

LA-UR 10-05188



IMPLEMENTATION & COMPARISON OF RDMA OVER ETHERNET

Students:

Lee Gaiser, Brian Kraus, and James Wernicke

Mentors:

Andree Jacobson, Susan Coulter, Jharrod LaFon, and Ben McClelland

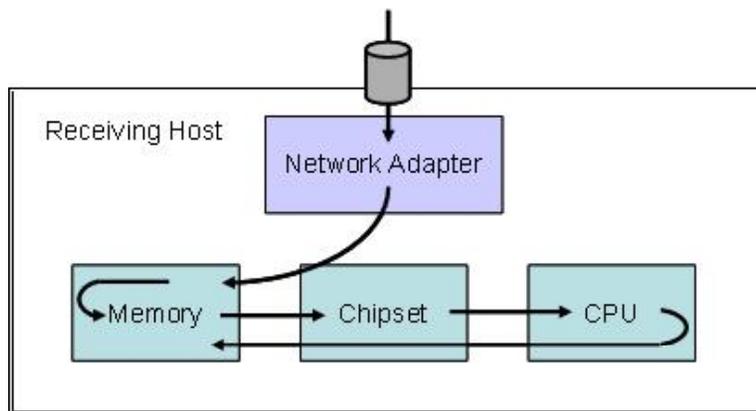
Summary



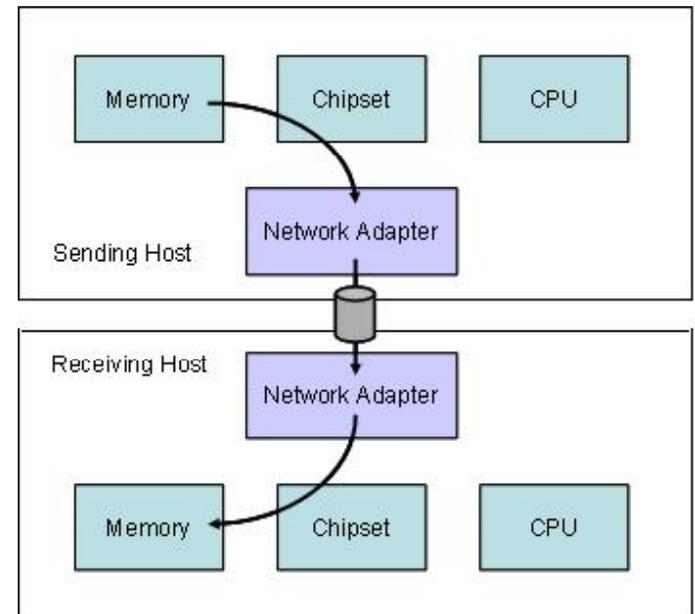
- Background
- Objective
- Testing Environment
- Methodology
- Results
- Conclusion
- Further Work
- Challenges
- Lessons Learned
- Acknowledgments
- References & Links
- Questions

Background : Remote Direct Memory Access (RDMA)

- RDMA provides high-throughput, low-latency networking:
 - ▣ Reduce consumption of CPU cycles
 - ▣ Reduce communication latency



Memory Copy



Zero-copy flow

Images courtesy of <http://www.hpcwire.com/features/17888274.html>

Background : InfiniBand



- Infiniband is a switched fabric communication link designed for HPC:
 - High throughput
 - Low latency
 - Quality of service
 - Failover
 - Scalability
 - Reliable transport
- How do we interface this high performance link with existing Ethernet infrastructure?

Background : RDMA over Converged Ethernet (RoCE)

- Provide Infiniband-like performance and efficiency to ubiquitous Ethernet infrastructure.
 - ▣ Utilize the same transport and network layers from IB stack and swap the link layer for Ethernet.
 - ▣ Implement IB verbs over Ethernet.
- Not quite IB strength, but it's getting close.
- As of OFED 1.5.1, code written for OFED RDMA auto-magically works with RoCE.

Objective



We would like to answer the following questions:

- What kind of performance can we get out of RoCE on our cluster?
- Can we implement RoCE in software (Soft RoCE) and how does it compare with hardware RoCE?

