CISL Computational & Information Systems Laboratory

Performance Analysis of MPI over InfiniBand on Yellowstone

Zhengyang Liu

Mentor: Dr. John Dennis

Collaborators:
Prof. Malathi Veeraraghavan (University of Virginia)
Prof. Robert D. Russell (University of New Hampshire)
Fabrice Mizero (SIParCS)
Patrick MacArthur (University of New Hampshire)

Aug 1, 2013

NCAR NSF UNIVERSITY of VIRGINIA

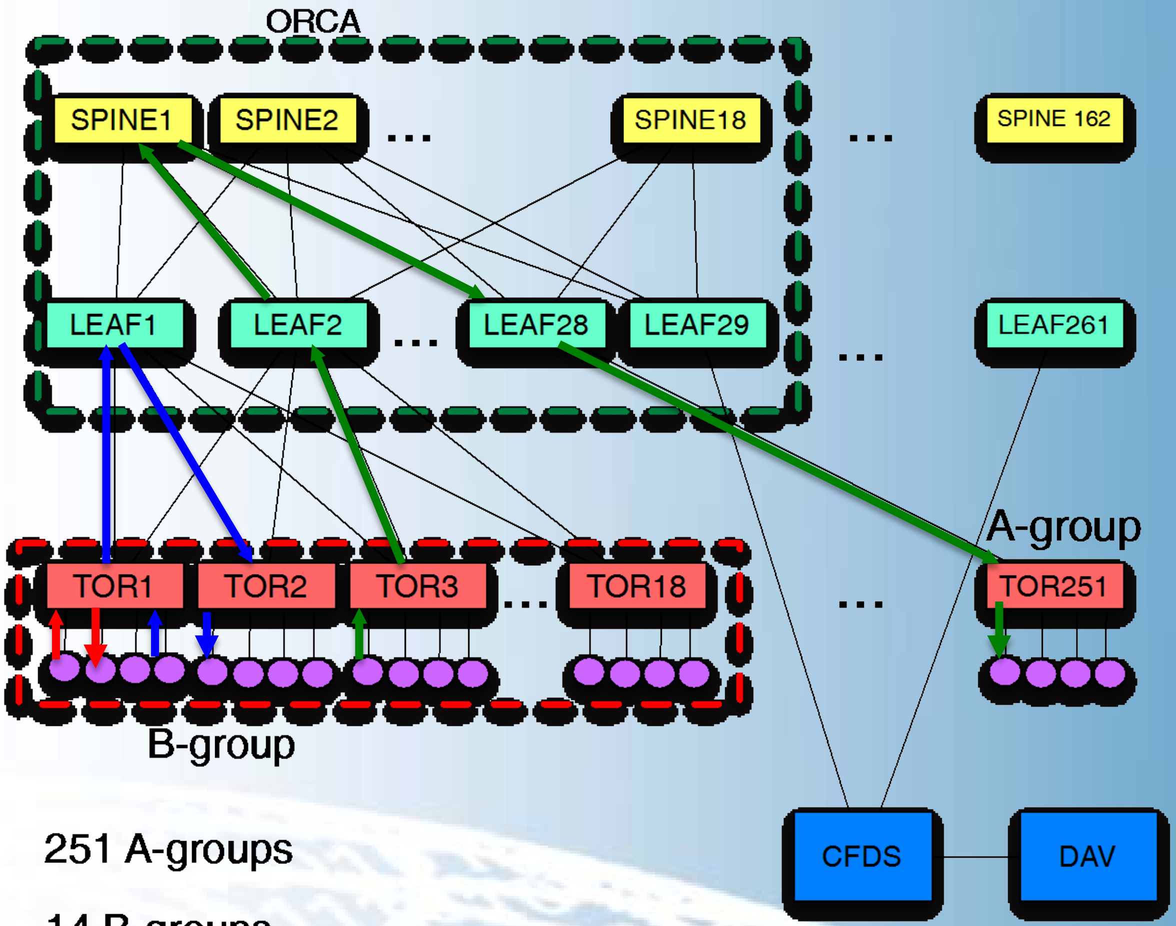
Big Picture

- **Understanding the causes of poor performance of CESM on Yellowstone: a 5-step approach**
  - Experimental execution and data collection
  - CESM trace analysis
  - IBMgtSim: routing study
  - Network simulation
  - Integrated simulation

