3D IC Technology: the Perfect Storm

Philip Garrou

Microelectronic Consultants of NC, Research Triangle Park NC, 27709
Ph: (919) 248-9261, philgarrou@att.net

Abstract

IC technology, which has traditionally been dominated by dimensional scaling, is facing several technical and economic hurdles as it moves forward. Low K insulation has not been able to meet performance projections, copper traces are becoming more and more resistive, clock rates have been constrained due to thermal issues and multicore processors are demanding major increases in bandwidth and decreases in latency. Economic constraints will also begin limiting the number of IC companies able to develop leading-edge IC designs. Moving past 45 nm digital CMOS scaling will no longer guarantee lower cost and higher performance. All of these issues have crated a "perfect storm scenario" for the widespread adoption of 3D IC technology.

Key words: 3D IC, low K, bandwidth,

Introduction

IC technology has, up to now, been driven by "Device Scaling", i.e. shrinking gate dimensions and decreasing operating voltage, to improve gate switching delay and thus device performance. This has resulted in continued increases in speed and device density, both of which have been described by "Moore's Law" which states that chip performance will double every 18 – 24 months.

Unfortunately, interconnect shrinkage has also had a negative effect on device performance since smaller cross section wire dimensions have increased resistance, and tighter pitches can raise capacitance, resulting in an overall increase in RC delay.

Transistor Gate Delay

"CMOS scaling" has faced speed vs power trade-off limitations since the 65nm node. Strained Si [1] and high-k gate dielectrics [2] have continued to boost performance down to the 45nm node although adding significant process complexity. However, slow-down of transistor performance scaling is being observed in 45-nm node MOSFETs [3]. Fig. 1 shows the scaling trend of the intrinsic delay of high performance transistors. Starting from the 45 nm technology generation, the intrinsic delay is expected to stop decreasing and to exhibit counter-scaling thereafter [3].

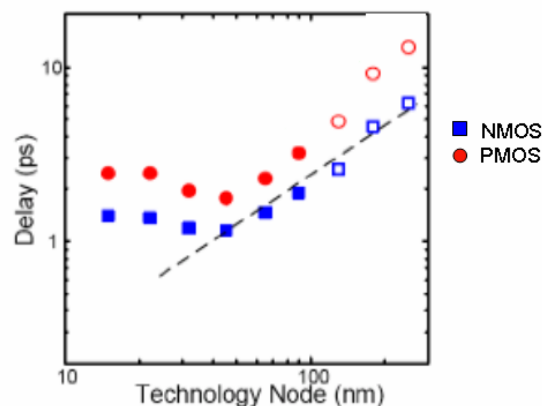

Fig. 1 Transistor Delay as a Function of Technology Node [3]

Interconnect Scaling

In 1995 Intel's Bohr concluded that the increasing interconnect RC delay caused by reductions in interconnect pitch would not support the performance requirements of future VLSI circuits. He suggested that interconnect material changes such as Cu and low-k dielectrics would help. This triggered the widespread introduction of copper metallization (led by IBM in 1997) and the search for a manufacturable low-K dielectric.

Downloaded from http://meridian.allenpress.com/ism/article-pdf/2010/1/000001/2252742/ism-2010-ta1-paper1.pdf by guest on 24 June 2024

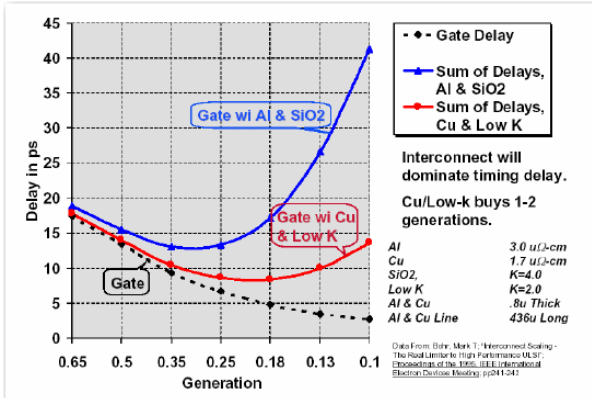

Fig. 2 The Interconnect Scaling Dilemma [4]

