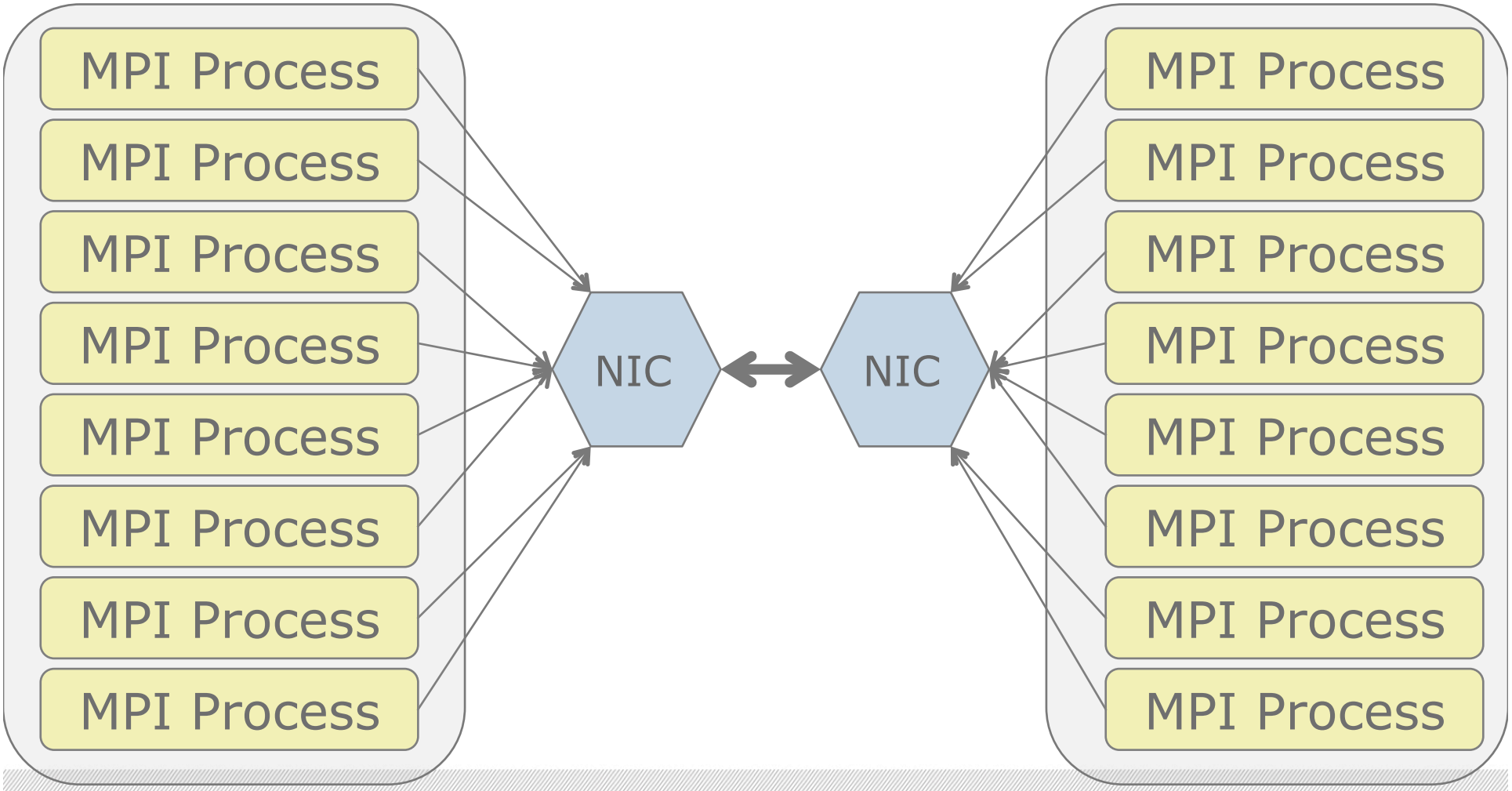
Is the Pingpong Communication Model Still Relevant?

