Abstract: Computer system performance has improved due to creatively using more transistors (Moore’s Law) in parallel via bit-, instruction-, thread-, and data-level parallelism. With the slowing of technology scaling, the only known way to further improve computer system performance under energy constraints is to employ hardware accelerators. Each accelerator is a hardware component that executes a targeted computation class faster and usually with (much) less energy. Already today, many chips in mobile, edge and cloud computing concurrently employ multiple accelerators in what we call accelerator-level parallelism (ALP).

This talk develops our hypothesis that ALP will spread to computer systems more broadly. ALP is a known way to dramatically improve power-performance to enable broad, future use of deep AI, virtual reality, self-driving cars, etc. To this end, we review past parallelism levels and the ALP already present in mobile systems on a chip (SoCs). We then aid understanding of ALP with the Gables model and charge computer science researchers to develop better ALP “best practices” for: targeting accelerators, managing accelerator concurrency, choreographing inter-accelerator communication, and productively programming accelerators. This joint work with Vijay Janapa Reddi of Harvard.

Bio: Mark D. Hill is John P. Morgridge Professor and Gene M. Amdahl Professor of Computer Sciences at the University of Wisconsin-Madison, where he also has a courtesy appointment in Electrical and Computer Engineering. His research interests include parallel-computer system design, memory system design, and computer simulation. He received the 2019 Eckert-Mauchly Award and is a fellow of IEEE and the ACM. He serves as Chair of the Computer Community Consortium (2018-19) and served as Wisconsin Computer Sciences Department Chair 2014-2017. Hill has a PhD in computer science from the University of California, Berkeley.