The 1.5 decade long delay in the implementation of low-K, is depicted in Fig. 3. The 1999 ITRS roadmap anticipated that K=3.0 dielectric materials would be introduced at the 130 nm node and we would be approaching K = 1.5 by 2005. Each subsequent roadmap pushed the low K timing out further. Many trade press and technical articles have documented the issues such as CTE and fracture toughness that arose when trying to integrate highly porous low-K dielectrics [5-7].

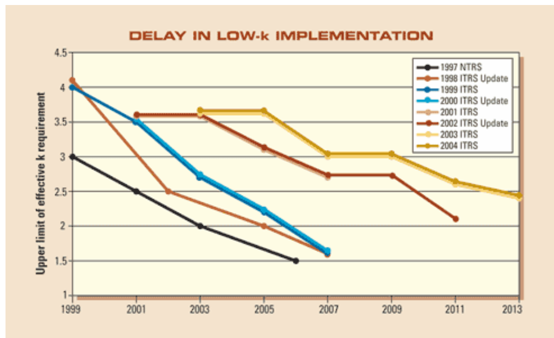

Fig. 3 ITRS Roadmap Delay in Low K Implementation [8]

The 2009 ITRS roadmap reports that low K adoption into manufacturing has actually followed the following timeline.

	90 nm	65 nm	45 nm
K	3.0	3.0	2.7-2.8

Table 1 Low K Adoption [9]

As the dense carbon doped oxides (K = 2.8) attempt to evolve into porous carbon doped oxides with K < 2.5 (ULK) there have been widely reported problems in the areas of manufacture, test, assembly and package of these fragile chips. In fact the 2009 ITRS roadmap now points to a “red brick wall” when attempting to go past K=2.5 [9].

Thus, while transistor switching performance has continued to improve down to the 45 nm node, on chip interconnect has continued to deteriorate in performance. RC delay has in fact becomes the dominant factor over gate delay.

Cu Interconnect Resistance

As interconnect cross sections have continued to shrink, line resistivity and capacitance have become a problem even for Cu lines since smaller cross section wire dimensions have increased resistance as shown in Figure 4 [10].

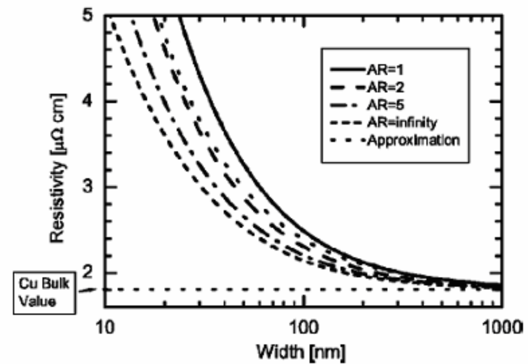

Fig. 4 Cu Resistivity: Effect of Line Width Scaling [10]

Barrier layer thickness, required to prevent migration of copper ions in the insulator materials, does not scale and thus becomes a increasing fraction of the allowable interconnect cross-sectional area which increases the effective resistivity even further as shown in Fig 5. At 32 nm interconnect will be 6X more resistive than they were at 130 nm [11-12].

