The Road Ahead

When is 3D 2B?

Andrew B. Kahng
University of California, San Diego

A KEY FACET of "More than Moore" scaling comes under the heading of "3D"—the stacking of multiple integrated-circuit dies using through-silicon vias (TSVs). Stacked-die products reached the marketplace decades ago, but with peripheral I/O on individual dies and interchip interconnect on the "side" of the stack. TSVs are a game changer, with diameters and pitches that permit thousands of through-silicon interconnects per square millimeter of die area. The ITRS (*International Technology Roadmap for Semiconductors*; http://www.itrs.net) now roadmaps TSV technology in its "Interconnect" and "Assembly and Packaging" chapters.

Today, TSV-based CMOS image sensors are already in mass production. The DRAM roadmap has taken on a "vertical" hue as well, since further horizontal scaling conflicts with cell capacitance and aspect ratio requirements. It is likely that by next year, we will see TSV-based DRAM, flash, wireless, and TV products.

The promises of 3D integration are compelling. First, new and exciting form factors—imagine a teraflop in a cubic centimeter! Second, heterogeneous multitechnology integration—imagine logic, memory, *and* RF integrated without any sacrifice of circuit quality due to the constraints of having only one manufacturing process! Third, unprecedented I/O performance—imagine how processor architectures can change when off-chip memory access requires 2 nanoseconds instead of 30! With nascent wide I/O interfaces and very conservative 10-micron-diameter TSV technology, off-chip interfaces can achieve equivalent capacitance as low as tens to hundreds of femtofarads, without ESD protection. Fourth, improved scaling of interconnect—imagine that systems can escape the "Flatland" of 2D integration, so that global wires need not grow unnecessarily long! (A 3D-embedded system can have a Rent parameter of up to $p = 0.67$ without "dilation" of wire lengths; in 2D this limit is only $p = 0.5$.) These considerations potentially make 3D a low-power, low-cost, enabling technology for applications that range from immortal sensor nodes to exascale supercomputing, from green energy harvesting to implanted medical diagnostics.

Whenever I get a bit pessimistic about looming capex, variability, and throughput costs of, say, triple-patterned 15-nm logic ICs—not to mention the nonscaling of analog circuits and supply voltages and synchronization paradigms—I tell myself that mainstream TSV-based 3D has to be ("2B") just around the corner. One day, integration levels will reach a level such that it will be difficult to find suppliers who have all the IP blocks needed by a given system enabled for a single-die solution. One day, cost-reliability-performance trade-offs will make 3D an obvious first choice for some applications. One day, system implementers will think 3D first, and then see whether 2D is really needed for the solution. But before "one day" dawns, a number of obstacles must be overcome.

Test and yield issues include the "known good die" problem: that is, how to ensure that each piece of silicon in the system can be tested, both functionally and parametrically, in isolation. Wafer-to-wafer integration strategies face the challenge of compounded yield losses from each tier of the 3D stack.

Thermal and thermomechanical issues stem from the absence of good thermal pathways by which heat can be removed. Peak temperatures and temperature gradients significantly worsen with 3D stacked-die integration. This raises the specter of increased reliability risks, ranging from cracked vertical connections due to differences in thermal expansion

 0740-7475/10/$26.00 © 2010 IEEE Copublished by the IEEE CS and the IEEE CASS

between adjacent dies, to accelerated electromigration and other failure mechanisms (if not actual thermal runaways). Several mitigations have been proposed, including use of TSVs or carbon nanotube bundles as thermal vias, as well as liquid cooling with microfluidic channels—but are far from mainstream. Accuracy and scalability of multidomain (electrical, thermal, mechanical) modeling and simulation are challenged by the severity of 3D thermal issues. How to partition responsibility for thermal management on future multicore architectures—across virtualization, task migration, application, and hardware control mechanisms—is also an open question. Given that DRAM refresh times can change by a factor of two with every 10°C, thermal issues could be showstoppers for such hoped-for applications as memory stacked on top of processors.

Power delivery issues stem from the need to deliver high-quality power to all dies in the stack. In light of thermal issues, the hottest, highest-performing die will likely be placed at either end of the stack so that they can be nearest to the heat sinks. As a result, however, these are the dies through which power delivery to other dies must be achieved. High-quality power and ground distribution has growing cost due to higher frequencies and peak currents, increased variabilities and hence lower margins, and so on. And while the cost feasibility of 3D power delivery (in terms of wiring layers and routing congestion) remains an open question, it seems clear that drilling lots of holes in a high-end 28-nm processor chip will be an expensive proposition.

Design enablement issues begin with modeling and simulation (frequency-dependent TSV electrical modeling, the aforementioned thermo-electromechanical simulation, for instance) and end with "true 3D" physical implementation tools and methodologies. The 2009 ITRS "Design" chapter specifies a timeline for evolution from "pseudo-3D" (where 2D chip and chip-package implementation tools are bolted together after manually partitioning the system across multiple dies) to "true 3D" tool chains. Future editions of the ITRS will amplify the design technology requirements. At the same time, the EDA providers who enable 3D design will hopefully avoid replaying the experiences from multichip module and die-package design tool initiatives.

While each of these issues is daunting by itself, many observers would agree that the main obstacle facing TSV-based 3D integration isn't any of these—it is the identification of compelling, dominating 3D solutions for high-value applications. *How will 3D reach the marketplace?*

The first wave of 3D products could be those for which heterogeneous integration enables new form factors and/or cost reductions. The lead example has been CMOS image sensors; other integrations of sensors + analog + signal processing are likely to follow. The second wave could bifurcate into cost-driven and performance-driven product types. Cost reductions result from power reductions, which in turn result from shorter interconnects and smaller driving gates. Nonscaling blocks on the SoC, such as analog circuits, take up a greater proportion of the die area in new processes. Hence, stacking such blocks implemented in an older process, on top of logic implemented in a leading (newer) process, may reduce cost as well. Performance improvements will be enabled by faster, less power-hungry, and wider-bandwidth connections between processor and off-chip memory—at least, in low-power products that permit stacking of memory on logic with no thermal issues.

Personally, I am still trying to think through what will *compel* the adoption of 3D integration for particular products. At the technology and device level, will TSVs and stacking become compelling when the associated device mobility changes and keep-off distances in layout drop below certain thresholds, or when new stacking-friendly memory technologies emerge with less temperature dependence? At the system and architecture level, will TSVs and stacking become compelling when I/O bandwidths, bandwidth densities (Gbps per mm^2), and/or energy (pJ/bit) cross certain thresholds? Or does mainstream adoption depend on the emergence of an implementation tool chain? If you have any thoughts on this, please feel free to contact me.

To conclude: When is 3D 2B? We'll "C"—on The Road Ahead. ■

■ Direct questions and comments about this column to Andrew B. Kahng, University of California at San Diego, Dept. of Computer Science and Engineering, 9500 Gilman Dr., MC-0404, La Jolla, CA 92093-0404; abk@ucsd.edu.

Selected CS articles and columns are also available for free at http://ComputingNow.computer.org.