

Bandwidth-Aware Design of Large-Scale Clusters for Scientific Computations

Mitsuhisa Sato

Center for Computational Sciences, University of Tsukuba
msato@cs.tsukuba.ac.jp

Abstract. The bandwidth of memory access and I/O, network is the most important issue in designing a large-scale cluster for scientific computations. We have been developing a large scale PC cluster named PACS-CS (Parallel Array Computer System for Computational Sciences) at Center for Computational Sciences, University of Tsukuba, for wide variety of computational science applications such as computational physics, computational material science, etc. For larger memory access bandwidth, a node is equipped with a single CPU which is different from ordinary high-end PC clusters. The interconnection network for parallel processing is configured as a multi-dimensional Hyper-Crossbar Network based on trunking of GigabitEthernet to support large scale scientific computation with physical space modeling. Based on the above concept, we are developing an original mother board to configure a single CPU node with 8 ports of Gigabit Ethernet, which can be implemented in the half size of 19 inch rack-mountable 1U size platform. PACS-CS started its operation on July 2006 with 2560 CPUs and 14.3 TFlops of peak performance. Recently, we have newly established an alliance to draw up the specification of a supercomputer, called Open Supercomputer Specification. The alliance consists of three Japanese universities: University of Tsukuba, University of Tokyo, and Kyoto University (T2K alliance). The Open Supercomputer Specification defines fundamental hardware and software architectures on which each university will specify its own requirement to procure the next generation of their supercomputer systems in 2008. This specification requests the node to be composed of commodity multicore processors with high aggregated memory bandwidth, and the bandwidth of internode communication to be 5 GB/s or more in physical link level and 4 GB/s or more in MPI level with Link aggregation technology using commodity fabric. We expect several TFlops in each system. In order to support scalable scientific computations in a large-scale cluster, the bandwidth-aware design will be important.