Downloaded from http://meridian.allenpress.com/ism/article-pdf/2010/1/000001/2252742/isom-2010-ta1-paper1.pdf by guest on 24 June 2024

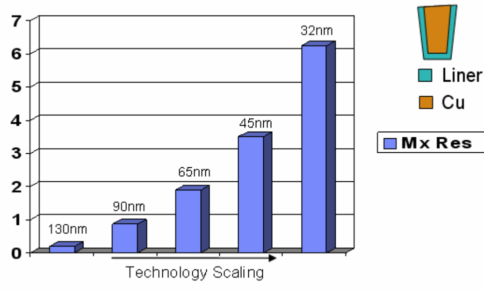

Fig 5 Total Metal Resistance vs Scaling [12]

Clock Rates

Clock rate, the rate at which a processor can complete a processing cycle, is constrained by CPU power dissipation. Processors dissipate energy by both the action of switching and energy lost in the form of heat due to the resistivity of the circuit. Power consumption in a chip is given by Eq. 1. where P is power, C is the capacitance being switched per clock cycle, V is voltage, and F is the processor frequency (cycles per second).

Eq. 1 $$P = C \times V^2 \times F$$

Increases in frequency thus increase the amount of power used and heat generated in a processor. Because of these resistivity / power issues, clock rates for high speed processors saturated around 2004 as is depicted in Fig 6.

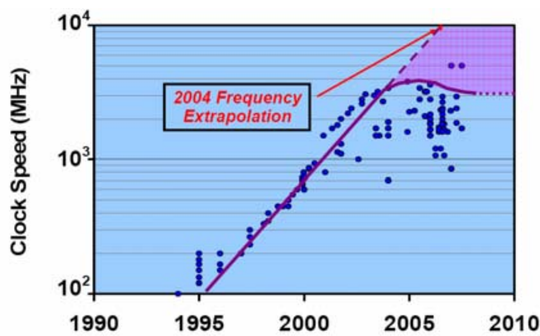

Figure 6 Processor Clock Speed [13]

Since air cooling is limited to ~ 120 – 130 W, processor frequencies had to be restrained.

To compensate for this problem, processor manufacturers began incorporating multiple cores onto one die. So called "multicore processors" have enabled performance increases to continue. Combining CPUs on a single die significantly improves the performance since signals between the CPUs travel shorter distances, and therefore degrade less. These higher quality signals allow more data to be sent in a given time period since individual signals can be shorter and do not need to be repeated as often.

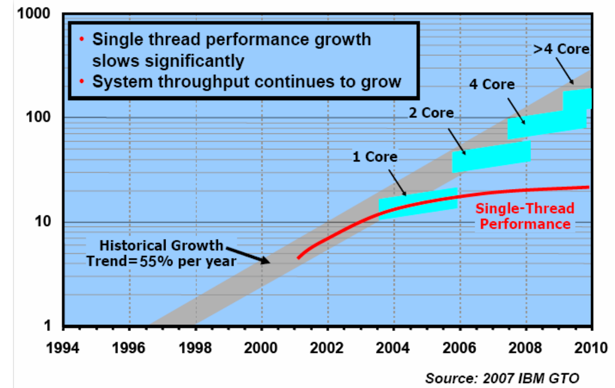

Fig 7 Multicore Processors Maintain Performance Improvements [13]

But, multicores can become starved for data, requiring more main memory and larger caches. With numerous cores on a single chip there is an enormous need for increased bandwidth and lower latency memory.

Bandwidth & Latency

Bandwidth is the amount of data bits transferred per second. Memory latency measures the number of cycles needed to access main (off chip) memory. Multicore processors require both increased bandwidth and decreased latency.

The Future Economics of Scaling

In the future, only a few IC companies may be able to develop leading-edge IC designs. At the 45-nm node, a new 300-mm fab reportedly costs about $3 billion, process technology R&D runs $2.4 billion and a "mask set" is up to $9 million. Much of the data is uncertain for the 32-nm node and beyond, but some say that by then, a new 300-mm fab could go for $10 billion, as process R&D costs reach $3 billion and design costs $75 million as shown in Fig. 2-14 [14].

