Sustained Petascale: The Next MPI Challenge

Al Geist

Oak Ridge National Laboratory,
PO Box 2008,
Oak Ridge, TN 37831-6016
gst@ornl.gov
http://www.csm.ornl.gov/ geist

Abstract. The National Science Foundation in the USA has launched an ambitious project to have a sustained petaflop computer in production by 2011. For applications to run at a sustained petaflop, the computer will need to have a peak performance of nearly 20 PF and millions of threads. This talk will address the challenges MPI must face to be used in sustained petascale applications.

The first significant challenge for MPI will be the radical change in supercomputer architectures over the next few years. The architectures are shifting from using faster processors to using multi-core processors. It is also speculated that the processors will be heterogeneous with the cores on a single processor having different functions. This change in architecture is as disruptive to software as the shift from vector to distributed memory supercomputers 15 years ago. That change required complete restructuring of scientific application codes and gave rise to the message passing programming paradigm that drove the popularity of PVM and MPI. Similarly, will these new architectures drive the creation of a new programming paradigm or will MPI survive? Or perhaps MPI becomes part of a new hybrid paradigm. These questions will be addressed in this talk.

The configuration of these sustained petascale systems will require applications to exploit million-way parallelism and significant reductions in the bandwidth and amount of memory available to a million cores. Most science teams have no idea how to efficiently scale their applications to this level and those teams that have thought about it believe that MPI may not be the right approach. Several talks at this conference describe potential features and capabilities that may allow MPI to be more effective for the reduced bandwidth and increased parallelism. The talk will point out these features and their associated talks.

The third significant challenge for MPI will be fault tolerance and fault recovery. The amount of memory in sustained petascale systems makes writing out checkpoints impractical. While the need to restart an application when 900,000 CPUs are still working fine but one has failed is an inefficient use of resources. This talk will describe the latest ideas for fault recovery in MPI and will describe a new idea called holistic fault tolerance that is being investigated.

F. Cappello et al. (Eds.): EuroPVM/MPI 2007, LNCS 4757, pp. 3–4, 2007.

Finally the talk will describe productivity issues for sustained petascale application performance and their implications for MPI. The productivity of scientists and engineers is based on how easy and how fast they can solve a new science problem. Issues such as debugging, performance tuning, scalability, validation, and knowledge discovery all play a part. The talk will address the challenges MPI has in these areas.