Big Picture

- **Understanding the causes of poor performance of CESM on Yellowstone: a 5-step approach**
  - Experimental execution and data collection
  - CESM trace analysis
  - IBMgtSim: routing study
  - Network simulation
  - Integrated simulation

Yellowstone network



251 A-groups

14 B-groups

9 ORCAs

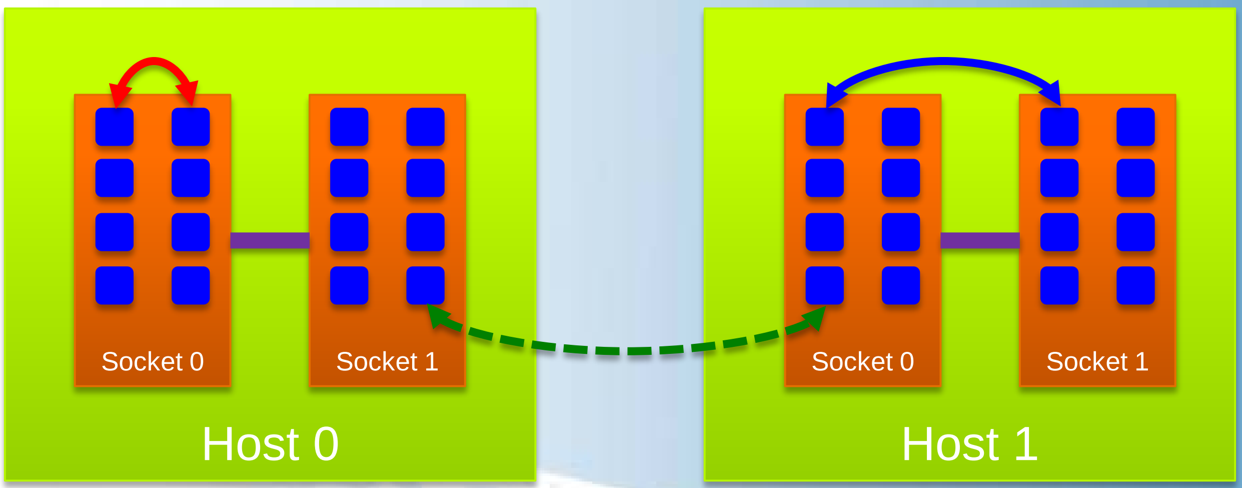
→ 2-hop

→ 4-hop

→ 6-hop

*Credit: Dr. John Dennis

Communication Patterns



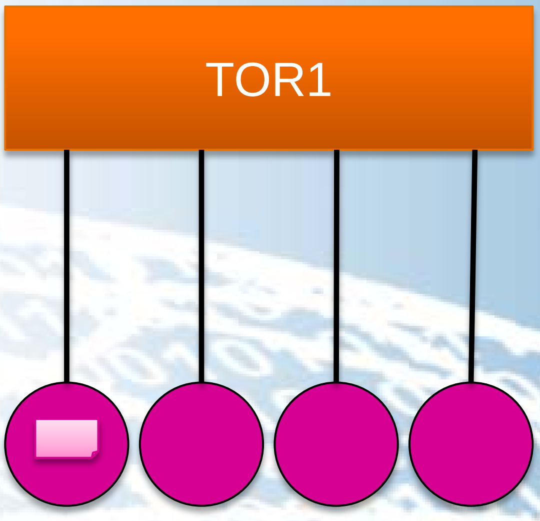
⟷ Intra-socket (via shared cache/memory)

⟷ Inter-socket (via shared memory over QuickPath Interconnect)