Testing Environment

Hardware:

- HP ProLiant DL160 G6 servers
- Mellanox MNPH29B-XTC 10GbE adapters
- 50/125 OFNR cabling

Operating System:

- CentOS 5.3
- 2.6.32.16 kernel

Software/Drivers:

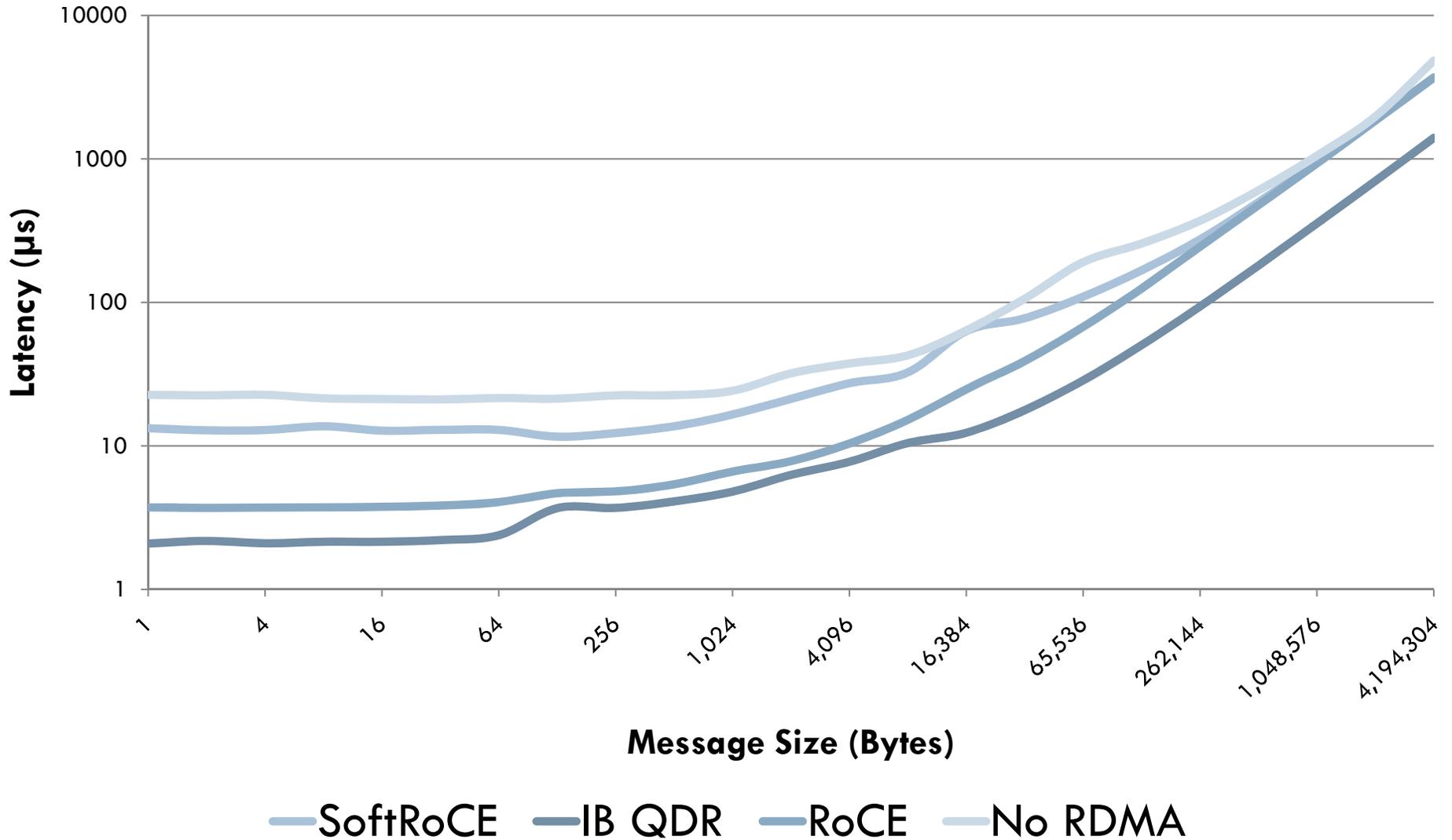
- Open Fabrics Enterprise Distribution (OFED) 1.5.2-rc2 (RoCE) & 1.5.1-rxe (Soft RoCE)
- OSU Micro Benchmarks (OMB) 3.1.1
- OpenMPI 1.4.2



