davies, J., Kaya, S., and Slavcheva, G. Simulation of Intrinsic Parameter Fluctuations in Decananometer and Nanometer-Scale MOSFETs. *IEEE Trans. Elec. Dev, 50* (Sept. 2003), 1837-1852.

[2] Cea, S. M., *et al*. Front End Stress Modeling for Advanced Logic Technologies. In *IEDM Tech. Dig*. (San Francisco, Dec., 2004). IEEE, 2004, 963-966.

[3] Colgan, E.G., *et al*. A Practical Implementation of Silicon Microchannel Coolers for High Power Chips. In *Proc. 21st Semiconductor Thermal Measurement and Management Symp.,* (San Jose, CA, March 15, 2005). IEEE, 2005, 1-7.

[4] Frank, D. J. Power Constrained Device and Technology Design for the End of Scaling. In *IEDM Tech. Dig*. (San Francisco, Dec., 2002). IEEE, 2002, 643-6.

[5] Frank, D. J., Haensch, W., Shahidi, G., and Dokumaci, O. Optimizing CMOS Technology for Maximum Performance. *IBM J. Res. Dev., 50,* 4/5 (July/Sept 2006).

[6] Guarini, K. W., *et al*. Electrical Integrity of State-of-the-Art 0.13um SOI CMOS Devices and Circuits Transferred for Three-Dimensional (3D) Integrated Circuit (IC) Fabrication. In *IEDM Tech. Dig*. (San Francisco, Dec., 2002). IEEE, 2002, 943-945.

[7] Narayanan, V., *et al.* Band-Edge High-Performance High-κ /Metal Gate n-MOSFETs using Cap Layers Containing Group IIA and IIIB Elements with Gate-First Processing for 45 nm and Beyond. In *Symp VLSI Tech., Dig. Tech. Papers*, (Honolulu, June 2006). IEEE, 2006, 224-225.

[8] Shah, N. *Stress Modelling of Nanoscale MOSFET*. Ph.D. Thesis, University of Florida, 2005.

[9] Stork, J.M.C. Balancing SoC Design and Technology Challenges at 45nm, VLSI Technology Symposium 2006.

[10] Bernstein, K. *et al.* Design and CAD Challenges in sub-90nm Technology, ICCAD 2003 129-137.

[11] Pham, D., *et al.* The design and implementation of first generation Cell Processor, ISSCC 2005, 184-185.

[12] Puri, R., *et al.* Pushing ASIC performance in a power envelope, DAC 2003, 788-793.

[13] Puri, R., *et al.* Keeping hot chips cool, DAC 2005, 285-288.

[14] Visweswariah, C. Death, Taxes, and Failing Chips, DAC 2003, 343-347.

[15] Cho, M., *et al.* Wire density driven global routing for CMP variation and timing, ICCAD 2006.

[16] Naffziger, S., *et al.* The implementation of a 2-Core multi-threaded Itanium family processor, ISSCC 2005, 182-183.

[17] Nowak, E**.**, *et al.* Turning Silicon on its edge, IEEE Circuits and Devices Magazine, Jan-Feb 2004, 20-31.

[18] Sapatnaker, S., *et al.* Physical Design Automation Challenges for 3D ICs, IEEE International Conference on IC Design and Technology, 2006.