Phloem MPI Benchmarks

Summary Version
1.3

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 and Examples
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 example
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 example
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 example

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

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

MPI point-to-point intra-node latency example

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

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

MPI point-to-point inter-node latency example

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 example

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

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

MPI Messaging Rate example

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 2 MPI processes receiving messages from 2 MPI processes. This should be run with 2 process per node over 3 nodes.

```
mpirun –n 6 ./sqmr --num_cores=2 --num_nbors=2
```

Measure the aggregate messaging rate for 4 MPI processes on a node receiving messages from multiple MPI processes. This should be run with 4 processes per node over 5 nodes.

```
mpirun -n 20 ./sqmr --num_cores=4 --num_nbors=4
```

Results of Interest

If reporting benchmark results, please provide the following information:

- Messaging rate
  - With one target node and 1 or more paired nodes, run MPI processes per node of powers of 2 from 1 to the max messaging rate for a message size of 8 bytes.
- Point-to-point latency
  - On one node, run MPI processes per node of powers of 2 from 2 to the number of cores for message sizes of powers of 2 up to 4KB.
  - On two neighbor nodes, run MPI processes per node of powers of 2 from 1 to the number of cores for message sizes of powers of 2 up to 4KB.
  - On two worst-case nodes, run MPI processes per node of powers of 2 from 1 to the number of cores for message sizes of powers of 2 up to 4KB.
- Aggregate Bidirectional Bandwidth
  - On one node, run MPI processes per node of powers of 2 from 2 to the maximum bandwidth achieved for message sizes of powers of 2 up to 4MB.
  - On two neighbor nodes, run MPI processes per node of powers of 2 from 1 to the maximum bandwidth achieved for message sizes of powers of 2 up to 4MB.
  - On two worst-case nodes, run MPI processes per node of powers of 2 from 1 to the maximum bandwidth achieved for message sizes of powers of 2 up to 4MB.
- Bisection Bandwidth
  - Over the entire system with worst-case process pairing, run MPI processes per node of powers of 2 from 1 to the maximum bandwidth achieved for message sizes of powers of 2 up to 4MBs.
- Collective Latency
  - Over the entire system, run MPI processes per node of powers of 2 from 1 to the number of cores with message sizes of 8B, a message size near the bandwidth latency product, and the smallest message size that results in the maximum inter-node bandwidth.
  - If necessary, message sizes may be reduced based on the available memory.
- When reporting results, please use minimum values for presta latency and mpiBench results. Provide maximum results for presta bandwidth and SQMR results.
- Provide mpiBench collective latency results for Barrier, Bcast, Allreduce and Alltoall.