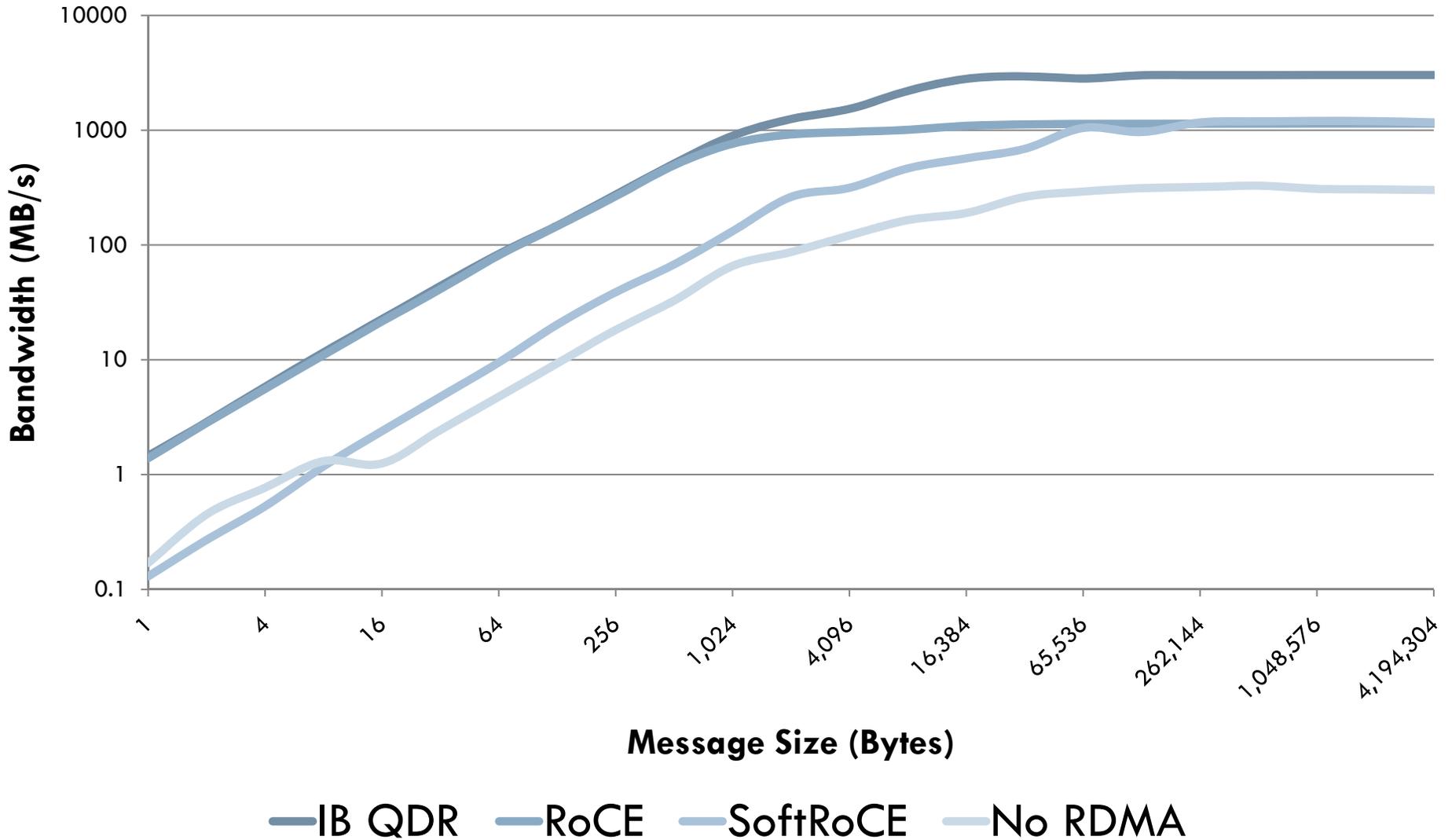
Methodology

- Set up a pair of nodes for each technology:
 - ▣ IB, RoCE, Soft RoCE, and no RDMA
- Install, configure & run minimal services on test nodes to maximize machine performance.
- Directly connect nodes to maximize network performance.
- Acquire latency benchmarks
 - ▣ OSU MPI Latency Test
- Acquire bandwidth benchmarks
 - ▣ OSU MPI Uni-Directional Bandwidth Test
 - ▣ OSU MPI Bi-Directional Bandwidth Test
- Script it all to perform many repetitions