William Gropp, Luke Olson and Philipp Samfass

Hormozd was one of my first students at Illinois, and his enthusiasm, curiosity, and drive made him a joy to work with. His excitement in developing new ways to think about performance for HPC algorithms convinced me to revisit this area, and his determination was inspirational.

Hormozd will be missed but never forgotten.

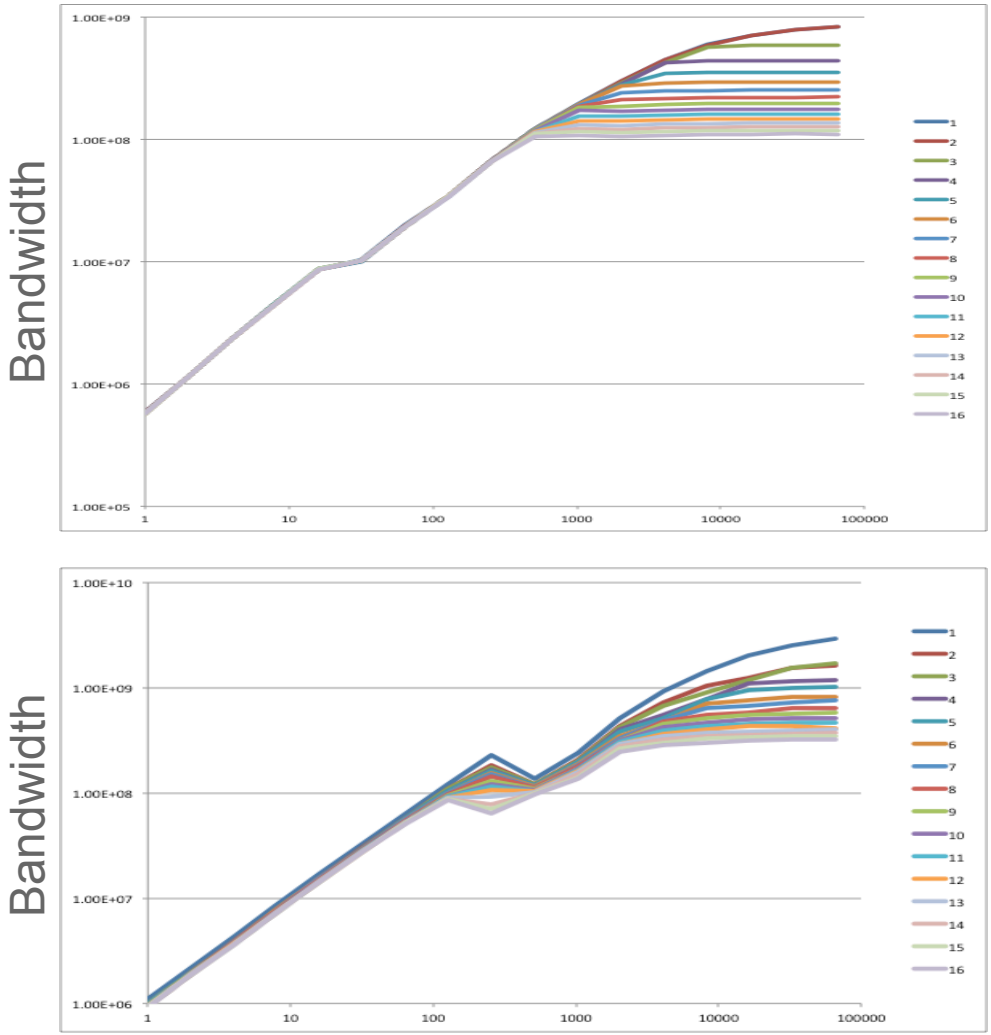
SMP Nodes: One Model



Classic Performance Model

- s + r n
  - Sometimes called the “postal model”
- Model combines overhead and network latency (s) and a single communication rate 1/r for n bytes of data
- Good fit to machines when it was introduced
- But does it match modern SMP-based machines?
  - Let’s look at the the communication rate per process with processes communicating between two nodes

