ALCF MPI Benchmarks

Summary Version

1.0

Purpose of Benchmark

The purpose of the ALCF MPI benchmark suite us to benchmark the bandwidth and latency of basic communication operations.

Characteristics of Benchmark

ALCF MPI benchmarks suite consists of five independent programs: mmps, pingpong, aggregate, bisection, and collectives, which measure and report, correspondingly, a zero length messaging rate, point-to-point communication latency with ping-pong benchmark, single node aggregate communication bandwidth, bisection bandwidth, and the latency of certain collective operations.

Mechanics of Building Benchmark:

Each benchmark is build by modifying the Makefile according to the target platform. Benchmark-specific tuning can be done according to the notes given below.

Messaging rate “mmps”: To maximize the rate, the offeror may choose appropriate number of communicating neighbors XNBORS, the number of MPI tasks placed on the reference node, and the window size - the parameter, which determines the number of messaging sent simultaneously to each neighbor. The placement of the neighbors is defined by proper modification of the “getranks.c” file according to specifics of the interconnect subsystem.

Point-to-point communication latency is measured by a “pingpong” benchmark for three cases: intranode, nearest neighbor, and farthest path. The offeror can modify the “getranks.c” file to specify the communicating tasks according to specifics of the interconnect. For intranode benchmark, the communicating tasks must reside on the same node. For nearest neighbor benchmark, the communicating tasks must be placed on two distinct nodes, connected by the nearest path, provided by the interconnect subsystem. For the farthest path benchmark, the two communicating tasks must be placed on two distinct nodes with the longest interconnect path between them.

Single node aggregate bandwidth benchmark measures and reports the total aggregate interconnect bandwidth, available to a task by aggregating the bandwidth of all available links on the node. The benchmark utilizes point-to-point communications. To maximize the bandwidth, the offeror may specify “N” - the number of communicating tasks. The offeror may also modify the “getranks.c” file to place the communicating tasks according to specifics of their proposed interconnect subsystem.

Bi-section bandwidth benchmark is measures and reports the aggregate bandwidth between the worse-case bi-section of the partition. The offeror should generate the bisection by modifying the “getranks.c” file according to specifics of their interconnect subsystem.

The latency of collective operations, specifically Barrier, Broadcast, and Allreduce, is measured and reported by a “collective” benchmark. The measurements are performed across entire partition via World communicator, as well as the two non-overlapping sub partitions, dividing the World in two equal parts. The offeror may specify the division of the World into parts by modifying the “split.c” file, according to their preference and using specifics of the interconnect subsystems.

Mechanics of Running Benchmark

Other then the number of MPI processes, the benchmarks do not have specific arguments.

Verification of Results

N/A