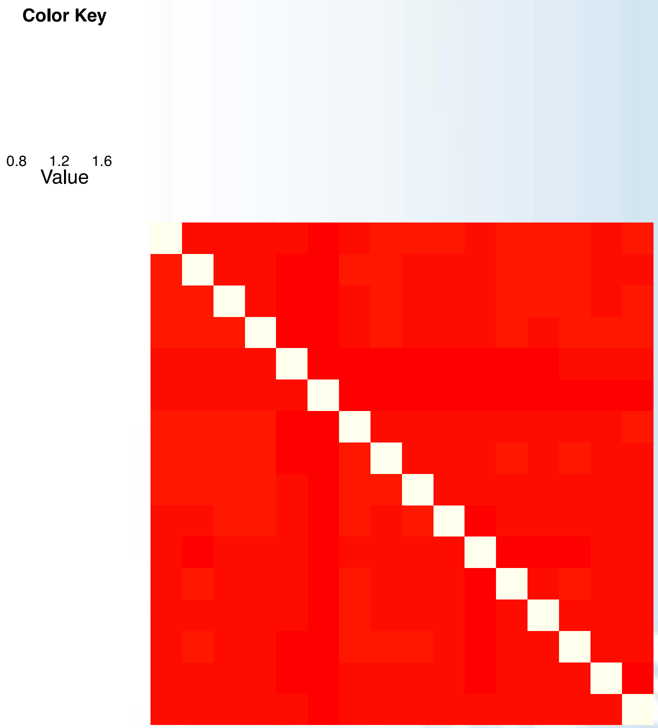
⇠⇢ Inter-node (via InfiniBand)

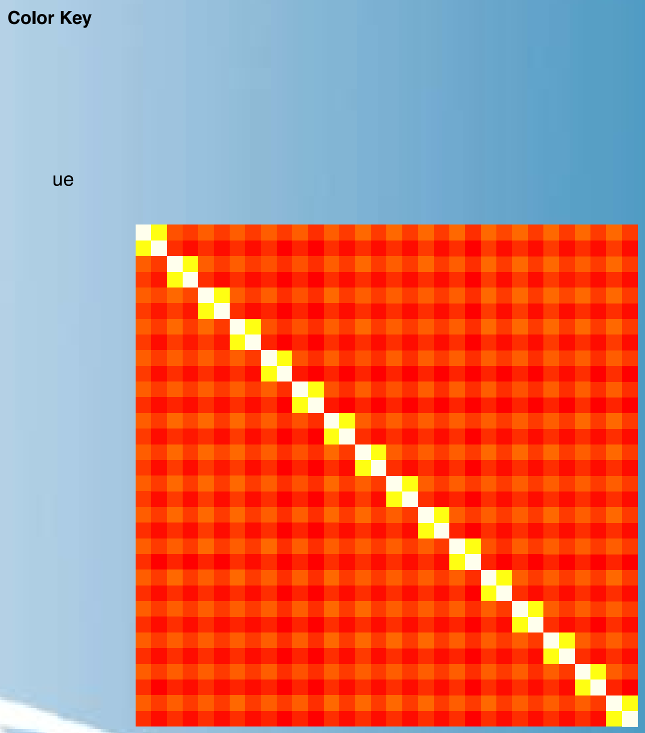
Latency Benchmark: mpi_pingpong

- **Approximate one-way latency by measuring round-trip latency**
- **Results represent ideal latencies between nodes**