Downloaded from http://meridian.allenpress.com/ism/article-pdf/2010/1/000001/2252742/ism-2010-ta1-paper1.pdf by guest on 24 June 2024

	45 nm	32 nm
Fab cost	$3B	$5-10B
ProcessR&D cost	$2.4B	$3B
Design Cost	$20-50MM	$75MM
Mask Cost	$9MM	NA

Table 2 Cost of Future Technology Nodes [14]

Fewer IDMs / Foundries will be able to afford to build fabs and few will be able to achieve a real return on investment (ROI) for such significant R&D expenditures. Fig 9 shows current and future (projected) logic fabs by technology node [15] This leads us to conclude that there will only be a limited number of logic, memory and foundry fabs in the future (Fig 10)

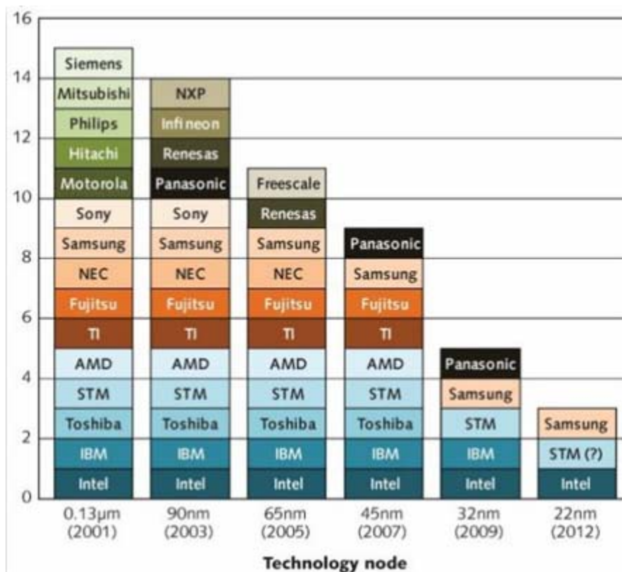

Fig. 9 Logic Fabs by Technology Node [16]

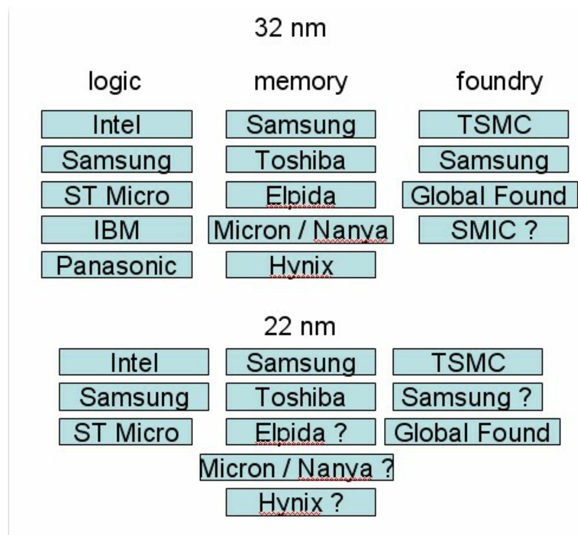

Fig. 10 Projected IC Fabs at the 32 and 22 Nodes

"The Perfect Storm"

In 2001 a seminal paper in the Proceedings of IEEE, professors from Stanford, MIT and Ga. Tech. predicted that chip interconnect would "…decelerate or halt the historical progression of the semiconductor industry…" the authors concluded that 3D integration "…should be rigorously explored to help alleviate interconnect delay and density problems…and reduce chip area" [16,17] .

In the intervening decade, many experts have concluded that CMOS device shrinkage, as we know it, will come to an end somewhere around the 22 nm node depending on device and structure and that 3D IC integration is one of the few technologies available to alleviate this problem until an alternative to CMOS technology is adopted .

3D IC Integration

3-dimensional integration (3D IC) is a system level architecture/technology wherein multiple layers of planar devices are stacked and vertically interconnected through the silicon substrate (or other semiconductor material) in the Z direction as shown in Fig 11.