Rates Per MPI Process

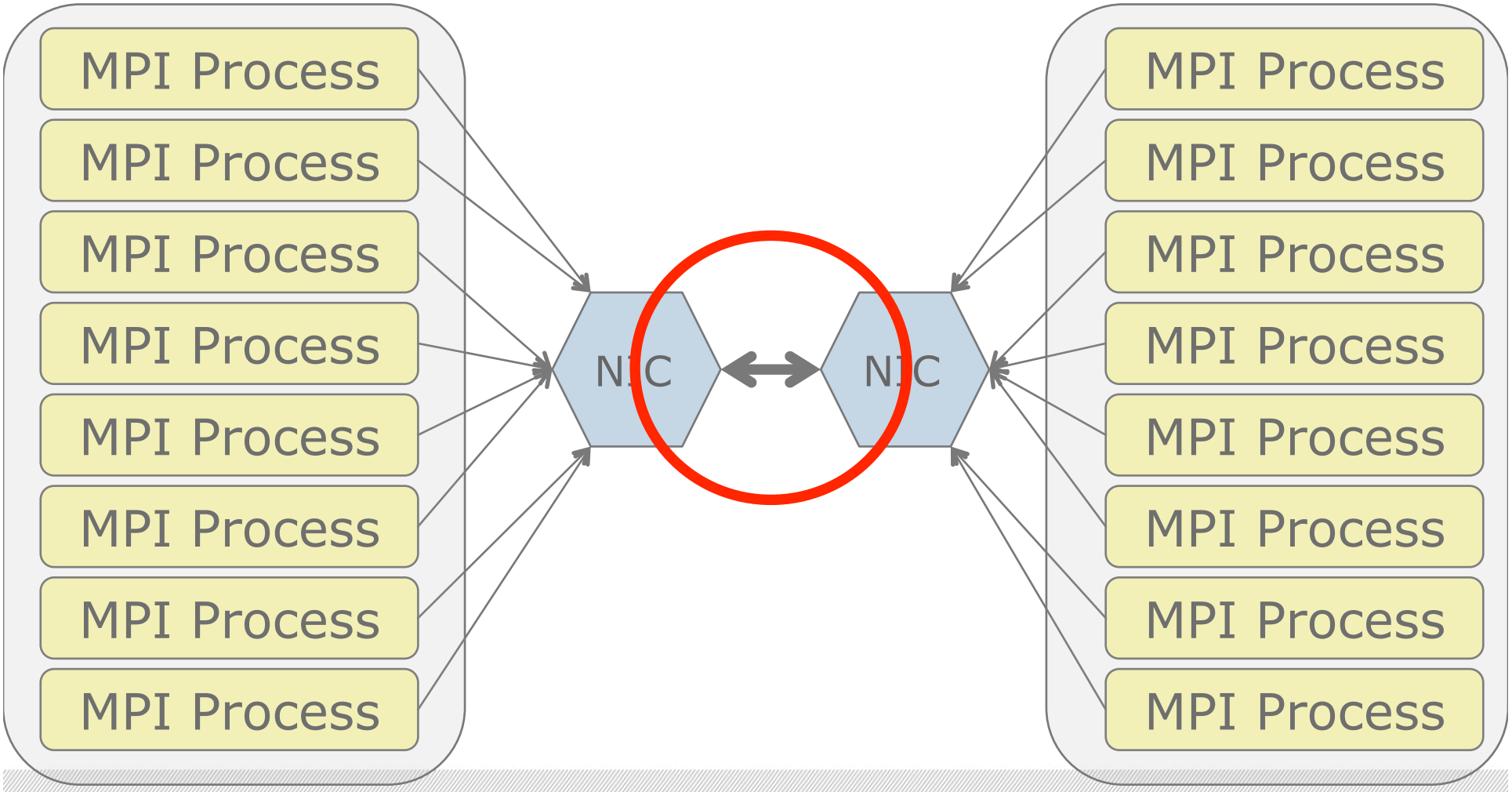


- Ping-pong between 2 nodes using 1-16 cores on each node
- Top is BG/Q, bottom Cray XE6
- "Classic" model predicts a single curve – rates independent of the number of communicating processes

Why this Behavior?