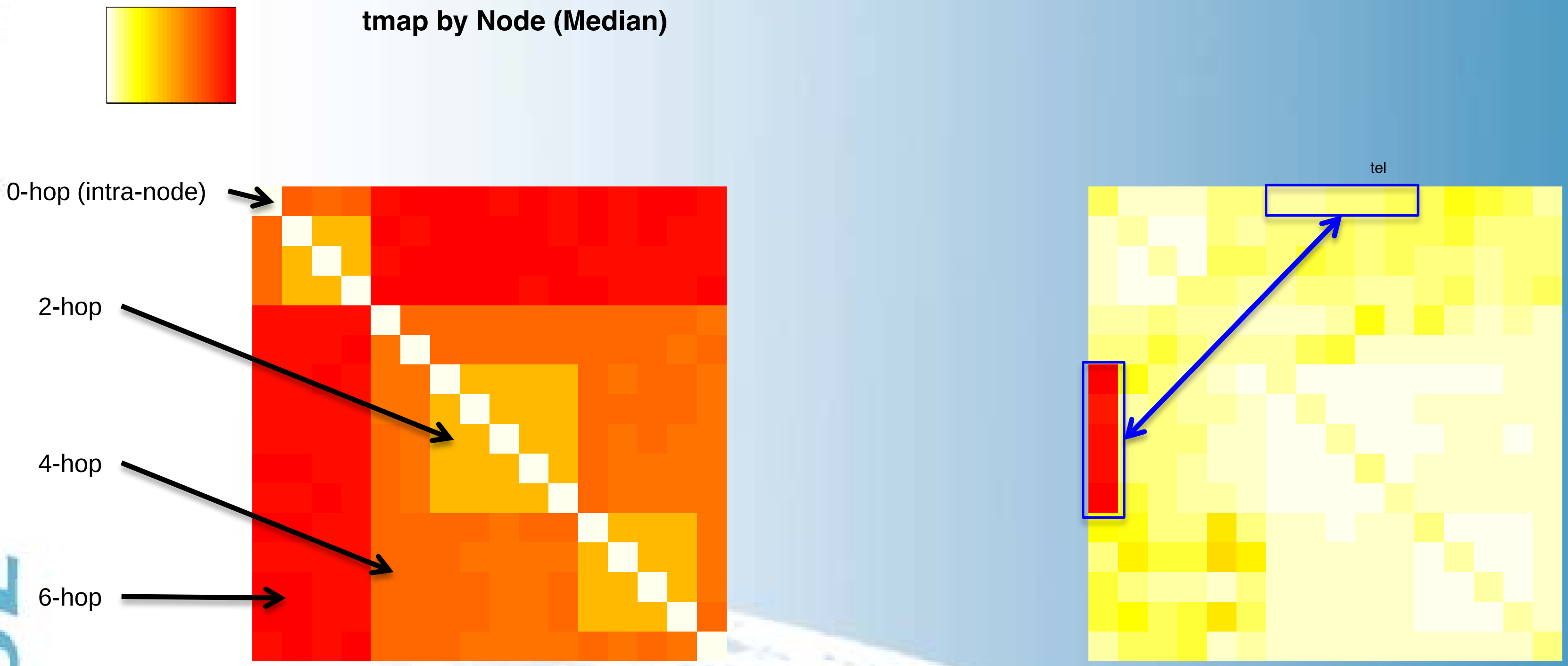
Jellystone Results





0-byte messages; all hosts connected to the same TOR; no competing processes

Yellowstone Results



0-byte messages

256 KB messages

