The Road Ahead

Roadmapping Power

Andrew B. Kahng
University of California, San Diego

In their purest form, technology roadmaps give "precompetitive" statements of future technical requirements and challenges. The goal is to ensure that potential solutions are identified, investigated, pruned, productized, standardized, and delivered to the marketplace in time to ensure the continued stream of technology benefits. To give the flavor of the technology roadmapping task, consider the question: "What signaling styles, voltage levels and per-pin bit rates, and what number and areal density of off-chip interfaces, with what cost and power limits, must design and test technology handle in the year 2026 for product companies to be successful?" The 15-year horizon is a consequence of necessary lead times for funding and execution of basic research, publication and confirmation of initial results, down-selection of technology options, and the entire commercial R&D life cycle (e.g., from start-ups looking for funding to production-worthy technologies that are viable from a sourcing and cost-of-ownership perspective).

It's not too difficult to "drive by the rear-view mirror": plot data points from recent conferences, product announcements, and data sheets; perform a linear regression; and project into the future. This style of roadmapping can have value, but one downside is that clearly absurd conclusions ("the number of [layout, test, verification, software, . . .] engineers will exceed the population of the planet by the year X", or "we will need one nuclear reactor per 1,000 microprocessors by the year Y", etc.) can result. It's also not too difficult to look up to five years into the future: if a technology solution isn't on the drawing board somewhere today, it's unlikely to be in the marketplace five years from now. So, roadmapping folks talk a lot to VCs, CTOs, and analysts. Additional clues are derived by, for example, tracking the incidence of key phrases ("gallium arsenide," "thermoelectric cooling," etc.) in patent filings and the open literature: a rising number of occurrences can indicate a long-term hope; a falling number can mean a transition to productization or a dead end. Unfortunately, these basic techniques don't address the core challenge in devising a technology roadmap, namely, the prediction of long-term requirements and solutions.

In recent years, power has been a very thorny area for the International Technology Roadmap for Semiconductors (ITRS). The functionality and cost of servers and datacenters, mobile and home-networked products, and embedded sensors have been power-limited for many years. This is seen in the recent power-driven revisions to the microprocessor roadmap, as well as the balancing act between channel lengths, mobilities, and leakage currents in the low-power and high-performance transistor roadmaps. How will we advance from today's technology to the enablement of hundreds of teraoperations per square centimeter of silicon, and hundreds of terabits per second transmitted both intra- and inter-chip, delivering the desired user experiences while always staying within market-determined power limits (tens of watts for a server chip, sub-watt for mobile)? This is where a "power roadmap" comes into play.

The initial step toward a power roadmap is to understand how we've gotten to where we are today, especially for key product classes such as those roadmapped in the ITRS's "System Drivers" chapter. There is a veritable catalog of design and test technology innovations that have helped to keep IC and system power in check up to now. For instance:

- When did architectural clock gating start to impact mainstream design? (In 1996, at the

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

250-nm node.) How did it impact dynamic power? (Up to a 3× reduction.) And how did it impact static power? (Not significantly.)
- What about multi-V_{TH} design? (Mainstream in 2000 at the 180-nm node, no impact on dynamic power, up to a 3× reduction in static power.)
- Multicore processor architectures? (2003, 90-nm node, a 2× reduction in dynamic power, a negative [2×] impact on static power.)
- Power gating? (2004, 90-nm node, slight negative impact on dynamic power, up to a 10× reduction in static power.)
- Multi-V_{DD} design? (2006, 65-nm node, slight negative impact on dynamic power, up to a 1.5× reduction in static power.)
- Dynamic voltage and frequency scaling? (2006, 65-nm node, tradeoff between a 1.5× reduction in dynamic power and a 3× reduction in static power.)

Obviously, many other techniques are in this catalog: adaptive body-biasing, resilience mechanisms, heterogeneous multicore architectures and special-purpose hardware accelerators, introduction of richer low-power and sleep states, and so on. Beyond determining where a given technique falls in terms of chronology and power impacts, it is also necessary to understand its impacts on design flow complexity as well as on design brittleness when engineering change orders are made. This is not easy; witness the persistent cloud of uncertainty regarding clock skewing for IR drop reduction, useful skew for performance improvement, adoption of clock mesh topologies for skew reduction, or the choice of buffers versus inverters in clock distribution.

It is also essential to understand "legs" and "sweet spots": is a given technique a multinode solution, at what node is its sweet spot, and at what point does it run out of steam? Following are three illustrative examples.

For one example: In power gating methodologies, when do the area and layout complexity advantages of footswitches over headswitches disappear due to changing p and n device mobilities, or the scaling of ESD risks? And at what node will the use of both headswitches and footswitches provide what advantages?

For a second example: How much power savings, margin, and yield does split-rail (logic separate from SRAM) power distribution deliver in advanced nodes, in light of SRAM V_{CC}-min scaling? And what are the costs of a split-rail methodology in terms of metal and bumping resources, bill of materials, design flow complexity, or architectural and floorplanning constraints?

For a final example: How should gate-level leakage optimization best exploit multi-channel length versus multi-V_{TH} cell swapping? And how does the answer to this question change with adoption of FinFET transistor architectures, or spacer-based multiple-patterning lithography solutions?

There are obvious competitive benefits to being able to answer the above questions. But of course, that is just the "easy" part of developing a roadmap for power. The hard part is to look into the future, which brings two fundamental roadmapping challenges. First, how will existing techniques be applied in the future, for example, to affect a greater percentage of the system or chip? We must model how architects, designers, and CAD engineers will enable more voltage islands, more power gating, more clock gating, and so on: more of everything, to achieve less power without harming time to market. Second, what new power mitigation techniques will see deployment between now and 2026?

First, there will be contributions from the manufacturing technologists, such as through-silicon vias and thinned-wafer processes to enable 3D integration; reduced variability, implying reduced guardbands and overdesign; microfluidic cooling; and lower-permittivity dielectrics.

Second, there will be contributions from the device and circuit technologists, such as vertical transistors (FinFETs) and near-threshold circuits for logic; phase-change RAM and resistive RAM for memory; low-swing and hybrid optical-electrical signaling for interconnect; and improved memory interfaces and DC-DC conversion.

Third, there will be contributions from system-level designers, such as reconfigurable and heterogeneous architectures; exploitation of error-tolerance that occurs naturally in applications; development platforms for concurrent software and algorithms; and energy-efficient partitioning of computation between sensors and servers.

Finally, there will be contributions from design and test technologists, such as asynchronous design flows; margin-saving adaptivity mechanisms that span online self-testing and on-chip variability monitors; more power-intelligent cell library, synthesis and

signoff flows; better power analysis and estimation tools; and new design space exploration or "path-finding" tools.

All of the above—system architectures, design methodologies, process-device-circuit innovations, tools for design and test—must co-evolve as they continually advance to enable more (innovation) with less (resources). Certainly, there is no shortage of important challenges for the entire technology community.

Here's the punch line. Current ITRS efforts, led by the ITRS's Design Technology Working Group, are targeting development of a new "roadmap for power." In this column, I've tried to give some flavor of both the promises and the challenges inherent in this very ambitious undertaking. My request to *Design & Test* readers: tell me your thoughts on what should be included in the roadmap for power. I look forward to hearing from you at the address given below. ■

■ 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.