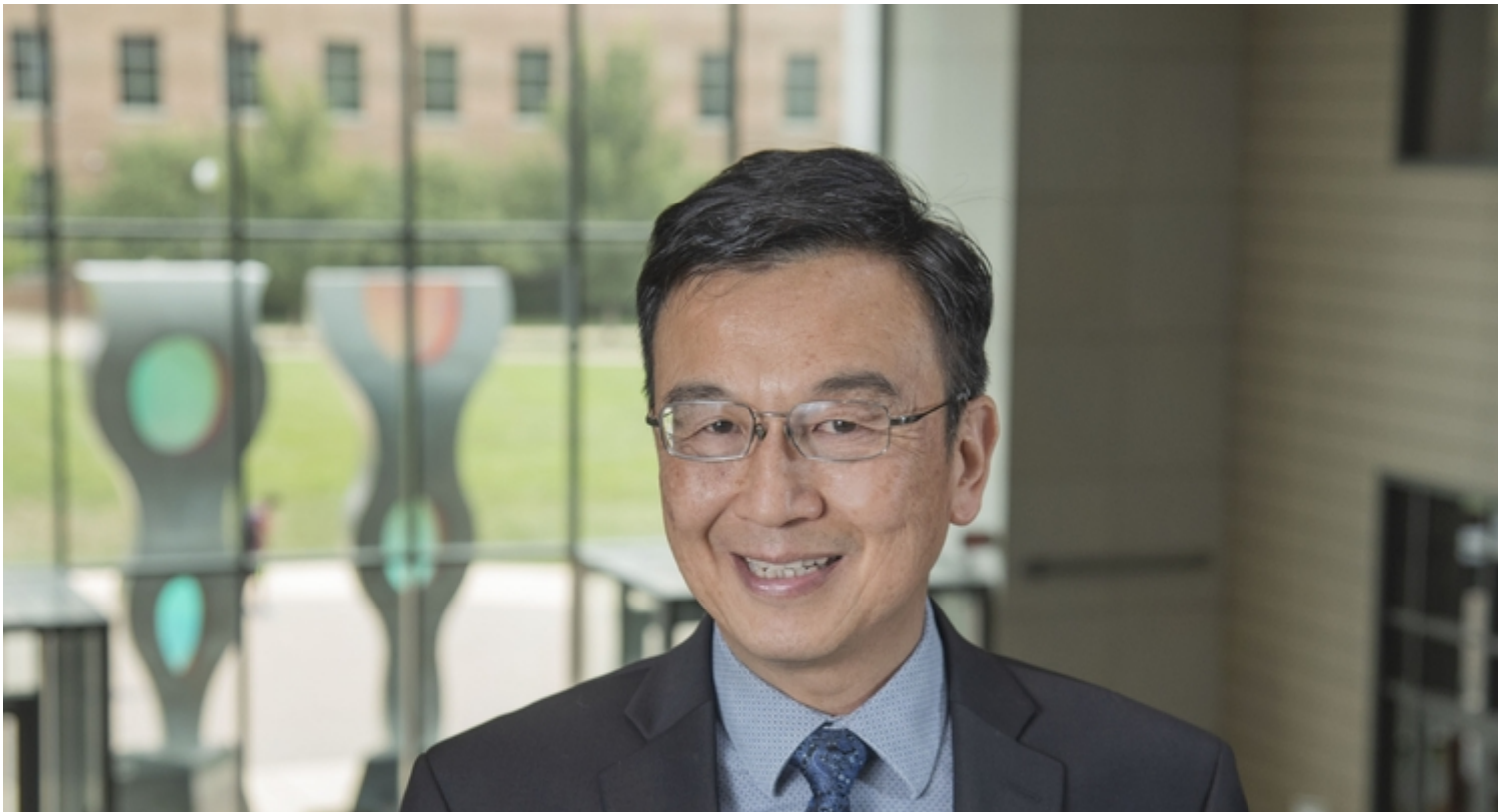
SORS: BaM: A Parallel System Architecture and Software Stack for Accelerating Compute-Directed Access to Massive Datasets

Objectives

Abstract: Compute Devices have traditionally relied on CPU OS services to bring data into the memory in bulk before performing algorithmic computation on individual data-structure elements. For example, Graphics Processing Units (GPUs) have relied on the host CPU OS services to bring chunks of storage data into its device memory for use by compute kernels. This approach is well-suited for small data sets that fits into the physical memory or applications with known data access patterns that enable partitioning of their dataset to be processed in a pipelined fashion. However, trending applications such as graph and data analytics, recommender systems, and graph neural networks, require data-dependent and sparse access to vast feature vectors and embedding datasets. CPU OS services are unsuitable for these applications due to high CPU-GPU synchronization overheads, I/O traffic amplification, and low CPU software throughput. GPU-initiated access avoids these overheads by removing the CPU from the storage control path and, thus, can potentially support these applications at much higher speed. However, there is a lack of system architecture and software stack that enable efficient GPU-initiated storage access for applications today. I will present a vision for enabling fast, compute-directed sparse access to massive datasets, the BaM system architecture to realize this vision, and the BaM software stack that efficiently supports emerging applications on existing and upcoming GPUs.



: Wen-mei W. Hwu is a Senior Distinguished Research Scientist and Senior Director of Research at NVIDIA. He is also a Professor Emeritus and the Sanders-AMD Endowed Chair Emeritus of ECE at the University of Illinois at Urbana-Champaign. His research is in the architecture, algorithms, and infrastructure software for data intensive and computational intelligence applications. He served as the Illinois director of the IBM-Illinois Center for Cognitive Computing Systems Research Center (c3sr.com) from 2016 to 2020. He was a PI of the NSF Blue Waters supercomputer project. He received the ACM SigArch Maurice Wilkes Award, the ACM Grace Murray Hopper Award, the IEEE Computer Society Charles Babbage Award, the ISCA Influential Paper Award, the MICRO Test-of-Time Award, the IEEE Computer Society B. R. Rau Award, the CGO Test-of-Time Award, numerous best paper awards, numerous teaching awards, and the Distinguished Alumni Award in CS of the University of California, Berkeley. He is a Fellow of IEEE and ACM.

Speakers

Speaker: Wen-mei Hwu, NVIDIA and University of Illinois at Urbana-Champaign
Host: Toni Peña, Accelerators for High Performance Computing Group Manager, CS, BSC

Barcelona Supercomputing Center - Centro Nacional de Supercomputación

---