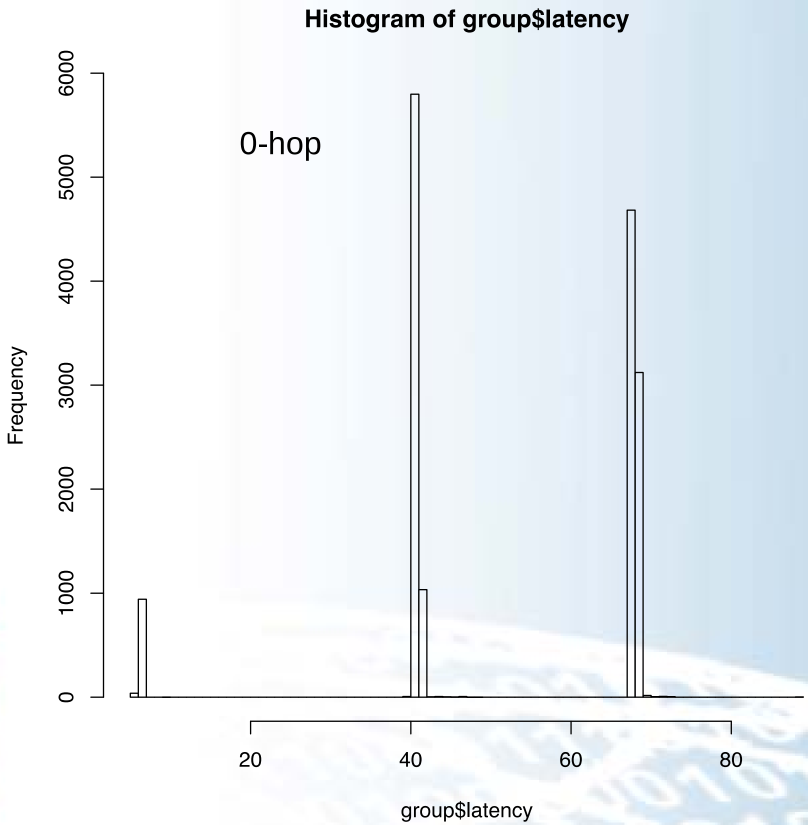
Latency vs. # of Hops

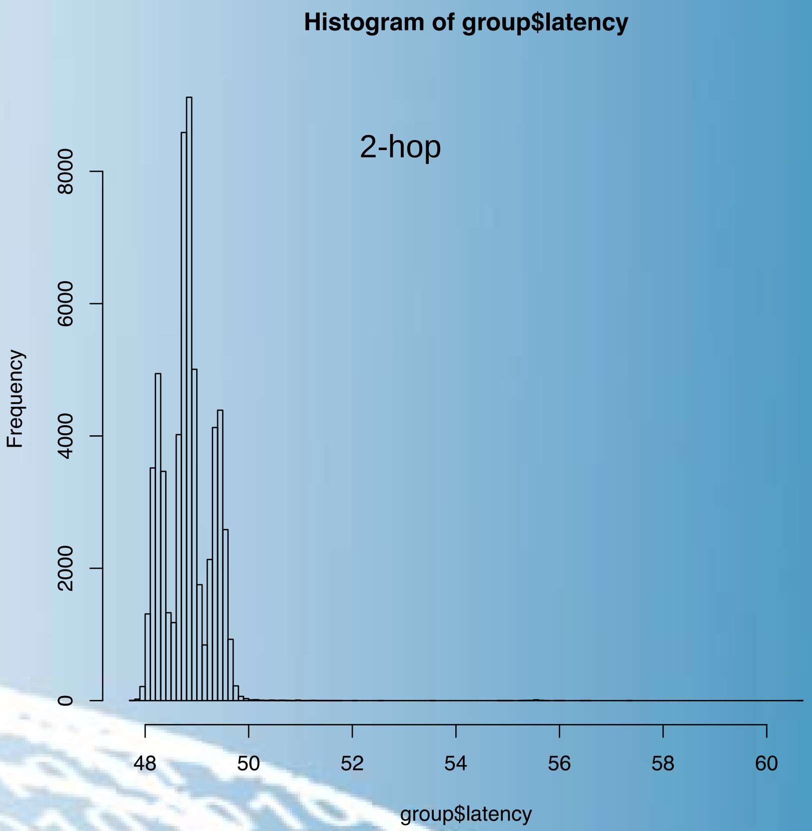
- **Experiment:**
  - mpi_pingpong on 1024 cores
  - 1,048,576 communication pairs*
  - 256 KB messages

