MPICH2: A New Start for MPI Implementations

William Gropp

Mathematics and Computer Science Division
Argonne National Laboratory
Argonne, IL
gropp@mcs.anl.gov
http://www.mcs.anl.gov/~gropp

Abstract. This talk will describe MPICH2, an all-new implementation of MPI designed to support both MPI-1 and MPI-2 and to enable further research into MPI implementation technology. To achieve high-performance and scalability and to encourage experimentation, the design of MPICH2 is strongly modular. For example, the MPI topology routines can easily be replaced by implementations tuned to a specific environment, such as a geographically dispersed computational grid. The interface to the communication layers has been designed to exploit modern interconnects that are capable of remote memory access but can be implemented on older networks. An initial, TCP-based implementation, will be described in detail, illustrating the use of a simple, communication-channel interface. A multi-method device that provides TCP, VIA, and shared memory communication will also be discussed.

Performance results for point-to-point and collective communication will be presented. These illustrate the advantages of the new design: the point-to-point TCP performance is close to the raw achievable latency and bandwidth, and the collective routines are significantly faster than the "classic" MPICH versions (more than a factor of two in some cases). Performance issues that arise in supporting `MPI_THREAD_MULTIPLE` will be discussed, and the role of a proper choice of implementation abstraction in achieving low-overhead will be illustrated with results from the MPICH2 implementation.

Scalability to tens or hundreds of thousands of processors is another goal of the MPICH2 design. This talk will describe some of the features of MPICH2 that address scalability issues and current research targeting a system with 64K processing elements.