Moshe Y. Vardi

DOI:10.1145/2018396.2018397

Is Moore's Party Over?

The retirement of the U.S. space shuttle fleet brought a *Wall Street Journal* columnist to lament that “Mankind Nears the End of the Age of Speed.” The article also mentioned

the retirements of the supersonic Concorde and the SR-71 Blackbird spy plane, as well as Boeing's abandonment of its concept airliner, the Sonic Cruiser. “The human race is slowing down,” complained the author.

Reading that article made me pause to reflect on the slowdown of computing. For almost 50 years we have been riding Moore's Law's exponential curve. Oh, what a ride it has been! No other technology has ever improved at a geometric rate for decades. It has been nothing short of a wild party. But exponential trends always slow down eventually, and the end of “Moore's Party” may be near.

I am not betting here against “Moore's Law.” Such a bet is a well-known sucker bet. But Moore's Law is often “over-stood.” One often reads how Moore's Law predicts the ongoing improvement in microprocessor speed or performance. But Moore's Law says nothing about speed or performance; Moore's 1965 paper was strictly about the exponential increase in transistor density on a chip. How can increased transistor density be translated to improved compute performance? After all, it has really been the improvement in performance that has changed the world around us dramatically since the beginning of the computer age. Indeed, over the past 50 years the computer industry faced a dual challenge. First, it had to keep marching to the drum of Moore's Law, which turned from an astute observation to a self-fulfilling prophecy. Second, it had to translate an increase in transistor density to an increase in compute performance.

This translation was accomplished in two ways. First, Moore's Law is underlain by the continued scaling down of transistor size, postulated by IBM researchers in 1974. This enabled transistors to be switched faster and faster, increasing microprocessor frequency. Second, and crucially important, has been the ability of computer architects to harness the power of transistor parallelism to speed up the execution of sequential programs by using bit-level and inter-instruction parallelism.

This unstoppable march hit a wall in May 2004, when Intel canceled its Tejas and Jayhawk microprocessor projects because of heat problems caused by high power consumption. Thus, just as the world economy is struggling with the energy crisis, the computer industry is struggling with its own energy crisis. Dealing with this crisis has been the major challenge for the industry for the last few years. A July 2008 *Communications*' article by Mark Oskin entitled “The Revolution Inside the Box” pointed out that the performance curve of microprocessors almost flattened in 2004, and concluded, “No longer is the road ahead clear for microprocessors.” A May 2011 article “The Future of Microprocessors,” by Shekhar Borkar and Andrew Chien, declared that “Energy efficiency is the new fundamental limiter of processor performance,” and asserted that “Moore's Law continues but demands radical changes in architecture and software.”

There are those, however, who argue that neither architecture nor software can be the solution. Provocatively

titled “Dark Silicon and the End of Multicore Scaling,” an ISCA'11 paper by H. Esmaeilzadeh et al. argues that energy is the fundamental barrier. Ultimately, improved performance requires more transistors to work faster in parallel, consuming more and more power. The paper predicts that as we continue to increase transistor density on a chip, an increased fraction of these transistors will have to be powered down and stay “dark.” This means that even for highly parallel workloads we may see performance improvements lower than 20% per product generation.

While these predictions are a matter of ongoing debate, it is not too early, I believe, to start reflecting on their implications. For decades, the IT industry's business model has been predicated on double-digit annual performance improvements. I believe the next trend, which has already begun, is the commoditization of compute cycles. This will put inexorable pressure on profit margins of hardware vendors, bringing tremendous change to the computer industry, but it will make computing cheaper and more ubiquitous. The explosion of mobile devices, faced with their own energy challenges, is evidence of the force of this trend.

Peering further into the future, new physical phenomena, such as graphene and plasmonics, will replace today's dominant CMOS technology, unleashing a new age of compute-performance improvements. Remind me to write about this in 2020!

Moshe Y. Vardi, EDITOR-IN-CHIEF