- The T = s + r n model predicts the *same* performance independent of the number of communicating processes
  - What is going on?
  - How should we model the time for communication?

SMP Nodes: One Model

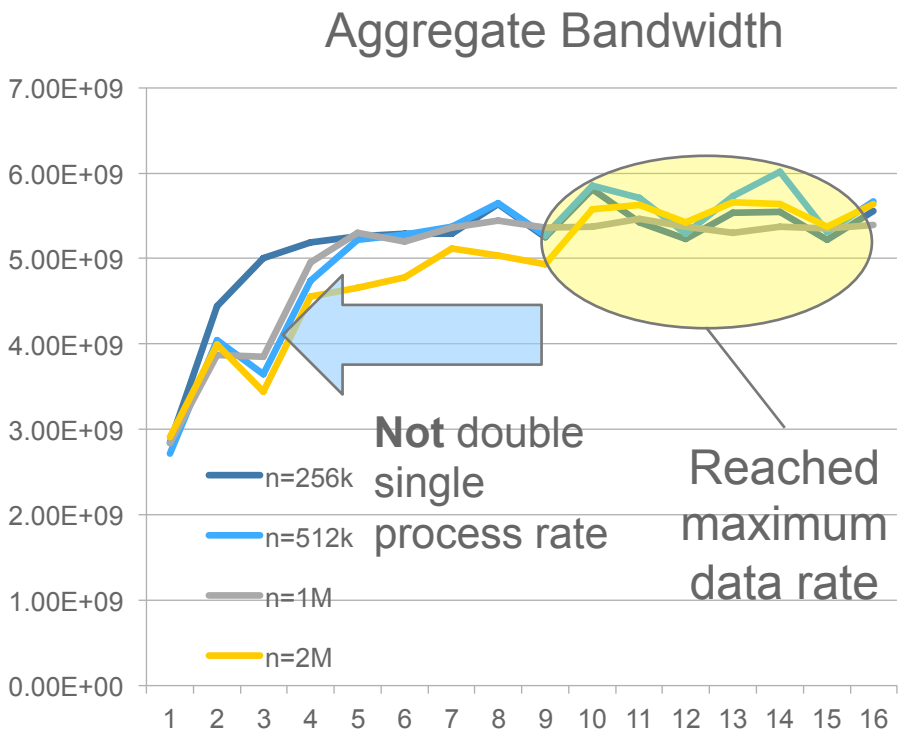


Modeling the Communication

- Each link can support a rate r_L of data
- Data is pipelined (Logp model)
  - Store and forward analysis is different
- Overhead is completely parallel
  - k processes sending one short message each takes the same time as one process sending one short message

A Slightly Better Model

- Assume that the sustained communication rate is limited by
  - The maximum rate along any shared link
    - The link between NICs
  - The aggregate rate along parallel links
    - Each of the “links” from an MPI process to/from the NIC



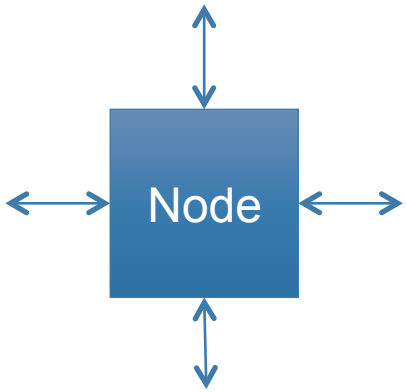
A Slightly Better Model

- For k processes sending messages, the sustained rate is
  - $\min(R_{NIC\text{-}NIC}, k\ R_{CORE\text{-}NIC})$
- Thus
  - $T = s + k\ n/\min(R_{NIC\text{-}NIC}, k\ R_{CORE\text{-}NIC})$
- Note if $R_{NIC\text{-}NIC}$ is very large (very fast network), this reduces to
  - $T = s + k\ n/(k\ R_{CORE\text{-}NIC}) = s + n/R_{CORE\text{-}NIC}$

