

Abstract

Critical Factors and Key Directions for Petaflops-scale Supercomputers

Even as the high performance computing community approaches 100 Teraflops Linpack performance, challenges to supercomputer hardware and software design may impede further progress and limit scalability and performance to cost. The impact of latency, overhead, contention, and starvation are already limiting delivered efficiencies on the world's largest machines to a few percent for some critical applications. The consequence of these factors may be reduced impact of supercomputing on science, technology, industry, commerce, national security, and society even as its capability approaches important new levels. While commodity microprocessor and DRAM based clusters and MPPs have dramatically advanced the scale of high end computing over the last decade, reliance on devices expressly designed for consumer electronics and commercial enterprise computing may impose severe limitations on future expansion of those same capabilities. Fortunately, exploration of innovations in parallel architecture and methods has revealed advanced but practical strategies that aggressively attack these sources of performance degradation and may deliver future systems that efficiently perform a wide range of challenging algorithms across the trans-Petaflops performance regime. This presentation will diagnose the causes of current inefficiencies on conventional systems and describe the seminal innovations that are likely to circumvent their deficiencies. Included in this discussion will be such novel but near term concepts as streaming, the Data Vortex network, processor in memory, multithreading, and parcel-driven transaction processing among others; some taken from the speaker's own research.