Phloem MPI Benchmarks

Summary Version

1.2

Purpose of Benchmark

The purpose of the Phloem MPI benchmark suite us to benchmark the bandwidth, latency, and messaging rate of basic communication operations.

Characteristics of Benchmark

The Phloem MPI benchmark suite consists of three independent benchmarks: Presta, mpiBench, and SQMR.

The Presta benchmark com measures ping-pong latency and aggregate bandwidth for 1 or more pairs of MPI processes to provide intra- and inter-node aggregate bandwidth as well as bisection bandwidth.

The mpiBench benchmark measures collective latency for several blocking and non-blocking collectives for MPI_COM_WORLD and sub-communicators.

The SQMR benchmark measures messaging rate for MPI point-to-point operations.

Mechanics of Building Benchmark

Build flags can be modified in the top-level Makefile.inc file. The default make target should build the appropriate benchmarks with "make".

Mechanics of Running Benchmark

Please see the Phloem top-level README as well as individual benchmark READMEs for more information regarding the benchmarks. The following items identify measurements of interest and example commands for running the benchmarks.

MPI point-to-point intra-node aggregate bandwidth

Run the presta/com benchmark with the number of cores MPI processes on a single node.

```
mpirun -n 16 ./com -m bw.message.sizes  # on 1 node
```

MPI point-to-point inter-node aggregate bandwidth

Run the presta/com benchmark with MPI processes equal to 2x the number of cores per node over two nodes.

```
mpirun -n 32 ./com -m bw.message.sizes  # over 2 nodes
```

MPI bi-section bandwidth

Run the presta/com benchmark with MPI processes equal to Nx the number of cores per node over all nodes, where N is the number of nodes in the system.

```
mpirun -n N ./com -m bw.message.sizes  # over all nodes
```

MPI point-to-point intra-node latency

Run the presta/com benchmark with the number of cores MPI processes on a single node.

```
mpirun -n 16 ./com -m latency.message.sizes -w Latency # on 1 node
```

MPI point-to-point inter-node latency

Run the presta/com benchmark with MPI processes equal to 2x the number of cores per node over two nodes.

```
mpirun -n 32 ./com -m latency.message.sizes -w Latency  # over 2 nodes
```

MPI Collective Latency

Run the mpigraph/mpiBench benchmark with MPI processes equal to Nx the number of cores per node over all nodes, where N is the number of nodes in the system.

```
mpirun -n 32 ./mpiBench -d 2 -p 2
```

MPI Messaging Rate

Measure the messaging rate for a single MPI process receiving messages from a single remote MPI processes. This should be run with 1 process per node over 2 nodes.

```
mpirun –n 2 ./sqmr --num_cores=1 --num_nbors=1
```

Measure the messaging rate for a single MPI process receiving messages from multiple MPI processes. This should be run with 1 process per node over 4 nodes.

```
mpirun –n 4 ./sqmr --num_cores=1 --num_nbors=3
```

Measure the aggregate messaging rate for MPI processes on a node receiving messages from multiple MPI processes. This should be run with multiple processes per node over multiple nodes. Example for 4 processes per node over 4 nodes:

```
mpirun -n 16 ./sqmr --num_cores=4 --num_nbors=3
```

Results of Interest

If reporting benchmark results, please provide the following information:

- Message Sizes
  - For each point-to-point latency benchmark
    - Provide message size results for powers of 2 to 4KB
  - For each point-to-point bandwidth benchmark
    - Provide message size results for powers of 2 to 4MB.
  - Provide SQMR messaging rate results for a message size of 8 bytes.
  - Provide Bcast and Allreduce results for 3 message sizes:
    - 8 bytes
    - a message size near the bandwidth latency product
    - a message size that demonstrates the maximum bandwidth for this test configuration
  - Provide Alltoall results for
    - 8 bytes
    - a message size near the bandwidth latency product
- Benchmark job sizes
  - Provide presta and mpiBench results for
    - 1 node
    - 2 nodes
    - the full system
  - Provide sqmr results for 1 target node and 1 or more neighbor nodes. See the examples above.
  - Provide results with the following process counts per node:
    - For single-node tests: powers of two from 2 to the number of cores
    - For latency tests: powers of two from 1 to the number of cores
    - For bandwidth tests: powers of two from 1 to number at which the maximum bandwidth is achieved
- When reporting results, please use minimum values for presta latency and mpiBench results. Provide maximum results for presta bandwidth and SQMR results.
- For presta latency and bandwidth results not run over a single node or the entire system, please provide optimal/neighbor results and worst-case results.
- Provide mpiBench results for the collectives: Barrier, Bcast, Allreduce and Alltoall.