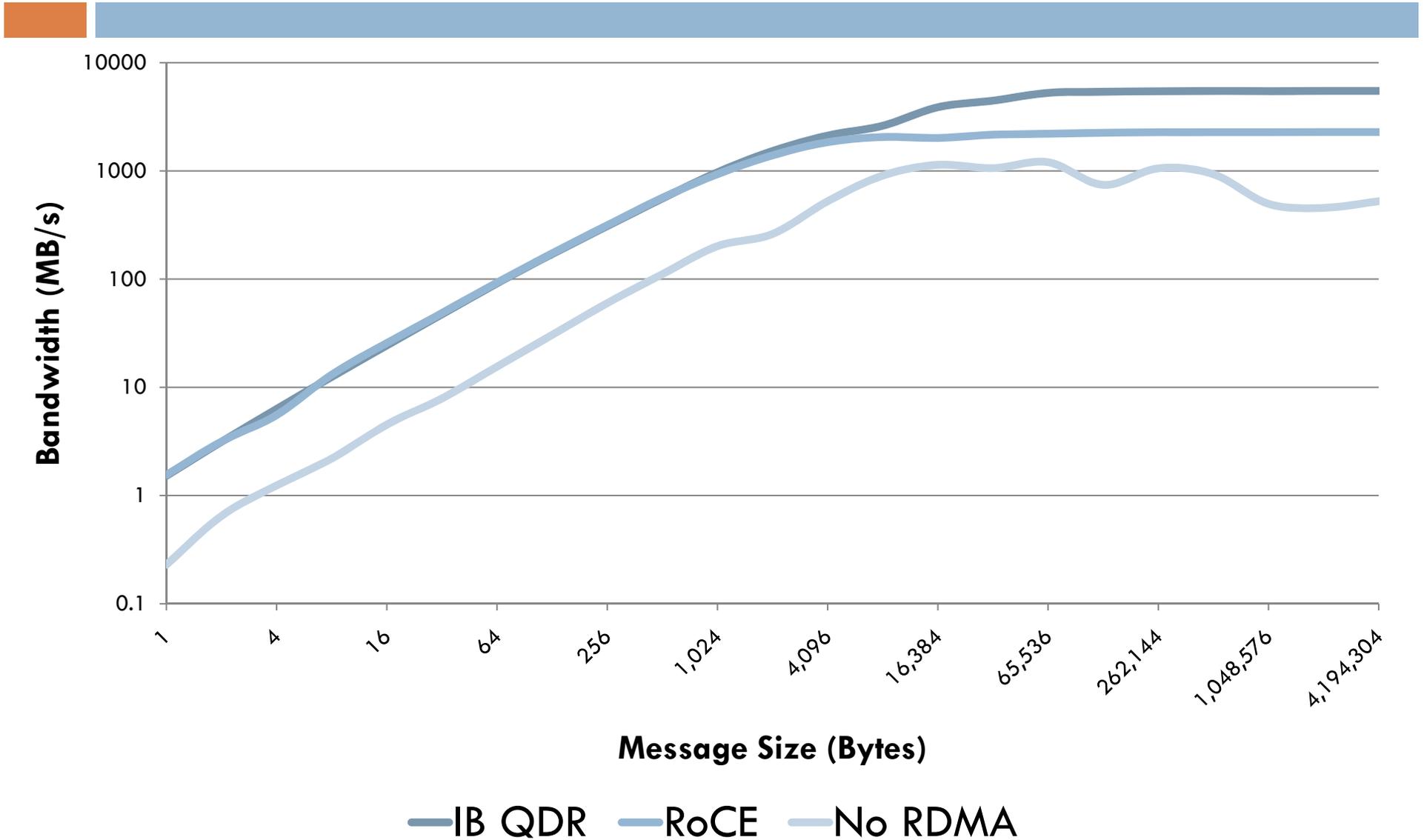
Results : Latency



Results : Uni-directional Bandwidth



Results : Bi-directional Bandwidth



Results : Analysis

Peak Values	IB QDR	RoCE	Soft RoCE	No RDMA
Latency (μ s)	1.96	3.7	11.6	21.09
One-way BW (MB/s)	3024.8	1142.7	1204.1	301.31
Two-way BW (MB/s)	5481.9	2284.7	-	1136.1

- RoCE performance gains over 10GbE:
 - Up to 5.7x speedup in latency
 - Up to 3.7x increase in bandwidth
- IB QDR vs. RoCE:
 - IB less than 1 μ s faster than RoCE at 128-byte message.
 - IB peak bandwidth is 2-2.5x greater than RoCE.

Conclusion



- RoCE is capable of providing near-Infiniband QDR performance for:
 - ▣ Latency-critical applications at message sizes from 128B to 8KB
 - ▣ Bandwidth-intensive applications for messages <1KB.
- Soft RoCE is comparable to hardware RoCE at message sizes above 65KB.
- Soft RoCE can improve performance where RoCE-enabled hardware is unavailable.