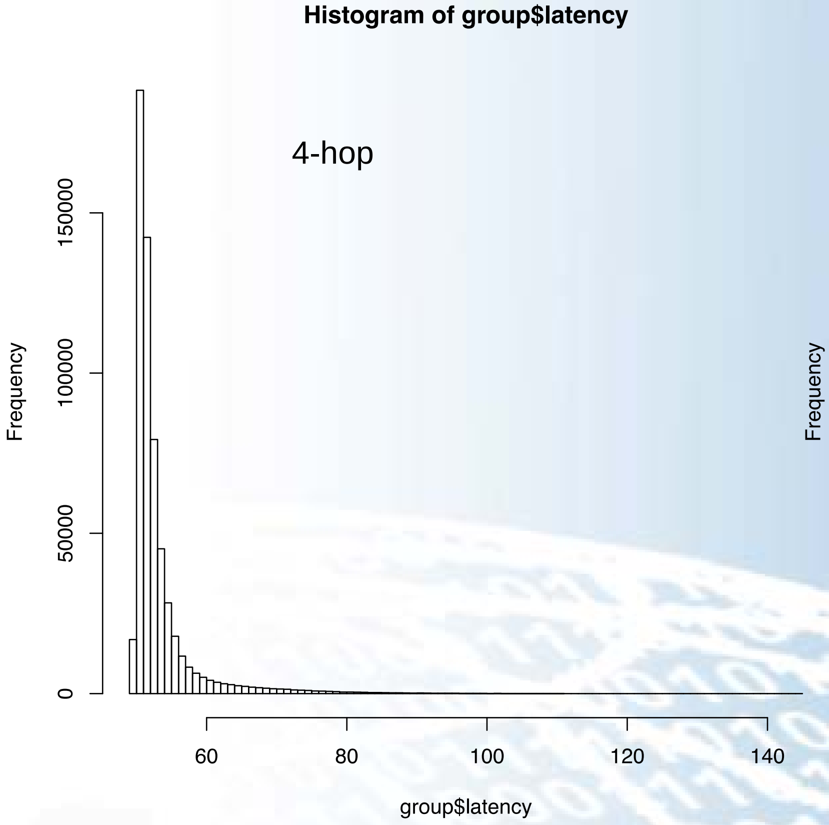
Unit: μs

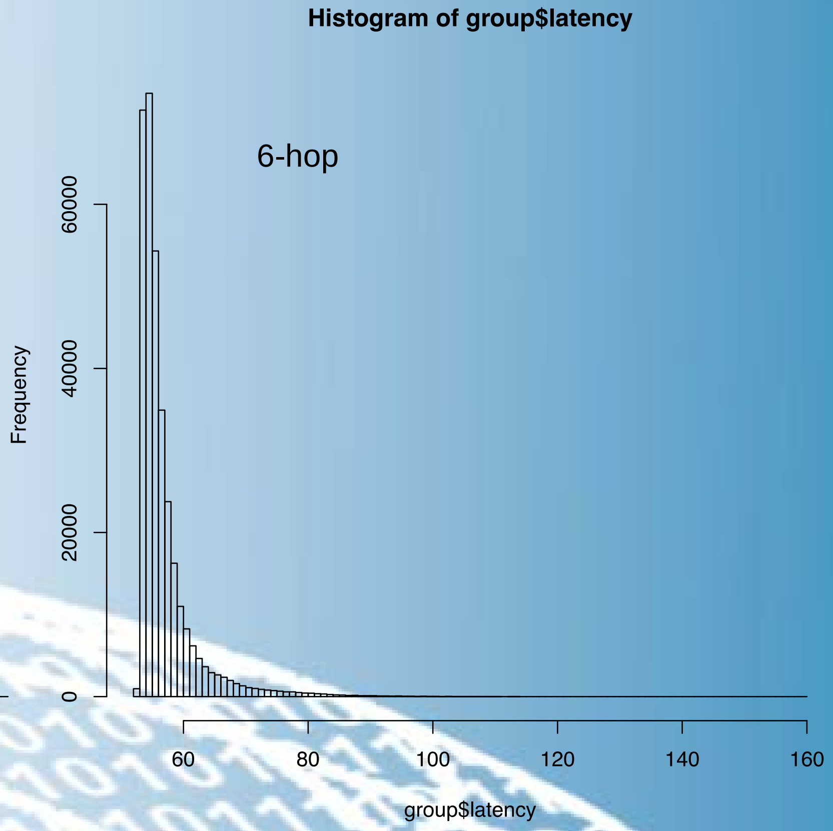
# of Hops	# of pairs	Min.	Avg.	Max.
0	15,680	5.95	52.20	88.29
2	59,904	47.72	48.83	60.64
4	588,736	49.55	53.30	114.10
6	332,016	52.75	56.92	159.30

*routing data available for 996,336 pairs


Histogram of group$latency
0-hop
Frequency
0
1000
2000
3000
4000
5000
6000
20
40
60
80
group$latency


Histogram of group$latency
2-hop
Frequency
0
2000
4000
6000
8000
48
50
52
54
56
58
60
group$latency


Histogram of group$latency
4-hop
Frequency
0
50000
100000
150000
60
80
100
120
140
group$latency


Histogram of group$latency
6-hop
Frequency
0
20000
40000
60000
60
80
100
120
140
160
group$latency