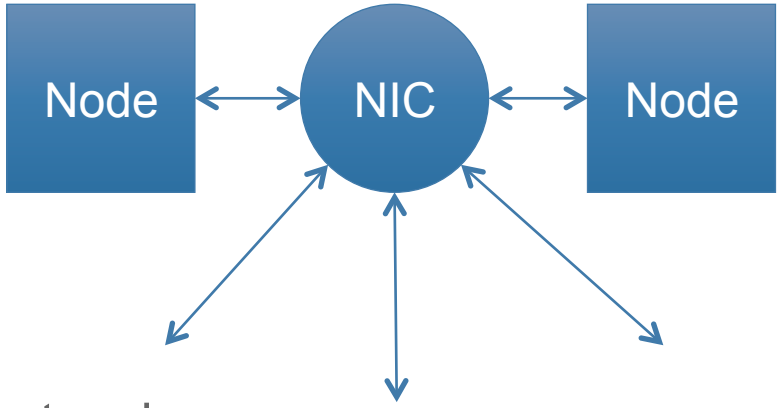
Two Examples

- Two simplified examples:

Blue Gene/Q



Cray XE6



- Note differences:
  - BG/Q : Multiple paths into the network
  - Cray XE6: Single path to NIC (shared by 2 nodes)
  - Multiple processes on a node sending can exceed the available bandwidth of the single path

