Point-to-Point Message Performance on Blue Waters

William Gropp
www.cs.illinois.edu/~wgropp

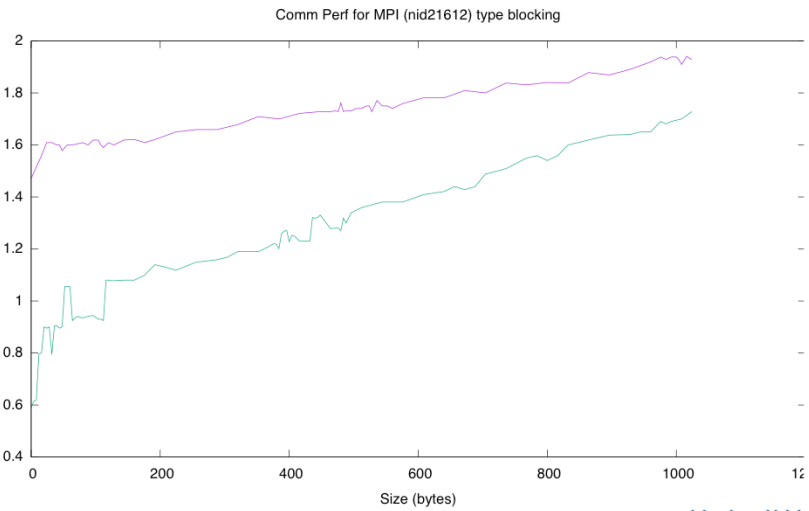
The Test Code

- mpptest
  - "Reproducible Measurements of MPI Performance Characteristics" http://link.springer.com/chapter/10.1007/3-540-48158-3_2
  - http://www.mcs.anl.gov/research/projects/mpi/mpptest/
- Key Features
  - Strives for reproducible results (care in measuring times; sequence of measurements designed to avoid some common disturbances
  - Adaptive message length to find transitions in performance
  - Many other options for different patterns, routines, etc.

The Tests

- Short messages (between 0-1024 bytes), both within a node and between two nodes
- Mid-size messages – extends to 4096 bytes
- Rendezvous – Extends to 16k bytes, enough to trigger rendezvous protocol on Blue Waters, internode only

Short Messages

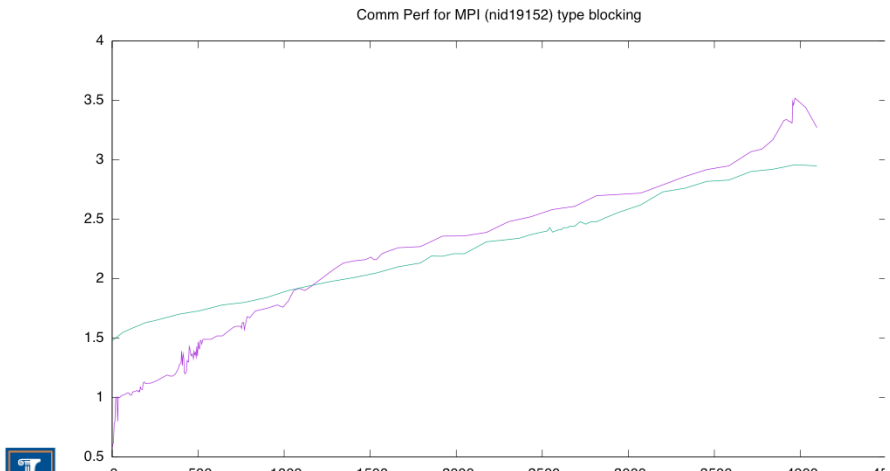


4

PARALLEL@ILLINOIS

Top line is internode, bottom intra (within) node.
Note differing slopes
Note very low latency for very short messages.
Note that getting a good fit for the T(n)= s+rn model requires avoiding the first 50-200 bytes, where the timing behavior is clearly different.

Mid-sized Messages

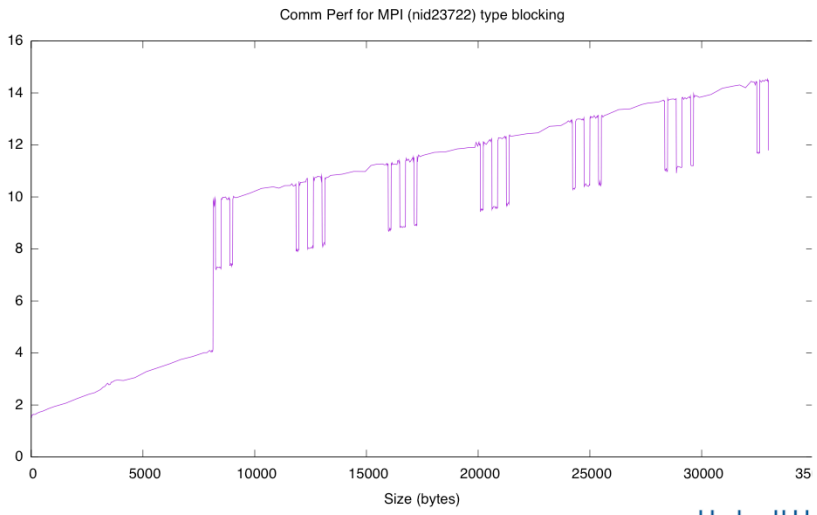


5

PARALLEL@ILLINOIS

Same colors, order as short. Note cross-over but similar slope after about 1200 bytes

Rendezvous Messages



6

PARALLEL@ILLINOIS

Note large jump at 8k. This is the rendezvous threshold
Note different slope before, after 8k. Likely due to change in how data is moved.
"Strange" behavior after 8k is due to "bug" in mpptest – adaptive size control sometimes picks sizes that are odd or not a multiple of 8; these are significantly more costly (over 2usec). Left in plot to illustrate the impact of such message sizes.