The Road Ahead

Product Futures

Andrew B. Kahng
University of California, San Diego

■ **Often in this** space I write about underlying technologies and fabrics of semiconductors: 3D integration, new nonvolatile memories, low-power design technologies, process variability, and so on. But what of the future semiconductor-based products and systems that incorporate these technologies? In the *International Technology Roadmap for Semiconductors* (*ITRS*; http://www.itrs.net), the chapter on system drivers contains simple, center-line roadmaps for microprocessor (MPU), SoC, and other key product classes and fabrics that drive industry and technology. The *ITRS* identifies product roadmaps up to the level of packages and modules. Another prominent effort, the International Electronics Manufacturing Initiative (iNEMI; http://www.nemi.org), identifies product roadmaps at the level of boards and boxes. Using the *ITRS*'s MPU roadmap as an example, let's explore how product roadmaps drive technology roadmaps, as well as opportunities for improved product roadmapping.

Product and technology roadmaps

The *ITRS* MPU roadmap is a simple projection of die area, number of transistors and cores, clock frequency, power dissipation, and other chip-level attributes. But via the "Overall Roadmap Technology Characteristics," found in the *ITRS* executive summary, the MPU roadmap influences almost every other technology in the semiconductor supply chain. Here are a few examples that illustrate how MPU attributes affect nearly all of the technologies in the semiconductor roadmap.

- *Number of transistors and die area*. These attributes determine the metal and poly pitches, and the layout pattern flexibility and density, that lithography and interconnect technologies must achieve. Associated challenges include extreme ultraviolet, multiple patterning, and mask cost. If die area increases, lithography maximum field size and mask reduction factors can be affected.
- *Maximum on-chip clock frequency and power dissipation*. These attributes tell us the transistor switching speeds, as well as the on- and off-currents, that must be achieved by process integration, devices, and structures (PIDS). Associated challenges arise with the transition to ultra-thin-body buried-oxide SOI or FinFET devices, and with process variability and signoff margins. Power dissipation also tells us the package maximum power limits and the local power densities that assembly and packaging must handle; this brings a plethora of issues such as through-silicon via and 3D stacking technologies, wafer thinning, microfluidic cooling, and power delivery.
- *Number of cores and size of onboard memories*. These attributes determine allowable defect densities (in front-end processes and yield enhancement), automated-test efficiency and design-for-test overheads (in test and design), and embedded memory requirements (for emerging research devices, PIDS, and system drivers). Associated challenges come from DFT and verification reuse, parallel test, and the embedding or stacking of new nonvolatile memories.
- *MPU characteristics in the long-term years of the roadmap (8-15 years out).* These attributes tell us what technology roadmaps for emerging research devices, and for modeling and simulation, must deliver—including the electrical and reliability models for everything from FinFETs to carbon nanotube FETs to resistive RAMs.

ITRS MPU model

Since 2001, the *ITRS* MPU model has sought to reflect recent product trends and provide direct bridges

0740-7475/11/$26.00 © 2011 IEEE Copublished by the IEEE CS and the IEEE CASS

from products to basic technology parameters such as metal-1 (M1) half-pitch, in the context of a simple, center-line projection out to a 15-year horizon. The MPU is broken down into three elements: logic, SRAM, and overhead. Layout density—the number of logic gates or SRAM bits per unit area—is expressed in terms of the *ITRS*'s scaling heartbeat, M1 half-pitch. Density calculations are based on canonical layouts of 2-input NAND gates and 6-transistor SRAM bitcells. On-chip capacitance is then estimated based on chip dimensions, layout efficiencies for both gates and interconnects, transistor sizes, and so on. Static and dynamic power components follow from capacitance calculations, supply voltage, and transistor on- and off-currents.

Power management challenges exposed by the MPU model have influenced the *ITRS*'s scaling of on-chip supply voltage, as well as requirements for off-state device leakage. To control dynamic power, recent projections have slammed the brakes on frequency scaling—from 17% per year, to 13% per year, to 8% per year, and most recently to just 4% per year. This reflects current product trajectories and permits a device roadmap with slower improvement of the CV/I time constant, which in turn permits devices to be more favorably situated along leakage-speed tradeoff curves. Even with these accommodations from the device roadmap, MPU power still rises above the fixed (~130W) platform limit imposed by both marketplace and regulatory bodies. Thus, the model relies on a "design effort factor" that reduces switching activity by 5% per year—a reminder that continuous low-power design technology innovation is needed for the duration of the semiconductor roadmap.

Improving product roadmaps

Given that roadmaps for products determine roadmaps for technologies, it is natural to ask whether there are opportunities for improvement. The current MPU model simply doubles the transistor count with each successive technology node: the total number of 6T SRAM bits doubles, and the number of logic cores and the number of transistors per core each increase by a factor of 1.4, so that the total number of logic transistors doubles as well. With respect to this starting point, many potential model enhancements can be identified, such as the following.

- The shift from 6T to 8T SRAM bitcells will change the modeling of bitcell area density.
- Modern cores have L1 and L2 cache sizes roughly flat at 32 Kbytes and ~512 Kbytes each (further, some have suggested that on-die last-level cache sizes beyond ~80 Mbytes will not have benefits); this suggests revisiting the 2× scaling of SRAM bits per process node.
- Emerging nonvolatile memory technologies will change architectural, area, and power modeling.
- The present dichotomy of SRAM + logic does not capture the modern trichotomy of core + last-level cache + "uncore," where core consists of logic and L1/L2 caches, and uncore includes memory controller, graphics engine, and so on. (Indeed, a new model for how core, last-level cache, and uncore scale forward from today's roughly equal portions of die area is targeted for future *ITRS* editions.)
- Assumptions of homogeneous multicore organization must change in light of power limits that induce more "dark silicon" (i.e., not all of the chip can be active at the same time), with utility being regained through special-purpose heterogeneity.
- Even the relevance of the current MPU model to future mobile platforms, or to future distributed computing, cloud computing, exascale computing, and other architectures, is still largely unknown.

Certainly, there are many directions in which the *ITRS* MPU model, and its relevance to the future of MPU products, can be improved. Similar opportunities exist for the *ITRS* SoC product models, as well. (Readers, please send your thoughts and suggestions!)

Beyond materials, processes, and devices, and beyond design and test technologies, there is a road ahead for "products" that drive the basic *ITRS* technologies. And beyond "products" there is a road ahead for "information technology and society," the subject of a companion column, next time. As always, I welcome your feedback. ■

■ 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.