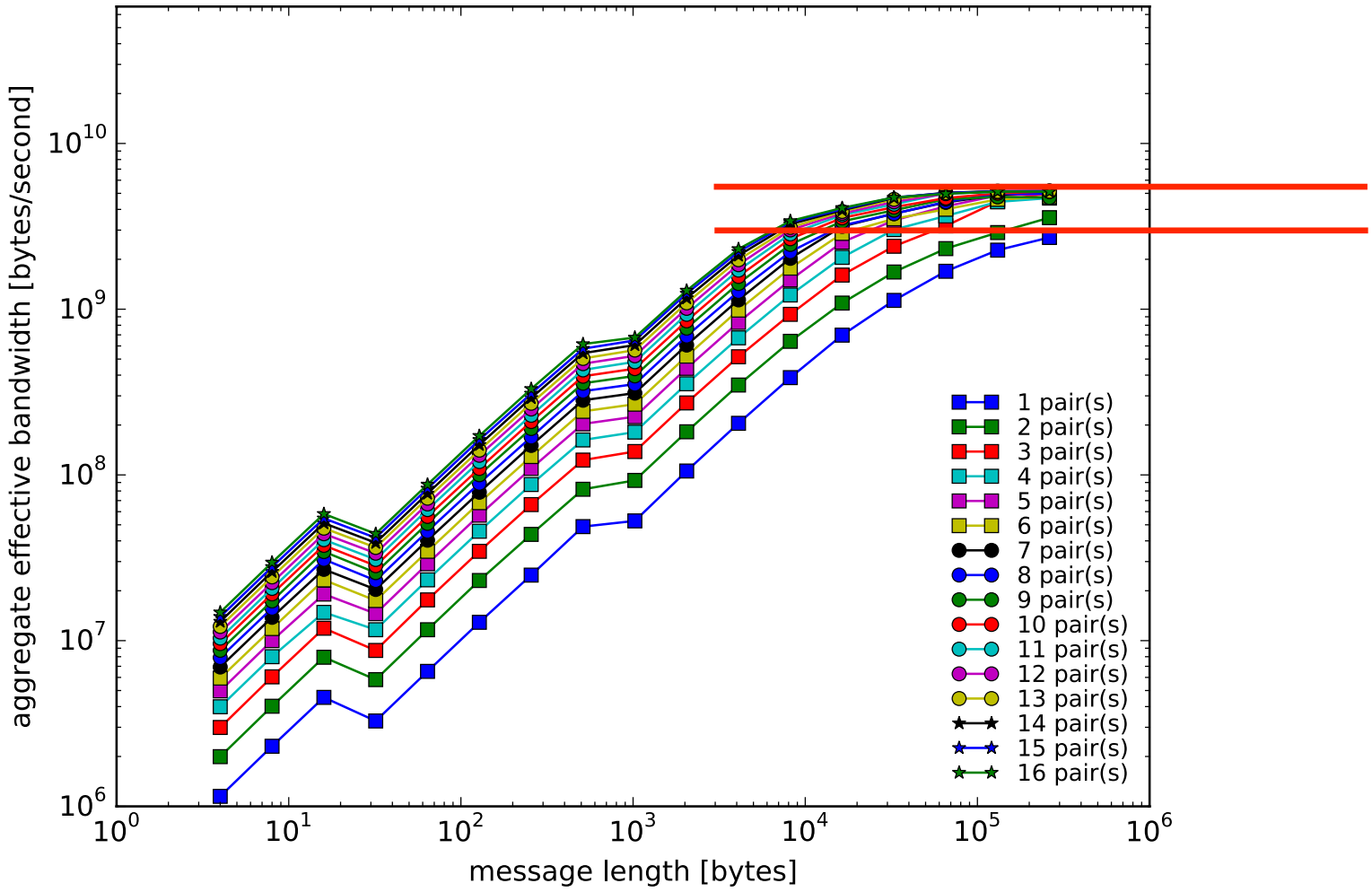
The Test

- Nodecomm discovers the underlying physical topology
- Performs point-to-point communication (ping-pong) using 1 to # cores per node to another node (or another chip if a node has multiple chips)
- Outputs communication time for 1 to # cores along a single channel
  - Note that hardware may route some communication along a longer path to avoid contention.
- The following results use the code available soon at
  - https://bitbucket.org/william_gropp/baseenv