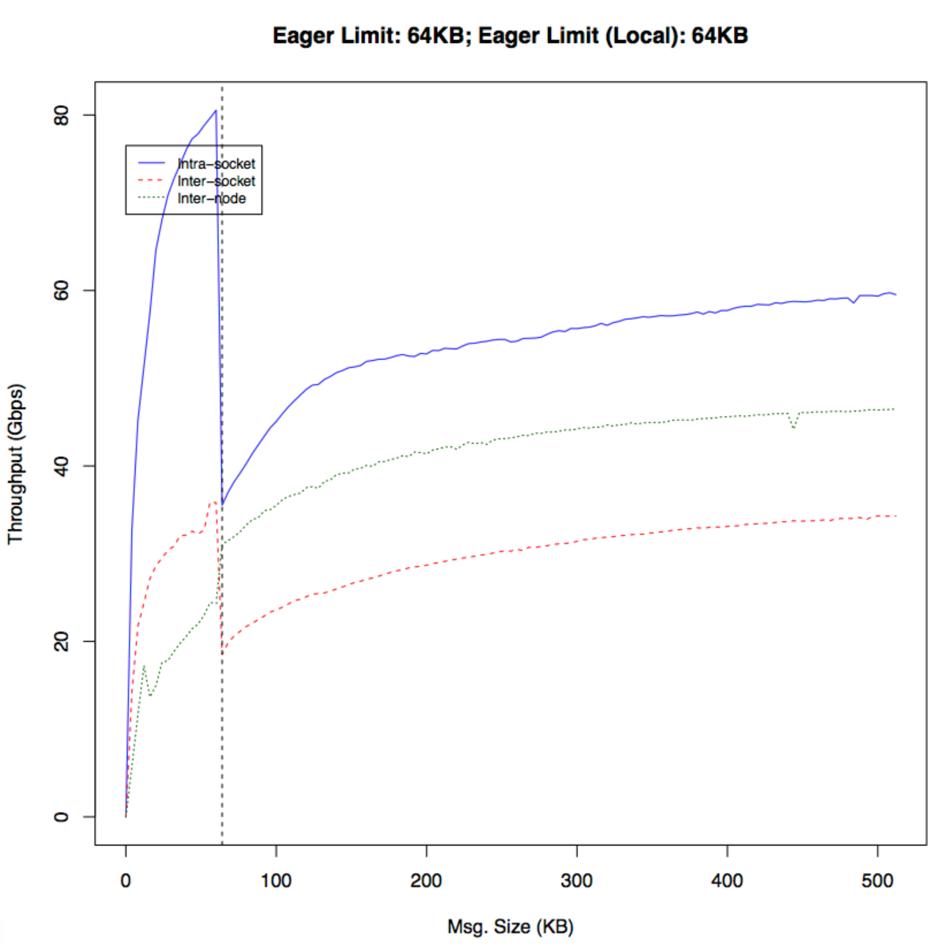
Bandwidth Benchmark: mpi_bw

- **Measures throughput between two MPI ranks**
- **3 communication patterns:**
  - Intra-socket
  - Inter-socket
  - Inter-node
- **2 communication protocols:**
  - Eager protocol
  - Rendezvous protocol

Communication Protocols

- Rendezvous Protocol: buffer negotiation before sending
- Eager Protocol: send directly without confirming available buffer space
- InfiniBand: Eager protocol uses SEND/RECV verbs (two-sided communication); Rendezvous protocol uses WRITE/READ verbs (one-sided communication)
- Eager Limit: threshold below which Eager protocol is used

Jellystone Results



- Intra-node throughput decreases when msg. size > eager limit
- Inter-node throughput increases when msg. size > eager limit
- Inter-node communication faster than inter-socket communication: RDMA vs shared memory

Summary

- Identified contention through mpi_pingpong benchmarks
- Studied effect of different communication patterns/protocols on throughput

Future Work

- **Analyses of larger data sets**
  - > 500 million data points
  - Analysis needs to be parallelized
- **Study interaction between MPI and InfiniBand**
  - Open-source MPI implementations
  - Network sniffing

CISL Computational & Information Systems Laboratory

Thank You

Zhengyang Liu
zl4ef@virginia.edu

NCAR NSF UNIVERSITY of VIRGINIA