Further Work & Questions



- How does RoCE perform over collectives?
- Can we further optimize RoCE configuration to yield better performance?
- Can we stabilize the Soft RoCE configuration?
- How much does Soft RoCE affect the compute nodes ability to perform?
- How does RoCE compare with iWARP?

Challenges



- Finding an OS that works with OFED & RDMA:
 - Fedora 13 was too new.
 - Ubuntu 10 wasn't supported.
 - CentOS 5.5 was missing some drivers.
- Had to compile a new kernel with IB/RoCE support.
- Built OpenMPI 1.4.2 from source, but wasn't configured for RDMA; used OpenMPI 1.4.1 supplied with OFED instead.
- The machines communicating via Soft RoCE frequently lock up during OSU bandwidth tests.

Lessons Learned



- ❑ Installing and configuring HPC clusters
- ❑ Building, installing, and fixing Linux kernel, modules, and drivers
- ❑ Working with IB, 10GbE, and RDMA technologies
- ❑ Using tools such as OMB-3.1.1 and netperf for benchmarking performance

Acknowledgments



Andree Jacobson

Susan Coulter

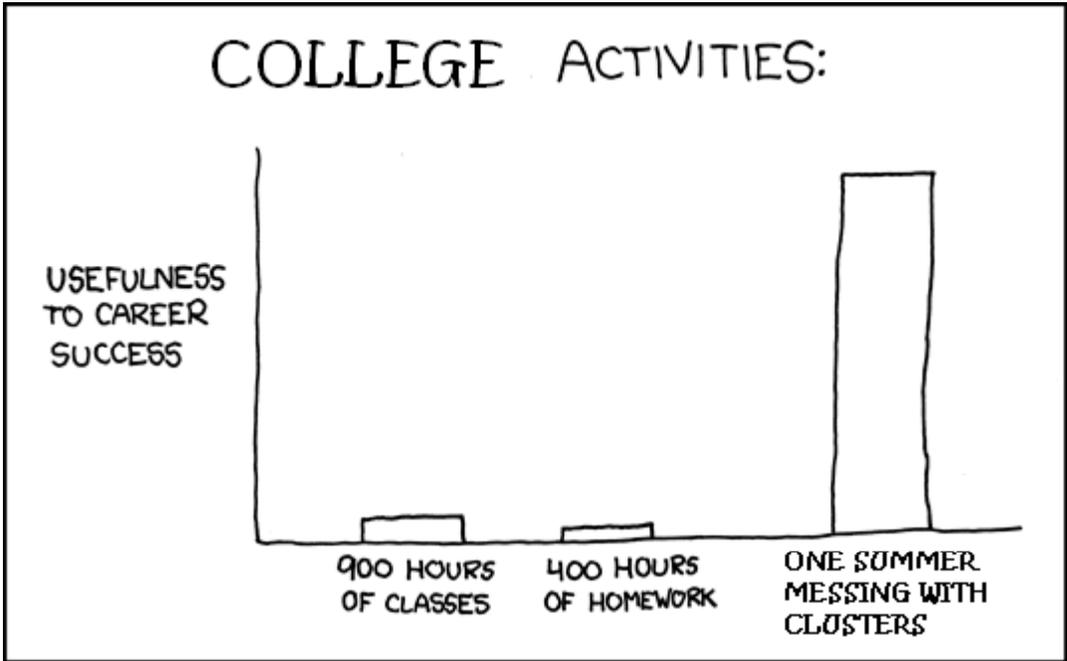
Jharrod LaFon

Ben McClelland

Sam Gutierrez

Bob Pearson

Questions?



References & Links

- Submaroni, H. et al. **RDMA over Ethernet – A Preliminary Study**. OSU.
<http://nowlab.cse.ohio-state.edu/publications/conf-presentations/2009/subramoni-hpidc09.pdf>
- Feldman, M. **RoCE: An Ethernet-InfiniBand Love Story**. HPCWire.com.
April 22, 2010.
<http://www.hpcwire.com/blogs/RoCE-An-Ethernet-InfiniBand-Love-Story-91866499.html>
- Woodruff, R. **Access to InfiniBand from Linux**. Intel. October 29, 2009.
<http://software.intel.com/en-us/articles/access-to-infiniband-from-linux/>
- OFED 1.5.2-rc2
<http://www.openfabrics.org/downloads/OFED/ofed-1.5.2/OFED-1.5.2-rc2.tgz>
- OFED 1.5.1-rxe
<http://www.systemfabricworks.com/pub/OFED-1.5.1-rxe.tgz>
- OMB 3.1.1
<http://mvapich.cse.ohio-state.edu/benchmarks/OMB-3.1.1.tgz>