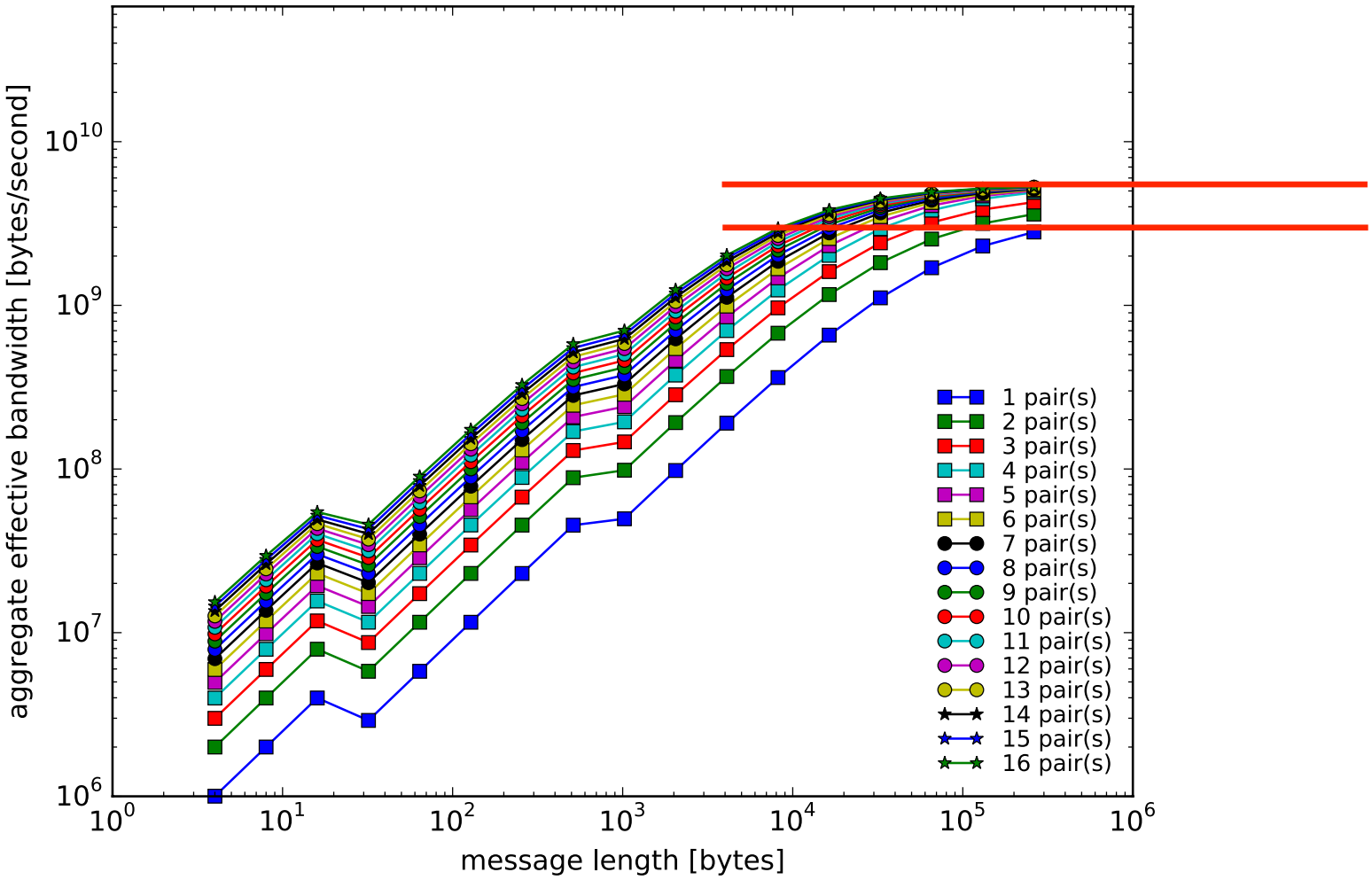
How Well Does this Model Work?

- Tested on a wide range of systems:
  - Cray XE6 with Gemini network
  - IBM BG/Q
  - Cluster with InfiniBand
  - Cluster with another network
- Results in
  - Modeling MPI Communication Performance on SMP Nodes: Is it Time to Retire the Ping Pong Test
    - W Gropp, L Olson, P Samfass
    - Proceedings of EuroMPI 16
    - https://doi.org/10.1145/2966884.2966919
- Cray XE6 results follow

