Remove the Memory Wall:
From performance modeling to architecture optimization

Xian-He Sun
Department of Computer Science
Illinois Institute of Technology
sun@iit.edu

Data access is a known bottleneck of high performance computing (HPC). The prime sources of this bottleneck are the performance gap between the processor and memory storage and the large memory requirements of ever-hungry applications. Although advanced memory hierarchies and parallel file systems have been developed in recent years, they only provide high bandwidth for contiguous, well-formed data streams, performing poorly for accessing small, noncontiguous data. Unfortunately, many HPC applications make a large number of requests for small and noncontiguous pieces of data, as do high-level I/O libraries such as HDF-5. The problematic memory wall remains after years of study and, in fact, is becoming the most important issue of HPC. We propose a new I/O architecture for HPC. Unlike traditional I/O designs where data is stored and retrieved by request, our architecture is based on a novel "Server-Push" model in which a data access server proactively pushes data from a file server to the compute node's memory or to it's cache directly based on the architecture design. Simulation results show that with the new approach the cache hit rates increase well above 90% for various benchmark applications that are notorious for poor cache performance.

Performance evaluation is the driven force of the push-based model. Mechanisms of performance modeling, evaluation, and optimization are applied to data access pattern identification, prefetching algorithm design, data replacement strategy development, and architecture optimization to enable the "Server-Push" model. Our current success illustrates the power and unique role of performance evaluation in computing.

Dr. Xian-He Sun is a professor of Computer Science at the Illinois Institute of Technology (IIT). He received his BS in Mathematics in 1982 from Beijing Normal University, P.R. China, and completed his MS in Mathematics, MS and Ph.D. in Computer Science in 1985, 1987, and 1990, respectively, all from Michigan State University. He was a post-doctoral researcher at the Ames National Laboratory, a staff scientist at the ICASE, NASA Langley Research Center, an ASEE fellow at the US Navy Research Laboratories, and was an associate professor and the founding director of the Scalable Computing Software laboratory in the Department of Computer Science, Louisiana State University before he joined the Computer Science Department, IIT in August 1999. Currently he is a professor at IIT, a guest faculty in the Mathematics and Computer Science Division at the Argonne National Laboratory, and the director of the Scalable Computing Software laboratory at IIT. His research interests include parallel

1-4244-0054-6/06/$20.00

and distributed processing, pervasive computing, performance evaluation, and scientific computing.