Retrospective on Amdahl's Law in the Multicore Era

Mark D. Hill, University of Wisconsin–Madison
Michael R. Marty, Google

The authors of a 2008 Computer *article reflect on their corollaries to Amdahl's law almost 10 years later.*

FROM THE EDITOR

As part of our 50th anniversary celebration, this special feature revisits influential *Computer* articles from the past. This month, the original authors of "Amdahl's Law in the Multicore Era" reflect on their July 2008 article, which offered a corollary to complement Amdahl's law when applied to multicore hardware resources.
–Ron Vetter, Editor in Chief Emeritus

Making the common case fast is one of the great ideas in computer architecture. Amdahl's law formalized that great idea by providing the maximum theoretical speedup when improving only part of the system. The law also says that, for example, if a program spends 75 percent of its time in a function that can be made infinitely faster, the maximum speedup of the whole program is still no better than four. The law was originally based on an argument Gene Amdahl made in 1967 for focusing on improving processors themselves rather than shifting focus to multiprocessor systems—based on the diminishing returns in multiprocessor performance from serial bottlenecks.

After nearly 40 years of incredible improvements to processing speed, major processor vendors started hitting a technology wall in the mid-2000s and saw diminishing returns in how fast they could make a single processor. The response was something Amdahl sought to avoid: a move toward multiprocessor computer systems. In fact, today's processors (cores) are about 20 times slower than if they had continued to double in performance every 18 months since 2003.

As computer architects tussled with the rich design space of multicore processors, we found appeal in applying simple models to gain clarity and intuition. Process

0018-9162/17/$33.00 © 2017 IEEE

EDITOR **RON VETTER**
University of North Carolina Wilmington; vetterr@uncw.edu

technology afforded designers a fixed quantity of resources (whether it be transistor count or the amount of power available on a chip), and single-core performance typically saw diminishing returns with more resources applied—an observation known as Pollack's rule. Turning to Amdahl's law, we came up with corollaries to reason about the performance of multicore processors. We considered three different multicore organizations—homogenous, heterogeneous, and dynamic—and wrote about them in our widely cited paper ("Amdahl's Law in the Multicore Era," *Computer*, vol. 41, no. 7, 2008, pp. 33–38).

In the homogenous model, all cores are the same and chip designers can choose how many resources (and therefore how performant) the cores are. The model predicted that with more chip resources, the number of homogeneous cores would grow and that each would be more performant unless parallelism was almost perfect, in which case many minimal cores are preferred.

The model was correct in the sense that most multicore chips, to this day, still seek better cores. And multicore chips—especially server chips—have considerably increased core counts. Today's Intel Xeon Skylake chips boast 28 cores (56 threads) compared to dual-core chips from a decade ago, which is a considerable increase, but not at the pace of Moore's law. A 2006 MacBook Pro came with a two-core processor, whereas the 2017 MacBook Pro comes with only a four-core (eight-thread) processor. The lack of core scaling is, in part, due to the recent end of Dennard scaling, where new process technology not only doubled the number of transistors, but each used only half the power, resulting in roughly constant power per chip (unless one was greedy, as was often the case). With the end of Dennard scaling, more active transistors now means more power and that no use case—cloud, server, desktop, or mobile—can economically tolerate a repeated doubling of power.

The heterogeneous model predicted that with more chip resources, there would be increasing value in chips comprised of different types of cores. For the Amdahl's law model, this suggested one big core and many small cores. This model has been implemented for cores of the same instruction-set architecture, notably the ARM big.LITTLE product line, where lower-performance cores are used in low-demand situations to conserve power. Perhaps the heterogeneous model most accurately predicts chips with one or a few CPUs together with a GPU with many cores. A recent AMD Accelerated Processing Unit (APU) chip, for instance, consists of four CPU cores and 512 GPU cores. While GPU cores accelerate one aspect of computation, we underestimated the trend to many accelerators, especially for the consumer market, such as with Apple's A5-A8 chips.

The dynamic model considered chips in which resources could be moved between serial and parallel work at runtime. Perhaps the dynamic model's most prominent realization is the Turbo mode found in most modern Intel processors, where the clock frequency of one core can be boosted to higher speeds (thereby consuming more power) when other cores are idle. More generally, "dark silicon" refers to chips shifting power among types of work by turning off resources when not in use. This approach makes accelerators even more attractive, as they can be turned off. The ultimate success of accelerator-focused chips—especially beyond the consumer market—might be programmability.

Our 2008 article got some things right, but others wrong—especially our underestimation of the effects of Dennard scaling's demise. More importantly, we made people think. Our article used models that, while encouraging readers to question assumptions and look for insights, were so simple that we thought they might not be publishable. This contrasts with using simulation, where, all too often, people just trust fidelity and plot data. After almost a decade and more than 1,000 citations (according to Google Scholar), we continue to take inspiration from

ARCHIVED ARTICLES

The authors' original paper remains very popular, as indicated by the number of downloads it receives from the IEEE Computer Society Digital Library. All of the original articles mentioned in this special column are free to view at www.computer.org/computer-magazine/from-the-archives-computers-legacy.

statistician George E.P. Box, who said in 1987: "Essentially, all models are wrong, but some are useful."

ACKNOWLEDGMENTS

Amdahl was an alumnus of the University of Wisconsin–Madison. Hill was named the Gene M. Amdahl Professor of Computer Sciences in 2013, with Amdahl's blessing. Marty met Amdahl at an alumni gathering shortly after the original article was published, and they discussed the emergence of multicore chips. Amdahl was pleased to meet Marty, but not with multicore developments. Amdahl passed away on 10 Nov. 2015, at the age of 92. May he rest in peace.

MARK D. HILL is the Gene M. Amdahl Professor of Computer Sciences, John P. Morgridge Professor, and Computer Sciences Department Chair at the University of Wisconsin–Madison. His research interests include parallel-computer system design, memory system design, and computer simulation. Hill received a PhD in computer science from the University of California, Berkeley. He is a Fellow of IEEE and ACM, and serves as the Computer Community Consortium Vice Chair. Contact him at markhill@cs.wisc.edu.

MICHAEL R. MARTY is a senior staff engineer at Google, where he works on the design of Google's computing platform. His research interests include computer architecture, systems, and networking. He received a PhD in computer science from the University of Wisconsin–Madison. Contact him at mike.marty@gmail.com.