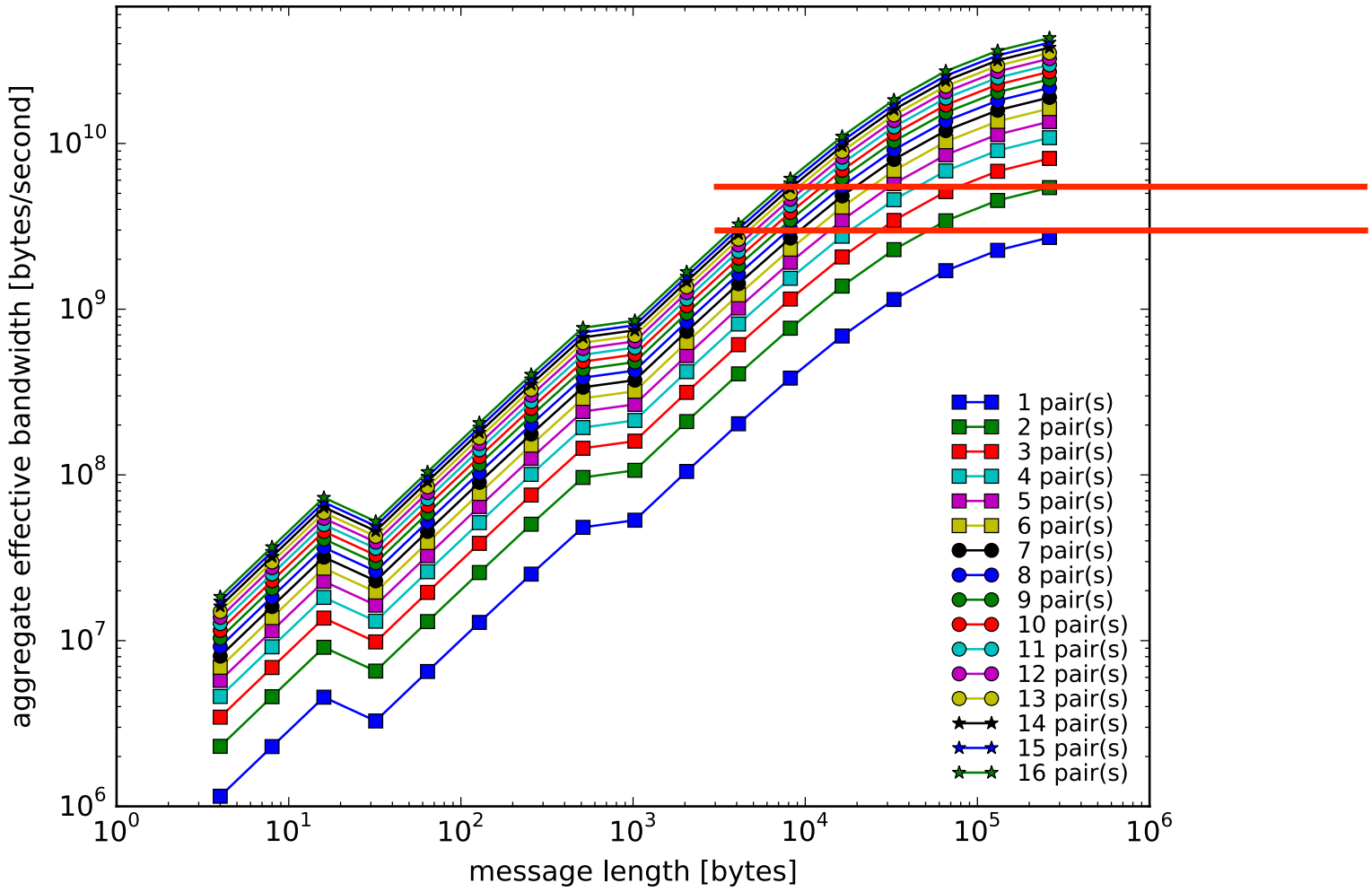
Cray: Measured Data



Cray: 3 parameter (new) model



Cray: 2 parameter model



Notes

- Both Cray XE6 and IBM BG/Q have inadequate bandwidth to support each core sending data along the same link
  - But BG/Q has more independent links, so it is able to sustain a higher effective “halo exchange”

Modeling Communication

- For k processes sending messages concurrently from the same node, the correct (more precisely, a much better) time model is
  - $T = s + k\, n/\min(R_{NIC\text{-}NIC}, k\, R_{CORE\text{-}NIC})$
- Further terms improve this model, but this one is sufficient for many uses

Conclusion

- Yes, it is time to retire (or at least augment) the pingpong test
- Fortunately, a single additional parameter significantly improves the value of the communication performance model
- For algorithm and code designers, an additional message
  - Distribute communication in time so that off-node communication is less of a bottleneck

Thanks!

- ExxonMobile Upstream Research
- Blue Waters Sustained Petascale Project, supported by the National Science Foundation (award number OCI 07–25070) and the state of Illinois.
- Cisco Systems for access to the Arcetri UCS Balanced Technical Computing Cluster