Downloaded from http://meridian.allenpress.com/ism/article-pdf/2010/1/000001/2252742/isom-2010-ta1-paper1.pdf by guest on 24 June 2024

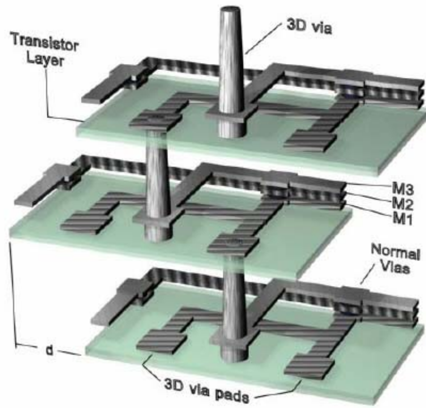

Fig. 11 3D IC Integration (U Alberta)

3D IC integration alleviates interconnect delay problems by reducing global interconnect wiring length and at the same time reducing chip area. The large number of the long interconnects needed in 2D structures can be replaced in a 3-D structure by short vertical interconnect which greatly enhances circuit performance and reduces the total wiring length required for a given system configuration as shown in Fig 12.

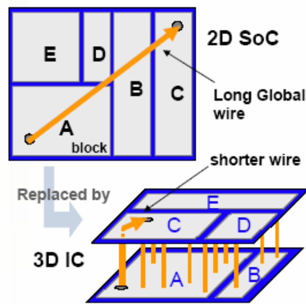

Fig. 12 3D IC Brings Reduction in Global Wiring and Chip size

Keeping wire lengths short keeps power use down by reducing the average load capacitance and resistance and decreasing the number of repeaters needed for such long wires. This in turn reduces the heat generated by the circuit. Having shorter signal paths between die thus makes it possible to reduce the system's power consumption. Stacking and interconnecting devices in the third dimension thus promises less power dissipation, higher clock rates and higher integration density.

Functional blocks (IP macros) such as memory, analog and logic can be pre-designed, verified and subsequently reused. These functional blocks can be fabricated using optimized processes on separate strata and combined to form custom circuit products. Reuse of such macros in new applications, results in faster and more reliable design work as well as reducing developmental risks [18].

In its ultimate manifestation individual functions, or groups of functions that are process compatible, would be fabricated in different strata, thinned, diced and vertically bonded together to form the final functioning circuit as shown in Fig 13 .

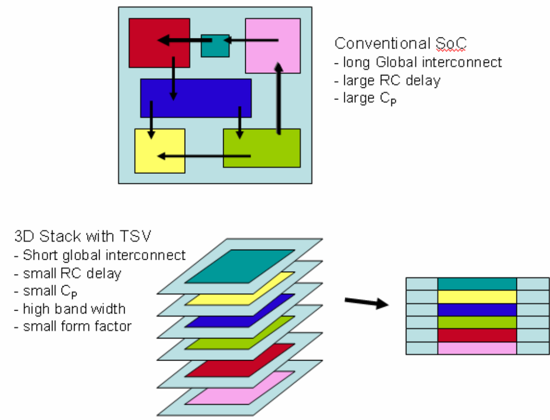

Fig. 13 Ultimate 3D IC

2D vs 3D comparisons that have been made so far show that 3D IC gains the user 2 generations of performance as shown in Fig 14. [19,20]

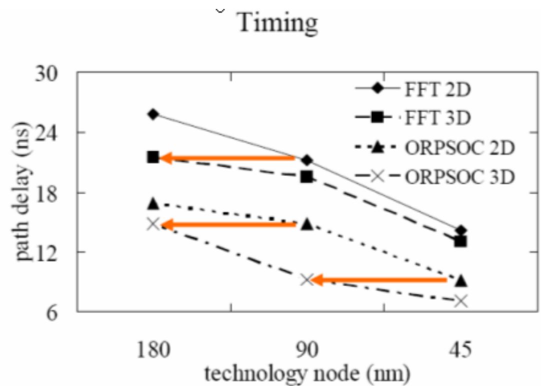

Fig. 14 3D Equals 2 Generations of Scaling [20]

Downloaded from http://meridian.allenpress.com/ism/article-pdf/2010/1/000001/2252742/isom-2010-ta1-paper1.pdf by guest on 24 June 2024

The Era of 3D IC

The Microelectronics industry usually introduces technology at a evolutionary NOT a revolutionary pace. In the next few years we will begin seeing 3D IC introduced into every aspect of our IC manufacturing infrastructure [21].

