Open MPI: A High-Performance, Fault-Tolerant Message-Passing Interface

Brian Barrett, Ralph Castain, Galen Shipman, CCS-1

Open MPI is a mature message-passing interface (MPI) implementation developed by Los Alamos National Laboratory (LANL) in collaboration with a number of academic, industry, and national-laboratory partners. Open MPI is designed to run at the large scales currently found within the Advanced Simulation and Computing program. Recently, Open MPI successfully demonstrated its performance at scale at Sandia National Laboratories, where it was an integral part in boosting the high-performance linpack (HPL) performance to 53.00 teraflops, 84.7% of peak performance up from 38.27 teraflops and 71.4% of peak a year earlier. Open MPI is the production MPI at LANL supporting several clusters including the Roadrunner base system.

The Open MPI project was originally started by LANL; Indiana University, Bloomington; and the University of Tennessee, Knoxville. Involvement in the project has since grown to include other national laboratories, universities, and commercial vendors. Support for a number of interconnects, including InfiniBand (MVAPI and Open Fabrics), Myrinet (GM and MX), Cray Portals, TCP, and shared memory, is provided in a single unified release. The Open Run-Time Environment, developed in unison with the Open MPI project, provides a scalable run-time environment with support for process launching utilizing rsh/ssh, PBS/Torque's TM interface, BProc, and SLURM.

The performance of Open MPI is competitive with other high-performance MPI implementations such as MPICH/MX and MVAPICH. Figure 1 illustrates

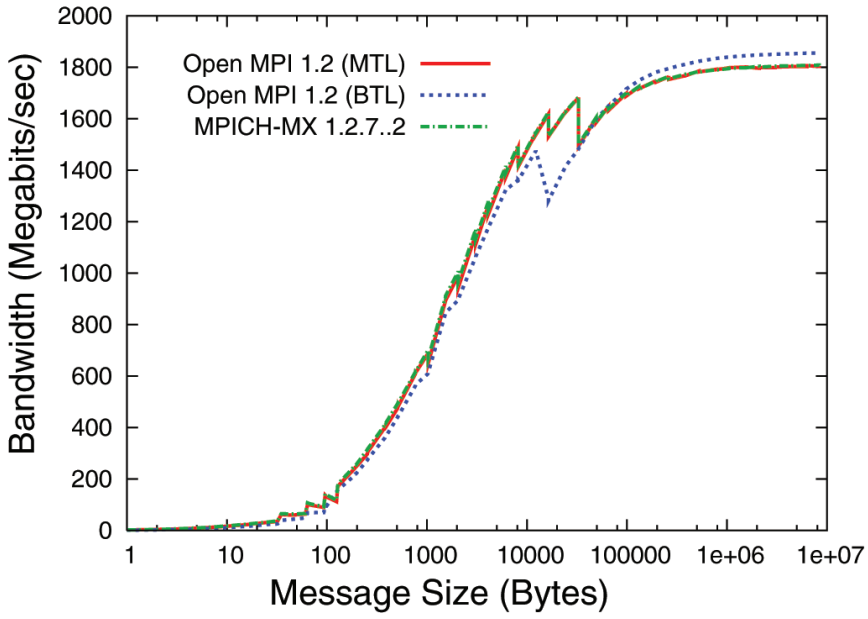

Fig. 1.
The performance of Open MPI over Myrinet MX, which is comparable to MPICH/MX.

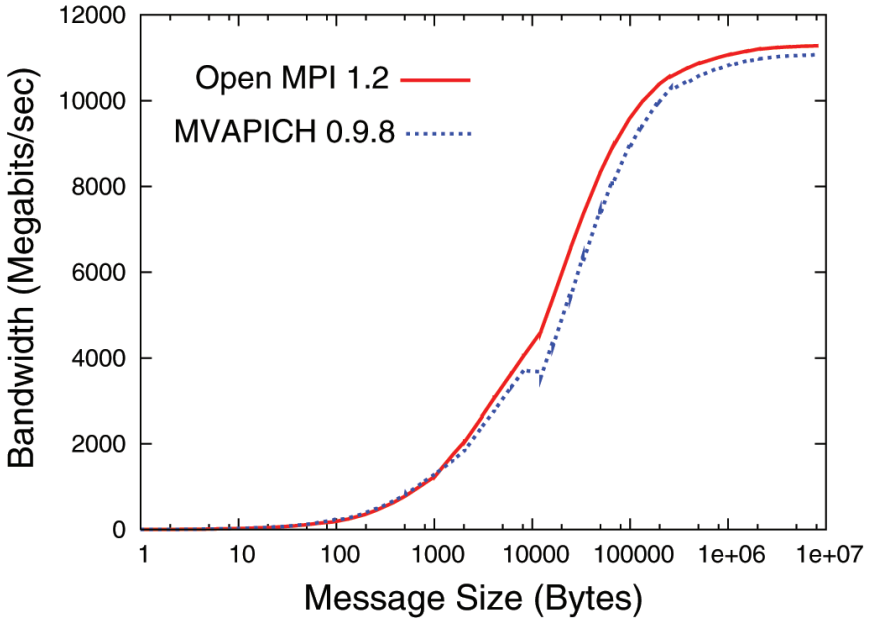

Fig. 2.
The performance of Open MPI over Mellanox InfiniBand, which is comparable to MVAPICH.

the performance of Open MPI over Myrinet MX, which is comparable to MPICH/MX. Figure 2 illustrates the performance of Open MPI over Mellanox InfiniBand, which is comparable to MVAPICH. Open MPI provides 1-byte latency of 3.11 μs using Mellanox InfiniBand vs 3.15 μs for MVAPICH, and 3.15 μs using Myrinet MX vs 2.97 μs for MPICH/MX. These results indicate that a single MPI implementation can provide good performance across multiple architectures while providing the scalability and stability of a production-grade MPI implementation.

The Open MPI team at Los Alamos is currently researching a number of issues related to running in large-scale InfiniBand environments. These include adaptive resource allocation, protocol selection and recovery from spurious host channel adapter (HCA) resets that are seen in large-scale InfiniBand clusters. In addition to generic large-scale InfiniBand clusters, the Los Alamos team is also researching issues related to the full-scale Roadrunner configuration. These include optimizations for heterogeneous data transfer, run-time support for dynamic processes across heterogeneous processors, and efficient usage of InfiniBand in nonuniform memory architectures.

For more information contact Galen Shipman at gshipman@lanl.gov.

Funding Acknowledgements
NNSA's Advanced Simulation and Computing (ASC), System Software and Support, and Tools and Capabilities programs.