References:

[1] S. Thompson et. a., "Uniaxial-process-induced strained-Si: extending the CMOS roadmap", IEEE Trans Electron Devices, Vol. 53, 2006, p. 1010.

[2] M. Bohret. A., "The High K Solution", IEEE Spectrum Vol. 44, Oct. 2007 p.29 - 35

[3] A. Khakifirooz, "The Future of High-Performance CMOS: Trends and Requirements", 38th European Solid-State Device Research Conference, Sept. 2008

[4] M. Bohr, "Interconnect Scaling - The Real Limiter to High Performance ULSI"; Proceed. 1995, IEEE Int. Electron Devices Meeting; p.241

[5] L. Peters, "Industry confronts sub-100nm Challenges", Semiconductor International, Jan 2003.

[6] D. Lammers, "Worries Dull SiLK's Sheen at IBM Micro", EE Times, April 21st, 2003.

[7] J. Cataldo , D. Lammers, "Altera Pounces as Xilinx becomes latest to abandon low-K", EE Times, March 17th 2003.

[8] A. Braun, "Low-k bursts into the Mainstream…Incrementally", Semiconductor International, May 2005.

[9] www.itrs.net/Links/2009ITRS/Home2009

[10] W.Steinhogl et. al., "Size-Dependent Resistivity of Metallic Wires in the Mesoscopic Range", Physical Rev B 66, 2002, p 75414 .

[11] P. Kapur et. al., "Technology and Reliability Constrained Future Copper Interconnects—Part I: Resistance Modeling", IEEE Trans on Electron Devices, Vol. 49, 2002, p. 590

[12] R. Puri, "3D IC Design and CAD Challenges", Sematech Workshop on Thermal and Design Issues in 3D ICs", Albany NY, Oct 2007

[13] M. Ignatowski, "Challenges and Opportunities for Exploiting 3D Technology in System Designs", Sematech Workshop on Thermal and Design Issues in 3D ICs", Albany NY, Oct 2007.

[14] M. LaPedus, "Cost Casts ICs into Darwinian Struggle", EE Times, 3/30/2007

[15] H. Jones, "Cost of Participation in the Semiconductor Industry increasing: Impact after 32/28nm", SEMI Industry Strategy Symp (ISS), Jan 2010.

[16] J. Meindl, "Interconnect Opportunities for Gigascale Integration", IEEE Micro , Vol. 23, Issue 3, May-June 2003 Page 28

[17] K. Banerjee et. al., "3D ICs: A Novel Chip Design for Improving Interconnect Performance and System on Chip Integration", Proceed. IEEE, V. 89, 2001, p. 602

[18] M. Shapiro, "3D Technology: Applications and Requirements" 3-D Architectures for Semiconductor Integration and Packaging, Burlingame, 2008

[19] K. Nomura et. al., "Hierarchial Cache System for 3D Multicore Processors in sub 90 nm CMOS", Design, Automation, Test Conf. Europe, Nice, 2009

[20] P. Franzon et. al., "Design and CAD for 3D Integrated Circuits" Design Automation Conference (DAC), 2008

[21] P. Garrou, Handbook of 3D Integration: Technology and Applications of 3D IC" P. Garrou, C. Bower, P. Ramm Eds., Wiley-VCH 2008.

Downloaded from http://meridian.allenpress.com/ism/article-pdf/2010/1/000001/2252742/isom-2010-ta1-paper1.pdf by guest on 24 June 2024